Statistical and Machine Learning Methods for Handling Selectivity in Non-Probability Samples

Doctoral Thesis

approved by the Department IV of the University of Trier
in partial fulfillment of the requirements for the degree of

Dr. rer. pol.

by

Simon Jonas Lenau, M.Sc.

Trier

Date of submission: June 07, 2021
Date of defence: December 02, 2022
Published: February 2023

Supervisors:
Prof. Dr. Ralf Münnich (Trier University)
Prof. Dr. Silvia Biffignandi (University of Bergamo)

Acknowledgements

First of all, I would like to thank my first supervisor, Prof. Ralf Münnich, for the cooperation, support and supervision. I am thankful for the confidence he placed in me and my work, which enabled me to contribute and pursue my own research ideas. I appreciate the working environment and research infrastructure he established, which created a productive and pleasant working atmosphere.

I also want to thank Prof. Silvia Biffignandi for agreeing to be my second supervisor and for providing valuable comments and suggestions on my work. I appreciate the interesting discussion we had during the thesis defense.

My work was supported by the European Union's Seventh Framework Programme for Research, Technological Development and Demonstration under Grant No. 312691 (InGRID) and the European Union's Horizon 2020 research and innovation programme under Grant No. 730998 (InGRID-2). I am grateful for this support.

I want to express my sincere gratitude to the entire team at the Economic and Social Statistics Department at Trier University for the friendly and cooperative collaboration. In various discussions and constructive comments, they had an substantial impact on this work.

Finally, I warmly thank my family for accompanying and helping me through this project, just as in all other situations in life. I am particularly grateful for the support and patience my girlfriend Wanja Patzke provided in the final phase of my doctorate.

Abstract

Non-probability sampling is a topic of growing relevance in national statistical institutes as well as academic and applied research, especially due to its occurrence in the context of new emerging data sources like web surveys and Big Data. Although it offers vast opportunities, especially for relatively cheap, fast and easy data collection, non-probability sampling also poses substantial statistical challenges because the respective sampling mechanisms are typically neither controlled nor known by the researchers. Well-established principles that facilitate reliable estimation and inference in case of probability sampling, thus, do not hold for non-probability sampling. Consequently, the use of non-probability samples as sources of statistical information raises substantial concerns in terms of data quality and representativity. To address these concerns and potential remedies, the overarching aims of this thesis are to discuss, expand and evaluate various methods for tackling the specific issues of non-probability samples in a common framework.

The main problem arising from the use of non-probability sampling is that the dependencies between sample inclusion and variables of interest are unknown. Therefore, the first methodological challenge in this context is to operationalize and quantify the selectivity (non-representativity) of the respective sampling mechanism. The second challenge is to account for potential selectivity and resulting biases in point estimation and inference.

Various pre-existing methods to tackle these core challenges posed by non-probability samples are identified and summarized. Particular attention is paid to the mathematical and algorithmic foundations required to implement and apply these methods. For quantifying selectivity, dependencies between sample inclusion and variables of interest are examined using suitable auxiliary information. Methods considered for this purpose are manual comparisons, statistical tests, matching, representativity indicators and a strategy to calculate intervals for the MSE of design linear estimators. To account for potential selectivity and biases of non-probability samples in estimation and inference, two broader paradigms can be distinguished. The *model-based paradigm* predicts information about the variables of interest outside the non-probability sample, a purpose for which various statistical and machine learning models are considered. In the *pseudo-design-based paradigm*, pseudo-design weights are estimated to mimic the design weights in probability sampling. The methods studied in this regard are propensity and calibration weighting as well as sub-sampling. A bandwidth of strategies to achieve a synthesis of both paradigms is discussed as well.

The two methodological novelties proposed and implemented in the scope of this thesis contribute to either of the two paradigms outlined above. The first proposal introduces semi-parametric artificial neural networks as prediction models, which integrate B-spline layers with an optimal knot positioning strategy in the general structure and fitting procedure of artificial neural networks. Extending and complementing these ideas, the second proposal introduces calibrated semi-parametric artificial neural networks to determine pseudo-design weights for non-probability samples. The rationale behind this approach is to establish propensity models of adaptable complexity that describe non-probability sample selection while incorporating soft and exact calibration constraints for estimates of totals, covariances and correlations. These two proposals constitute integrations and extensions of estimation methods that are commonly used for non-probability samples

and provide further possibilities and increased flexibility to utilize different types of auxiliary information for estimation. Complementing the theoretical foundation of these new methods, custom-made computational implementations are developed for fitting (calibrated) semi-parametric artificial neural networks by means of (stochastic) gradient descent, `BFGS` and sequential quadratic programming algorithms.

The performance of all the discussed methods with regard to the challenges posed by non-probability sampling is evaluated and compared, considering a bandwidth of scenarios in terms of sample selection mechanisms and available auxiliary information. This is done by means of a Monte Carlo simulation study as well as an application to a real non-probability sample, the WageIndicator web survey. Potentials and limitations of the different methods for handling the specific challenges of non-probability samples under various circumstances are highlighted by the theoretical and empirical discussion. Due to the heterogeneity in nature and purposes of different non-probability samples, no method is found to be suitable under all circumstances. The best strategy to use for non-probability samples rather depends on the particular selection mechanism, research interest and available auxiliary information. Nevertheless, the findings in this thesis show that existing methods as well as the newly proposed (calibrated) semi-parametric artificial neural networks can be used to ease or even fully counterbalance the issues of non-probability samples and highlight the conditions under which this is possible.

Contents

List of Figures

List of Tables

List of Algorithms

List of Symbols

General functions and symbols

General symbols

$a,\ b,\ c,\ d,\ h,$ $s,\ u,\ v$	Arbitrary numbers, defined in the respective context where they are used
$\boldsymbol{a},\ \boldsymbol{b},\ \boldsymbol{c},\ \boldsymbol{d},\ \boldsymbol{u},\ \boldsymbol{v}$	Arbitrary vectors, defined in the respective context where they are used
$\boldsymbol{A},\ \boldsymbol{B},\ \boldsymbol{C},\ \boldsymbol{D},$ $\boldsymbol{U},\ \boldsymbol{V}$	Arbitrary matrices, defined in the respective context where they are used
$\boldsymbol{L_a},\ \boldsymbol{U_a}$	Lower and upper boundary for parameters $\boldsymbol{a}$, of same dimension as $\boldsymbol{a}$
$\mathbb{R}^h_{>0}, \mathbb{R}^h_{\geq 0}$	Sets of all positive and non-negative real-valued vectors of size h
$\boldsymbol{\xi},\ \boldsymbol{\xi}^+,\ \boldsymbol{\xi}^-,\ \boldsymbol{\xi}^*$	Vectors of slack-variables

General functions

$[\cdot]_i$	i-th element of a vector
$[\cdot]_{ij}$	Element in row i, column j of a matrix
$\|\|\cdot\|\|_p$	p-norm of vectors
$\|\|\cdot\|\|_{\mathrm{F}}$	Frobenius norm of matrices
$\|\cdot\|$	Cardinality of a set
$\mathrm{Abs}\,(\cdot)$	Absolute value function – element-wise for vectors / matrices
$\delta\,(\cdot)\,,\ \tilde{\delta}\,(\cdot)$	General distance (or loss) functions
$\det(\cdot)$	Determinant
$\mathbf{diag}\,(\cdot)$	Diagonal-Operator: if argument is a matrix: the vector of diagonal elements if argument is a vector: a diagonal matrix with main diagonal equal to the argument
$\dim(\cdot)$	Dimension / length of a vector
$\exp(\cdot)$	Exponential function – element-wise for vectors / matrices
$\mathrm{F}\,(\cdot)$	Arbitrary function, if necessary defined in the respective context where it is used
$\mathbb{I}\,(\cdot)$	Indicator-function, being one iff the argument is true and zero else – applied element-wise for vectors / matrices
$\mathrm{tr}\,(\cdot)$	Trace of a matrix
$\mathscr{L}\,(\cdot)$	Log-likelihood function, the logarithm of a likelihood function
$\ell\,(\cdot)$	Logarithm of a density function
$\mathrm{L}\,(\cdot)$	Likelihood function
$\log(\cdot)$	Natural logarithm – element-wise for vectors / matrices
$\mathrm{Max}\,(\cdot)$	Maximum value function
$\mathrm{Min}\,(\cdot)$	Minimum value function
$\mathrm{ncol}\,(\bullet)$	Number of columns in argument matrix
$\mathrm{nrow}\,(\bullet)$	Number of rows in argument matrix
$\mathbf{pinv}\,(\cdot)$	Pseudo inverse function

$\mathbf{Rowmax}\,(\cdot)$	Row-wise maximum function, returning the vector of maxima for each row of the argument
$\mathrm{sign}\,(\cdot)$	Sign (or signum) function – element-wise for vectors / matrices
$\mathbf{softmax}\,(\cdot)$	Softmax function
$\mathrm{Sup}\,(\cdot)$	Supremum
$\mathbf{vec}\,(\cdot)\,, \mathbf{vec}^{-1}\,(\cdot)$	Vectorizing function that transforms a matrix into a vector of unique elements and its inverse function

Operators

$:=$	Definition operator
$\overset{!}{=}$	‚Must be equal' operator
$\leftarrow$	Assignment operator
$\circ$	Element-wise multiplication for vectors and matrices
$\oslash$	Element-wise division for vectors and matrices
$\otimes$	Kronecker product
$^{\circ}$	Element-wise exponentiation for vectors and matrices, e.g. $\boldsymbol{A}^{\circ 2}$
$^{\mathsf{T}}$	Transposition
$\wedge$	Logical conjunction

Special matrices

$\mathbf{0}_{p\times q}$	$p \times q$ null matrix
$\mathbf{1}_{p\times q}$	$p \times q$ matrix of ones
$\mathbf{I}_p$	$p \times p$ identity matrix
$\boldsymbol{L}, \boldsymbol{U}$	Lower and upper triangular part for the LU-factorization
$\boldsymbol{P}$	Pivoting matrix for the LU-factorization

Indices & index sets

$a,\ b,\ c,\ i,\ j,\ k,\ l,\ m$	(Running) indices
$\mathcal{I},\ \mathcal{J},\ \mathcal{M}$	Index sets, used for subsetting vectors and matrices
$\mathcal{I}_j$	j-th element of $\mathcal{I}$

Optimization framework

General functions and symbols

$\boldsymbol{\alpha},\ \boldsymbol{\alpha}^{+},\ \boldsymbol{\alpha}^{-},$ $\boldsymbol{\lambda},\ \boldsymbol{\lambda}^{*},\ \boldsymbol{\Lambda}$	Vectors of Lagrange multipliers – symbol α has different meaning in other contexts
$\underset{x}{\mathrm{argmax}}\,(\cdot)$	Function for maximizing argument-function w.r.t. x
$\underset{x}{\mathrm{argmin}}\,(\cdot)$	Function for minimizing argument-function w.r.t. x
$\boldsymbol{\Delta}_{\boldsymbol{\Theta}}$	Step direction for updating $\boldsymbol{\Theta}$
ϵ	Epsilon for one-sided limit calculations
$\bar{\mathbf{g}}\,(\cdot)$	General function representing equality constraints
$\tilde{\mathbf{g}}\,(\cdot)$	General function representing inequality constraints
$\mathbf{H}_{\mathrm{F}}\,(\cdot)$	Hessian matrix of F
$\mathbf{J}_{\mathrm{F}}\,(\cdot)$	Jacobian matrix of F

$\mathrm{L}(\cdot)$	Lagrange function of an optimization problem
$\lim_{x \to 0}(\cdot)$	Limit of argument-function for x going to 0
$\max_{x}(\cdot)$	Maximum value of argument-function w.r.t. x
$\min_{x}(\cdot)$	Minimum value of argument-function w.r.t. x
$\boldsymbol{\Theta},\ \boldsymbol{\Theta}^{(a)},\ \boldsymbol{\Theta}(\cdot)$	Vector / Matrix of optimization parameters and its value at iteration a – may be expressed as a function of penalty or shrinkage parameters

(Sequential) quadratic programming

$\mathcal{A}^{(a)}$	Working-set of inequality constraints active at iteration a
$\bar{\boldsymbol{G}},\ \widetilde{\boldsymbol{G}},\ \boldsymbol{G}$	Left-hand side multipliers of the linear equality, inequality and combined constraints
$\varphi(\cdot)$	Merit function for determining the step size
Ψ	Combined vector of step-direction and slack-variables
$\boldsymbol{Q}$	Quadratic distance multiplier matrix in a quadratic optimization problem
ς	Penalty-parameter for the slack-variables in the `SQP`-algorithm
$\bar{\varsigma}$	Updating-constant for ς in the `SQP`-algorithm
$\bar{\boldsymbol{t}},\ \widetilde{\boldsymbol{t}},\ \boldsymbol{t}$	Right-hand side of the linear equality, inequality and combined constraints

The `BFGS`-update

$\widetilde{\mathbf{B}}$	(Damped) `BFGS`-approximation of the inverse of a Hessian matrix
$\widetilde{\mathbf{H}}$	(Damped) `BFGS`-approximation of a Hessian matrix, $\widetilde{\mathbf{H}} \coloneqq \widetilde{\mathbf{B}}^{-1}$
$\boldsymbol{\eta},\ \theta,\ \boldsymbol{s},\ \boldsymbol{y}$	Vectors and parameters used for (modified) `BFGS` updating of $\widetilde{\mathbf{H}}$ or $\widetilde{\mathbf{B}}$ – symbols may have different meanings in other contexts

Distributions, related functions and parameters

General symbols related to distributions

$\perp\!\!\!\perp$	Independence
$\mathrm{CV}(\cdot)$	Coefficient of variation
$\mathrm{CV}(\cdot \mid x)$	Conditional coefficient of variation given x
$\mathrm{E}(\cdot),\ \widehat{\mathrm{E}}(\cdot)$	Expected value and its estimate
$\mathrm{E}(\cdot \mid x),\ \widehat{\mathrm{E}}(\cdot \mid x)$	Conditional expectation given x and its estimate
$\mathrm{F}_{\boldsymbol{Y}}(\cdot),\ \widehat{\mathrm{F}}_{\boldsymbol{Y}}(\cdot)$	(Cumulative) distribution function of $\boldsymbol{Y}$ and its estimate
$\mathrm{f}_{\boldsymbol{Y}}(\cdot)$	Density function of $\boldsymbol{Y}$
$\mathrm{f}_{\boldsymbol{Y}}(\cdot \mid x)$	Conditional density of $\boldsymbol{Y}$ given x
$\mathrm{P}(\cdot),\widehat{\mathrm{P}}(\cdot)$	Probability of an event, ant its estimate
$\mathrm{P}(\cdot \mid x),\ \widehat{\mathrm{P}}(\cdot \mid x)$	Conditional probability of an event given x and its estimate
$\mathrm{V}(\cdot),\ \widehat{\mathrm{V}}(\cdot)$	Variance and its estimate
$\mathrm{V}(\cdot \mid x),\ \widehat{\mathrm{V}}(\cdot \mid x)$	Conditional variance given x and its estimate

The family of exponential distributions

$\mathbf{a}(\phi)$, $\mathbf{b}(\boldsymbol{\theta})$, $\mathbf{c}(\cdot, \phi)$	Component functions of the family of exponential distributions
$\boldsymbol{\theta}$	Parameter of the family of exponential distributions

Binomial distribution and related symbols

$\binom{h}{\boldsymbol{B}}$	Binomial coefficients – element-wise for vectors / matrices $\boldsymbol{B}$
$\mathrm{B}(n, p)$	Binomial distribution with mean n trials and success probability p
$\boldsymbol{p}$	Vector of success probabilities of the binomial distribution

Normal distribution and related symbols

$\mathrm{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\Phi(\cdot)$	(Cumulative) distribution function of the standard normal distribution
$\Phi^{-1}(\cdot)$	Quantile function of the standard normal distribution
$\Phi'(\cdot)$	Density function of the standard normal distribution

Other distributions and related symbols

df	Degrees of freedom
$\mathsf{F}_{\mathrm{df}_1,\mathrm{df}_2}$	F-distribution with df_1 and df_2 degrees of freedom
t_{df}	t-distribution with df degrees of freedom
χ^2_{df}	χ^2-distribution with df degrees of freedom

Sampling, data sets and estimation

General symbols

$\boldsymbol{b}(\cdot)$	Weight rescaling function for re- and sub-sampling methods
$\mathrm{D}(\cdot)$	Sampling design
$\mathbb{S}$	Set of all possible samples

Data sets and identifiers

cal, $\mathcal{S}^{\mathsf{cal}}$	Reference data for calibration targets and corresponding set of unique identifiers
nps, $\mathcal{S}^{\mathsf{nps}}$	General non-probability sample and corresponding set of unique identifiers
P, $\mathcal{S}^{\mathrm{P}}$	Target population and corresponding set of unique identifiers
ps, $\mathcal{S}^{\mathsf{ps}}$	General probability sample and corresponding set of unique identifiers
res, $\mathcal{S}^{\mathsf{res}}$	Reference data to estimate a response model and corresponding set of unique identifiers
$\mathcal{S}^{(i)}$	Set of unique identifiers for i-th sub-group of the population, defined in the respective context where it is used
s, $\mathcal{S}^{\mathsf{s}}$, t, $\mathcal{S}^{\mathsf{t}}$	Arbitrary data sets and corresponding sets of unique identifiers
$\bar{\mathsf{s}}$, $\mathcal{S}^{\bar{\mathsf{s}}}$	Population P without data set s: $\left(\mathcal{S}^{\bar{\mathsf{s}}} \coloneqq \mathcal{S}^{\mathrm{P}} \setminus \mathcal{S}^{\mathsf{s}}\right)$
u, $\mathcal{S}^{\mathsf{u}}$	General union of two or more data sets and corresponding set of unique identifiers

Information in data sets

fr^{s}	Sampling fraction for $\mathcal{S}^{\mathsf{s}}$: $fr^{\mathsf{s}} \coloneqq n^{\mathsf{s}}/\mathrm{N}$
N	Number of elements / observations in the population $\left(\left\lvert\mathcal{S}^{\mathrm{P}}\right\rvert\right)$
n^{s}	Number of elements / observations in a data set s ($\lvert\mathcal{S}^{\mathsf{s}}\rvert$)
$\tilde{n}^{\mathsf{s}}$	Number of elements in a resample of s
o	Number of $\boldsymbol{Y}$-variables
$\boldsymbol{\pi}^{\mathsf{s}}$	Vector of inclusion or participation probabilities for elements being in $\mathcal{S}^{\mathsf{s}}$
$\boldsymbol{p}^{\mathsf{s}}, \widehat{\boldsymbol{p}}^{\mathsf{s}}$	Vector of inclusion or participation propensities for elements being observed in $\mathcal{S}^{\mathsf{s}}$ and its estimate
p	Number of $\boldsymbol{X}$-variables
q	Number of $\boldsymbol{Z}$-variables
$\boldsymbol{r}^{\mathsf{s}},\ \widetilde{\boldsymbol{r}}^{\mathsf{s}}$	Vector of inclusion indicators for a data set s and a modified (weighted) version thereof
r	Number of covariances or correlations
$\boldsymbol{w}^{\mathsf{s}}$	Vector of weights in data set s
$\boldsymbol{X}, \boldsymbol{X}^{\mathsf{s}}$	Matrices of auxiliary variables in the population and data set s, used for calibration and prediction methods
$\boldsymbol{X}^{\mathsf{s}}_{\mathcal{I}\cdot}$	Submatrix of $\boldsymbol{X}^{\mathsf{s}}$ consisting of all rows indexed by $\mathcal{I}$
$\boldsymbol{X}^{\mathsf{s}}_{\mathcal{J}\cdot}$	Submatrix of $\boldsymbol{X}^{\mathsf{s}}$ consisting of all columns indexed by $\mathcal{J}$
$\boldsymbol{X}^{\mathsf{s}}_{\mathcal{I}\mathcal{J}}$	Submatrix of $\boldsymbol{X}^{\mathsf{s}}_{\mathcal{I}\cdot}$ consisting of all columns indexed by $\mathcal{J}$
$\boldsymbol{x}^{\mathsf{s}}_{i\cdot}$	i-th row/observation of matrix $\boldsymbol{X}^{\mathsf{s}}$
$\boldsymbol{x}^{\mathsf{s}}_{\cdot j}$	j-th column/variable of matrix $\boldsymbol{X}^{\mathsf{s}}$
x^{s}_{ij}	Element in row i, column j of matrix $\boldsymbol{X}^{\mathsf{s}}$
$\boldsymbol{Y}, \boldsymbol{Y}^{\mathsf{s}}$	Matrix of target variables in the population and data set s
$\boldsymbol{Z}, \boldsymbol{Z}^{\mathsf{s}}$	Matrix of design or response model variables in the population and data set s

Estimation and testing

α	Significance levels for statistical tests – symbol $\boldsymbol{\alpha}$ has different meaning in other contexts
$\mathrm{CI}\,(\cdot)$	Confidence interval of an estimator
$\mathrm{CIR}\,(\cdot)$	Confidence interval coverage rate of an estimator
$\mathbf{e}\,(\cdot)$	Function expressing the difference from the weighted mean
$\boldsymbol{\mu}_{\boldsymbol{Y}},\ \widehat{\boldsymbol{\mu}}_{\boldsymbol{Y}}\,(\boldsymbol{w}^{\mathsf{s}})$	Row-vector of $\boldsymbol{Y}$-means and their estimates using weights $\boldsymbol{w}^{\mathsf{s}}$
$\mathbf{MSE}\,(\cdot)$	Mean squared error of an estimator
$\widehat{\mathrm{N}}\,(\boldsymbol{w}^{\mathsf{s}})$	Estimated population size when using $\boldsymbol{w}^{\mathsf{s}}$, sum of weights $\boldsymbol{w}^{\mathsf{s}}$
$\boldsymbol{\rho}_{\boldsymbol{Y}},\ \widehat{\boldsymbol{\rho}}_{\boldsymbol{Y}}\,(\boldsymbol{w}^{\mathsf{s}})$	Correlation matrix of $\boldsymbol{Y}$ and its estimate using weights $\boldsymbol{w}^{\mathsf{s}}$
$\widehat{\mathrm{R}}\,(\widehat{\boldsymbol{p}})$	Representativity indicator estimated from $\widehat{\boldsymbol{p}}$
$\mathbf{RBias}\,(\cdot)$	Relative bias of an estimator
$\mathbf{RRMSE}\,(\cdot)$	Relative root mean squared error of an estimator
$\boldsymbol{\Sigma}_{\boldsymbol{Y}},\ \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}\,(\boldsymbol{w}^{\mathsf{s}}),$ $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}\,(\boldsymbol{w}^{\mathsf{s}})$	Covariance matrix of $\boldsymbol{Y}$, its ML and unbiased estimates using weights $\boldsymbol{w}^{\mathsf{s}}$

$\boldsymbol{\tau}_{\boldsymbol{Y}},\ \widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right)$	Row-vector of $\boldsymbol{Y}$-totals and their estimates using weights $\boldsymbol{w}^{\mathsf{s}}$
$\boldsymbol{\vartheta},\ \widehat{\boldsymbol{\vartheta}}$	General statistic(s) and corresponding estimators / estimates
$\widehat{T},\ \widehat{\boldsymbol{T}}$	(Estimated) test statistic and a vector of such statistics
$\widehat{\mathbf{V}}_{\mathbf{b}}\left(\cdot\right),\ \widehat{\mathbf{V}}_{\mathbf{w}}\left(\cdot\right),$ $\widehat{\mathbf{V}}_{\mathbf{t}}\left(\cdot\right)$	Estimated between, within and total variance. Defined in the respective context where they are used

Weighting framework

$\mathbf{C}_{\boldsymbol{a}}$	Centering constant for parameter(s) $\boldsymbol{a}$, of same dimension as $\boldsymbol{a}$
$\boldsymbol{g}, \boldsymbol{g}\left(\cdot\right)$	Vector of correction weights and function that is used to generate it in the functional form approach
$\boldsymbol{\epsilon}, \boldsymbol{\varepsilon}$	Vectors of multiplicative error terms
$\mathbf{u}, \mathbf{v}, \mathbf{w}$	Vectors of importance weights
$\widetilde{\boldsymbol{w}}$	Vector of output / pseudo-design weights
$\boldsymbol{\omega}$	Vector of weighting model parameters

Prediction models

General symbols

$\boldsymbol{\beta}$	(Regression) coefficients used to predict $\boldsymbol{Y}$. Dimensions depend on the respective prediction model
$\mathrm{cv}\left(\cdot\right), \mathrm{gcv}\left(\cdot\right)$	Cross-validation and generalized cross-validation criterion
$\boldsymbol{E}, \boldsymbol{E}^{\mathsf{s}}$	Matrix of residuals in the population and data set s
$\mathbf{m}\left(\boldsymbol{X}, \boldsymbol{\Theta}\right)$	Arbitrary model, depending on independent variables $\boldsymbol{X}$ and model-parameters $\boldsymbol{\Theta}$
$\boldsymbol{p}\left(\cdot\right)$	Penalty function for shrinkage methods
R^2	Coefficient of determination
$\boldsymbol{t}\left(\cdot\right),\ \boldsymbol{t}_{\mathrm{k}}\left(\cdot\right)$	General transformation functions, transforming its input matrix into a matrix of possibly different shape
$\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{X}}^{\mathsf{s}}$	General transformed predictor matrix in the population and data set s
$\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{Y}}^{\mathsf{s}}$	Matrix of predictions for $\boldsymbol{Y}$ in the population and data set s
$\widetilde{\boldsymbol{Y}}, \widetilde{\boldsymbol{Y}}^{\mathsf{s}}$	Matrix of adjusted dependent variable in the population and data set s

Generalized linear & additive models

$\boldsymbol{\eta},\ \boldsymbol{\eta}\left(\cdot\right),$ $\boldsymbol{\eta}^{\mathsf{s}},\ \boldsymbol{\eta}^{\mathsf{s}}\left(\cdot\right)$	Systematic component / combination of independent variables for generalized linear and additive models in the population and data set s – may be expressed as a function of parameters
$\boldsymbol{l}\left(\cdot\right), \boldsymbol{l}^{-1}\left(\cdot\right)$	Link and inverse link function, connecting systematic component $\boldsymbol{\eta}$ and conditional mean $\boldsymbol{\mu}$
$\boldsymbol{\kappa}$	Vector of parameters defining the transformation $\boldsymbol{t}\left(\boldsymbol{X}, \boldsymbol{\kappa}\right)$ of $\boldsymbol{X}$
$\boldsymbol{\kappa}^{(i)}$	Sub-vector of $\boldsymbol{\kappa}$, defining the transformation $\boldsymbol{t}_{\mathrm{i}}\left(\boldsymbol{x}_{\cdot i}, \boldsymbol{\kappa}\right)$ of $\boldsymbol{x}_{\cdot i}$
$\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}$	Vector of expected values for each observation of $\boldsymbol{y}_{\cdot l}$ in data set s
$\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})},\ \boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}\left(\cdot\right)$	Covariance matrix of all observation of $\boldsymbol{y}_{\cdot l}$ in data set s. May be expressed as a function of dispersion parameters

ϕ	Dispersion parameter
$\boldsymbol{W}$	Weight-Matrix for iteratively reweighted estimation of generalized linear or additive (mixed) models

Generalized additive mixed models

$\boldsymbol{\phi},\ \boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})},\ \boldsymbol{\phi}^{(\boldsymbol{u})}$	Vectors of dispersion or variance (component) parameters
$\boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot l}}^{(\mathsf{s})}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right),$ $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{(\mathsf{s})}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)$	Variance components in data set s, modeled as functions of variance component parameters

(Multivariate adaptive regression) Splines

$\mathbf{B}_k^l\left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)$	k-th B-Spline basis function of order l for variable $\boldsymbol{x}_{\cdot j}$, depending on knot vector $\boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}$
$\boldsymbol{I}^{\boldsymbol{x}_{\cdot j}}, \widetilde{\boldsymbol{I}}^{\boldsymbol{x}_{\cdot j}}, \bar{\boldsymbol{I}}^{\boldsymbol{x}_{\cdot j}}$ $\boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}, \bar{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}$	Vectors of spline knots associated with variable $\boldsymbol{x}_{\cdot j}$
$\mathscr{K}^{\boldsymbol{x}_{\cdot j}}$	Set of possible node vectors $\boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}$ in the MARS algorithm associated with variable $\boldsymbol{x}_{\cdot j}$
$\mathscr{V}^{(a)}$	Set of variable associated with the a-th base function in the MARS algorithm

Support vector machines

e	Insensitivity bandwidth in support vector regression
$\mathrm{K}\left(\boldsymbol{t}\left(\boldsymbol{x}_{i\cdot}, \boldsymbol{x}_{j\cdot}\right)\right)$	Kernel function

Artificial neural networks

$\boldsymbol{d}_i^{(kl)}$	Coefficient for backpropagation, containing Jacobian information

Heckman model

$\boldsymbol{Y}^*$	Latent variable for response tendency
$\boldsymbol{\beta}^*$	Regression coefficients for modeling $\boldsymbol{Y}^*$
$\widetilde{\beta}_k$	Regression coefficients for modeling the selection bias

List of Abbreviations

Computational implementation and algorithms

Programming software & languages

C++	The C++ programming language
C	The C programming language
R	The programming software R

Libraries

Armadillo	A C++ library for linear algebra
BLAS	Library of Basic Linear Algebra Subprograms
Eigen	C++ libraries for linear algebra
LAPACK	A linear algebra package (which makes use of BLAS sub-programs)
LIBSVM	A library for support vector machines
NLOPT	A library (inter alia for C++) for nonlinear optimization
Open BLAS	An optimized BLAS library
RBLAS	BLAS library provided by the R-software
SuperLU	A library (inter alia for C++) for solving systems of linear equations
ViennaCL	A C++ library for linear algebra

R-packages

ann	Package for the proposed semi-parametric artificial neural networks
calmod	Package for the proposed weighting framework (calibrated semi-parametric artificial neural networks)
data.table	Package for efficient computation with large data sets
e1071	Package for *unweighted* SVMs
earth	Package for multivariate adaptive regression splines
glmnet	Package for penalized GLMs
Matching	Package for matching
mgcv	Package for generalized additive mixed models
MissMech	Package for testing Missing Completely at Random (MCAR)
mvtnorm	Package for computing multivariate normal and t-distributions
Rcpp	Package for integrating C++ and R
RcppArmadillo	Package for integrating Armadillo using Rcpp
RcppEigen	Package for integrating Eigen using Rcpp
rpart	Package for regression trees
sampleSelection	Package for integrated response and dependent variable models
sqp	Package for sequential quadratic programming
survey	Package for analysis of complex survey samples
TOSTER	Package for Two One-Sided Tests (TOST) equivalence testing
uniftest	Package for testing for uniformly distributed variables

Algorithms

`BFGS`	The Broyden-Fletcher-Goldfarb-Shanno-algorithm
`DFP`	The Davidon-Fletcher-Powell-algorithm
`QP`	The quadratic programming algorithm (with active-set strategy for inequality constraints)
`SLSQP`	A least-squares based `SQP`-algorithm
`SQP`	The sequential quadratic programming algorithm

General abbreviations

AMELI	Advanced Methodology for European Laeken Indicators
ANN	Artificial neural network
ANOVA	Analysis of variance
Bagging	Bootstrap aggregation
B-spline	Basis Spline
CAPI	Computer-assisted personal interviewing
CRAN	The Comprehensive R Archive Network
CI-rate	Confidence interval coverage rate
EU-SILC	European Union Statistics on Income and Living Conditions
GAM	Generalized additive models
GAMM	Generalized additive mixed models
GLM	Generalized linear models
GLMM	Generalized linear mixed models
GREG	Generalized regression estimator
HT-estimator	Horvitz-Thompson-estimator
i.i.d.	Independent and identically distributed
InGRID	Inclusive Growth Research Infrastructure Diffusion
IRWLS	Iteratively reweighted least squares
ISCED	International Standard Classification of Education
KKT-conditions	Karush-Kuhn-Tucker optimality conditions
LASSO	Least absolute shrinkage and selection operator
LU-factorization	Factorization of a matrix in lower and upper triangular part, for solving systems of linear equations
MAR	Missing at random
MARS	Multivariate adaptive regression splines
MC	Monte Carlo
MCAR	Missing completely at random
ML	Maximum likelihood (method)
MNAR	Missing not at random
MRP	Multilevel regression and post-stratification
MSE	Mean squared error
OLS	Ordinary least squares
P-spline	Penalized B-spline
PSU	Primary sampling unit

R-indicator	Representativity indicator
RBias	Relative bias
RRMSE	Residual (or restricted) maximum likelihood
RRMSE	Relative root mean squared error
SVM	Support vector machine
WI	WageIndicator (web survey)
w.r.t.	With respect to

1 Introduction

In modern societies, statistical information constitutes an essential foundation to address crucial societal, political, environmental and economic needs and challenges. The demand for such information is rising over the recent decades and commonly presumed to increase further in the years to come (cf. e.g. European Statistical System, 2014, pp. 6 ff; Statistisches Bundesamt, 2020, pp. 14 ff; United Nations, 2013, pp. 80 ff). A part of the data required to fulfill this demand may be obtained from full censuses, which include the whole set of elements (e.g. persons, households or companies) constituting the population of interest. However, the majority of statistical information is based on sample surveys, which are limited to a subset of observations from such a population. In comparison to full censuses, these surveys are typically cheaper, faster and easier to conduct. Under certain conditions, they may even provide more accurate results (cf. Fuller, 2009, p. 1; Lohr, 2010, pp. 17 f; Särndal, Swensson and Wretman, 1992, pp. 3 ff).

Although the concept of generalizing from observed elements to a larger set of similar units is a fundamental aspect of human cognition and therefore much older, the explicit use of sample surveys for social and economic statistics dates back to the work of Kiær (1895; 1897; cf. Kruskal and Mosteller, 1980, pp. 172 ff). Yet, the widespread acceptance of sampling as an accurate source of information in science, official statistics and general society is particularly due to the scientific contributions of Bowley (1925), Neyman (1934) as well as Hansen and Hurwitz (1943). These authors formally introduce and extend the fundamental ideas of *probability sampling*, establishing a particular form of known and controlled randomization mechanisms for sample selection. Probability samples provide reliable information that is reasonably generalizable to a finite population while only a subset of all relevant elements needs to be observed. Complementing the design of sample selection procedures, the second fundamental aspect of survey statistics is therefore to estimate population quantities from a sample and quantify the accuracy of the resulting estimates (cf. Kish, 1965, p. 4; Särndal, Swensson and Wretman, 1992, p. 29). In conjunction, probability sampling and corresponding estimation strategies constitute a gold standard for drawing valid conclusions about a finite population that is only partially observed. As a consequence, survey sampling in national statistical institutes, academic and applied research throughout the last decades is predominantly focused on probability sampling. Correspondingly, terms like 'representative' or 'scientific samples' that emphasize the quality of certain sources of information are used nearly synonymously with probability samples (cf. Baker et al., 2013b, pp. 90 ff; Elliott and Valliant, 2017, p. 249; Kruskal and Mosteller, 1979a, p. 15).

Especially in the recent years, however, a paradigm shift in survey statistics is becoming apparent. Obtaining proper probability samples is getting increasingly difficult and costly, e.g. due to rising non-contact or non-response rates and declining coverage of sampling frames. These developments are particularly severe for probability telephone samples but pose considerable threats to other forms of probability sampling as well, especially in light of growing demands to provide statistical information in a cost-efficient and timely manner. At the same time, the expansion of digitalization and (online) telecommunication leads to rapid increases in extent as well as availability of alternative and new data sources, of which Big Data and web surveys are prevalent examples. These data sources typically include only a subset of elements from a target population, but most commonly do not originate from a probability sampling mechanism. Therefore, they constitute cases of

non-probability sampling. Their most highlighted advantages are easier, cheaper and/or faster data collection in comparison to classical probability samples. Especially Big Data often results as by-product of other processes, such as administrative, mobile phone location or search engine data. Such types of "organic data" (Groves, 2011, p. 866) may be available as natural aspects of a modern society's continuous self-observation, potentially even in real-time. This can help to largely reduce the need to gather the information anew for specific research purposes, lower costs of data collection and reduce the response burden for elements in a target population. Consequently, official statistics as well as academic and applied researchers increasingly consider and use non-probability samples as sources of statistical information, either in addition or as an alternative to probability samples (cf. e.g. Baker et al., 2013b; Directors-General of the National Statistical Institutes, 2013; European Statistical System, 2014, pp. 22 ff; Münnich and Zwick, 2016; National Research Council of the United States, 2013; Statistisches Bundesamt, 2020; United Nations, 2014).

Yet, these developments do not imply that probability sampling is outdated and can easily be replaced by non-probability sampling. As the term non-probability sampling merely expresses a differentiation to probability sampling, it encompasses manifold different selection processes that are not limited to the new data sources outlined above. These heterogeneous processes may not have much in common, but an important characteristic that all of them share is the violation of well-established principles that facilitate estimation and generalization in case of probability sampling. The arbitrary, often unknown and uncontrolled nature of selection mechanisms raises substantial concerns in terms of data quality and representativity. In particular, it is often unclear whether non-probability samples predominantly or exclusively contain information about specific subgroups from a target population. This implies that parts of the population may be systematically excluded from or disproportionately represented in the sample, leading to coverage and selectivity issues. For example, web survey or search engine data completely omits non-users of the internet and over-represents frequent in comparison to rare users. Depending on the actual research interest, such systematic differences pose a considerable risk of biased estimates, thereby impairing accuracy and reliability of conclusions drawn from non-probability samples. Coverage and selection biases cannot even be compensated by the immense volume of Big Data, although the converse is sometimes argued. "Compensating for quality with quantity is a doomed game" (Meng, 2018, p. 695). Therefore, the use of non-probability samples for providing adequate statistical information raises important methodological questions and challenges to be addressed (cf. e.g. Biffignandi and Bethlehem, 2012; Enderle, Münnich and Bruch, 2013; European Statistical System, 2014, p. 25; Japec et al., 2015; Lohr and Raghunathan, 2017; Shlomo and Goldstein, 2015; Statistisches Bundesamt, 2020, pp. 16 f; United Nations, 2014, p. 2).

Despite all potential imperfections, non-probability samples are gaining prevalence and relevance in research and official statistics, a development which is largely caused by the growing abundance and availability of new digital data sources. The substantial potential benefits and challenges of these data sources render non-probability samples a topic that requires thorough statistical treatment. Development and evaluation of methods for quality assessment and estimation in non-probability samples is therefore a major current challenge for survey statistics. At the same time, the profession's strong and historical devotion to probability sampling is by no means obsolete. At least in the medium term, a combination and joint usage of upcoming non-probability and classical probability samples

is unanimously considered the most promising approach to the growing relevance of new digital data sources (cf. e.g. Baker et al., 2013b; Daas et al., 2015; European Statistical System, 2014, p. 25; Groves, 2011; Japec et al., 2015; Lohr and Raghunathan, 2017; Meng, 2018; Münnich and Zwick, 2016; United Nations, 2014, p. 2).

> "Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring the reliability of useful statistical information through the theory of probability."
> (Deming, 1950, p. 2)

Following this notion of sampling, the focus in this thesis lies on methods for assessing and compensating the problems provoked by non-probability sample selection. A variety of such methods is proposed in the relevant literature, ranging from approaches for examining a sample's selectivity to strategies for estimation and inference. However, there is considerable lack of a comprehensive joint discussion, evaluation and comparison of these methods in a common framework. So far, most publications consider only one or at best a few of the proposed methods (cf. Daas et al., 2015, p. 249; Elliott and Valliant, 2017, p. 262). To fill this gap, the first aim of the current work is to provide a unifying discussion of methods that are relevant for dealing with non-probability sampling.

In addition, proposing new approaches for estimation from non-probability samples constitutes the second aim of this thesis. The first proposal introduces semi-parametric artificial neural networks for prediction, which incorporate B-spline layers that perform knot optimization in a general neural network structure. The second methodological proposal in scope of this thesis establishes calibrated response models to provide weights for non-probability samples. These models are calibrated semi-parametric artificial neural networks that can be used to model complex non-probability sampling processes while incorporating soft and exact calibration (benchmarking) constraints for estimates of totals, covariances and correlations. Both proposals constitute integrations and extensions of well-established and frequently used estimation methods for non-probability samples. In that way, they provide further possibilities and flexibility to utilize auxiliary information in estimation, especially when jointly using non-probability and probability samples for this purpose. In addition to these content-wise similarities, both proposals are based on the flexibility of B-splines and artificial neural networks and use related optimization routines, which constitutes similarities also in terms of their mathematical and computational formulation. To allow for actual application of the newly proposed methods, custom-made computational implementations are developed in the context of this thesis as well.

The third aim of the current work is to evaluate and compare the performance of all considered methods with regard to the challenges posed by non-probability sampling. This is mainly done by means of Monte Carlo simulation studies, which allow examining a variety of scenarios for underlying population structure and sample selection mechanism. To still not exclusively rely on simulated populations and samples, the methods are additionally applied and evaluated with regard to a real non-probability sample.

To achieve these three research goals, this work is organized into eight chapters. Following the current introductory chapter 1, fundamental aspects in dealing with non-probability samples are discussed in chapter 2. The variety and relevance of non-probability samples in manifold research areas and kinds of analyses are summarized, paying particular attention

to new digital data sources and their important peculiarities. By referring to specific characteristics and benefits of probability sampling that do not hold for other forms of sample selection, a formal discussion of non-probability sampling and the core statistical challenges arising from its use is provided.

In chapter 3, approaches for operationalizing and quantifying selectivity of non-probability samples are presented. These include manual comparisons, statistical tests and matching for examining (non-)compliance of non-probability sample estimates with suitable auxiliary information. Furthermore, the use of propensity models for describing the non-probability sampling process, representativity indicators, and a framework for quantifying estimation error in non-probability samples based on the work of Meng (2018) and Schouten (2007) are discussed.

The mathematical and computational foundations required throughout the subsequent discussion are presented in chapter 4. The relevant methods to perform numerical optimization are discussed, including algorithms for linear programming as well as unconstrained and constrained non-linear optimization. The main purpose of these methods in the current thesis is to facilitate estimation from non-probability samples by solving optimization problems commonly arising in this context.

Based on these foundations, methods to perform estimation from non-probability samples are discussed in chapter 5. These methods can be partitioned into two major paradigms. The *model-based* paradigm focuses on predicting the target variable(s) or their characteristics of interest for observations outside the non-probability sample. A variety of statistical and machine learning models are considered relevant for this purpose in the scientific discourse and successively presented. These include matching, (generalized) linear and additive regression models and the corresponding mixed models. Further existing prediction methods discussed in this context are (multivariate adaptive) regression splines, regression trees, artificial neural networks and support vector machines. Semi-parametric artificial neural networks are developed as a novel alternative to these established prediction models. In contrast, the *pseudo-design-based* paradigm aims at generating surrogate weights to mimic design weights in probability sampling. Pre-existing methods discussed in this context include response propensity and calibration weighting as well as subsampling. Calibrated semi-parametric neural networks are proposed as a new approach to integrate and extend the ideas of propensity and calibration weighting. Strategies to achieve a synthesis of model- and pseudo-design-based methods are discussed as well, including the use of weighted prediction models, jointly modeling selection process and target variables, and methods for weighted aggregation of predictions. Approaches to perform inference from non-probability samples, which typically again refer to one of the above paradigms, are furthermore presented.

In chapter 6, the different methods' suitability to assess and compensate the potential selectivity of non-probability samples is evaluate and compared. An overview of the applied software packages that provide implementations of the considered methods is presented. In a preliminary Monte Carlo simulation study, the custom-made implementations of the proposed (calibrated) semi-parametric neural networks and corresponding numerical solvers are evaluated. The major Monte Carlo simulation designed to evaluate, compare and discuss the performance of all the methods for non-probability samples considered throughout the preceding chapters under a variety of scenarios is then presented.

To nevertheless not exclusively rely on simulations, an application to a real non-probability sample is discussed in chapter 7. The methods for assessing and compensating selectivity are applied and evaluated with regard to the WageIndicator volunteer web survey (cf. Tijdens et al., 2010).

The concluding chapter 8 summarizes and discusses the main findings, advantages and drawbacks of the research in scope of this thesis. Some topics for future research are outlined as well.

2 Issues and Challenges Regarding Non-probability Samples

Probability sampling constitutes a gold standard in many scientific fields, including economy, politics, sociology, health, forestry and official statistics in particular (cf. e.g. Barratt, Ferris and Lenton, 2015, p. 4; Berrens et al., 2003, p. 2; van den Brakel et al., 2017, p. 183; Citro, 2014, p. 137; Einstein and Baecher, 1983; Meng, 2018, p. 689; Münnich et al., 2016; Wooldridge, 2012, p. 6). This primacy is to a large extent due to the fact that probability sampling provides a known randomization mechanism based on sample selection. Using this property, statistical theory permits construction of various estimators that are unbiased and under certain conditions generalizable to a finite population, e.g. through variance estimates, confidence intervals and statistical tests. This is possible for any variable of interest, without making any assumptions about its distribution (cf. Breidt and Opsomer, 2017, p. 191; Kalton, 1983, p. 90). Although a more formal discussion of its properties and benefits is deferred to section 2.2, probability sampling is therefore an important tool for obtaining reliable estimates and measuring their precision based on a single sample (cf. Wolter, 2007, p. 1; Särndal, Swensson and Wretman, 1992, p. 33). Nevertheless, this gold standard is not an imperative in all occasions. Its necessity and applicability depend on the area of interest.

> "Great advances of the most successful sciences – astronomy, physics, chemistry – were, and are, achieved without probability sampling. [...] Probability sampling for randomization is not a dogma, but a strategy, especially for large numbers."
>
> (Kish, 1965, pp. 28 f)

Even disciplines strongly devoted to probability sampling often require research which cannot rely on its randomization for various reasons: in many cases of *observational studies* for causal inference, randomized experiments are not feasible. This may be for ethical reasons, e.g. in case of certain medical conditions and treatments (cf. Cochran and Chambers, 1965, p. 236; Mercer et al., 2017, p. 250), drug usage (cf. Barendregt, Van der Poel and Van de Mheen, 2005, p. 124; Heckathorn, 2002, p. 10) or influences of educational backgrounds (cf. Wooldridge, 2012, p. 14). Besides ethical issues, another cause to abstain from probability sampling can be timeliness, i.e. when it comes to evaluation of long-term processes, such as effects of child nutrition (cf. Rubin, 1974, p. 687) or certain social programs (cf. Cochran and Chambers, 1965, p. 240). In addition, the *lack of a valid sampling frame* to select elements from prevents the practicability of probability sampling for certain areas of interest (cf. Citro, 2014, p. 141). This is especially the case when studying rare or hidden populations that are hard to locate. Purposive or respondent-driven sampling can be advisable in such circumstances (cf. Lohr, 2010, p. 517). Examples include research on individuals suffering from certain health conditions (cf. Feild et al., 2006), exhibiting illegal or socially stigmatized behavior (cf. Barratt, Ferris and Lenton, 2015, p. 4; Frank and Snijders, 1994, p. 53; Salganik and Heckathorn, 2004), or being experts in highly specific fields (cf. Tongco, 2007, p. 147). Reasons of cost and practicability furthermore lead researchers across various scientific fields to rely on *convenience samples* that are easy and/or inexpensive to obtain (cf. Berrens et al., 2003, p. 2; Nielsen et al., 2017, p. 31; Särndal, Swensson and Wretman, 1992, p. 529).

In fields traditionally relying on probability sample surveys, the difficulty to achieve the selected elements' participation is an additional issue of growing relevance. *Non-response* is rising over the years (cf. Citro, 2014, p. 142; de Heer, 1999; de Heer and de Leeuw, 2002) and often selective, i.e. related to certain characteristics of the sampled elements. Various potential reasons for this trend are studied by Groves and Couper (1998). Even for a sample from a probability sampling design, degree and selectivity of the occurring non-response often raise concerns about whether the theoretical ideal of a probability sample is still met. Hence, this is considered a threat to important theoretical benefits of probability sampling (cf. Baker et al., 2013b, p. 91; Elliott and Valliant, 2017, p. 250; Little, 1988b, p. 287; Lohr, 2010, p. 534). Although partially established as a remedy for low response rates, similar arguments apply to *access panels* because these are typically defined as a sampling frame that is constituted by volunteerism rather than a population register or probability sample (cf. Amarov and Rendtel, 2013, p. 103; Enderle, Münnich and Bruch, 2013, p. 92; Loosveldt and Sonck, 2008, p. 93).

All of these examples can constitute non-probability sampling since this term is merely defined in differentiation to probability sampling. Although violating the requirements of known and controlled randomization that are used in established probability sampling theory, the above outline illustrates that non-probability samples are regularly used in academic and applied research as well as official statistics. Rather than being neglected for their issues and imperfections, they therefore constitute a topic that needs to be discussed and handled in survey statistics (cf. sections 2.2 and 2.3; Baker et al., 2013b; Buelens, Burger and van den Brakel, 2015, p. 7; Elliott, 2009, p. 1; Japec et al., 2015, p. 860). Some of the above examples are already tackled thoroughly in statistical literature, in particular non-response (cf. e.g. van Buuren, 2018; Little and Rubin, 2019; Rubin, 1987; Särndal and Lundström, 2005) and observational studies (cf. e.g. Cochran, Moses and Mosteller, 1983; Rosenbaum, 2010; Rubin, 2006).

Especially through the recent rise and quantity of *new data sources*, further efforts to address non-probability samples from a statistical point of view are nevertheless demanded because a unified framework for estimation and inference from such samples does not exist (cf. e.g. Baker et al., 2013b, p. 93; Beręsewicz, 2015, p. 54; Daas et al., 2015, p. 249; Gelman et al., 2016a, pp. 102 f; Japec et al., 2015, p. 860; Pfeffermann, 2015, pp. 427 ff; Shlomo and Goldstein, 2015, p. 787). This is additionally motivated in the following section 2.1, where a literature review on characteristics and application examples of new data sources is given. Important concepts of probability sampling, to which non-probability sampling is generally defined to be the negation (cf. e.g. Mercer et al., 2017, p. 251), are summarized in section 2.2. Combining this introduction to probability sampling with the preceding literature review, the formal discussion of statistical issues arising from non-probability samples in section 2.3 is fostered.

2.1 Literature Review: Characteristics and Usage of New Data Sources

In light of declining response rates, coupled with increasing efforts to reduce costs and provide more information in greater detail, the growing availability and scope of new data sources are often seen as a promising trend for statistics (cf. Beręsewicz, 2015, p. 45; Buelens, Burger and van den Brakel, 2018, p. 326; Daas et al., 2015, p. 249; Groves, 2011, p. 868; Holt, 2007, pp. 1 ff; United Nations, 2014, p. 2). These data sources commonly emerge with the expansion of digitalization and the internet. Important examples are *Big Data* and *web surveys*, both being umbrella terms to manifold types of information and ways in which they are obtained (referred to as the selection processes or data generating mechanisms; cf. Buelens, Burger and van den Brakel, 2018, pp. 322 f; Greenacre, 2016, p. 397; Japec et al., 2015, p. 854). Similar as for the examples mentioned before, such new kinds of data often provide information that is otherwise not available. In addition, they may arise as by-products of other processes and can therefore be used for reducing cost and response-burden as well as to circumvent non-response and increase timeliness (cf. Bethlehem and Biffignandi, 2012, p. 55; Buelens, Burger and van den Brakel, 2018, p. 322; Dever, Rafferty and Valliant, 2008, p. 47; Münnich and Zwick, 2016, p. 74). At the same time, these aspects tremendously differentiate many new and promising data sources from classical survey data obtained through probability sampling. Such differences concern size and structure, but also coverage of target populations and selectivity of the gathered data. As a consequence, new data sources often violate the well-established principles of probability sampling, casting doubts about quality and generalizability of the conclusions that are drawn from such data (cf. e.g. Buelens, Burger and van den Brakel, 2018, pp. 322 f; Pfeffermann, 2015, p. 430; Valliant and Dever, 2011, pp. 108 ff).

As the first important type of new data reinforcing the relevance and discussion of non-probability samples, Big Data is a term of vague and various definitions. A useful comprehensive overview is given by the National Research Council of the United States (2013). Descriptions and definitions of Big Data typically refer to the so-called three V's: *volume*, characterizing the huge amount of data, *velocity*, referring to the speed with which such data occurs and is collected, as well as *variety*, indicating the multitude of formats, structures and sources (cf. Laney, 2001). Often, further aspects are emphasized in form of additional V's, such as *variability* and *veracity*, referring to data quality, accuracy and trustworthiness (cf. Japec et al., 2015, pp. 841 f; Pfeffermann, 2015, pp. 427 ff). Volume, velocity and variety are challenging aspects with regard to data storage and processing up to estimation (cf. e.g. Govindaraju, Raghavan and Rao, 2015; Lynch, 2008; National Research Council of the United States, 2013). Various and ongoing efforts in hardware and algorithmic advancement, parallel computation as well as data compression and sub-sampling tackle these issues from computer and (partially) from statistical sciences (cf. e.g. Govindaraju, Raghavan and Rao, 2015; Lynch, 2008; National Research Council of the United States, 2013). Variability and veracity, on the other hand, are challenging with respect to generalizability, especially when it comes to assessing the selection process and accounting for its effects (cf. Japec et al., 2015, p. 849; National Research Council of the United States, 2013, p. 166). The sheer volume of information in Big Data does not ensure precision of estimates, especially when it comes to selection bias (cf. Baker et al., 2013a, p. 27; Buelens, Burger and van den Brakel, 2018, p. 327; Lazer et al., 2014). These are core topics that need to be addressed by statistical science if such data is to

be used as addition and (partial) replacement of traditional survey data (cf. Daas et al., 2015, p. 249; Meng, 2018; Pfeffermann, 2015, p. 430; Shlomo and Goldstein, 2015; United Nations, 2014).

Being that broadly defined, Big Data encompasses diverse data sources and has various statistical applications. For example, satellite images are used for estimating timber reserves (cf. Münnich et al., 2016) and indicators of poverty and social exclusion (cf. Jean et al., 2016). To estimate the latter on local levels, mobile phone data (cf. Blumenstock, Cadamuro and On, 2015) and GPS tracking information for vehicles (cf. Marchetti et al., 2015) are furthermore used. Social media (mainly Twitter) posts are evaluated to measure consumer confidence (cf. van den Brakel et al., 2017) and stock prices (cf. Ranco et al., 2015), quantify happiness (cf. Dodds et al., 2011) and well-being (cf. Luhmann, 2017), evaluate electoral campaigns (cf. Hong and Nadler, 2012), voting behavior (cf. Allcott and Gentzkow, 2017; Ceron et al., 2014) and epidemics (cf. Signorini, Segre and Polgreen, 2011). Search engine data is also used for modeling epidemics (cf. Ginsberg et al., 2009), as well as to predict private consumption (cf. Vosen and Schmidt, 2011) and unemployment (cf. Fondeur and Karamé, 2013; Xu et al., 2013), while online food service reviews may help to assess food safety (cf. Nsoesie, Kluberg and Brownstein, 2014).

Examples for estimators applied in this context cover averages and inequality indicators (cf. e.g. Dodds et al., 2011), but fitting prediction models to Big Data is far more common. The latter include e.g. linear regression (cf. Allcott and Gentzkow, 2017; Ginsberg et al., 2009), time series (cf. Hong and Nadler, 2012; Vosen and Schmidt, 2011) and linear mixed (small-area) models (cf. Marchetti et al., 2015) as well as support vector machines (cf. Signorini, Segre and Polgreen, 2011), matching, regression trees (cf. Buelens, Burger and van den Brakel, 2018) and artificial neural networks (cf. Xu et al., 2013). These selected examples are by no means comprehensive for all the topics, methods and estimates relevant in the context of Big Data, but merely demonstrate the increasing scientific relevance of such data (cf. Rodríguez-Mazahua et al., 2016, p. 3081). More detailed overviews are, for example, given by Baker et al. (2010), Govindaraju, Raghavan and Rao (2015), Japec et al. (2015) and Rodríguez-Mazahua et al. (2016).

The second important new data source challenging and supplementing classical probability samples are web surveys. These are somewhat closer to the classical data sources of survey users and may in certain cases simply constitute a new mode of interviewing. This is the case when the questionnaire in a probability sample is answered on the web (cf. e.g. Bianchi, Biffignandi and Lynn, 2017; Cole, 2005). In most cases, however, web surveys are not selected through probability sampling (cf. Baker et al., 2010, p. 3; Baker et al., 2013a, p. 34; Couper, 2000, pp. 477 ff) since some form of volunteerism and/or effects of unknown covariates are typically incorporated in the selection process (cf. Baker et al., 2013a, p. 34; Mercer et al., 2017, p. 251). The degree of self-selection can range from participation in access panels upon request when the recruitment group is a valid probability sample (cf. Enderle, Münnich and Bruch, 2013; Loosveldt and Sonck, 2008) up to fully relying on respondents that actively attempt to participate. The latter case occurs when a survey is publicly available on the web, such that anyone can respond – in principle even multiple times (cf. Bethlehem and Biffignandi, 2012, p. 422; Bethlehem, 2008b, p. 25; Steinmetz et al., 2014, p. 274). Since survey participation is typically influenced by motivational and cognitive characteristics of the respondents in relation to the survey's topic (cf. Baker et al., 2010, p. 38; Groves, Presser and Dipko, 2004), advertising may expand the range of such surveys, but does not necessarily reduce self-

selection issues. This is especially the case when advertisement channels are highly related to characteristics that influence participation, e.g. when publishing on websites where visits are related to the survey variable(s) of interest (cf. Faas and Schoen, 2006, p. 180). Nevertheless, many publications stress advantages of web surveys as an alternative and extension to traditional surveys, mostly referring to aspects of cost and practicability but also to the potential for reducing social desirability and interviewer effects (cf. e.g. Sanders et al., 2007, p. 258).

Web surveys are used in various areas. Examples include studies of wage distributions (cf. Tijdens and Steinmetz, 2016; Tijdens et al., 2014), associations between enterprises (cf. Biffignandi and Pratesi, 2000; 2002), road traffic (cf. Posawang et al., 2010) and market research (cf. Chiang, Zhang and Zhou, 2006; Roster et al., 2004). Further uses can be found in evaluating health-care (cf. Bethell et al., 2004) and end-of-life treatment preferences (cf. Feild et al., 2006) as well as studies on personality (cf. Buchanan and Smith, 1999) and psychopathology (cf. Kendler et al., 2009). Political attitudes and emotions are studied using online surveys (cf. Masch and Gabriel, 2020), just as voting decisions and turnout (cf. Faas and Schoen, 2006; Wang et al., 2015).

Similarly as for Big Data, statistical methods used in the context of web surveys often include models, mainly (generalized) linear regression (cf. e.g. Bethell et al., 2004; Hitchman et al., 2015; Yan and Tourangeau, 2008) and mixed models (cf. e.g. Ganesh et al., 2017; Gelman et al., 2016a; Wang et al., 2015), but artificial neural networks (cf. Chiang, Zhang and Zhou, 2006; Posawang et al., 2010) and regression trees (cf. Kern, Klausch and Kreuter, 2019, p. 81) are used as well. Nonetheless, estimators for means and proportions (cf. e.g. Barratt, Ferris and Lenton, 2015; Ryzin, 2008; Chiang, Zhang and Zhou, 2006) as well as correlations (cf. e.g. Faas and Schoen, 2006; McCabe, 2008) are slightly more present in this case than for Big Data. Again, these examples are not meant to be exhaustive, but illustrate the rising usage of web surveys (cf. Greenacre, 2016, p. 399) as well as the increasing number of scientific publications based thereon (cf. Lehdonvirta et al., 2020, p. 3). More detailed overviews and discussions of web surveys are, for example, given by Bethlehem and Biffignandi (2012) as well as Sue and Ritter (2012).

Big Data and web surveys constitute two major types of new data sources finding their way into statistical science, its applications and even official statistics (cf. Buelens, Burger and van den Brakel, 2018; Citro, 2014; Meng, 2018; Pfeffermann, 2015; Tam and Clarke, 2015). As both are inherently related to digitalization and information technology, joint efforts from statistical and computer science are required to establish reasonable and feasible ways for their utilization, particularly with regard to Big Data (cf. National Research Council of the United States, 2013, p. 4). Currently, scientific collaborations of both disciplines concerning Big Data are often focused on computability of prevalent prediction models. The considered aspects especially regard the computational challenges of volume, velocity and variety, somewhat neglecting the data quality aspects described by variability and veracity (cf. Buelens, Burger and van den Brakel, 2018, p. 326; National Research Council of the United States, 2013, pp. 93 ff; Daas et al., 2015, p. 249; Japec et al., 2015, p. 860). These core statistical issues coincide for the two emerging sources of information summarized above since these sources are often based on uncontrolled data generating processes and selection mechanisms. Such data quality aspects are more present in statistical publications about web surveys, where some focus is laid on the evaluation and compensation of unknown data generating processes (cf. e.g. Bethlehem

and Biffignandi, 2012; Chen, Valliant and Elliott, 2019; Elliott and Valliant, 2017; Lee, 2004; Pratesi et al., 2004). These prevalent topics shape much of the growing relevance and upcoming challenges of non-probability samples in general and new data sources in particular (cf. Japec et al., 2015, p. 863). To foster a more formal discussion of these issues in section 2.3, fundamentals of probability sampling and corresponding estimation methods are briefly summarized in the following section 2.2.

2.2 Overview of Probability Sampling and Design-based Estimation

To understand obstacles in using non-probability samples, it is important to refer to crucial characteristics and benefits of probability sampling that are invalid for other forms of sample selection. Because these specific advantages concern the construction of estimators that are unbiased and facilitate quantification of precision from a single sample (cf. Breidt and Opsomer, 2017, p. 191; Kalton, 1983, p. 90), fundamental aspects of sampling and design-based estimation are summarized in the current section 2.2. The given summary is focused on aspects that are relevant to establish the subsequent discussion, based on and in line with the works of Fuller (2009), Lohr (2010), Särndal, Swensson and Wretman (1992) and Wolter (2007). These are advisable references for a comprehensive overview of (probability) sampling, estimation and inference that goes beyond the scope of this thesis.

Design-based survey sampling and estimation considers a fixed finite population P of size $\mathrm{N} \in \mathbb{N}$, determined by a finite set of identifiers

$$\mathcal{S}^{\mathrm{P}} \coloneqq \{1, \ldots, \mathrm{N}\} \tag{2.1}$$

that each uniquely represent one unit (or element) of the population. In contrast to sampling *with replacement*, where the same element can be sampled multiple times, sampling *without replacement* allows each element to occur at most once in a sample. For different reasons, the without replacement scenario is typically the predominant case in survey sampling. First of all, duplication of elements does not add any new information when sampling from a finite population. Furthermore, uniqueness of elements facilitates a set-theoretical representation of sampling and estimation in correspondence to definition 2.1. Last but not least, the theoretical framework of without replacement sampling even allows obtaining estimators for the with replacement scenario (cf. Cochran, 1977, p. 30; Fuller, 2009, p. 26; Särndal, Swensson and Wretman, 1992, pp. 48 ff, 110 ff; Wolter, 2007, p. 32). Therefore, sampling without replacement is considered exclusively throughout this thesis. In this case, a general sample denoted by s is defined by a subset $\mathcal{S}^{\mathsf{s}} \subseteq \mathcal{S}^{\mathrm{P}}$ of size $n^{\mathsf{s}} \coloneqq |\mathcal{S}^{\mathsf{s}}| \leq \mathrm{N}$. Each unit's index $i = 1, \ldots, \mathrm{N}$ can occur in $\mathcal{S}^{\mathsf{s}}$ at most once. Consequently, the sample can be described by a vector of inclusion indicators $\boldsymbol{r}^{\mathsf{s}} = [r_1^{\mathsf{s}}, \ldots, r_{\mathrm{N}}^{\mathsf{s}}]^{\mathsf{T}} \in \{0,1\}^{\mathrm{N}}$ with elements

$$r_i^{\mathsf{s}} \coloneqq \mathbb{I}\left(i \in \mathcal{S}^{\mathsf{s}}\right) = \begin{cases} 1\,, & \text{if} \quad i \in \mathcal{S}^{\mathsf{s}} \\ 0 & \text{else} \end{cases} \tag{2.2}$$

for all $i \in \mathcal{S}^{\mathrm{P}}$, where $\mathbb{I}$ is the indicator function. The set of all 2^{N} possible samples from the population (or values of $\boldsymbol{r}^{\mathsf{s}}$) is denoted by $\mathbb{S} \coloneqq \left\{\mathcal{S}^{\mathsf{s}} : \mathcal{S}^{\mathsf{s}} \subseteq \mathcal{S}^{\mathrm{P}}\right\}$. A sampling design is defined as a function $\mathrm{D} : \mathbb{S} \to [0,1]$, assigning each possible sample a probability of

being drawn, with $\sum_{\mathcal{S}^{\mathsf{s}} \in \mathbb{S}} \mathrm{D}(\mathcal{S}^{\mathsf{s}}) = 1$. In a fixed and finite population setting, D is the only probabilistic component in sampling (cf. Cochran, 1977; Fuller, 2009, p. 3; Lohr, 2010, pp. 28 ff; Särndal, Swensson and Wretman, 1992, p. 27). An overview of possible computational implementations for sampling designs is provided by Tillé (2006).

With $\boldsymbol{r}^{\mathsf{s}}$ uniquely defining the sample, its stochastic behavior is the key element to describe and account for randomness in this framework. In compliance with $\mathrm{D}(\mathcal{S}^{\mathsf{s}})$ being the probability of a sample $\mathcal{S}^{\mathsf{s}}$ being drawn, the chance of element i being sampled under a given design is

$$\pi_i^{\mathsf{s}} \coloneqq \mathrm{E}(r_i^{\mathsf{s}}) = \mathrm{P}(i \in \mathcal{S}^{\mathsf{s}}) = \sum_{\mathcal{S}^{\mathsf{s}} \in \mathbb{S}} \mathrm{D}(\mathcal{S}^{\mathsf{s}}) \cdot \mathbb{I}(i \in \mathcal{S}^{\mathsf{s}}) \tag{2.3}$$

and called its *inclusion probability*. A *probability sampling design* is coherently defined by two properties of D:

a) The sampling design is known, such that each sample has an identifiable probability of being selected, and

b) all elements in the population have a chance of being sampled, such that $\pi_i^{\mathsf{s}} > 0$ for all $i \in \mathcal{S}^{\mathrm{P}}$, which implies full coverage of the target population $\mathcal{S}^{\mathrm{P}}$.

The terms probability sampling and random sampling are typically used synonymously (cf. also chapter 3), and *simple random sampling* is defined by $\pi_i^{\mathsf{s}} = n^{\mathsf{s}}/\mathrm{N}$ for all $i \in \mathcal{S}^{\mathrm{P}}$. In analogy to equation 2.3, *second order inclusion probabilities* indicate the probability that two elements i and j are simultaneously part of the sample. They are defined as

$$\pi_{ij}^{\mathsf{s}} \coloneqq \mathrm{E}\left(r_i^{\mathsf{s}} \cdot r_j^{\mathsf{s}}\right) = \mathrm{P}(i \in \mathcal{S}^{\mathsf{s}} \wedge j \in \mathcal{S}^{\mathsf{s}}) = \sum_{\mathcal{S}^{\mathsf{s}} \in \mathbb{S}} \mathrm{D}(\mathcal{S}^{\mathsf{s}}) \cdot \mathbb{I}(i \in \mathcal{S}^{\mathsf{s}}) \cdot \mathbb{I}(j \in \mathcal{S}^{\mathsf{s}}) \quad , \tag{2.4}$$

where $\pi_{ii}^{\mathsf{s}} = \pi_i^{\mathsf{s}}$ and $\pi_{ij}^{\mathsf{s}} \leq \mathrm{Min}\left(\pi_i^{\mathsf{s}}; \pi_j^{\mathsf{s}}\right)$ (cf. Fuller, 2009, p. 2; Kish, 1965, p. 20; Lohr, 2010, pp. 28 ff; Särndal, Swensson and Wretman, 1992, p. 32).

The main purpose of sampling is to perform estimation based on the collected data (cf. Kish, 1965, p. 8; Särndal, Swensson and Wretman, 1992, p. 3). A sample survey is conducted to obtain information about a general unknown population (or sub-population) statistic of interest, denoted by $\boldsymbol{\vartheta} \in \mathbb{R}^{h \times u}$ for given $h, u \in \mathbb{N}$. This statistic is defined with regard to some *target variables* $\boldsymbol{Y}$. These represent a vector of o characteristics associated with each element i in the population $\mathcal{S}^{\mathrm{P}}$, which is denoted by $\boldsymbol{y}_{i\cdot} = \begin{bmatrix} y_{i1} & \dots & y_{io} \end{bmatrix} \in \mathbb{R}^{1 \times o}$. The combination of these variables for all N elements in the population constitute a matrix

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & \dots & y_{1o} \\ \vdots & \ddots & \vdots \\ y_{\mathrm{N}1} & \dots & y_{\mathrm{N}o} \end{bmatrix} \in \mathbb{R}^{\mathrm{N} \times o} \tag{2.5}$$

(cf. Wolter, 2007, p. 8). Since the population is not fully observed, $\boldsymbol{Y}$ and hence $\boldsymbol{\vartheta}$ are typically unknown. Therefore, a sample consisting of units $i \in \mathcal{S}^{\mathsf{s}}$ is drawn, for which variables $\boldsymbol{y}_{i\cdot}$ are measured. Consequently, a sub-matrix of $\boldsymbol{Y}$ is observed, which is denoted by

$$\boldsymbol{Y}^{\mathsf{s}} \coloneqq \boldsymbol{Y}_{\mathcal{S}^{\mathsf{s}}\cdot} \in \mathbb{R}^{n^{\mathsf{s}} \times o} \quad , \tag{2.6}$$

where $\boldsymbol{Y}_{\mathcal{S}^{\mathsf{s}}\cdot}$ denotes the rows of $\boldsymbol{Y}$ indexed by $\mathcal{S}^{\mathsf{s}}$. Estimation is required because the remaining part of $\boldsymbol{Y}$ is considered unknown. To that end, an *estimator* for $\boldsymbol{\vartheta}$ is defined

as a function $\widehat{\boldsymbol{\vartheta}} : \mathbb{S} \to \mathbb{R}^{h \times u}$ to obtain an *estimate* $\widehat{\boldsymbol{\vartheta}}\left(\mathcal{S}^{\mathsf{s}}\right)$ for an actual sample. Estimators may use $\boldsymbol{Y}^{\mathsf{s}}$ alone, i.e. exclusively rely on information from the sample. Since $\boldsymbol{Y}^{\mathsf{s}}$ is a matrix (cf. equation 2.5), this definition of $\widehat{\boldsymbol{\vartheta}}$ readily includes multivariate statistics, such as ratios, covariances, correlations and regression coefficients (cf. Särndal, Swensson and Wretman, 1992, pp. 38 ff, 176 ff). Yet, additional variables are frequently used in $\widehat{\boldsymbol{\vartheta}}$, which provide information outside the sample and therefore differ from $\boldsymbol{Y}$. They are referred to as *auxiliary variables* and denoted by

$$\boldsymbol{X} \;=\; \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{\mathrm{N}1} & \dots & x_{\mathrm{N}p} \end{bmatrix} \;\in\; \mathbb{R}^{\mathrm{N} \times p} \qquad , \tag{2.7}$$

with corresponding observations $\boldsymbol{X}^{\mathsf{s}} \coloneqq \boldsymbol{X}_{\mathcal{S}^{\mathsf{s}}\cdot}$ for the sample in accordance with equation 2.6. The crucial difference to the target variables is that $\boldsymbol{X}^{\mathsf{s}}$ is observed through the sample and some additional information external to s is available about $\boldsymbol{X}$. Such information about auxiliary variables may refer to the whole population (e.g. from a register) or only a subset of it (e.g. some other sample), on aggregated or individual level (cf. Breidt and Opsomer, 2017, p. 192; cf. Burgard, Münnich and Rupp, 2019, p. 1). Some examples for sources and types of such variables are given in section 3.2. The availability of external information makes such auxiliary variables valuable, e.g. for prediction and calibration techniques described in chapter 5 (cf. Cochran, 1977, pp. 189 ff; Fuller, 2009, p. 5; Särndal, Swensson and Wretman, 1992, pp. 33 ff, 219 ff). More details regarding the underlying general theory and definition of estimators are given by Dekking et al. (2005, pp. 42 ff) and Witting (1985, pp. 17 ff).

In addition, complex probability sampling designs commonly make use of *design variables*, denoted by

$$\boldsymbol{Z} \;=\; \begin{bmatrix} z_{11} & \dots & z_{1q} \\ \vdots & \ddots & \vdots \\ z_{\mathrm{N}1} & \dots & z_{\mathrm{N}q} \end{bmatrix} \;\in\; \mathbb{R}^{\mathrm{N} \times q} \tag{2.8}$$

and $\boldsymbol{Z}^{\mathsf{s}} \coloneqq \boldsymbol{Z}_{\mathcal{S}^{\mathsf{s}}\cdot}$ as before. These variables determine the sample selection. They have direct impact on the inclusion probabilities and have to be considered when constructing an estimator and determining its quality. Common examples for such variables being used in complex designs include identifiers used for stratification or clustering. In these cases, units of the population are grouped depending on $\boldsymbol{Z}$ and elements from all (stratification) or a subset of groups (clustering) are selected into the sample. Furthermore, size variables used for sampling with unequal probabilities can be part of $\boldsymbol{Z}$ (cf. Fuller, 2009, pp. 18 ff, 72 ff; Lohr, 2010, pp. 73 ff, 165 ff; Särndal, Swensson and Wretman, 1992, pp. 61 ff, 124 ff). Note that there can be some overlap between the sets of variables $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$, e.g. when design variables are used as auxiliaries or when estimating correlations between target and auxiliary variables. Nevertheless, differentiating these three types of variables as above simplifies notation throughout the following chapters.

With the sampling design being the only random component in a finite population setting, the expected value of $\widehat{\boldsymbol{\vartheta}}$ is given by

$$\mathbf{E}\left(\widehat{\boldsymbol{\vartheta}}\right) \;=\; \sum_{\mathcal{S}^{\mathsf{s}} \in \mathbb{S}} \mathrm{D}\left(\mathcal{S}^{\mathsf{s}}\right) \cdot \widehat{\boldsymbol{\vartheta}}(\mathcal{S}^{\mathsf{s}}) \qquad , \tag{2.9}$$

where $\widehat{\boldsymbol{\vartheta}}(\mathcal{S}^{\mathsf{s}})$ is the specific value of $\widehat{\boldsymbol{\vartheta}}$ for sample $\mathcal{S}^{\mathsf{s}}$. Important measures for the quality of an estimator concern its accuracy in terms of the variance

$$\mathbf{V}\left(\widehat{\boldsymbol{\vartheta}}\right) \;=\; \sum_{\mathcal{S}^{\mathsf{s}}\in\mathbb{S}} \mathrm{D}\left(\mathcal{S}^{\mathsf{s}}\right)\cdot\left(\widehat{\boldsymbol{\vartheta}}(\mathcal{S}^{\mathsf{s}}) - \mathbf{E}\left(\widehat{\boldsymbol{\vartheta}}\right)\right)^2 = \mathbf{E}\left(\widehat{\boldsymbol{\vartheta}}^2\right) - \left(\mathbf{E}\left(\widehat{\boldsymbol{\vartheta}}\right)\right)^2 \quad , \tag{2.10}$$

the bias

$$\mathbf{Bias}\left(\widehat{\boldsymbol{\vartheta}}\right) \;=\; \mathbf{E}\left(\widehat{\boldsymbol{\vartheta}}\right) - \boldsymbol{\vartheta} \tag{2.11}$$

and the mean squared error (MSE)

$$\mathbf{MSE}\left(\widehat{\boldsymbol{\vartheta}}\right) \;=\; \mathbf{E}\left(\left(\widehat{\boldsymbol{\vartheta}}(\mathcal{S}^{\mathsf{s}}) - \boldsymbol{\vartheta}\right)^2\right) = \mathbf{V}\left(\widehat{\boldsymbol{\vartheta}}\right) + \left(\mathbf{Bias}\left(\widehat{\boldsymbol{\vartheta}}\right)\right)^2 \quad . \tag{2.12}$$

An estimator $\widehat{\boldsymbol{\vartheta}}$ is *unbiased* if $\mathbf{Bias}\left(\widehat{\boldsymbol{\vartheta}}\right) = \mathbf{0}_{h\times u}$, in which case $\mathbf{MSE}\left(\widehat{\boldsymbol{\vartheta}}\right) = \mathrm{V}\left(\widehat{\boldsymbol{\vartheta}}\right)$ holds (cf. Särndal, Swensson and Wretman, 1992, pp. 34 ff; Wolter, 2007, p. 9).

Design weights are an important concept to construct unbiased estimators for probability samples. The design weight for element i in sample s is defined as the inverse of its inclusion probability, denoted by

$$w_i^{\mathsf{s}} \;\coloneqq\; \frac{1}{\pi_i^{\mathsf{s}}} \;\in \mathbb{R}_{>0} \qquad \text{for all} \quad i \in \mathcal{S}^{\mathsf{s}} \qquad , \tag{2.13}$$

such that $\boldsymbol{w}^{\mathsf{s}} = \begin{bmatrix} w_1^{\mathsf{s}} & \dots & w_{n^{\mathsf{s}}}^{\mathsf{s}} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}_{>0}^{n^{\mathsf{s}}}$ is the vector of design weights for the sample. This definition is central for estimation from probability samples, where inclusion probabilities are identifiable at least for the sampled units and strictly positive for the whole population (cf. Särndal, Swensson and Wretman, 1992, p. 32). For example, consider the case where the population statistic of interest is the total of $\boldsymbol{Y}$,

$$\boldsymbol{\tau}_{\boldsymbol{Y}} \;\coloneqq\; \sum_{i\in\mathcal{S}^{\mathrm{P}}} \boldsymbol{y}_{i\cdot} = \mathbf{1}_{\mathrm{N}\times 1}^{\mathsf{T}}\boldsymbol{Y} \qquad . \tag{2.14}$$

The corresponding estimator for sample s is defined by

$$\widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right) \;\coloneqq\; \sum_{i\in\mathcal{S}^{\mathsf{s}}} w_i^{\mathsf{s}}\cdot\boldsymbol{y}_{i\cdot} = \left(\boldsymbol{w}^{\mathsf{s}}\right)^{\mathsf{T}}\boldsymbol{Y}^{\mathsf{s}} \qquad . \tag{2.15}$$

Horvitz and Thompson (1952) introduced this estimator, which is therefore called the *Horvitz-Thompson estimator* (HT-estimator). If s is a probability sample, such that $\pi_i^{\mathsf{s}} > 0$ for all $i \in \mathcal{S}^{\mathrm{P}}$, the estimator is unbiased as a consequence of definitions 2.2, 2.3 and 2.13:

$$\mathrm{E}\left(\sum_{i\in\mathcal{S}^{\mathsf{s}}} w_i^{\mathsf{s}}\cdot\boldsymbol{y}_{i\cdot}\right) \;=\; \mathrm{E}\left(\sum_{i\in\mathcal{S}^{\mathrm{P}}} r_i^{\mathsf{s}}\cdot w_i^{\mathsf{s}}\cdot\boldsymbol{y}_{i\cdot}\right) = \sum_{i\in\mathcal{S}^{\mathrm{P}}} \frac{\mathrm{E}\left(r_i^{\mathsf{s}}\right)}{\mathrm{E}\left(r_i^{\mathsf{s}}\right)}\cdot\boldsymbol{y}_{i\cdot} \;=\; \sum_{i\in\mathcal{S}^{\mathrm{P}}} \boldsymbol{y}_{i\cdot} \tag{2.16}$$

(cf. Fuller, 2009, pp. 6 ff; Särndal, Swensson and Wretman, 1992, pp. 42 ff). A more general but less compact and intuitive motivation of design weights is given by Pfeffermann and Sverchkov (1999, p. 185; cf. appendix B.2).

Further statistics that are relevant in the context of this thesis include means

$$\boldsymbol{\mu}_{\boldsymbol{Y}} \coloneqq \mathrm{N}^{-1} \cdot \boldsymbol{\tau}_{\boldsymbol{Y}} \quad \in \quad \mathbb{R}^{o} \qquad , \tag{2.17a}$$

covariances

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}} \coloneqq \mathrm{N}^{-1} \cdot \left(\boldsymbol{Y}^{\mathsf{T}} \boldsymbol{Y}\right) - \boldsymbol{\mu}_{\boldsymbol{Y}}^{\mathsf{T}} \, \boldsymbol{\mu}_{\boldsymbol{Y}} \quad \in \quad \mathbb{R}^{o \times o} \tag{2.17b}$$

and correlations

$$\boldsymbol{\rho}_{\boldsymbol{Y}} \coloneqq \boldsymbol{\Sigma}_{\boldsymbol{Y}} \oslash \left(\mathbf{diag}\left(\boldsymbol{\Sigma}_{\boldsymbol{Y}}\right) \left(\mathbf{diag}\left(\boldsymbol{\Sigma}_{\boldsymbol{Y}}\right)\right)^{\mathsf{T}}\right)^{\circ \frac{1}{2}} \quad \in \quad \mathbb{R}^{o \times o} \qquad . \tag{2.17c}$$

Using $\boldsymbol{c} \in \{0,1\}^{\mathrm{N}}$ with elements $c_j \coloneqq \mathbb{I}\left(\left(y_{j1} \leq y_{i1}\right) \wedge \cdots \wedge \left(y_{jo} \leq y_{io}\right)\right)$, a distribution function $\mathrm{F}_{\boldsymbol{Y}} : \mathbb{R}^{1 \times o} \to [0,1]$ can furthermore be written as

$$\mathrm{F}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot}\right) \coloneqq \mathrm{P}\left(\boldsymbol{c} = 1\right) = \boldsymbol{\mu}_{\boldsymbol{c}} \tag{2.17d}$$

(cf. Galassi et al., 2009, pp. 263 ff; Särndal, Swensson and Wretman, 1992, pp. 181 ff). Here and in general, $\circ$, $\oslash$ and $^{\circ}$ respectively denote Hadamard (element-wise) product, division and power (cf. Reams, 1999).

It is convenient that the population statistics defined in equations 2.17 can be written as functions of totals. Therefore, corresponding design-based estimators are derived as functions of estimated totals (cf. Breidt and Opsomer, 2017, p. 190; Opsomer, 2009, p. 7). Considering that $\mathrm{N} = \sum_{i \in \mathcal{S}^{\mathrm{P}}} 1$, an estimate of the population size N is given by

$$\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right) \coloneqq \sum_{i \in \mathcal{S}^{\mathsf{s}}} w_i^{\mathsf{s}} = \left(\boldsymbol{w}^{\mathsf{s}}\right)^{\mathsf{T}} \mathbf{1}_{n^{\mathsf{s}} \times 1} = \left|\left|\boldsymbol{w}^{\mathsf{s}}\right|\right|_1 \qquad , \tag{2.18a}$$

resulting in

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right) \coloneqq \left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)^{-1} \cdot \widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right) \tag{2.18b}$$

as estimator for the population mean $\boldsymbol{\mu}_{\boldsymbol{Y}}$. In the same manner, the estimators

$$\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right) \coloneqq \left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)^{-1} \cdot \left(\left(\boldsymbol{Y}^{\mathsf{s}}\right)^{\mathsf{T}} \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{s}}\right) \boldsymbol{Y}^{\mathsf{s}}\right) - \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)^{\mathsf{T}} \widehat{\boldsymbol{\mu}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right) \tag{2.18c}$$

for covariances $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ and

$$\widehat{\boldsymbol{\rho}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right) \coloneqq \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right) \oslash \left(\mathbf{diag}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right) \left(\mathbf{diag}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)\right)^{\mathsf{T}}\right)^{\circ \frac{1}{2}} \tag{2.18d}$$

for correlations $\boldsymbol{\rho}_{\boldsymbol{Y}}$ are determined by successively plugging in these estimates. A corresponding estimate $\widehat{\mathrm{F}}_{\boldsymbol{Y}} : \mathbb{R}^{1 \times o} \times \mathbb{R}^{n^{\mathsf{s}}} \to [0,1]$ of the distribution function is given by

$$\widehat{\mathrm{F}}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot}, \boldsymbol{w}^{\mathsf{s}}\right) \coloneqq \widehat{\boldsymbol{\mu}}_{\boldsymbol{c}}\left(\boldsymbol{w}^{\mathsf{s}}\right) \qquad , \tag{2.18e}$$

with $\boldsymbol{c}$ defined as for equation 2.17d (cf. Galassi et al., 2009, pp. 263 ff; Särndal, Swensson and Wretman, 1992, pp. 186 ff).

It can be shown that the maximum likelihood (ML) estimator $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})$ is biased for $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ even under probability sampling where the other presented estimators are unbiased (cf. appendix B.1; Dekking et al., 2005, pp. 292 f). This bias is caused by the fact that definition 2.18c does not account for the covariance of mean estimates, denoted by $\boldsymbol{\Sigma}_{\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})\right)}$:

$$\mathbf{Bias}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})\right) = -\boldsymbol{\Sigma}_{\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})\right)} \approx (1-\nu(\boldsymbol{w}^{\mathsf{s}}))\cdot\boldsymbol{\Sigma}_{\boldsymbol{Y}} \quad , \tag{2.18f}$$

where

$$\nu(\boldsymbol{w}^{\mathsf{s}}) \coloneqq \widehat{\mathrm{N}}\left((\boldsymbol{w}^{\mathsf{s}})^{\circ 2}\right) \Big/ \left(\widehat{\mathrm{N}}(\boldsymbol{w}^{\mathsf{s}})\right)^2 \quad . \tag{2.18g}$$

Approximation 2.18f is commonly used for bias correction since it does not depend on the second order inclusion probabilities defined in equation 2.4. The resulting corrected covariance estimator is

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}}) \coloneqq \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})\cdot(1-\nu(\boldsymbol{w}^{\mathsf{s}}))^{-1} \quad . \tag{2.18h}$$

In case of constant weights, this adjustment reduces to Bessel's correction (using $n-1$ instead of n in the denominator of the sample variance) and is therefore exact. Estimation of correlations (equation 2.18d) does not require adjustment because numerator and denominator are multiplied by the same factor. As an alternative but hardly ever used way to adjust for the bias in $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})$, the estimates for $\boldsymbol{\Sigma}_{\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})\right)}$ described below may be used for correction (cf. e.g. Galassi et al., 2009, p. 266; Lumley, 2004; R Core Team, 2018; Särndal, Swensson and Wretman, 1992, pp. 186 f).

For a probability sampling design D, first and second order inclusion probabilities are known. If both are strictly positive for samples coming from such a design, i.e. $\pi^{\mathsf{s}}_{ij} > 0$ for all $i,j \in \mathcal{S}^{\mathrm{P}}$ is fulfilled additionally to the positivity of π^{s}_{i} required for probability sampling, D is called *measurable*. In this case, the quality of an estimator with respect to (w.r.t.) the sampling design can be quantified, and valid inference (e.g. calculation of variance estimates, confidence intervals and statistical tests) based on a sample can be carried out (cf. Fuller, 2009, p. 11; Särndal, Swensson and Wretman, 1992, p. 33). In case of a probability design with only a single stage of random selection, the variance of the total estimator (cf. definition 2.15) due to linearity can be written as

$$\mathbf{V}\left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})\right) = \boldsymbol{\Sigma}_{\left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})\right)} = \sum_{i\in\mathcal{S}^{\mathrm{P}}}\sum_{j\in\mathcal{S}^{\mathrm{P}}}\left(\frac{\pi^{\mathsf{s}}_{ij}}{\pi^{\mathsf{s}}_{i}\cdot\pi^{\mathsf{s}}_{j}}-1\right)\cdot\boldsymbol{y}_{i\cdot}\boldsymbol{y}_{j\cdot}^{\mathsf{T}} \quad . \tag{2.19}$$

Horvitz and Thompson (1952, p. 670) derive an unbiased estimator for 2.19, which is given by

$$\widehat{\mathbf{V}}_{HT}\left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})\right) \coloneqq \sum_{i\in\mathcal{S}^{\mathsf{s}}}\sum_{j\in\mathcal{S}^{\mathsf{s}}}\frac{1}{\pi^{\mathsf{s}}_{ij}}\cdot\left(\frac{\pi^{\mathsf{s}}_{ij}}{\pi^{\mathsf{s}}_{i}\cdot\pi^{\mathsf{s}}_{j}}-1\right)\cdot\boldsymbol{y}_{i\cdot}\boldsymbol{y}_{j\cdot}^{\mathsf{T}} \quad . \tag{2.20a}$$

A more stable version for sampling designs with fixed sample sizes is proposed by Sen (1953, cited in Särndal, Swensson and Wretman, 1992, p. 54) as well as Yates and Grundy (1953, p. 257), which is defined by

$$\widehat{\mathbf{V}}_{SYG}\left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}}(\boldsymbol{w}^{\mathsf{s}})\right) \coloneqq \sum_{i\in\mathcal{S}^{\mathsf{s}}}\sum_{j\in\mathcal{S}^{\mathsf{s}}}\left(\frac{\pi^{\mathsf{s}}_{i}\cdot\pi^{\mathsf{s}}_{j}}{\pi^{\mathsf{s}}_{ij}}-1\right)\cdot\left(\frac{\boldsymbol{y}_{i\cdot}}{\pi^{\mathsf{s}}_{i}}-\frac{\boldsymbol{y}_{j\cdot}}{\pi^{\mathsf{s}}_{j}}\right)\left(\frac{\boldsymbol{y}_{i\cdot}}{\pi^{\mathsf{s}}_{i}}-\frac{\boldsymbol{y}_{j\cdot}}{\pi^{\mathsf{s}}_{j}}\right)^{\mathsf{T}} \quad . \tag{2.20b}$$

For complex survey designs with multiple stages of selection, the variance and its estimate are determined by recursive application of equations 2.19 and 2.20 for each stage while adding the expected value of variances from subsequent stages. This is, for example, relevant when drawing groups (clusters) of elements in a first step, from which a subset of elements are drawn in one or more subsequent steps (cf. Bruch, Münnich and Zins, 2011, p. 3; Cochran, 1977, pp. 274 ff; Durbin, 1953, p. 263). Non-linear estimators, like those defined in equations 2.18b to 2.18h, usually require some sort of approximation to estimate their variance (cf. Cochran, 1977, pp. 318 ff; Kovar, Rao and Wu, 1988, p. 25). For example, Taylor linearization can be used to get an approximate variance estimator in the form of equation 2.20 (cf. Cochran, 1977, p. 319; Münnich and Zins, 2011; Särndal, Swensson and Wretman, 1992, pp. 172 ff; Wolter, 2007, pp. 226 ff).

As an alternative to linearization techniques for variance estimation, resampling methods can be used for non-linear estimators and do work for linear ones as well. In the general design-based context, these methods attempt to model the repeated sample distribution of any statistic $\widehat{\boldsymbol{\vartheta}}$ by iterative draws from an artificial population, which often is the sample s itself. Being an aspect of the distribution of $\widehat{\boldsymbol{\vartheta}}$, the variance (equation 2.10) or in certain cases the MSE (equation 2.12) are then estimated from the approximated distribution of $\widehat{\boldsymbol{\vartheta}}$. Assuming that the artificial population is defined by the sample identifiers $\mathcal{S}^{\mathsf{s}}$, Chipperfield and Preston (2007) describe various resampling methods by the following algorithm 1:

Algorithm 1: General resampling algorithm

1: **Input:** $\mathcal{S}^{\mathsf{s}} \in \mathbb{S}$; $\widehat{\boldsymbol{\vartheta}} : \mathbb{S} \to \mathbb{R}^{h \times u}$; $\boldsymbol{b} : \mathbb{N}^{n^{\mathsf{s}}} \to \mathbb{R}^{n^{\mathsf{s}}}$; $\tilde{n}^{\mathsf{s}}, a \in \mathbb{N}$

2: **for** $j = 1, \ldots, a$ **do**

3: Draw a sub-sample of size $\tilde{n}^{\mathsf{s}}$ from $\mathcal{S}^{\mathsf{s}}$ with a corresponding vector $\boldsymbol{c} \in \mathbb{N}^{n^{\mathsf{s}}}$, where c_i indicates the number of times that element i occurs in the current sub-sample for all $i = 1, \ldots, n^{\mathsf{s}}$

4: Calculate weights

$$\boldsymbol{w}^{\mathrm{bt}(j)} := \boldsymbol{w}^{\mathsf{s}} \circ \boldsymbol{b}\left(\boldsymbol{c}\right) \quad , \tag{2.21}$$

depending on $\boldsymbol{c}$ by some prespecified function $\boldsymbol{b}$

5: Calculate $\widehat{\boldsymbol{\vartheta}}^{(j)}$ using weights $\boldsymbol{w}^{\mathrm{bt}(j)}$

6: **end for**

7: **Return:**

$$\widehat{\mathbf{V}}\left(\widehat{\boldsymbol{\vartheta}}\right) := \frac{1}{a-1} \sum_{j=1}^{a} \left(\widehat{\boldsymbol{\vartheta}}^{(j)} - \left(\frac{1}{a} \sum_{k=1}^{a} \widehat{\boldsymbol{\vartheta}}^{(k)} \right) \right)^2 \quad \in \quad \mathbb{R}^{h \times u} \tag{2.22}$$

Depending on the choice of the sub-samples and corresponding weights, this general form includes jackknife, bootstrap and balanced repeated replication methods (cf. Chipperfield and Preston, 2007, p. 168). Note that other characteristics of the distribution of $\widehat{\boldsymbol{\vartheta}}$ can be estimated via resampling as well. This is achieved by replacing equation 2.22 with the corresponding estimate, e.g. with quantiles to directly determine a confidence interval. In some of these cases, additional assumptions are required, for example when estimating the MSE (cf. Efron, 1981; Rao and Wu, 1988, p. 232).

Bootstrapping introduced by Efron (1979) is a common choice when it comes to resampling methods because it can be applied for smooth as well as non-smooth estimators under general sampling designs (cf. Kovar, Rao and Wu, 1988; Lohr, 2010, p. 386; Rao and Wu, 1988; Wolter, 2007). Therefore, two variants of the bootstrap are considered in the following: as the original form, the *Monte Carlo bootstrap* is based on simple random sampling with replacement and subsamples of size $\tilde{n}^{\mathsf{s}} = n^{\mathsf{s}}$. The weights are simply multiplied by the number of times an element is drawn, such that $\boldsymbol{b}(\boldsymbol{c}) = \boldsymbol{c}$. An obvious limitation in the present context is the with replacement strategy, resulting in biased variance estimates for non-negligible sampling fractions $fr^{\mathsf{s}} \coloneqq n^{\mathsf{s}}/\mathrm{N}$ (cf. Wolter, 2007, p. 200). In addition, simple random sampling is a rarely realistic approximation when it comes to a general sampling design D that can include one or multiple stages of stratification, clustering and/or unequal probability sampling (cf. Chipperfield and Preston, 2007, p. 167; Rao and Wu, 1988). The *rescaling bootstrap*, originally proposed by Rao and Wu (1988) and modified by Rao, Wu and Yue (1992) as well as Chipperfield and Preston (2007) and Preston (2009), tackles these issues and allows for sampling without replacement for complex multi-stage designs (cf. Preston, 2009). For a single-stage sample, $\boldsymbol{b}$ is defined by

$$\boldsymbol{b}(\boldsymbol{c}) \; = 1 + \sqrt{\left(1 - \frac{n^{\mathsf{s}}}{\mathrm{N}}\right) \cdot \frac{\tilde{n}^{\mathsf{s}}}{n^{\mathsf{s}} - \tilde{n}^{\mathsf{s}}}} \cdot \left(\frac{n^{\mathsf{s}}}{\tilde{n}^{\mathsf{s}}} \cdot \boldsymbol{c} - \mathbf{1}_{n^{\mathsf{s}} \times 1}\right) \qquad , \tag{2.23}$$

with straightforward extensions to multiple stages (cf. Chipperfield and Preston, 2007; Preston, 2009). Different other strategies for dealing with the drawbacks of the original Monte Carlo bootstrap exist, as well as further alternatives to iteratively estimate the variances of general estimators $\widehat{\boldsymbol{\vartheta}}$. An overview is given by Bruch, Münnich and Zins (2011), Efron and Tibshirani (1998) or Wolter (2007).

The above summary of probability sampling and design-based estimation serves as a foundation for methods to assess and deal with non-probability samples (cf. chapters 3 and 5). As a first step to that end, it facilitates a discussion of challenges arising from non-probability samples in the following section 2.3.

2.3 Challenges in Dealing with Non-probability Samples

In section 2.2, important concepts of sampling are highlighted. In principle, general samples are considered, for which central advantages hold if they are selected through a (measurable) probability sampling design. For example, these advantages allow for unbiased estimates and valid inference. To contrast non-probability with probability sampling, a general probability sample is now denoted by ps. A design generating such a sample by definition provides known probabilities for each possible sample $\mathcal{S}^{\mathsf{ps}}$, implying that the inclusion probabilities π_i^{ps} are known at least for the sampled units. Furthermore, it ensures full coverage of the target population, such that $\pi_i^{\mathsf{ps}} > 0$ for all $i \in \mathcal{S}^{\mathrm{P}}$ (cf. Särndal, Swensson and Wretman, 1992, p. 32).

Non-probability sampling is generally characterized in differentiation to probability sampling. A *non-probability sample*, denoted by nps and identified by a set of indices $\mathcal{S}^{\mathsf{nps}} \subseteq \mathcal{S}^{\mathrm{P}}$ is therefore defined by

a) unknown selection probabilities for at least some possible samples $\mathcal{S} \in \mathbb{S}$, resulting in unidentifiable inclusion probabilities π_i^{nps} and/or

b) incomplete coverage of the target population, resulting in $\pi_i^{\mathsf{nps}} = 0$ for some $i \in \mathcal{S}^{\mathrm{P}}$.

Both conditions are directly related to properties of the vector of inclusion probabilities $\boldsymbol{\pi}^{\mathsf{nps}} \in [0,1]^{\mathrm{N}}$ (cf. Biffignandi and Pratesi, 2003; Buelens, Burger and van den Brakel, 2018, p. 326; Jacoby and Handlin, 1991, p. 170; Lohr, 2010, p. 5; Pfeffermann, 2015, p. 431; Valliant and Dever, 2011, p. 108) and may in principle be tackled by modifying sample selection towards probability sampling.

However, non-probability samples are usually used in situations where adequate data obtained from a probability design is neither available nor collectable, e.g. for measuring real-time dynamics in unemployment (cf. Fondeur and Karamé, 2013) or road traffic (cf. Buelens, Burger and van den Brakel, 2018, pp. 327 f). In this regard, 'adequate' may be defined with respect to required precision, cost or time constraints (cf. Buelens, Burger and van den Brakel, 2018, p. 322), especially when it comes to small target populations and/or sensitive information (cf. Barendregt, Van der Poel and Van de Mheen, 2005, p. 124; Feild et al., 2006, p. 566). At the same time, non-probability samples often result from processes not primarily aiming at collecting this particular data (cf. section 2.1). Therefore, adjustments in the selection of a non-probability sample are highly dependent on the specific data context and usually not or only gradually feasible (cf. Blumenstock, Cadamuro and On, 2015; Beręsewicz, 2015, p. 46; Japec et al., 2015, p. 843).

Consequently, the limitations discussed above commonly have to be addressed for non-probability samples that are already collected, or at least for a given selection mechanism that cannot be revised as probability sampling. In such situations, the definition of non-probability sampling causes difficulties in estimation. Due to part ***a)*** of the above definition, design weights $\boldsymbol{w}^{\mathsf{nps}}$ are usually not at all obtainable for non-probability samples (cf. definition 2.13). But even when inclusion probabilities are hypothetically considered known, part ***b)*** prohibits general statements of unbiasedness (cf. equation 2.16). Therefore, both conditions imply that such data can be subject to biased estimates $\widehat{\boldsymbol{\vartheta}}$ for $\boldsymbol{\vartheta}$ because valid design weights in the sense of the previous section 2.2 are lacking. Potentially even more violated are assumptions underlying design-based quality assessments for $\widehat{\boldsymbol{\vartheta}}$

from a single sample since unbiased variance estimation additionally requires known and positive second-order inclusion probabilities. Furthermore, the MSE no longer equals the variance (cf. equation 2.11), and the bias component can usually only be assessed using information that is external to sample (cf. Bethlehem, 2008b, p. 21; Buelens, Burger and van den Brakel, 2018, p. 327; Lohr, 2010, p. 529). Overall, attempting to use non-probability samples for obtaining point estimates that are reasonably generalizable to the target population is a challenging task and measuring the quality of such estimates even more. Any of these objectives is rarely achievable without making assumptions about the data generating process and/or the variable of interest (cf. Baker et al., 2013a, p. 107; Buelens, Burger and van den Brakel, 2015, p. 2; Lohr, 2010, p. 7; Valliant and Dever, 2011, p. 106).

To formalize these issues, the population density of $\boldsymbol{Y}$ can be written as

$$\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot}\right) = \frac{\mathrm{P}\left(r_i^{\mathsf{nps}}=1\right)}{\mathrm{P}\left(r_i^{\mathsf{nps}}=1 \mid \boldsymbol{y}_{i\cdot}\right)} \cdot \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \mid r_i^{\mathsf{nps}}=1\right) \tag{2.24}$$

by means of Bayes' theorem, where $\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \mid r_i^{\mathsf{nps}}=1\right)$ is the density of $\boldsymbol{Y}$ in the non-probability sample. This is a basic framework for relating sample and population distribution in the context of informative sampling discussed by Pfeffermann and Sverchkov (1999) as well as Pfeffermann (2011) and closely related to the work of Smith (1983). In case of a probability sample ps, inclusion probabilities $\boldsymbol{\pi}^{\mathsf{ps}}$ are known and strictly positive. Under this condition, the population's distribution can be estimated from the sample using design weights because it holds that $\mathrm{P}\left(r_i^{\mathsf{ps}}=1\right)/\mathrm{P}\left(r_i^{\mathsf{ps}}=1 \mid \boldsymbol{y}_{i\cdot}\right) = \mathrm{E}\left(w_i^{\mathsf{ps}} \mid \boldsymbol{y}_{i\cdot}, r_i^{\mathsf{ps}}=1\right)/\mathrm{E}\left(w_i^{\mathsf{ps}} \mid r_i^{\mathsf{ps}}=1\right)$ (cf. equation 2.13; appendix B; Pfeffermann and Sverchkov, 1999, p. 185). This is basically a slightly different justification for the use of design weights than that discussed in the previous section 2.2.

In non-probability sampling, however, inclusion probabilities and design weights are typically unknown. As long as sample inclusion $\boldsymbol{r}^{\mathsf{nps}}$ and target variables $\boldsymbol{Y}$ are independent, it holds that $\mathrm{P}\left(r_i^{\mathsf{nps}}=1 \mid \boldsymbol{y}_{i\cdot}\right) = \mathrm{P}\left(r_i^{\mathsf{nps}}=1\right)$, such that the selection mechanism generating $\boldsymbol{r}^{\mathsf{nps}}$ can be ignored. Lending on the terminology of missing data adjustments, this is often referred to as the *missing-completely-at-random (MCAR)* case and e.g. occurs in simple random sampling. Yet, this scenario is highly unlikely for non-probability samples and generally not verifiable in reality (cf. van Buuren, 2018, p. 37; Mercer et al., 2017, p. 257). If $\boldsymbol{r}^{\mathsf{nps}}$ and $\boldsymbol{Y}$ are dependent, the population and sample distribution in equality 2.24 differ, and the sample selection process must thus be accounted for. This is possible if some auxiliary variables $\boldsymbol{X}$ ensure conditional independence of $\boldsymbol{Y}$ and $\boldsymbol{r}^{\mathsf{nps}}$, which is commonly called the *missing-at-random (MAR)* scenario and includes MCAR as a special case. As introduced in section 2.2, such auxiliary variables $\boldsymbol{X}$ are observed in the non-probability sample, and additional information about them external to the sample is available. Methods for utilizing these variables in estimation are discussed in chapter 5. If $\boldsymbol{X}$ cannot be determined to fulfill the conditions for MAR, dealing with the remaining dependency between $\boldsymbol{Y}$ and $\boldsymbol{r}^{\mathsf{nps}}$ is only possible under strong assumptions. This setting is labeled *missing-not-at-random (MNAR*; cf. appendix B; Bethlehem, 2010; Buelens, Burger and van den Brakel, 2018; Forster and Smith, 1998; Pfeffermann, 2011; 2015; Pfeffermann and Sikov, 2011; Rubin, 1976).

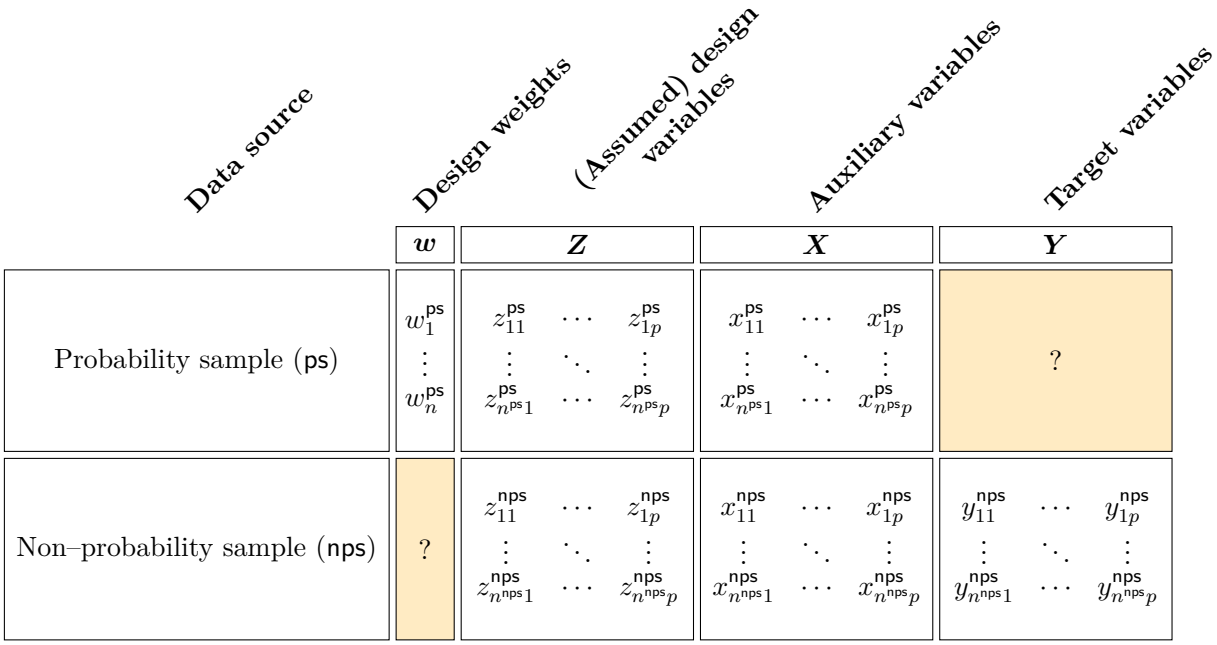


Figure 2.1: Schematic comparison of information in a probability and non-probability sample (cf. Yang and Kim, 2018, p. 3)

To illustrate the previous and subsequent discussion, figure 2.1 provides a schematic comparison of the relevant information inherent to a non-probability and a probability sample. The data columns are split in a block-wise manner, corresponding to the classification of variables introduced in section 2.2. Note that these blocks do not necessarily need to be mutually exclusive and that the probability sample can by definition coincide with the whole population (cf. Pfeffermann, 2011, p. 117). This representation solely simplifies graphical and notational descriptions throughout this and the following chapters.

In probability sampling, it is clear which variables belong to the set of design variables $\boldsymbol{Z}$ that determines sample inclusion probabilities and design weights. Even the values of these variables are often considered known already before selecting a sample, e.g. for determining stratification and cluster structures in the sampling frame. But even if their values are known only after sampling, it is still predetermined which variables influence the sample selection in a probability sampling design, such that inclusion probabilities and design weights can be determined by data collected during the survey (cf. e.g. Cochran, 1977, p. 89; Fuller, 2009, p. 28; Lohr, 2010, p. 3). In contrast, it is usually not even clear which variables have an impact on the selection processes of non-probability samples. At best, such variables can be assumed, usually based on some auxiliary information external to the sample (e.g. using the probability sample ps, cf. chapter 3). Even in the rather hypothetical scenario where the set of design variables was known, knowledge about their respective influence could still be lacking. Therefore, design weights are typically not obtainable. These, together with full coverage of the population, would be required for unbiased design-based estimation from a non-probability sample (cf. Baker et al., 2013a, p. 34; Lohr, 2010, p. 529). On the other hand, design weights and full coverage of the target population are given in case of the probability sample. However, target variables are not (adequately) observed here, otherwise the non-probability sample would not be required (cf. section 2.1). Although the latter does not allow for generalization without assumptions, it therefore provides some vital information to obtain $\widehat{\boldsymbol{\vartheta}}$ that is not available for the probability sample (cf. equation 2.24; Blumenstock, Cadamuro and On, 2015,

p. 1073; Dever, Rafferty and Valliant, 2008, p. 47; Japec et al., 2015, p. 866; Yang and Kim, 2018, p. 3). As a consequence, none of the samples alone is suitable to obtain generalizable estimates $\widehat{\boldsymbol{\vartheta}}$. Possible remedies can be to relax the constraints that render probability sampling infeasible, to ignore the selection mechanism that leads to the non-probability sample (and acknowledge the resulting limitations), or to assess this sampling process and account for its impact.

As indicated by the relation between sample and population density presented in equation 2.24 and the related discussion, assessment of such selection mechanisms is hardly possible when using only the non-probability sample. If parts of the population are systematically excluded from the sample, this data source alone does not provide any information about them. Investigating such under-coverage is only possible using some external information about the non-covered part of the population. Similar arguments apply even in cases where examination is not about coverage issues: determining whether certain elements are under- or over-represented in the non-probability sample – and may therefore differ with respect to their inclusion probabilities – is only feasible when the sample distribution can be compared with external benchmarks. The dependencies between sample inclusion and other variables are, therefore, not analyzable from the non-probability sample alone, because there is no variation of $\boldsymbol{r}^{\mathsf{nps}}$ in the non-probability sample (cf. definition 2.2; Bethlehem, 2008b, p. 31; Buelens et al., 2014, p. 6; Lohr, 2010, p. 529; Loosveldt and Sonck, 2008, p. 93).

In summary, the variables of interest for many areas of research are not observed in scope of probability samples. In contrast, these target variables can often be observed in non-probability samples where sampling mechanisms are uncontrolled and/or unknown. Strategies proposed for dealing with the challenges of non-probability samples are, therefore, typically based on additionally utilizing some reference data set. This approach results in scenarios as outlined in figure 2.1, where two (or more) data sets are used complementarily to deal with the non-probability selection mechanism.

There are two main paradigms that utilize the overlapping information between both data sets to perform estimation in such settings. On the one hand, the probability sample can be supplemented. Modeling the target variables using the auxiliaries, for example, allows for imputation in the reference data set. The challenge here is to find a well-performing combination of auxiliary variables and prediction method for each and every variable of interest. On the other hand, the auxiliary and/or assumed design variables are often used to construct surrogate weights for the non-probability sample. Yet, mimicking the important properties of classical design weights is challenging. For each specific non-probability sample, auxiliary variables that adequately describe the unknown selection process must be identified and observed. This may not always be possible, e.g. when the selection process depends on the target variables themselves. An in-depth discussion of these general estimation paradigms is provided in chapter 5.

In either case, a careful examination of the data generating process and the potential selectivity of a non-probability sample is required to perform estimation. As a first step for capturing these issues, concepts of representativity and methods for assessing it in the context of non-probability sampling are discussed in the following chapter 3.

3 Representativity and Selectivity

Known and strictly positive inclusion probabilities that allow for unbiased estimation and quantifiable accuracy with regard to many important statistics of interest (cf. section 2.2) often lead researchers to equate probability sampling with *representativeness* (cf. e.g. Buelens et al., 2014, p. 4; Little, 1988b, p. 287; Loosveldt and Sonck, 2008, p. 96; Pfeffermann, 2015, p. 448). In turn, terms like *non-representative* or *selective* are commonly and synonymously used to characterize non-probability samples (cf. e.g. Gelman et al., 2016b, p. 117; Pfeffermann, 2015, p. 444; Steinmetz et al., 2014, p. 275; Valliant and Dever, 2011, p. 105). In this interpretation, non-representativity is seen as cause for the issues related to non-probability sampling discussed in the previous section 2.3 (cf. Japec et al., 2015, p. 864; Meng, 2018, p. 688). To analyze these issues as far as possible, it is therefore important to consider the meaning, operationalization and assessment of representativity.

However, despite the importance of the phrase "representative method" (Kiær, 1895; 1897, cited in Kruskal and Mosteller, 1980, pp. 172 ff; Neyman, 1934, p. 559) in the history of statistics, the terms 'representative' and 'selective' are often used in a rather broad sense and, hence, subject to various different interpretations (cf. Bethlehem, 2010, p. 169; Kish, 1965, p. 26; Schouten, Cobben and Bethlehem, 2009, p. 102). An overview of how these terms can be understood to describe non-probability samples is given in the following section 3.1. Representativity and selectivity are frequently applied descriptions for both a sampling design as well as a single sample realized from it (cf. Kruskal and Mosteller, 1979a, p. 15). However, the data generating process is typically unknown for non-probability samples. As a consequence, selectivity arguments are mostly applied to a single realized sample (cf. e.g. Buelens et al., 2014, p. 6; Steinmetz et al., 2014, p. 288). To assess such features of a non-probability sample, auxiliary information external to the sample is required (cf. section 2.3; Baker et al., 2013a, p. 90), sources of which are summarized in section 3.2.

In the subsequent sections, approaches to operationalize and quantify the selectivity of non-probability samples are discussed. Different strategies are proposed and applied in the relevant literature (cf. e.g. Baker et al., 2013a, pp. 34 ff, 93 ff; Bethlehem and Biffignandi, 2012, pp. 303 ff, 385 ff; Schonlau et al., 2009; Meng, 2018; Petrucci and Rocco, 2019). These include comparison and statistical tests for revealing (dis-)similarities in auxiliary variables (sections 3.3 and 3.4), which can help to identify variables valuable for characterizing the non-probability selection mechanism (sections 3.5 to 3.7). These considerations can provide some guidance to assess the possible discrepancy between estimated and true statistic of interest (section 3.8). Furthermore, they foster approaches for estimation from non-probability samples, which are discussed in chapter 5.

3.1 Concepts of Representativity

In a series of articles, Kruskal and Mosteller (1979a,b,c; 1980) review history and usage of the concept of 'representative sampling' in non-scientific, general scientific and scientific statistical literature. Although it is possible to distinguish between representativity of either a single sample or a sample selection mechanism, this differentiation seems to get lost in the variety of meanings attributed to representativity itself (cf. Kruskal and Mosteller, 1979a, p. 15). Elaborating on the various different connotations of representativity, the authors provide an overview and discussion that is still relevant in contemporary research within and beyond the context of non-probability sampling (cf. e.g. Beręsewicz, 2015, p. 48; Schouten, Cobben and Bethlehem, 2009, p. 101; Tillé, Wilhelm et al., 2017, p. 7; Zhang, Thomsen and Kleven, 2013, p. 276).

The findings of Kruskal and Mosteller (1979a,b,c; 1980) can be summarized as follows. A first usage of representativity is as a rhetorical *"general acclaim"* (Kruskal and Mosteller, 1979a, p. 15) in a rather vague sense. In this case, representativity is used to praise data without further elaboration. A second sense implies the *"absence of selective forces"* (Kruskal and Mosteller, 1979a, p. 16). This meaning frequently appears in statistical science and is closely related to the third idea of *"coverage of the population's heterogeneity"* (Kruskal and Mosteller, 1979c, p. 254) in a sample. This leads to a fourth perception of representativity, implying that a sample is *"typical of the population"* (Kruskal and Mosteller, 1979a, p. 14). In certain cases, this refers to samples of typical units, which even may be of size one (single units) if in some sense lying near the center of a distribution of interest. If typicality is interpreted such that a sample as a whole should be typical for the population, with regards not only to the center but also (e.g.) the variability of a distribution, this meaning reverts to the idea of covering the population's heterogeneity. The ideal thereof is a sample as *"mirror or miniature of the population"* (Kruskal and Mosteller, 1979b, p. 111). Further meanings, which are mainly found in the statistical literature, refer to representative sampling as a *"specific sampling method"*, *"permitting good estimation"* or *"good enough for a particular purpose"* (Kruskal and Mosteller, 1979c, p. 245). These interpretations are usually rather explicitly regarding a particular context, such as predefined estimators and research questions. Sometimes, the terms 'representative sample' and 'probability sample' are simply defined as congruent (cf. Kruskal and Mosteller, 1979b, p. 111). This leads back to the depiction of non-probability samples as non-representative or selective, which is discussed at the beginning of the current chapter 3.

Beyond such equating definitions, considering the connotations of (non-)representativity can help to describe the challenges posed by non-probability sampling. While interpretations referring to specific sampling methods or purposes are highly situational, the remaining ones are quite general and help to operationalize (non-)representativity in a rather broad context. On the one hand, the above differentiation suggests that a representative sample should be typical for a population, fully covering its heterogeneity, and ideally a mirror thereof. Hence, distributions of relevant variables in the sample should coincide with that of the population to at least some precision (depending on the specific context). On the other hand, a sample should be selected without (untreated) selective forces, which otherwise can prevent covering the full variability of a population or lead to a distorted mirror image.

In summary, except for the vague rhetorical use, all of these views on representativity and selectivity emphasize relational aspects. Selective forces are usually defined as under- or over-representation of certain values of (e.g.) variables $\boldsymbol{Z}$ in the non-probability sample. On the one hand, this implies that $\boldsymbol{Z}$ and $\boldsymbol{\pi}^{\mathsf{nps}}$ are correlated, on the other hand, the distributions of $\boldsymbol{Z}^{\mathsf{nps}}$ and $\boldsymbol{Z}$ are typically different (cf. Kruskal and Mosteller, 1979c, pp. 261 ff). These considerations are therefore inherently linked to the relation of population and sample density stated in equation 2.24. Employing those relational aspects of representativity can help to assess and potentially compensate selectivity issues in non-probability samples: first of all, comparing distributions with external benchmarks is one way to assess whether some variables are related to the mechanism generating the non-probability sample. These ideas are the basis for sections 3.3 and 3.4, where comparisons and statistical tests for auxiliary variables are introduced. As extensions or alternatives, further approaches are based on evaluating the potential of certain variables to describe the non-probability sampling process. These are considered in sections 3.6 to 3.8. To obtain benchmarks and/or evaluate the dependencies between variables and the sample selection mechanism, some source of auxiliary information external to the non-probability sample is needed. A short overview on types of auxiliary data that are commonly considered in such contexts is provided in section 3.2.

3.2 Auxiliary Information: Para- and Reference Data

As discussed in section 3.1, representativity and selectivity are commonly defined with regards to relational aspects of sample and population distributions (cf. equation 2.24). Consequently, nearly all methods proposed for assessing the potential selectivity of non-probability samples aim at examining this relationship, although in different ways. For some variables of interest $\boldsymbol{Y}^{\mathsf{nps}}$ observed in a non-probability sample, this would only be feasible if the sample distribution $\mathrm{f}_{\boldsymbol{Y}^{\mathsf{nps}}}(\boldsymbol{y}_{i\cdot})$ could be compared to that of the population, $\mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y}_{i\cdot})$. However, this is essentially an ideal case that hardly ever occurs in real applications of non-probability samples. If the distribution of $\boldsymbol{Y}$ was known for the entire population, it could be used to compute the population statistics of interest. In this case, no sample would be needed at all.

Therefore, common strategies rely on auxiliary variables to obtain some proxy information about a sample's selectivity with regard to the target variables $\boldsymbol{Y}$ (cf. e.g. Bethlehem, 2008b, p. 31; Loosveldt and Sonck, 2008, p. 93; Schouten et al., 2016, p. 732). As introduced in section 2.2, auxiliary variables denoted by $\boldsymbol{X}$ are observed in the non-probability sample, and some external information about $\boldsymbol{X}$ outside this sample is available. As far as these auxiliaries are related to the target variables, representativity with regard to $\boldsymbol{X}$ can be considered an indication that the same holds regarding $\boldsymbol{Y}$, although this is neither a necessary nor sufficient condition (cf. Kruskal and Mosteller, 1979c, p. 263). As discussed before, the sets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ may even be overlapping.

Availability of information about $\boldsymbol{X}$ may mean that the population distribution $\mathrm{f}_{\boldsymbol{X}}(\boldsymbol{x}_{i\cdot})$ itself or some of its characteristics (e.g. the means $\boldsymbol{\mu}_{\boldsymbol{X}}$) are considered known. This distribution or its characteristics are used as a 'gold standard' or benchmark for assessing representativity and therefore must be based on external information about $\boldsymbol{X}$ that is of high quality. This suggests the use of an ideal probability *reference sample* with no or ignorable non-response, in which the auxiliary variables are measured in exactly

the same manner as in the non-probability sample (cf. sections 2.3 and 2.2; Biffignandi and Bethlehem, 2012, p. 370).[1] By definition, such a probability reference sample can encompass the whole *population*. For certain target populations, this kind of information is available and frequently used, e.g. in form of population or business registers. In other cases, estimates for $f_{\boldsymbol{X}}(\boldsymbol{x}_{i\cdot})$ or its characteristics from a probability sample are used as benchmarks (cf. e.g. Bethlehem, 2010, p. 174; Daas et al., 2015, p. 257; Kreuter et al., 2010, pp. 391 f; Särndal and Lundström, 2005, pp. 10 f; Schouten, Shlomo and Skinner, 2009, p. 13). Sometimes, reference surveys are conducted specifically for the purpose of comparison with the non-probability sample, in which case exactly coinciding questionnaires and modes of interviewing are applied. However, this is mainly done when the primary research goal is to evaluate the quality of non-probability samples (cf. e.g. Spijkerman et al., 2009, p. 1642). Reference samples that are already available but not specifically designed for the purpose of comparison are far more common in most other scenarios of non-probability sampling (cf. e.g. Barratt, Ferris and Lenton, 2015, p. 10; Bethell et al., 2004, p. 5; Ryzin, 2008, p. 246). In these latter cases, differences in survey modes and questionnaire designs interfering with the comparisons of non-probability and reference sample can be an additional issue (cf. Schonlau, Van Soest and Kapteyn, 2007, p. 9; Yeager et al., 2011, pp. 712 ff).

A further type of auxiliary information besides population and reference sample data is *para-data* (cf. e.g. Buelens, Burger and van den Brakel, 2015, p. 28; Shlomo, Skinner and Schouten, 2012, p. 210; Steinmetz et al., 2014, p. 287). Although this term itself is not uniquely defined, para-data is commonly constituted by information gathered during the sampling process but external to the survey itself (cf. Kreuter and Olson, 2013, pp. 2 f). For example, characteristics of the environment (e.g. residential area), the person that was contacted (e.g. gender) or the type of initial contact (e.g. questions asked to the interviewer) may be recorded as para-data in case of interviewer-administered surveys. Further information can either be gathered by the interviewer or, especially when using computer-aided interviewing, automatically collected. Examples include the frequencies and timestamps of certain events when sampling and surveying, e.g. the number and time of (attempted) contacts or provided answers. Automation can provide further information especially for data that is gathered online, such as "device-type para-data and questionnaire navigation para-data" (Callegaro, 2013, p. 262). For example, this data can include information about the device, browser and system language used to visit a (survey) page, as well as keystroke and mouse click frequencies (cf. Callegaro, 2013; Durrant, D'Arrigo and Müller, 2013; Olson and Parkhurst, 2013).

"In most real life studies, auxiliary variables are available" (Schouten, 2018, p. 33), and all of the above examples can be seen as constituting some kind of *reference data set* containing information about elements which are not (necessarily) part of the non-probability sample. In context of the current chapter 3, such auxiliary information is central to the different strategies for investigating the potential selectivity and biases of non-probability samples, for which different methods are discussed in the following sections. Introducing rather informal comparisons of reference and non-probability sample in section 3.3, this leads to more refined strategies in the subsequent discussion. The use of auxiliary information to compensate for selectivity is discussed in chapter 5.

[1] The assumption that non-response in the reference sample is ignorable is not always valid, but typically itself not testable without even better reference data (cf. e.g. Barratt, Ferris and Lenton, 2015, p. 5; Biffignandi and Bethlehem, 2012, p. 370; Ghitza and Gelman, 2013, p. 769; Pasek, 2016, p. 286).

3.3 Comparing Auxiliary Variables to Assess Representativity

Combining the discussion in sections 2.2 to 3.2, sample and population distribution of the auxiliary variables $\boldsymbol{X}$ are linked through the non-probability sampling process by

$$\mathrm{f}_{\boldsymbol{X}}\left(\boldsymbol{x}_{i\cdot}\right) = \frac{\mathrm{P}\left(r_i^{\mathsf{nps}} = 1\right)}{\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{x}_{i\cdot}\right)} \cdot \mathrm{f}_{\boldsymbol{X}^{\mathsf{nps}}}\left(\boldsymbol{x}_{i\cdot}\right) \quad , \tag{3.1}$$

where $\mathrm{f}_{\boldsymbol{X}^{\mathsf{nps}}}\left(\boldsymbol{x}_{i\cdot}\right) = \mathrm{f}_{\boldsymbol{X}}\left(\boldsymbol{x}_{i\cdot} \,|\, r_i^{\mathsf{nps}} = 1\right)$ and $\mathrm{f}_{\boldsymbol{X}}\left(\boldsymbol{x}_{i\cdot}\right)$ respectively denote the density of $\boldsymbol{X}$ in the sample and the population. If the vector of sample inclusion indicators $\boldsymbol{r}^{\mathsf{nps}}$ and variables $\boldsymbol{X}$ are independent, these two densities do almost everywhere not differ in expectation because $\mathrm{P}\left(r_i^{\mathsf{nps}} = 1\right) = \mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{x}_{i\cdot}\right)$ (cf. Pfeffermann and Sverchkov, 1999; Pfeffermann, 2011; Smith, 1983). Therefore, one way to assess representativity of a sample with respect to $\boldsymbol{X}$ is by comparing $\mathrm{f}_{\boldsymbol{X}^{\mathsf{nps}}}\left(\boldsymbol{x}_{i\cdot}\right)$ and $\mathrm{f}_{\boldsymbol{X}}\left(\boldsymbol{x}_{i\cdot}\right)$ (cf. e.g. Braver and Bay, 1992, p. 927; Gelman et al., 2016b, pp. 91 f). This approach is inherently related to the concept of representativity as covering the full heterogeneity of a population up to creating a miniature of this population (cf. section 3.1). If the two densities in equality 3.1 coincide, this is an indication that the sample does not exhibit over- or under-coverage regarding values of $\boldsymbol{X}$. If there is a large difference, the sample is prone to be selective with respect to $\boldsymbol{X}$, although the quantification of 'large' is highly situational in this context. As discussed in section 3.2, a comparison for $\boldsymbol{Y}$ itself in terms of equation 3.1 is not feasible when $\boldsymbol{Y}$ is known solely for the non-probability sample, but an overlap between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is possible.

A popular strategy to assess potential selectivity of non-probability samples is therefore to compare the resulting distribution of $\boldsymbol{X}$ with that of some reference data (cf. e.g. Bethlehem, 2008b, p. 31; Buelens et al., 2014, p. 4; van den Brakel et al., 2017, p. 184). For real applications, this is mostly done by graphical or tabulated comparisons of frequency distributions. In many cases, such comparisons are used exclusively with respect to socio-demographic variables for which reference data of high quality is available, such as gender, age groups or education (cf. Lohr, 2010, p. 529; Weisberg, 2005, pp. 172 ff). In some cases, cross-classifications corresponding to multivariate distributions are compared (cf. e.g. Steinmetz et al., 2014, p. 279), but in the more typical case, only marginal distributions are used for this purpose (cf. e.g. Cumming, 1990, pp. 134 f; Feild et al., 2006, p. 577; Gelman et al., 2016b, p. 92; Loosveldt and Sonck, 2008, p. 98; Schonlau et al., 2009, p. 305). Continuous variables are less common but still occasionally used for such comparisons. In these cases, specific aspects of the distributions are typically compared. As before, this mainly concerns assessments of similarity in univariate distributional aspects, such as means and, in case of longitudinal data, trends (cf. e.g. Beręsewicz, 2015, p. 46; van den Brakel et al., 2017, p. 184; Lehdonvirta et al., 2020, p. 13). Multivariate characteristics, such as regression coefficients, are considered only occasionally (cf. e.g. Bethell et al., 2004, p. 9; Sanders et al., 2007, p. 278; Steinmetz, Tijdens and Pedraza, 2009, p. 32).

Graphical or tabulated evaluations of frequency distributions or means are common examples when non-probability and reference samples are compared. Yet, such contrasts do not always yield a clear picture and can therefore lead to ambiguous interpretations. Even if it is selected by means of an ideal probability sampling design, a single realized sample can exhibit huge deviations from some benchmark data due to sampling variance

alone (cf. Särndal, Swensson and Wretman, 1992, p. 41). This is especially true when the benchmark itself is subject to sampling or non-sampling errors (cf. Baker et al., 2013a, p. 90; Biffignandi and Bethlehem, 2012, p. 370; Steinmetz et al., 2014, p. 288). Partially relaxing these limitations, some authors adopt the use of statistical tests as a way to formalize the comparisons discussed in the current section 3.3. This strategy for assessing selectivity of non-probability samples is presented in the following section 3.4.

3.4 Testing for Selectivity

To formally compare data sets in order to assess deviations of the non-probability sample from some benchmark data source, statistical tests are an apparent way (cf. e.g. Kruskal and Mosteller, 1979c, pp. 261 ff; Little, 1988a; Särndal, 2011). An advantage when applying these tests for selectivity assessment is that sampling errors of the reference data set can be incorporated, a topic that is rarely considered for the simple comparisons discussed in the previous section 3.3 (cf. Baker et al., 2013a, p. 90; Steinmetz et al., 2014, p. 288). A possible limitation of such tests in the present context is that general non-probability samples do not necessarily fulfill the assumptions concerning the randomness of the data generating process that are required for the validity of these tests (cf. e.g. Hawkins, 1981, p. 107; Stephens, 1976, p. 357). Nevertheless, statistical tests are de facto commonly used to assess selectivity of non-probability samples, which may be the case for exploratory or naive reasons or due to the lack of methodologically more adequate alternatives (cf. e.g. Barratt, Ferris and Lenton, 2015, p. 7; Drabble et al., 2018, p. 7; Loosveldt and Sonck, 2008, p. 96; Schillewaert and Meulemeester, 2005, p. 166; Smyk, Tyrowicz and Van der Velde, 2021, pp. 438 ff; Yeager et al., 2011, p. 718). It is therefore important to discuss and evaluate the usability of such tests for selectivity assessment. In the following paragraphs, a number of selected parametric and non-parametric tests are introduced for this purpose, which can be used to compare distributions or their moments (means and variances).

In a general framework, S arbitrary separate data sets $\mathcal{S}^s$ of sizes n^s for $s = 1, \ldots, S$ are considered for comparison by means of a statistical test. The combination of these data sets is defined by

$$\mathcal{S}^{\mathsf{u}} \coloneqq \bigcup_{s=1}^{S} \mathcal{S}^s \tag{3.2}$$

for the combined data set u of size n^{u}. For example, $\mathcal{S}^{\mathsf{u}} = \mathcal{S}^{\mathsf{nps}} \cup \mathcal{S}^{\mathsf{ps}}$ may be the combination of a non-probability and a reference sample. Denoting the matrices of (partially) common variables for each data set s by

$$\boldsymbol{A}^s = \begin{bmatrix} \boldsymbol{X}^s & \boldsymbol{Z}^s & \boldsymbol{Y}^s \end{bmatrix} \in \mathbb{R}^{n^s \times v} \tag{3.3}$$

for notational simplicity, the matrix for the combined data set is

$$\boldsymbol{A}^{\mathsf{u}} = \begin{bmatrix} \boldsymbol{A}^1 \\ \vdots \\ \boldsymbol{A}^S \end{bmatrix} \in \mathbb{R}^{n^{\mathsf{u}} \times v} \quad . \tag{3.4}$$

Partial but not necessarily complete overlap in the variables then implies that some columns $\boldsymbol{a}^{\mathsf{u}}_{\cdot j}$ are observed for all of the data sources. Other variables may be unknown for some data sets, leading to missing values in certain rows of $\boldsymbol{A}^{\mathsf{u}}$.

In the context of missing data, various tests are used for testing whether the missing values follow a MCAR or MAR pattern (cf. van Buuren, 2018, p. 37). As outlined above and depicted in figure 2.1, missingness also occurs for the combination of non-probability and reference samples, where missing values arise in data sets where $\boldsymbol{Y}$ is not observed. Consequently, tests for missing data patterns can be adapted to assess selectivity of non-probability samples (cf. section 2.3; Mercer et al., 2017, p. 253). Jamshidian and Jalal (2010) discuss and evaluate a range of such tests, which are included as part of the subsequent discussion.

Note that to allow for weighting in the considered tests, a vector of weights

$$\boldsymbol{w}^{\mathsf{u}} \;=\; \left[\left(\boldsymbol{w}^{1}\right)^{\mathsf{T}} \;\; \ldots \;\; \left(\boldsymbol{w}^{S}\right)^{\mathsf{T}} \right]^{\mathsf{T}} \quad \in \quad \mathbb{R}^{n^{\mathsf{u}}} \tag{3.5}$$

in analogy to equation 3.4 is assumed for the combined data set. Since classical design weights are typically unknown for the non-probability sample (cf. section 2.3), an arbitrary vector $\widetilde{\boldsymbol{w}} \in \mathbb{R}^{n^{\mathsf{nps}}}$ is used for this sample. The simplest approach is to set these weights to ones, implying that all observations are given the same relevance (cf. e.g. Steinmetz et al., 2014, p. 279). Techniques for constructing more sophisticated weights for non-probability samples are discussed in section 5.2, e.g. to attribute more relevance to cases that appear under-represented in the non-probability sample.

Several tests for response patterns are based on variance decomposition. The main idea in this regard is to break down the total variability around the mean into parts corresponding to variance within and between the different data sets. The variation within each of the sub-data sets $s = 1, \ldots, S$ is that around the estimated mean $\widehat{\boldsymbol{\mu}}_{\boldsymbol{A}}\left(\boldsymbol{w}^{s}\right)$ for this data set. It is calculated from the centered variables for all $i = 1, \ldots, n^{s}$ observations, denoted by

$$\mathbf{e}\left(\boldsymbol{a}_{i\cdot}^{s}\right) \;\coloneqq \boldsymbol{a}_{i\cdot}^{s} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{A}}\left(\boldsymbol{w}^{s}\right) \qquad . \tag{3.6a}$$

The variability between groups is based on the differences between means of the sub- and combined data set, expressed by

$$\mathbf{e}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{A}}\left(\boldsymbol{w}^{s}\right)\right) \;\coloneqq\; \widehat{\boldsymbol{\mu}}_{\boldsymbol{A}}\left(\boldsymbol{w}^{s}\right) - \widehat{\boldsymbol{\mu}}_{\boldsymbol{A}}\left(\boldsymbol{w}^{\mathsf{u}}\right) \qquad . \tag{3.6b}$$

Little (1988a) proposes testing the equality of means for all sub-data data sets (groups) to deduce whether a MCAR pattern can be rejected. The suggested test assumes that variables $\boldsymbol{A}^{s}$ for all $s = 1, \ldots, S$ data sources come from a population that follows a v-variate normal distribution, i.e.

$$\boldsymbol{A}^{s} \;\sim\; \mathrm{N}\left(\boldsymbol{\mu}_{\boldsymbol{A}}^{(s)}, \boldsymbol{\Sigma}_{\boldsymbol{A}}\right) \qquad , \tag{3.7}$$

or that sample sizes are sufficiently large to employ the central limit theorem (cf. Lindeberg, 1922; Little, 1988a, p. 1200). In this formulation, $\boldsymbol{\mu}_{\boldsymbol{A}}^{(s)}$ is the true mean vector in the target population of data set s, which may be different for the S data sets. The test's null hypothesis of equal means is

$$\mathrm{H}_0 : \; \boldsymbol{\mu}_{\boldsymbol{A}} = \boldsymbol{\mu}_{\boldsymbol{A}}^{(1)} \;=\; \ldots = \boldsymbol{\mu}_{\boldsymbol{A}}^{(S)} \qquad , \tag{3.8}$$

where $\boldsymbol{\mu}_{\boldsymbol{A}}$ is mean in the combined population. The covariances $\boldsymbol{\Sigma}_{\boldsymbol{A}}$ are assumed to be the same across all populations. Without loss of generality, assume the first c_s variables, corresponding to columns of $\boldsymbol{A}$ indexed by $\mathcal{J}^s = \{1, \ldots, c_s\}$, are observed for the s-th data source. The employed test-statistic $\widehat{T} \in \mathbb{R}_{\geq 0}$ is then given by

$$\widehat{T} \;=\; \sum_{s=1}^{S} \widehat{\mathrm{N}}\left(\boldsymbol{w}^s\right) \cdot \left[\mathbf{e}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{A}}\left(\boldsymbol{w}^s\right)\right)\right]_{\mathcal{J}^s} \left[\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{A}}\left(\boldsymbol{w}^{\mathsf{u}}\right)\right]^{-1}_{\mathcal{J}^s\mathcal{J}^s} \left(\left[\mathbf{e}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{A}}\left(\boldsymbol{w}^s\right)\right)\right]_{\mathcal{J}^s}\right)^{\mathsf{T}} \quad , \tag{3.9}$$

where $[\cdot]_{\mathcal{J}^s}$ denotes a sub-vector with elements indexed by $\mathcal{J}^s$ and $[\cdot]_{\mathcal{J}^s\mathcal{J}^s}$ is a sub-matrix with rows and columns indexed by $\mathcal{J}^s$. If null hypothesis 3.8 and normality assumption 3.7 hold, each of the single terms summed in equation 3.9 follows an F-distribution with 1 and $n^s - 2$ degrees of freedom. In this case, $\widehat{T}$ is asymptotically χ^2-distributed, with $||\boldsymbol{c}||_1 - v$ degrees of freedom, which can be used for testing the null hypothesis. Note that to calculate $\widehat{T}$ from equation 3.9, the complete data estimates $\widehat{\boldsymbol{\mu}}_{\boldsymbol{A}}\left(\boldsymbol{w}^{\mathsf{u}}\right)$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{A}}\left(\boldsymbol{w}^{\mathsf{u}}\right)$ for means and covariances are required (cf. definition 3.6b). When some variables $\boldsymbol{a}^s_{\cdot j}$ are unobserved, parts of the information required to directly estimate these quantities are missing. To obtain a complete data set in this case, imputation methods can be used. Little (1988a, p. 1200) suggests using the expectation-maximization algorithm for this purpose (cf. Dempster, Laird and Rubin, 1977; Orchard and Woodbury, 1972). Nevertheless, other imputation methods (cf. e.g. van Buuren, 2018; Fuller, 2009, pp. 288 ff) can likewise be used to obtain the required complete data estimates. This reasoning equally applies to the following tests whenever testing is performed for variables that are not observed in all S data sets (cf. Jamshidian and Jalal, 2010; Jamshidian, Jalal and Jansen, 2014). As discussed above, equation 3.9 assumes that the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{A}}$ is the same across all data sets. In terms of sensitivity and power, Little (1988a, p. 1202) argues against a possible extension to relax this assumption.

Nevertheless, Jamshidian and Jalal (2010) suggest testing the equality of covariances (homoscedasticity) as well as the feasibility of multivariate normality assumptions based on the work of Hawkins (1981). Again, this test assumes multivariate normality of the target population of each data set $s = 1, \ldots, S$, i.e.

$$\boldsymbol{A}^s \;\sim\; \mathrm{N}\left(\boldsymbol{\mu}^{(s)}_{\boldsymbol{A}}, \boldsymbol{\Sigma}^{(s)}_{\boldsymbol{A}}\right) \qquad . \tag{3.10}$$

In this case, the covariances $\boldsymbol{\Sigma}^{(s)}_{\boldsymbol{A}}$ may vary between populations, differentiating it from assumption 3.7. The null hypothesis is the equality of covariances in all S populations:

$$\mathrm{H}_0 : \; \boldsymbol{\Sigma}_{\boldsymbol{A}} = \boldsymbol{\Sigma}^{(1)}_{\boldsymbol{A}} \;=\; \ldots = \boldsymbol{\Sigma}^{(S)}_{\boldsymbol{A}} \qquad . \tag{3.11}$$

Denoted by $\boldsymbol{\Sigma}_{\boldsymbol{A}}$ is the combined data covariance, which is estimated using the pooled covariance estimator

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{A}} \;=\; \sum_{s=1}^{S} \frac{n^s - 1}{n^{\mathsf{u}} - S} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{A}}\left(\boldsymbol{w}^s\right) \qquad . \tag{3.12}$$

The test statistic is then given by $\widehat{\boldsymbol{T}} = \left[\widehat{T}^1_1 \;\; \ldots \;\; \widehat{T}^1_{n^1} \;\; \ldots \;\; \widehat{T}^S_{n^S}\right]^{\mathsf{T}} \in \mathbb{R}^{n^{\mathsf{u}}}$, with elements

$$\widehat{T}^s_i \;=\; \frac{\left((n^{\mathsf{u}} - S - v) \cdot \mathbf{e}\left(\boldsymbol{a}^s_{i\cdot}\right) \widehat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{A}} \left(\mathbf{e}\left(\boldsymbol{a}^s_{i\cdot}\right)\right)^{\mathsf{T}}\right)}{\left(v \cdot \left(\left(n^s - 1\right)\left(n^{\mathsf{u}} - S\right) - n^s \cdot \mathbf{e}\left(\boldsymbol{a}^s_{i\cdot}\right) \widehat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{A}} \left(\mathbf{e}\left(\boldsymbol{a}^s_{i\cdot}\right)\right)^{\mathsf{T}}\right)\right)} \tag{3.13}$$

for $i = 1, \ldots, n^s$ and $s = 1, \ldots, S$. Matrices of size 1×1 are treated as scalars for equation 3.13. Note that this constitutes $\widehat{\boldsymbol{T}}$ as a vector- rather than a scalar-valued test statistic, and essentially the same can occur for the following tests as well. Approaches to deal with vector-valued test statistics are therefore discussed after summarizing these tests. Hawkins (1981, p. 106) shows that under normality assumption 3.10 and null hypothesis 3.11, these test statistics are exactly F-distributed with v and $n-v-S$ degrees of freedom. This result can be used to test null hypothesis 3.11 and normality assumption 3.10, but only simultaneously (cf. Hawkins, 1981, p. 109).

Being based on ratios of variance components, both tests discussed so far are closely related to the F-test for the classical analysis of variance (ANOVA; cf. e.g. Faraway, 2002). The null hypothesis in this case is again the equality of means defined in equation 3.8, and a separate test statistic $\widehat{\boldsymbol{T}} \in \mathbb{R}^v_{\geq 0}$ is computed for each column (variable) $\boldsymbol{a}^{\mathsf{u}}_{\cdot j}$. Components of $\widehat{\boldsymbol{T}}$ are determined as ratios of between and within variances, defined by

$$\widehat{T}_j = \frac{\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{u}}\right) - S}{S-1} \cdot \frac{\sum\limits_{s=1}^{S} \widehat{\mathrm{N}}\left(\boldsymbol{w}^s\right) \cdot \left(\mathbf{e}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{a}^s_{\cdot j}}\left(\boldsymbol{w}^s\right)\right)\right)^2}{\sum\limits_{s=1}^{S} \widehat{\mathrm{N}}\left(\boldsymbol{w}^s\right) \cdot \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{a}^s_{\cdot j}}\left(\boldsymbol{w}^s\right)} \tag{3.14}$$

for all $j = 1, \ldots, v$ variables. Under null hypothesis 3.8 and normality assumption 3.7, each $\widehat{T}_j$ is F-distributed with $S-1$ and $n^{\mathsf{u}} - S$ degrees of freedom. The two-tailed t-test is a very commonly used and therefore noteworthy special case of the F-test for $S = 2$ (cf. Aspin and Welch, 1949; Blitzstein and Hwang, 2013, p. 442; Box, 1953, p. 320; Moser, Stevens and Watts, 1989, p. 3964; Satterthwaite, 1946; Welch, 1947, p. 32).

As outlined, the above tests assume normally distributed populations or sufficient sample sizes to employ the central limit theorem in case of Little's and t-test (cf. Klenke, 2013, pp. 320 ff). For cases where these requirements are not fulfilled, Jamshidian and Jalal (2010) propose applying non-parametric methods to test for inequality of distributions in different data sets. Common examples for such non-parametric tests are the Kolmogorov-Smirnov (cf. Kolmogorov, 1933 and Smirnov, 1936, cited in Birnbaum, 1952), Kruskal-Wallis (cf. Kruskal and Wallis, 1952) or Anderson-Darling test (cf. Anderson and Darling, 1952; Scholz and Stephens, 1987), which are introduced in the following.

In the present context, the null hypothesis of the Kolmogorov-Smirnov test is that distributions of $\boldsymbol{A}$ are equal in all target populations of the $s = 1, \ldots, S$ sub-data sets:

$$\mathrm{H}_0 : \mathrm{F}_{\boldsymbol{A}} = \mathrm{F}^{(1)}_{\boldsymbol{A}} = \ldots = \mathrm{F}^{(S)}_{\boldsymbol{A}} \quad , \tag{3.15}$$

where $\mathrm{F}^{(s)}_{\boldsymbol{A}}$ is the distribution in the target population of data set s, and $\mathrm{F}_{\boldsymbol{A}}$ is the combined distribution. This test is based on the vector $\widehat{\boldsymbol{T}} = \left[\widehat{T}^1_1 \;\; \ldots \;\; \widehat{T}^1_v \;\; \ldots \;\; \widehat{T}^S_v\right]^{\mathsf{T}} \in \mathbb{R}^{S \cdot v}$ of test statistics, with elements defined by

$$\widehat{T}^s_j = \max_{a \in \boldsymbol{a}^{\mathsf{u}}_{\cdot j}} \left(\mathrm{Abs}\left(\widehat{\mathrm{F}}_{\boldsymbol{a}_{\cdot j}}\left(a, \boldsymbol{w}^{\mathsf{u}}\right) - \widehat{\mathrm{F}}_{\boldsymbol{a}_{\cdot j}}\left(a, \boldsymbol{w}^s\right)\right)\right) \tag{3.16}$$

for $j = 1, \ldots, v$ and $s = 1, \ldots, S$. These represent the maximum absolute difference in the estimated distribution functions between sub- and combined data set for all variables and sub-data sets (cf. definition 2.18e). Kolmogorov (1933, cited in Birnbaum, 1952, p. 425)

derives the asymptotic distribution

$$\widehat{\mathrm{F}}_{\left(\widehat{T}_j^s/n^s\right)}\left(\frac{\widehat{T}_j^s}{n^s}\right) = 1 - 2 \cdot \sum_{i=1}^{\infty} (-1)^{(i-1)} \cdot \exp\left(-2 \cdot i^2 \cdot \left(\widehat{T}_j^s\right)^2\right) \tag{3.17}$$

for $\widehat{T}_j^s$ (cf. Feller et al., 1948; Darling, 1957). Smirnov (1936, cited in Birnbaum, 1952, p. 425) presents critical values for testing hypothesis 3.15, which can be computed by means of recursive formulae (cf. Birnbaum, 1952; Lilliefors, 1967; Massey, 1950).

As an alternative, the Anderson-Darling test likewise tests hypothesis 3.15 based on differences between empirical distribution functions. Applied to multiple data sets (or groups), as proposed by Scholz and Stephens (1987), its test statistics $\widehat{\boldsymbol{T}} \in \mathbb{R}^v$ is defined by elements

$$\widehat{T}_j = \sum_{s=1}^{S} n^s \cdot \int\limits_{a \in \boldsymbol{a}_{\cdot j}^{\mathsf{u}}} \frac{\left(\widehat{\mathrm{F}}_{\boldsymbol{a}_{\cdot j}}(a, \boldsymbol{w}^s) - \widehat{\mathrm{F}}_{\boldsymbol{a}_{\cdot j}}(a, \boldsymbol{w}^{\mathsf{u}})\right)^2}{\widehat{\mathrm{F}}_{\boldsymbol{a}_{\cdot j}}(a, \boldsymbol{w}^{\mathsf{u}}) \cdot \left(1 - \widehat{\mathrm{F}}_{\boldsymbol{a}_{\cdot j}}(a, \boldsymbol{w}^{\mathsf{u}})\right)} \, d\, \widehat{\mathrm{F}}_{\boldsymbol{a}_{\cdot j}}(a, \boldsymbol{w}^{\mathsf{u}}) \tag{3.18}$$

for all $j = 1, \ldots, v$. As before, $\widehat{\mathrm{F}}_{\boldsymbol{a}_{\cdot j}}$ denotes the weighted empirical distribution function of column j in $\boldsymbol{A}$. Under null hypothesis 3.15 to be tested, the distribution of each $\widehat{T}_j$ converges to a weighted sum of independent χ^2-distributed variables with $S - 1$ degrees of freedom (cf. Anderson and Darling, 1952). The result is a "strange distribution function" (Marsaglia and Marsaglia, 2004, p. 2), for which a number of authors provide approximations and tabulated values for hypothesis testing, together with computational simplifications for non-continuous variables (cf. e.g. Anderson and Darling, 1954; Giles, 2001; Lewis, 1961; Marsaglia and Marsaglia, 2004; Scholz and Stephens, 1987; Sinclair and Spurr, 1988).

The test proposed by Kruskal and Wallis (1952) applied to the outlined setting is based on variance decomposition of the ranks for each column $\boldsymbol{a}_{\cdot j}^{\mathsf{u}}$ in $\boldsymbol{A}^{\mathsf{u}}$. Its test statistics $\widehat{\boldsymbol{T}} \in \mathbb{R}_{\geq 0}^v$ is defined by ratios of between and total variance of these ranks for all $j = 1, \ldots, v$, i.e.

$$\widehat{T}_j = \sum_{s=1}^{S} \widehat{\mathrm{N}}(\boldsymbol{w}^s) \cdot \left(\mathbf{e}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{d}_{\cdot j}}(\boldsymbol{w}^s)\right)\right)^2 \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}_{\cdot j}}(\boldsymbol{w}^{\mathsf{u}})\right)^{-1} \quad , \tag{3.19}$$

where $\boldsymbol{d}_{\cdot j}$ represents the ranks of variable $\boldsymbol{a}_{\cdot j}^{\mathsf{u}}$. Under null hypothesis 3.15, $\widehat{T}_j$ is asymptotically χ^2-distributed with $S - 1$ degrees of freedom (cf. Breslow, 1970; Kruskal, 1952). Different methods to obtain the exact small sample distribution of this test statistic are presented and compared by Iman and Davenport (1976) as well as Choi et al. (2003).

All tests presented so far are *difference tests*. These are frequently used in real applications, even though the research goal commonly is to show that there is no difference between data sources. If the null hypothesis is not rejected, this is considered an indication that the distribution of relevant variables does not differ between data sets. At least with respect to these variables, selectivity of the non-probability is then rejected (cf. e.g. Barratt, Ferris and Lenton, 2015, p. 13; Braunsberger, Wybenga and Gates, 2007, p. 761; Faas and Schoen, 2006, p. 183; Ryzin, 2008, p. 256; Yeager et al., 2011). Nevertheless, difference tests are based on using equalities, e.g. of means, covariances or distributions, as their respective null hypothesis (cf. hypotheses 3.8, 3.11 and 3.15). As the decision imposed by a statistical test is to reject or not reject its null hypothesis, difference tests are

therefore improper tools to accept hypotheses of equality (cf. Schuirmann, 1987, p. 659; Storey, 2002, p. 479). Hence, they have to be used and interpreted very carefully in the outlined context, and it seems important to discuss *equivalence tests* as a remedy for these obstacles. Such tests use equalities as alternative hypothesis, formalized as an interval in which the difference(s) $\widehat{\boldsymbol{T}} \in \mathbb{R}^v$ of the respective characteristics of $\boldsymbol{A}^{\mathsf{u}}$ (e.g. as above, means, variances or values of the distribution functions) are considered equal. The resulting null hypothesis is thus

$$\mathrm{H}_0 : \widehat{\boldsymbol{T}} \leq \boldsymbol{L}_{\widehat{\boldsymbol{T}}} \quad \text{or} \quad \widehat{\boldsymbol{T}} \geq \boldsymbol{U}_{\widehat{\boldsymbol{T}}} \tag{3.20}$$

for lower and upper interval boundaries $\boldsymbol{L}_{\widehat{\boldsymbol{T}}}, \boldsymbol{U}_{\widehat{\boldsymbol{T}}} \in \mathbb{R}^v$. A common strategy to perform an equivalence test is by testing the lower and upper part of hypothesis 3.20 separately. Each of them is used as null hypothesis in a one-sided test, with significance level corresponding to that of the whole test. If both tests are significant, hypothesis 3.20 can be rejected (cf. e.g. Berger, Hsu et al., 1996; Kirkwood and Westlake, 1981; Lakens, 2017; Schuirmann, 1987; Westlake, 1976). Alternatively, one can check whether the confidence interval for $\widehat{\boldsymbol{T}}$ corresponding to twice the significance level of the test is completely enclosed by $\boldsymbol{L}_{\widehat{\boldsymbol{T}}}$ and $\boldsymbol{U}_{\widehat{\boldsymbol{T}}}$ (cf. Kirkwood and Westlake, 1981, p. 593; Limentani et al., 2005, p. 223; Schuirmann, 1987, p. 661). When using equivalence tests, selecting $\boldsymbol{L}_{\widehat{\boldsymbol{T}}}$ and $\boldsymbol{U}_{\widehat{\boldsymbol{T}}}$ is the main difficulty. The choice has to be made with respect to the specific variable and research interest (cf. Dong et al., 2017; Schuirmann, 1987, p. 659; Tsong, Dong and Shen, 2017). In the scientific literature, the most relevant case of equivalence tests deals with equality of means, corresponding to hypothesis 3.8 as alternative hypothesis. Its realization is commonly based on t-tests, which are a special case of the ANOVA described in the context of equation 3.14 (cf. Lakens, 2017, p. 357; Schuirmann, 1987; Welch, 1951; Westlake, 1976).

As stressed before, all approaches apart from Little's (1988a) test presented throughout this section can result in a vector of test statistics rather than a single number. There are different approaches to apply tests for vector-valued test statistics. One option is to apply the theorem presented by Rosenblatt (1952). To that end, p-values are obtained from the respective distributions of the test statistics, and it is tested whether these p-values come from a uniform distribution. For example, this can be done by applying the test proposed by Neyman (1937; cf. Ledwina, 1994). It can either be applied to tests performed on $\boldsymbol{A}^{\mathsf{u}}$ after a Mahalanobis transformation (cf. Hawkins, 1981; Kessy, Lewin and Strimmer, 2018, p. 310) or when using multivariate distribution functions instead of univariate ones (cf. Justel, Peña and Zamar, 1997). An additional approach would be to perform multiple univariate tests, with corresponding adjustment of significance levels (cf. e.g. Holm, 1979). In this case, one has to consider that the research goal often is to provide evidence for *equivalence* of data sources. Therefore, significance adjustments have to be made such that it is harder to provide evidence for multiple variables than for a single one. In particular, suppose that one uses a difference test to conclude that a non-probability sample has a distribution that is similar to some reference data in case of a non-significant result. Lowering the significance level would make such a result much more likely, making the adjustment highly misleading – not to say wrong – for showing equivalence. Examples can be found in scientific publications (cf. e.g. Faas and Schoen, 2006, p. 183; Schillewaert and Meulemeester, 2005, p. 174) that motivate this word of caution.

To summarize the previous considerations, tests for selectivity provide decision rules for whether certain characteristics of variables in the non-probability sample coincide with those in some reference data set. As for the comparisons discussed in the previous section 3.3, this reflects the ideas of representativity as covering the full heterogeneity or producing mirror images of the target population (cf. section 3.1). In general, assumptions about differences or equalities can be used either as null or alternative hypotheses, but difference tests constitute the standard in applications even for showing equivalence. Assuming that there is no full overlap in the variable(s) of interest, tests may either rely on only the overlapping part or make use of methods for handling missing data (e.g. imputation; cf. Buelens, Burger and van den Brakel, 2018, p. 333; Kim et al., 2018). Applying these tests is of exploratory character in the context of non-probability sampling since they are developed assuming an underlying randomness in terms of the data generating process (cf. e.g. Hawkins, 1981, p. 107; Stephens, 1976, p. 357). This requirement may not generally hold in the context of non-probability samples. Results surely have to be used with caution, even though these tests are applied for missing data in the cited literature, where this limitation is in principle present as well (cf. e.g. section 3.8). An alternative to finding differences between non-probability and reference sample for assessing selectivity is to identify variables $\boldsymbol{Z}$ that are able to predict the selection mechanism. As a technique which is applicable for this purpose, matching is introduced in the following section 3.5.

3.5 Matching With Auxiliary Data

Approaches to assess representativeness in the sense of ideally yielding miniatures of the population are introduced in the previous sections 3.3 and 3.4. One benefit of these methods is that they can help to identify variables that are strongly related to $\boldsymbol{r}^{\mathsf{nps}}$ and, thus, able to represent the response process and its selectivity (cf. equation 3.1). Referring to section 2.2, and in correspondence with figure 2.1, variables used to describe the selection process are denoted by $\boldsymbol{Z}$ in the subsequent discussion. Again, there can be overlaps between $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$, such that the methods discussed above may e.g. be used to identify $\boldsymbol{Z}$ as a subset of columns from $\boldsymbol{X}$. In the ideal case, the variables $\boldsymbol{Z}$ perfectly describe the selection of a non-probability sample, such that the true inclusion probabilities $\boldsymbol{\pi}^{\mathsf{nps}}$ are known for observed $\boldsymbol{Z}$, just as in probability samples (cf. Schouten, Shlomo and Skinner, 2009, p. 11). Any other variable is then conditionally independent from $\boldsymbol{r}^{\mathsf{nps}}$ given $\boldsymbol{Z}$, e.g.

$$(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{r}^{\mathsf{nps}}) \,|\, \boldsymbol{Z} \quad . \tag{3.21}$$

In this case, it can be shown that

$$\mathrm{f}_{\boldsymbol{Y}^{\mathsf{nps}}}(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot}) \;=\; \mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot}) \qquad , \tag{3.22}$$

i.e. the conditional distributions of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ are the same for sample and population (cf. appendix B; Rosenbaum and Rubin, 1983, p. 44). Although linked to the previous sections, this relation 3.21 is closer to the notion of representativity as absence of selective forces discussed in section 3.1. However, conditional independence as required for equality 3.22 is highly dependent on the actual response process, available information and variables of interest. The outlined ideal case is usually hard to achieve in reality and typically not ultimately verifiable (cf. Mercer et al., 2017, p. 255).

Matching can be used to obtain some evidence whether assumption 3.21 is true since equality 3.22 should hold in that case. If it does, identical values in $\boldsymbol{Z}$ imply equal

distributions in $\boldsymbol{Y}$, regardless of whether an observation is in the non-probability sample or not (cf. Rosenbaum and Rubin, 1983, p. 45). Consequently, comparing observations with coinciding or similar values for $\boldsymbol{Z}$ in- and outside the non-probability sample can provide an indication for conditional independence. As before, some reference data set is required for comparison, such that selectivity assessment is again not feasible for variables $\boldsymbol{Y}$ when these are known exclusively for the non-probability sample. As a consequence, auxiliary variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ are required to use matching for selectivity assessment. Moreover, $\boldsymbol{X}$ must contain variables that are not included in $\boldsymbol{Z}$, which is e.g. the case when not all overlapping variables between a non-probability and reference sample are part of $\boldsymbol{Z}$ (cf. figure 2.1). If auxiliary variables $\boldsymbol{X}$ are related to $\boldsymbol{Y}$, they may be substituted as a proxy for checking equation 3.22. Hence, $\mathrm{f}_{\boldsymbol{X}^{\mathsf{nps}}}(\boldsymbol{x}_{i\cdot} \mid \boldsymbol{z}_{i\cdot})$ and $\mathrm{f}_{\boldsymbol{X}}(\boldsymbol{x}_{i\cdot} \mid \boldsymbol{z}_{i\cdot})$ should not differ under assumption 3.21. Since both $\boldsymbol{X}$ and $\boldsymbol{Z}$ are auxiliary variables for which information in- and outside the non-probability sample is available as before, the two densities can be evaluated and compared in this setting to check whether conditional independence holds (cf. Biffignandi and Pratesi, 2003, p. 5; Biffignandi and Bethlehem, 2012, p. 371; Buelens, Burger and van den Brakel, 2018, p. 325; Heckman et al., 1998, p. 1021; Mercer et al., 2017, p. 264).

Elements with coinciding or similar values for $\boldsymbol{Z}$ in- and outside the non-probability sample are determined for matching. The set of elements outside the non-probability sample that are matched to an element i in the non-probability sample is defined as

$$\mathcal{J}^{(i)} \coloneqq \{j \notin \mathcal{S}^{\mathsf{nps}} : \delta(\boldsymbol{z}_{i\cdot}, \boldsymbol{z}_{j\cdot}) \leq a_i\} \tag{3.23}$$

for all $i \in \mathcal{S}^{\mathsf{nps}}$, using some prespecified distance measure $\delta : \mathbb{R}^{1\times q} \times \mathbb{R}^{1\times q} \to \mathbb{R}_{\geq 0}$ and a vector of boundaries $\boldsymbol{a} = \begin{bmatrix} a_1 & \dots & a_{n^{\mathsf{nps}}} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}_{\geq 0}^{n^{\mathsf{nps}}}$ (cf. Stuart, 2010, pp. 5 ff).

When matches are exact, it holds that $\boldsymbol{z}_{i\cdot} = \boldsymbol{z}_{j\cdot}$ for all $j \in \mathcal{J}^{(i)}$. As a consequence, the distributions of $\boldsymbol{x}_{i\cdot}$ and corresponding matched rows $\boldsymbol{X}_{\mathcal{J}^{(i)}\cdot}$ are equal under conditional independence of $\boldsymbol{X}$ and $\boldsymbol{r}^{\mathsf{nps}}$ given $\boldsymbol{Z}$ (cf. equality 3.22). Since $\boldsymbol{x}_{i\cdot}$ and $\boldsymbol{X}_{\mathcal{J}^{(i)}\cdot}$ are both observed, their (dis-)similarities can be assessed to check whether they actually follow the same distribution. Such comparisons often concern the mean, e.g. when estimating treatment effects in observational studies (cf. Cochran and Chambers, 1965, pp. 244 f; Rosenbaum and Rubin, 1983, pp. 49 ff). In analogy, the bias of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$ in the non-probability sample can be approximated by

$$\begin{aligned} \mathbf{Bias}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right) &\approx \widehat{\mathrm{E}}\left(\widehat{\mathrm{E}}\left(\boldsymbol{x}_{i\cdot} \mid \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{nps}} = 1\right) - \widehat{\mathrm{E}}\left(\boldsymbol{x}_{i\cdot} \mid \boldsymbol{z}_{i\cdot}\right) \,\middle|\, r_i^{\mathsf{nps}} = 1\right) \\ &\approx \widehat{\mathrm{E}}\left(\boldsymbol{x}_{i\cdot} - \widehat{\mathrm{E}}\left(\boldsymbol{X}_{\mathcal{J}^{(i)}\cdot}\right) \,\middle|\, r_i^{\mathsf{nps}} = 1\right) \quad , \end{aligned} \tag{3.24}$$

assuming unbiasedness of the reference sample to estimate $\widehat{\mathrm{E}}\left(\widehat{\mathrm{E}}\left(\boldsymbol{x}_{i\cdot} \mid \boldsymbol{z}_{i\cdot}\right)\right)$ (cf. Biffignandi and Pratesi, 2003; Mercer et al., 2017, p. 264; Rivers, 2007). However, matching can also be used to compare further aspects of distributions, such as variances and measures of dependency for variables $\boldsymbol{X}$ (cf. Iacus, King and Porro, 2012, p. 2; Stuart, 2010, p. 11). In that regard, matching resembles the comparisons of auxiliary variables discussed in section 3.3 but considers differences conditional on $\boldsymbol{Z}$ rather than unconditional ones.

Exact matches are typically obtained by setting $a_i = 0$ since it holds for the most common choices of δ that $\delta(\boldsymbol{z}_{i\cdot}, \boldsymbol{z}_{j\cdot}) = 0$ iff $\boldsymbol{z}_{i\cdot} = \boldsymbol{z}_{j\cdot}$. However, the number of possible values for $\boldsymbol{z}_{i\cdot}$ and $\boldsymbol{z}_{j\cdot}$ grows tremendously with both the number of possible values for each variable

and the number of variables. Depending on the quantity and nature of variables $\boldsymbol{Z}$, it is therefore often not feasible to find identical matches (cf. Cochran, Moses and Mosteller, 1983, pp. 78 ff; Rosenbaum and Rubin, 1983, pp. 49 ff). As a consequence, various approaches exist to construct matches for pairs of values where $\boldsymbol{z}_{i\cdot} \approx \boldsymbol{z}_{j\cdot}$ rather than $\boldsymbol{z}_{i\cdot} = \boldsymbol{z}_{j\cdot}$. These approaches emerge from different choices of δ and $\boldsymbol{a}$ (cf. Biffignandi and Bethlehem, 2012, pp. 371 f; Mercer et al., 2017, p. 262). A short summary of selected methods is given below, Stuart (2010) provides an overview in greater detail.

As outlined above, *exact matching* arises from setting $\boldsymbol{a} = \boldsymbol{0}_{n^{\text{nps}} \times 1}$, as long as the choice of δ is such that $\delta\left(\boldsymbol{z}_{i\cdot}, \boldsymbol{z}_{j\cdot}\right) = 0$ iff $\boldsymbol{z}_{i\cdot} = \boldsymbol{z}_{j\cdot}$. The original idea of *caliper matching* is to use

$$\delta\left(\boldsymbol{z}_{i\cdot}, \boldsymbol{z}_{j\cdot}\right) \;=\; \text{Abs}\left(\boldsymbol{z}_{i\cdot} - \boldsymbol{z}_{j\cdot}\right) \tag{3.25}$$

in combination with some predefined vector $\boldsymbol{a}$, which often is a scalar multiple of a vector of ones (cf. Cochran and Rubin, 1973, pp. 420 f). Application of the absolute value and comparison with a_i has to be done element-wise. Therefore, this approach does not generally account for different scales of variables in $\boldsymbol{Z}$, which is why it is mainly used when $\boldsymbol{Z}$ contains a single column and/or in conjunction with other methods, such as propensity score matching (cf. section 3.6; Rosenbaum and Rubin, 1985, p. 37; Rubin and Thomas, 2000). One way for normalizing the distance for multiple variables in $\boldsymbol{Z}$ is to use *generalized Mahalanobis matching*, which is defined by

$$\delta\left(\boldsymbol{z}_{i\cdot}, \boldsymbol{z}_{j\cdot}\right) \;=\; \left(\boldsymbol{z}_{i\cdot} - \boldsymbol{z}_{j\cdot}\right) \left(\left(\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}^{\text{u}}} \left(\boldsymbol{w}^{\text{u}} \right) \right)^{0.5} \right)^{\mathsf{T}} \boldsymbol{C} \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}^{\text{u}}} \left(\boldsymbol{w}^{\text{u}} \right) \right)^{0.5} \right)^{-1} \left(\boldsymbol{z}_{i\cdot} - \boldsymbol{z}_{j\cdot} \right)^{\mathsf{T}} \quad . \tag{3.26}$$

Here, $\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}^{\text{u}}} \left(\boldsymbol{w}^{\text{u}} \right) \right)^{0.5}$ denotes the Cholesky decomposition of the covariance matrix that is estimated from the union of non-probability and reference data outlined in definition 3.2. Furthermore, $\boldsymbol{C} \in \mathbb{R}^{q \times q}$ is a diagonal weighting matrix to quantify the relevance of different variables in $\boldsymbol{Z}$. If it is the identity matrix, equation 3.26 reduces to the classical Mahalanobis distance. To include categorical variables, dummy coding can be applied (cf. Diamond and Sekhon, 2013, p. 934; Rubin, 1979, p. 319).

To sum up, if $\boldsymbol{Z}$ is valuable for providing conditional independence between the non-probability selection process and $\boldsymbol{X}$, the distribution of $\boldsymbol{X}$ given $\boldsymbol{Z}$ is the same for non-probability sample and population. This can be checked by means of matching, where elements in- and outside the non-probability sample that exhibit equal or similar values in $\boldsymbol{Z}$ are joined and compared. If $\boldsymbol{X}$ is related to $\boldsymbol{Y}$, differences in matched values of $\boldsymbol{X}$ can provide an indication for (non-)selectivity with regard to $\boldsymbol{Y}$ when controlling for $\boldsymbol{Z}$. As outlined above, finding similar units becomes very challenging with growing number of columns or presence of continuous variables in $\boldsymbol{Z}$, which is a major difficulty with the discussed methods (cf. Rosenbaum and Rubin, 1983, p. 49; Stuart, 2010, p. 6). Besides Mahalanobis matching, two further approaches deal with this challenge. The idea of *coarsened exact matching* is quite similar to exact matching and closely related to the concept of sub-classification (cf. Cochran and Chambers, 1965, pp. 243 f; Cochran, 1968). Observed values of $\boldsymbol{Z}$ are coarsened before matching, e.g. by constructing age classes or less detailed occupational categories. Exact matches are then determined based on these coarsened values (cf. Iacus, King and Porro, 2009; 2011; 2012).

The still most common matching methods are based on *propensity scores* (or *response propensities*) and likewise reduce the outlined dimensionality issue in advance (cf. e.g. King

and Nielsen, 2019, p. 435; Pearl, 2010, p. 114; Rosenbaum and Rubin, 1983). Response propensities are introduced in the following section 3.6.

3.6 Modeling the Participation Process

Propensity scores are commonly used to match similar elements without using a potentially large number of matching variables. These scores are defined as the conditional probability of element i being in the non-probability sample nps given its covariates $\boldsymbol{z}_{i\cdot}$:

$$p_i^{\text{nps}} \coloneqq \mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \mid \boldsymbol{z}_{i\cdot}\right) \quad \text{for all} \quad i \in \mathcal{S}^{\mathrm{P}} \qquad . \tag{3.27}$$

The vector of response propensities $\boldsymbol{p}^{\text{nps}} \in [0;1]^{\mathrm{N}}$ somewhat resembles the inclusion probabilities $\boldsymbol{\pi}^{\mathsf{ps}}$ that could be used in case of a random sample ps (cf. equation 2.3), but both need to be distinguished explicitly. Since inclusion probabilities are induced by the sampling design, the design variables influencing them are predefined, such that $\boldsymbol{\pi}^{\mathsf{ps}}$ is exactly known. Response propensities, on the other hand, are simply conditional probabilities given a set of variables $\boldsymbol{Z}$, which only in ideal (and mostly theoretical) cases are able to perfectly describe the non-probability sample selection process. In addition, the true response propensities are typically unknown and cannot be estimated from a non-probability sample alone (cf. Schouten, Cobben and Bethlehem, 2009, p. 105; Schouten, Shlomo and Skinner, 2011, p. 234). Due to these differences, the term 'propensity' is used for differentiation from any (assumed or manifest) true inclusion probability for the non-probability sample (cf. e.g. Shlomo et al., 2009a, p. 6). The naming differentiation within the class of propensity scores, e.g. between response or participation propensities (cf. e.g. Little, 1988b, p. 293; Lynn, 2014), is of rather minor importance for presenting the corresponding theory and application. Therefore, it is disregarded in the following, and the terms are used synonymously.

As for matching in general, the original application of propensity scores stems from observational studies. In this context, the propensities are used to analyze differences between two quasi-experimental groups, each exclusively being subject to *one* treatment, for which random assignment is not possible (cf. Rosenbaum and Rubin, 1983; Rubin, 1973; 1974; 1979; Rubin and Thomas, 1996). In the meanwhile, it is a widespread practice to use response propensities also for non-response adjustments in probability samples. The key interest here is the potential non-response bias caused by differences between respondents and non-respondents (cf. Buelens et al., 2012, p. 8; Little, 1986; 1988b; Kott, 2006, p. 141). A similar reasoning applies to non-probability samples, where differences between observed and unobserved values of $\boldsymbol{Y}$ are of interest. As discussed in chapter 2, non-response can be interpreted as a form of non-probability sampling, and response propensities are commonly used to account for other types of non-probability sampling as well (cf. e.g. Biffignandi and Bethlehem, 2012, p. 368; Baker et al., 2013b, p. 8; Enderle, Münnich and Bruch, 2013, p. 92; Isaksson and Lee, 2005, p. 3143; Loosveldt and Sonck, 2008, p. 93; Valliant and Dever, 2011, p. 115). Their importance is motivated by the fact that if $\boldsymbol{Y}$ and $\boldsymbol{r}^{\mathsf{nps}}$ are conditionally independent given $\boldsymbol{Z}$, this holds as well when conditioning on $\boldsymbol{p}^{\text{nps}}$ instead of $\boldsymbol{Z}$. Consequently, $\boldsymbol{Z}$ can be replaced by the propensity scores for the purpose of matching, which reduces the problem to a single matching variable (cf. appendix B; Rosenbaum and Rubin, 1983, p. 45).

However, true propensities are typically unknown, and one sample constitutes only a single realization of $\boldsymbol{r}^{\mathsf{nps}}$ (cf. Schouten, Cobben and Bethlehem, 2009, p. 105; Schouten, Shlomo

and Skinner, 2011, p. 234). Therefore, a common approach is to specify a statistical or machine learning model to obtain estimates

$$\widehat{p}_i^{\text{nps}} \;= \widehat{\mathrm{P}}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right) \tag{3.28}$$

for the true response propensities defined in equation 3.27 (cf. Little, 1988b, p. 293; Rosenbaum and Rubin, 1983; 1985). Various models and fitting strategies are available for predicting the probabilities of the binary variable $\boldsymbol{r}^{\mathsf{nps}}$ outlined in equality 3.28. Since such models are used in many ways for dealing with non-probability samples, a general overview of model formulation and estimation strategies is given in section 5.1. Briefly spoken, the most common way to model response propensities is through parametric models which are fit by maximum likelihood estimation, especially the binary logistic regression models discussed in section 5.1.3 (cf. e.g. Berkson, 1944; Isaksson and Lee, 2005, p. 3146; Rosenbaum and Rubin, 1983, p. 47; Schouten, Shlomo and Skinner, 2011, p. 238). Besides and beyond these, non-parametric and machine learning models, examples of which are presented in sections 5.1.4 to 5.1.9, are increasingly considered for estimation of propensity scores as well (cf. e.g. Brookhart et al., 2006, p. 1151; Buskirk and Kolenikov, 2015; Hirano, Imbens and Ridder, 2003, p. 1161; Lee, Lessler and Stuart, 2010, pp. 337 ff). Note that for fitting any of these models, it is required that there are observed values of $\boldsymbol{Z}$ in- and outside the non-probability sample, just as discussed and presumed in the previous section 3.5. Without such information, the observed outcome $\boldsymbol{r}^{\mathsf{nps}}$ to be modeled would not have any variance, and a model could, thus, not be fit (cf. Schonlau et al., 2009, p. 294; Schouten, Cobben and Bethlehem, 2009, p. 105). The outcome of the model that is relevant to the current context is the vector of estimated conditional probabilities $\widehat{\boldsymbol{p}}^{\text{nps}}$ corresponding to equation 3.28 (cf. also section 5.2.1).

In summary, by matching on such estimated response propensities rather than on $\boldsymbol{Z}$, the complexity is reduced from multiple to a single matching variable. Nevertheless, general response propensities as well as their predictions are continuous variables. Therefore, it is in many cases again infeasible to obtain exact matches. As before (cf. section 3.5), distance functions can be used to match values that are similar rather than equal, or some classification rule can be applied to the propensities to obtain a categorical variable for exact matching (cf. Rosenbaum and Rubin, 1983, pp. 51 f). As indicated in the previous section 3.5, an alternative that uses ideas similar to classification but does not rely on estimated propensities is *coarsened exact matching*. By matching exactly on coarsened values of $\boldsymbol{Z}$ rather than on estimated propensities that combine all columns of $\boldsymbol{Z}$ into a single vector, the reliance on model assumptions for $\boldsymbol{p}^{\text{nps}}$ is circumvented. Such assumptions are essential for nearly all applications of propensity score matching (cf. Cochran, 1968; Iacus, King and Porro, 2009; 2011; 2012; King and Nielsen, 2019). Despite criticism of overly depending on modeling assumptions, matching on estimated propensity scores is still the most common method for this purpose when $\boldsymbol{Z}$ contains a large number of columns (cf. e.g. King and Nielsen, 2019, p. 435; Pearl, 2010, p. 114).

Furthermore, response propensities are useful in a broader sense, e.g. for weighting non-probability samples (cf. section 5.2.1). Another application of these scores, which is relevant for assessing selectivity of non-probability samples, is the usage of propensities for estimating representativity indicators. These are introduced in the following section 3.7.

3.7 Representativity Indicators

As an alternative for measuring representativity in the context of non-response, Schouten, Cobben and Bethlehem (2009) as well as Skinner et al. (2009) introduce the *representativity indicators* (*R-indicators*). Although developed for this special case of missing data, the framework can be at least experimentally applied to non-probability sampling in general (cf. Petrucci and Rocco, 2019). R-indicators basically depend on the variance of response propensities, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{p}^{\text{nps}}}\left(\widetilde{\boldsymbol{w}}\right)$. Since true propensities $\boldsymbol{p}^{\text{nps}}$ are rarely known, estimated R-indicators are computed from model predictions $\widehat{\boldsymbol{p}}^{\text{nps}}$ (cf. section 3.6). The estimated overall R-indicator as a function $\widehat{\text{R}} : [0;1]^{n^{\text{nps}}} \to [0;1]$ is then defined by

$$\widehat{\text{R}}\left(\widehat{\boldsymbol{p}}^{\text{nps}}\right) \coloneqq 1 - 2 \cdot \left(\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{p}}^{\text{nps}}}\left(\widetilde{\boldsymbol{w}}\right)\right)^{\frac{1}{2}} \tag{3.29}$$

(cf. Schouten, Cobben and Bethlehem, 2009; Skinner et al., 2009).[2] As before, $\widetilde{\boldsymbol{w}}$ is an arbitrary vector of weights for the non-probability sample. The basic idea of definition 3.29 is that if there is no dependency between $\boldsymbol{r}^{\mathsf{nps}}$ and the variables $\boldsymbol{Z}$ used for the propensity model, the variance of the response propensities is zero (cf. equation 3.28; Schouten et al., 2016, pp. 730 f). Consequently, $\widehat{\text{R}}\left(\widehat{\boldsymbol{p}}^{\text{nps}}\right)$ being close to one is an indication for absence of selective forces with regard to $\boldsymbol{Z}$ as a whole (cf. section 3.1). As before, weights $\widetilde{\boldsymbol{w}}$ for the non-probability sample may all be ones or obtained from the more refined methods discussed in section 5.2.

In extension to the overall R-indicator, Schouten and Bethlehem (2009) as well as Schouten, Shlomo and Skinner (2011) define partial R-indicators, which are meant to assess the lack of representativity with respect to single variables. They propose unconditional as well as conditional partial R-indicators, which are respectively denoted by $\widehat{\text{R}}_{\text{u}}, \widehat{\text{R}}_{\text{c}} : [0;1]^{n^{\mathsf{nps}}} \times \mathbb{R}^{n^{\mathsf{nps}}} \to [0;0.5]$. Both are computed by means of variance decomposition, for $\boldsymbol{Z}$ consisting of categorical variables only. The estimated *unconditional* partial R-indicator of a categorical variable $\boldsymbol{z}_{\cdot j}^{\mathsf{nps}}$ is calculated from the propensity score's variation attributable to this variable. Denoting the K possible values of $\boldsymbol{z}_{\cdot j}^{\mathsf{nps}}$ by $\boldsymbol{a} \coloneqq \left[a_1, \ldots, a_K\right]^{\mathsf{T}}$, this R-indicator is defined by the between standard deviation of the propensities given $\boldsymbol{z}_{\cdot j}$:

$$\widehat{\text{R}}_{\text{u}}\left(\widehat{\boldsymbol{p}}^{\text{nps}}, \boldsymbol{z}_{\cdot j}^{\mathsf{nps}}\right) \coloneqq$$

$$\left(\sum_{k=1}^{K} \frac{\widehat{\text{N}}\left(\widetilde{\boldsymbol{w}} \circ \mathbb{I}\left(\boldsymbol{z}_{\cdot j}^{\mathsf{nps}} = a_k\right)\right)}{\widehat{\text{N}}\left(\widetilde{\boldsymbol{w}}\right)} \cdot \left(\widehat{\boldsymbol{\mu}}_{\widehat{\boldsymbol{p}}^{\text{nps}}}\left(\widetilde{\boldsymbol{w}} \circ \mathbb{I}\left(\boldsymbol{z}_{\cdot j}^{\mathsf{nps}} = a_k\right)\right) - \widehat{\boldsymbol{\mu}}_{\widehat{\boldsymbol{p}}^{\text{nps}}}\left(\widetilde{\boldsymbol{w}}\right)\right)^2\right)^{\frac{1}{2}}. \tag{3.30}$$

The estimated *conditional* partial R-indicator of variable $\boldsymbol{z}_{\cdot j}^{\mathsf{nps}}$ is conditioned on the other variables in $\boldsymbol{Z}$. It is basically the within standard deviation of the propensity scores given

[2] Note that due to using the bias corrected rather than the ML estimate of the variance, this definition by Schouten, Cobben and Bethlehem (2009, pp. 103 ff) and Skinner et al. (2009) can actually result in estimated global R-indicators below zero. The lowest theoretically possible estimate for the global R-indicator where it is still defined is $1 - 2 \cdot \sqrt{0.5} \approx -0.41$ (cf. equation 2.18f). This is ignored by the authors, who assume $\widehat{\text{R}}\left(\widehat{\boldsymbol{p}}^{\text{nps}}\right)$ to lie between zero and one. However, such values below zero typically do not occur for real samples since they require very few observations or extremely odd weights $\widetilde{\boldsymbol{w}}$. A similar reasoning applies to the partial R-indicators as well.

all other variables used to calculate them. Denote by $\boldsymbol{Z}_{\cdot\mathcal{J}}$ for $\mathcal{J} \coloneqq \{1, \ldots, q\} \setminus j$ the matrix of all $\boldsymbol{Z}$-variables except the j-th column. For $l = 1, \ldots, L$ unique possible values of $\boldsymbol{Z}_{\cdot\mathcal{J}}$ denoted by $\boldsymbol{B} \coloneqq \left[\boldsymbol{b}_{1\cdot}^{\mathsf{T}}, \ldots, \boldsymbol{b}_{L\cdot}^{\mathsf{T}}\right]^{\mathsf{T}} \in \mathbb{R}^{L \times (q-1)}$, the conditional partial R-indicator for $\boldsymbol{z}_{\cdot j}^{\mathsf{nps}}$ is defined as

$$\widehat{\mathrm{R}}_{\mathrm{c}}\left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}, \boldsymbol{z}_{\cdot j}^{\mathsf{nps}}\right) \coloneqq \left(\sum_{l=1}^{L} \frac{\widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}} \circ \mathbb{I}\left(\boldsymbol{Z}_{\cdot\mathcal{J}}^{\mathsf{nps}} = \boldsymbol{b}_{l\cdot}\right)\right)}{\widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}\right)} \cdot \widetilde{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{p}}^{\mathrm{nps}}}\left(\widetilde{\boldsymbol{w}} \circ \mathbb{I}\left(\boldsymbol{Z}_{\cdot\mathcal{J}}^{\mathsf{nps}} = \boldsymbol{b}_{l\cdot}\right)\right) \right)^{\frac{1}{2}}, \tag{3.31}$$

where the indicator function is applied row-wise. Shlomo et al. (2009b) show that equation 3.29 is biased for the corresponding population quantity based on the true response propensity $\boldsymbol{p}^{\mathrm{nps}}$ and propose a bias correction, which Heij, Schouten and Shlomo (2010) extend to the partial indicators defined in equations 3.30 and 3.31. However, the correction is derived with respect to a specific model for $\widehat{\boldsymbol{p}}^{\mathrm{nps}}$ and by using design-based estimates for means and variances (cf. section 2.2; Shlomo et al., 2009a, pp. 42 ff; Shlomo, Skinner and Schouten, 2012, p. 205; Heij, Schouten and Shlomo, 2010). Therefore, it is not generally applicable in the present context of non-probability sampling.

In the following paragraphs, potential minor modifications and extensions for the R-indicators are discussed, which especially concern their interpretability and applicability in the context of non-probability samples. In terms of interpretation, note that the ranges of R-indicators are restricted by the inequalities

$$\begin{aligned} 0 \;&\leq\; \widehat{\mathrm{R}}\left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}\right) \leq 1 \\ 0 \;&\leq\; \widehat{\mathrm{R}}_{\mathrm{c}}\left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}, \boldsymbol{z}_{\cdot j}^{\mathsf{nps}}\right), \widehat{\mathrm{R}}_{\mathrm{u}}\left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}, \boldsymbol{z}_{\cdot j}^{\mathsf{nps}}\right) \leq \frac{1 - \widehat{\mathrm{R}}\left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}\right)}{2} \;\leq 0.5 \end{aligned} \quad , \tag{3.32}$$

with higher values corresponding to more representativity in case of the overall R-indicator $\widehat{\mathrm{R}}\left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}\right)$. In contrast, larger values for the partial R-indicators imply more contribution of the respective variable to non-representativeness and, hence, less representativeness with regard to this variable (cf. Schouten, Cobben and Bethlehem, 2009, pp. 104 ff; Schouten, Shlomo and Skinner, 2011, pp. 236 f). This can be counterintuitive when interpreting the indicators. As a possible remedy to make direction and scale of the three types of indicators less ambiguous, it may be sensible to transform the partial indicators in the same manner as the overall one, i.e. by using $\widetilde{\mathrm{R}}_{\mathrm{k}}\left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}, \boldsymbol{z}_{\cdot j}^{\mathsf{nps}}\right) \coloneqq 1 - 2 \cdot \widehat{\mathrm{R}}_{\mathrm{k}}\left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}, \boldsymbol{z}_{\cdot j}^{\mathsf{nps}}\right)$ for $k \in \{u; c\}$. In that case, a value of one corresponds to perfect representativity with respect to the underlying propensity model for all R-indicators.

So far, R-indicators were proposed for categorical $\boldsymbol{Z}$-variables only. In such a setting, the indicators can be written as (functions of) total, between and within variance of groups defined by sub-matrices of $\boldsymbol{Z}$, whereas continuous variables are considered a future research topic (cf. Heij, Schouten and Shlomo, 2010, p. 8; Schouten and Bethlehem, 2009, p. 2; Schouten, Shlomo and Skinner, 2011, p. 233). Such continuous variables can play a role when assessing selectivity of non-probability samples. While the overall R-indicator is readily applicable for continuous variables, the framework for partial indicators needs to be slightly extended for considering such variables. Note that the idea of variance decomposition on which partial R-indicators are based is quite generally applicable: equations 3.30 and 3.31 are functions of explained and unexplained variability in an

ANOVA context. Such variance decompositions can be expressed by linear regression models, which are described in section 5.1.2 (cf. also equations 3.6 and 3.14; Faraway, 2002, p. 168). To assess selectivity for continuous variables, it, hence, seems sensible to express partial R-indicators in terms of explained and residual variances of linear models.

In summary, R-indicators can be used to evaluate (non-)representativity for the set of variables $\boldsymbol{Z}$ as a whole or for single variables $\boldsymbol{z}_{\cdot j}$. This is an advantage when aiming at aggregate measures of representativity that are not specific to the target variables. On the other hand, it constitutes a limitation since resulting R-indicators are different depending on the choice of variables $\boldsymbol{Z}$ (cf. Schouten, Shlomo and Skinner, 2011, p. 235) but not for different target variables $\boldsymbol{Y}$. Yet, a non-probability sample can exhibit differing degrees of selectivity for different target variables (cf. Bethlehem, 2008a, p. 10; Buelens et al., 2014, p. 4; Shlomo, Skinner and Schouten, 2012, p. 202). Therefore, Schouten, Cobben and Bethlehem, 2009, p. 107 as well as Shlomo, Skinner and Schouten (2012, p. 203) stress an additional use for R-indicators in determining the magnitude of possible biases in estimation with regard to $\boldsymbol{Y}$. This is discussed as part of the following section 3.8.

3.8 Quantifying the MSE

In the previous sections 3.3 to 3.7, methods for assessing selectivity of non-probability samples are discussed. These can help to identify variables that (partially) explain the selection process and quantify selectivity of non-probability samples. Based on some of these considerations, Meng (2018) and Schouten (2007) introduce a framework for representing the accuracy of an estimator by decomposing its error. To again allow for arbitrary weights $\widetilde{\boldsymbol{w}} \in \mathbb{R}^{n^{\mathsf{nps}}}$ in the non-probability sample, a weighted version $\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \in \mathbb{R}^{\mathrm{N}}$ of the sample inclusion indicator $\boldsymbol{r}^{\mathsf{nps}}$ is used. It is defined by

$$\widetilde{r}_i^{\,\mathsf{nps}} \;\coloneqq\; \begin{cases} \widetilde{w}_i & \text{if} \quad i \in \mathcal{S}^{\mathsf{nps}} \\ 0 & \text{else} \quad . \end{cases} \tag{3.33}$$

On this basis, many weighted estimators (e.g. for means, other moments, or distributions) can be written as design linear, i.e.

$$\widehat{\boldsymbol{\vartheta}} \;= \widehat{\boldsymbol{\vartheta}}(\mathcal{S}^{\mathsf{nps}}) = \left(\sum_{i \in \mathcal{S}^{\mathrm{P}}} \widetilde{r}_i^{\,\mathsf{nps}} \cdot \boldsymbol{t}\left(\boldsymbol{y}_{i\cdot}\right) \right) \Big/ \sum_{j \in \mathcal{S}^{\mathrm{P}}} \widetilde{r}_j^{\,\mathsf{nps}} \qquad , \tag{3.34}$$

by adequate choice of a transformation function $\boldsymbol{t} : \mathbb{R}^{1\times o} \to \mathbb{R}^{1\times h}$ for given $h \in \mathbb{N}$ and $\widetilde{\boldsymbol{w}}$. The deviation between estimated and true statistics $\widehat{\vartheta}_k$ and ϑ_k for all $k = 1, \ldots, h$ elements of $\widehat{\boldsymbol{\vartheta}} \in \mathbb{R}^h$ in such cases can then be expressed as

$$\widehat{\vartheta}_k - \vartheta_k \;=\; \boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}} \cdot \sqrt{\frac{1 - fr^{\mathsf{nps}}}{fr^{\mathsf{nps}}}} \cdot \sqrt{\mathrm{V}\left(\boldsymbol{t}_k\left(\boldsymbol{Y}\right)\right)} \cdot \sqrt{1 + \frac{\left(\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)\right)^2}{1 - fr^{\mathsf{nps}}}} \quad . \tag{3.35}$$

The coefficient of variation for weights applied to the non-probability sample is defined by

$$\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right) \;\coloneqq \sqrt{\mathrm{V}\left(\widetilde{\boldsymbol{w}}\right) \Big/ \left(\mathrm{E}\left(\widetilde{\boldsymbol{w}}\right)\right)^2} \qquad . \tag{3.36}$$

This coefficient is zero if the weights $\widetilde{\boldsymbol{w}}$ are constant over all sampled units. Furthermore, $fr^{\mathsf{nps}} \coloneqq n^{\mathsf{nps}}/\mathrm{N}$ is again the sampling fraction, and $\boldsymbol{t}_k(\boldsymbol{Y})$ denotes the k-th component or column of $\boldsymbol{t}(\boldsymbol{Y})$, for which the correlation with $\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}$ is given by $\boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}$ (cf. Fuller, 2009, p. 6; Meng, 2018, p. 702). Similar expressions for the context of non-response are proposed by Bethlehem (1988, p. 254), Särndal and Lundström (2005, p. 92) and Schouten (2007, p. 57).

As discussed by Meng (2018, p. 690), the components of equality 3.35 determine the estimation error in arbitrary samples, which is appealing in terms of interpretability. Referred to as *data quality* is the correlation between sample inclusion and target quantity $\boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}$, which is also called the *"data defect correlation"* (Meng, 2018, p. 691). The *data quantity* expressed by $\sqrt{(1 - fr^{\mathsf{nps}})/fr^{\mathsf{nps}}}$ depends on the sampling fraction. *Problem difficulty* is denoted by $\sqrt{\mathrm{V}\left(\boldsymbol{t}\left(\boldsymbol{Y}\right)\right)}$, which is the standard deviation for the quantity of interest. Since varying weights increase an estimator's variance and hence error, the additional factor $\left(1 + \left(\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)\right)^2 \big/ \left(1 - fr^{\mathsf{nps}}\right)\right)^{0.5}$ accounts for variation of weights. Despite inducing higher variability, the use of weights can often be sensible to reduce the data defect correlation (cf. section 5.2; Meng, 2018, pp. 690 ff).

Conditional on the sample size, only $\boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}$ and $\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)$ depend on the sample selection process, such that the MSE of an estimator $\widehat{\vartheta}_k$ can be written as

$$\begin{aligned} \mathrm{MSE}\left(\widehat{\vartheta}_k\right) &= \mathrm{E}\left(\left(\widehat{\vartheta}_k - \vartheta_k\right)^2\right) \\ &= \mathrm{E}\left(\boldsymbol{\rho}^2_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}} \cdot \left(1 + \frac{\left(\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)\right)^2}{1 - fr^{\mathsf{nps}}}\right)\right) \cdot \frac{1 - fr^{\mathsf{nps}}}{fr^{\mathsf{nps}}} \cdot \mathrm{V}\left(\boldsymbol{t}_k\left(\boldsymbol{Y}\right)\right) \quad , \end{aligned} \tag{3.37}$$

where expectation is with respect to the sampling mechanism (possible values of $\boldsymbol{r}^{\mathsf{nps}}$; cf. definition 2.12). From equation 3.37, it becomes evident that unless a probability sampling design is used to control $\boldsymbol{\rho}^2_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}$, this data defect correlation and – by definition of the sampling fraction – the population size N determine estimation quality far more than the sample size (cf. Meng, 2018, pp. 695 ff). Underlining the discussion in section 2.1, more observations alone do, hence, not guarantee better estimates when using non-probability samples.

As intuitive and appealing as equations 3.35 and 3.37 are, an obvious limitation with regards to realized non-probability samples is their dependency on $\boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}$. This population correlation coefficient between $\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}$ and a quantity of interest $\boldsymbol{t}_k\left(\boldsymbol{Y}\right)$ is generally unknown and difficult to estimate. Direct computation of $\boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}$ would require that $\boldsymbol{Y}$ was available for the whole population, in which case no sample would be needed at all. Even when using a reference data set, $\boldsymbol{Y}$ is typically known only for the non-probability sample, in which $\boldsymbol{r}^{\mathsf{nps}}$ is constant and adequate estimation of $\boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}$ thus not possible (cf. definition 3.33). Consequently, Schouten, Cobben and Bethlehem (2009, p. 107) as well as Shlomo, Skinner and Schouten (2012, p. 203) approximate $\boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}$ by the correlation between $\boldsymbol{t}_k\left(\boldsymbol{Y}\right)$ and the estimated response propensities (cf. equation 3.28). They use the worst-case scenario where this correlation is one in magnitude to obtain an upper limit for the possible absolute bias due to non-response as a function of the R-indicator. Going beyond this worst-case assumption, the fact that the correlation of two variables is bounded by their correlation with a third (cf. e.g. McCornack, 1956, p. 343; Yule, 1922, p. 250) is used by Schouten (2007, p. 57) as well as Schouten et al.

(2016, pp. 745 ff) to quantify the unknown correlation's magnitude in form of an interval:

$$\boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}} \;\in\; \boldsymbol{\rho}_{\boldsymbol{z}_{\cdot 1}\,\boldsymbol{t}_k(\boldsymbol{Y})} \cdot \boldsymbol{\rho}_{\boldsymbol{z}_{\cdot 1}\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}} \pm \left(\sqrt{1 - \boldsymbol{\rho}^2_{\boldsymbol{z}_{\cdot l}\,\boldsymbol{t}_k(\boldsymbol{Y})}} \cdot \sqrt{1 - \boldsymbol{\rho}^2_{\boldsymbol{z}_{\cdot l}\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}}\right) \quad . \tag{3.38}$$

Here, $\boldsymbol{\rho}_{\boldsymbol{z}_{\cdot 1}\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}$ and $\boldsymbol{\rho}_{\boldsymbol{z}_{\cdot 1}\,\boldsymbol{t}_k(\boldsymbol{Y})}$ denote the correlations between an auxiliary variable $\boldsymbol{z}_{\cdot l}$ and the weighted response indicator $\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}$ or the k-th component of $\boldsymbol{t}\,(\boldsymbol{Y})$, respectively.

When, as before, information about $\boldsymbol{Z}$ external to the non-probability sample is available, components of expression 3.38 can be estimated from the non-probability and reference data set. In particular, $\widehat{\boldsymbol{\rho}}_{\boldsymbol{z}_{\cdot 1}\,\boldsymbol{t}_k(\boldsymbol{Y})}$ is obtained from the non-probability sample and $\widehat{\boldsymbol{\rho}}_{\boldsymbol{z}_{\cdot 1}\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}}$ from the combined non-probability and reference data set. Similar as before (cf. section 3.6), conditional independence of $\boldsymbol{t}_k\,(\boldsymbol{Y})$ and $\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}$ given $\boldsymbol{z}_{\cdot l}$ is required for unbiased estimation of $\boldsymbol{\rho}_{\boldsymbol{z}_{\cdot 1}\,\boldsymbol{t}_k(\boldsymbol{Y})}$ in this setting. By plugging the boundaries obtained from relation 3.38 into equation 3.37, an estimated MSE-interval can be obtained. For this interval to be sufficiently narrow enough for a meaningful quantification of the MSE (cf. equation 3.37) based on realized samples, it is essential to find a variable $\boldsymbol{z}_{\cdot l}$ that is strongly correlated with quantity of interest $\boldsymbol{t}_k\,(\boldsymbol{Y})$ as well as the response process generating $\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}$. Schouten (2007, pp. 60 ff) reverses this argument to detect such variables by means of the interval's width.

Especially due to its intuitive character for representing the challenges (not only) of non-probability samples, the framework introduced in this section is an appealing case for design linear estimators in the form of equality 3.34. Although various important estimators are not themselves design linear, many of them can be written as functions of such estimators (cf. section 2.2). To make meaningful use of this framework for quantifying precision from a single realized sample, it is important to identify auxiliary variables $\boldsymbol{Z}$ that provide conditional independence of and are highly related to both the quantities of interest and the response process.

Determining such variables is one purpose of the approaches discussed throughout the current chapter 3. The overall objective of these methods is to use auxiliary variables for assessing possible errors in non-probability samples, in particular considering the issues of selectivity and bias. Underpinned by these evaluations, the apparent next step is to examine estimation approaches for non-probability samples that consider and – if possible – compensate potential deviations from representativity (cf. e.g. Bethlehem, 2008b, p. 31; Feild et al., 2006, p. 566; Loosveldt and Sonck, 2008, p. 96; Valliant and Dever, 2011, p. 106). Methods to perform estimation from non-probability samples are therefore presented in chapter 5. As most of those rely on a similar set of mathematical foundations, these required basics are introduced in the following chapter 4 to foster the subsequent discussion.

4 Mathematical and Computational Foundations

In this chapter, the mathematical and computational framework required to describe methods that deal with non-probability samples is introduced. The aim is to give a concise overview of selected concepts, approaches and algorithms that is oriented towards the subsequent chapters. To that end, the current chapter is rather technical and requires some prior computational and algebraic knowledge. More details regarding the underlying mathematical considerations are given in the referred literature. Especially Geiger and Kanzow (2002), Gill, Murray and Wright (1981) as well as Nocedal and Wright (1999) provide valuable and comprehensive overviews.

The presented fundamentals are frequently used in the context of survey statistics and non-probability samples (cf. e.g. Biffignandi and Pratesi, 2003, p. 8; Burgard, Münnich and Rupp, 2019; 2020; Deville, Särndal and Sautory, 1993, p. 1013; Folsom and Singh, 2000, p. 599; Nelder and Wedderburn, 1972, p. 373), and therefore central for discussing and implementing well-established statistical methods. In particular, almost all of the prediction and weighting models considered in the following chapter 5 are fit by means of methods that are presented in the current chapter 4. This also includes (calibrated) semi-parametric neural networks, which are introduced in sections 5.1.9 and 5.2.3, and for which development and implementation heavily rely on these foundations as well.

As a fundamental element, linear programming methods for computationally solving systems of linear equations are introduced in section 4.1. These are required for the methods discussed in section 4.2, which allow performing unconstrained and constrained non-linear optimization.

4.1 Linear Programming

For computationally solving tasks that commonly arise in the context of (non-probability) sampling and estimation, a fundamental component is to find the solution $\boldsymbol{c} \in \mathbb{R}^h$ to an exactly determined system of linear equations defined by

$$\begin{bmatrix} a_{11} & \cdots & a_{1h} \\ \vdots & \ddots & \vdots \\ a_{h1} & \cdots & a_{hh} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_h \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_h \end{bmatrix} \tag{4.1}$$

for a given matrix $\boldsymbol{A} \in \mathbb{R}^{h \times h}$ and a given vector $\boldsymbol{b} \in \mathbb{R}^h$. Various iterative and direct methods exist to solve equation 4.1 for $\boldsymbol{c}$ (cf. Hackbusch, 1994; Saad, 2003). One popular approach is *Gaussian elimination* (cf. Anderson et al., 1999; Demmel, 1997, pp. 38 ff; Li, 2005; Sanderson and Curtin, 2016), which is motivated in section 4.1.1 and formally introduced in section 4.1.2.

4.1.1 Solving Triangular Systems

Consider the special case of $\boldsymbol{A}$ being a lower triangular matrix, such that $a_{ij} = 0$ for all $j > i$. In this setting, system 4.1 can be written as

$$\begin{aligned} a_{11} \cdot c_1 &= b_1 \\ a_{21} \cdot c_1 + a_{22} \cdot c_2 &= b_2 \\ &\vdots \\ a_{h1} \cdot c_1 + a_{h2} \cdot c_2 + \cdots + a_{hh} \cdot c_h &= b_h \quad . \end{aligned} \tag{4.2}$$

The first equation can be solved directly by $c_1 = b_1/a_{11}$, and this result can then be used to successively solve each following equation for one more element of $\boldsymbol{c}$, which is called *forward substitution*.

If matrices $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{h \times s}$ are used instead of vectors $\boldsymbol{b}$ and $\boldsymbol{c}$ (e.g. for matrix inversion), the result constitutes s of such systems in form of equalities 4.2 that can be solved separately. This procedure is formalized in algorithm 2.

Algorithm 2: Forward / backward substitution

1: **Input:** $\boldsymbol{A} \in \mathbb{R}^{h \times h}$; $\boldsymbol{B} \in \mathbb{R}^{h \times s}$
2: Initialize $\boldsymbol{C} = \boldsymbol{0}_{h \times s}$
3: **for** $i = 1, \ldots, h$ **do**
4: **for** $j = 1, \ldots, s$ **do**
5: $c_{ij} \leftarrow \dfrac{b_{ij} - \boldsymbol{a}_{i\cdot}\boldsymbol{c}_{\cdot j}}{a_{ii}}$
6: **end for**
7: **end for**
8: **Return:** $\boldsymbol{C}$

In case of $\boldsymbol{A}$ being an upper triangular matrix, the solution is analogous, except that the sequence in step 3 has to be last-to-first row ($i = h, \ldots, 1$). This is referred to as *backward substitution* (cf. Gill, Murray and Wright, 1981, pp. 3 f).

4.1.2 Gaussian Elimination (LU-factorization)

If $\boldsymbol{A}$ is not a triangular but general non-singular square matrix, it is useful to represent it as a product of two triangular matrices. In that way, the original system 4.1 can be formulated as a case of solving triangular systems described in the previous section 4.1.1: when it holds that

$$\boldsymbol{L}\boldsymbol{U} \stackrel{!}{=} \boldsymbol{A} \tag{4.3}$$

for lower and upper triangular matrices $\boldsymbol{L}, \boldsymbol{U} \in \mathbb{R}^{h \times h}$, the original system 4.1 can be written as

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{L}(\boldsymbol{U}\boldsymbol{c}) = \boldsymbol{b} \quad , \tag{4.4}$$

which is again defined by triangular matrices. Hence, to solve equation 4.4, $\boldsymbol{v} := \boldsymbol{U}\boldsymbol{c} \in \mathbb{R}^h$ can be found as the solution of $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{b}$, based on which $\boldsymbol{U}\boldsymbol{c} = \boldsymbol{v}$ can be solved for $\boldsymbol{c}$ as the solution of equation 4.1.

To enable this strategy, Gaussian elimination can be used to obtain $\boldsymbol{U}$ and $\boldsymbol{L}$ in compliance with equation 4.3. Starting with $\boldsymbol{A}$, it is based on reducing the lower triangular part of $\boldsymbol{A}$ to zeros. This is formally achieved by subtracting a scalar multiple $(a_{ij}/a_{ii}) \cdot \boldsymbol{a}_{i\cdot}$ of row $\boldsymbol{a}_{i\cdot}$ from row $\boldsymbol{a}_{j\cdot}$, assuming that $a_{ii} \neq 0$. Since $a_{ij} - (a_{ij}/a_{ii}) \cdot a_{ii} = 0$, this subtraction results in the new a_{ij} being zero. The desired triangular matrices are constructed by iteratively applying this idea. Columns below the main diagonal of $\boldsymbol{A}$ are reduced to zeros, which is equivalent to finding $\boldsymbol{U} = \boldsymbol{L}^{-1}\boldsymbol{A}$. Determining $\boldsymbol{v} = \boldsymbol{L}^{-1}\boldsymbol{b}$ is equivalent to applying the corresponding changes to $\boldsymbol{b}$. This strategy to reduce system 4.1 to triangular form 4.2 is thus equivalent to the method that is used by most people to solve (small) linear systems by hand: the outlined multiplications of $\boldsymbol{A}$ and $\boldsymbol{b}$ with $\boldsymbol{L}^{-1}$ constitute the reformulations in left- and right-hand side of the original system 4.1 that are required to express one particular element c_j in the i-th equation as a linear combination of the other elements of $\boldsymbol{c}$ (cf. Gill, Murray and Wright, 1981, pp. 33 ff; Golub and Van Loan, 1996, pp. 94 ff).

An obvious problem with this approach occurs when $a_{ii} = 0$ since these preliminary considerations would then result in division by zero. A general solution for this issue is to interchange rows in both $\boldsymbol{A}$ and $\boldsymbol{b}$ to prevent a_{ii} from becoming zero. Such an order permutation is called partial (or row) *pivoting* and formally expressed as multiplication with a permutation matrix $\boldsymbol{P} \in \{0;1\}^{h \times h}$ that contains one exactly once in each row and column and zeros everywhere else. Through pivoting, it is assured that $\boldsymbol{L}$ and $\boldsymbol{U}$ fulfilling $\boldsymbol{PA} = \boldsymbol{LU}$ exist and are non-singular for any non-singular square matrix $\boldsymbol{A}$. Reordering of rows does not alter system 4.4 since the original order is restored when using $\boldsymbol{LUc} = \boldsymbol{Pb}$ to solve the problem for $\boldsymbol{c}$ (cf. Demmel, Gilbert and Li, 1999, pp. 38 ff; Gill, Murray and Wright, 1981, pp. 33 ff; Golub and Van Loan, 1996, pp. 94 ff). The iterative application of these ideas to find $\boldsymbol{L}$ and $\boldsymbol{U}$ is formalized in algorithm 3.

Algorithm 3: Gaussian elimination (LU-factorization)

1: **Input:** $\boldsymbol{A} \in \mathbb{R}^{h \times h}$
2: **Optional:** Use algorithm 4 to initialize $\boldsymbol{P}$ for partial pivoting, otherwise use $\boldsymbol{P} \coloneqq \mathbf{I}_h$
3: Initialize $\boldsymbol{U} \coloneqq \boldsymbol{PA}$ and $\boldsymbol{L}^{-1} \coloneqq \mathbf{I}_h$
4: **for** $i = 1$ to $(h-1)$ **do**
5: Define $\mathcal{J} \coloneqq \{i+1, \ldots, h\}$ and calculate

$$\boldsymbol{B} \coloneqq \mathbf{I}_h - u_{ii}^{-1} \cdot \begin{bmatrix} \mathbf{0}_{i\times 1} \\ \boldsymbol{U}_{\mathcal{J}i} \end{bmatrix} \begin{bmatrix} \mathbf{0}_{1\times(i-1)} & 1 & \mathbf{0}_{1\times(h-i)} \end{bmatrix}$$

6: Update $\boldsymbol{U} \leftarrow \boldsymbol{BU}$ and $\boldsymbol{L}^{-1} \leftarrow \boldsymbol{L}^{-1}\boldsymbol{B}$
7: **end for**
8: **Return:** $\boldsymbol{L}^{-1}$, $\boldsymbol{U}$ and $\boldsymbol{P}$

In the notation introduced in section 2.2, $\boldsymbol{U}_{\mathcal{J}i}$ in algorithm 3 corresponds to the column-vector defined as the sub-matrix of $\boldsymbol{U}$ with rows indexed by $\mathcal{J}$ and a single column indexed by i. Zero and identity matrices of the specified size are denoted by $\mathbf{0}$ and $\mathbf{I}$. Note that here and in the remaining parts of this thesis, certain row- or column-vectors are introduced and treated as matrices for notational and verbal simplicity. In particular, this eases general cases where matrices with one or more rows or columns can occur without requiring further specification. Each iteration of the for-loop in this algorithm 3 reduces

u_{ij} to zero for all rows $i = j+1, \ldots, h$, using the ideas described above. In that way, $\boldsymbol{U}$ is reduced to upper triangular form. The transformations used to achieve this are described by $\boldsymbol{L}^{-1}$ and hence can be inverted by pre-multiplication with $\boldsymbol{L}$, such that one obtains equation 4.3. Since $\boldsymbol{v}$ is found by $\boldsymbol{v} = \boldsymbol{L}^{-1}\boldsymbol{b}$, it is often not necessary to determine $\boldsymbol{L}$ itself. If nevertheless needed, the required inversion of $\boldsymbol{L}^{-1}$ can be based on algorithm 2 since this matrix is triangular (cf. Gill, Murray and Wright, 1981, pp. 33 ff; Trefethen and Bau, 1997, pp. 151 ff).

As outlined above, the LU-factorization works only for *pivoting elements* $u_{ii} \neq 0$, such that partial pivoting in step 2 of algorithm 3 may be required to achieve a valid solution. Therefore, the following algorithm 4 generates the pivoting matrix $\boldsymbol{P}$ to avoid these issues.

Algorithm 4: Partial (row) pivoting

1: **Input:** $\boldsymbol{A} \in \mathbb{R}^{h \times h}$
2: Initialize $\boldsymbol{P} \coloneqq \mathbf{I}_h$
3: **for** $i = 1$ to $(h-1)$ **do**
4: Determine the index of the largest absolute value in the subsequent rows of column i as $b \coloneqq \text{Max}\left(\left\{c : \text{Abs}\left(a^*_{ci}\right) = \text{Max}\left(\text{Abs}\left(\boldsymbol{A}^*_{\mathcal{J}i}\right)\right)\right\}\right)$, where $\mathcal{J} \coloneqq \{i, \ldots, h\}$ and $\boldsymbol{A}^* \coloneqq \boldsymbol{P}\boldsymbol{A}$
5: Reorder the rows of $\boldsymbol{P}$: **swap** $\boldsymbol{p}_{i\cdot}$ with $\boldsymbol{p}_{b\cdot}$
6: **end for**
7: **Return:** $\boldsymbol{P}$

By using the largest absolute entry in the respective column of $\boldsymbol{A}$, the pivoting elements are bounded as far away from zero as possible. Similar strategies can be applied for additional column-wise reordering of $\boldsymbol{A}$, e.g. to enhance sparsity and stability of the system. This combination is then termed full pivoting. By the use of LU-factorization with pivoting, algorithm 2 can be used to solve problem 4.1 (cf. Golub and Van Loan, 1996, pp. 94 ff; Li, 2005, p. 2). The main purpose of this strategy in the context of this thesis is to facilitate non-linear optimization, which is described in the following section 4.2.

4.2 Non-linear Optimization

Methods for solving exactly determined systems of linear equations of the form $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{b}$ for $\boldsymbol{c}$ are presented in section 4.1. Based on this foundation, *gradient methods* can be used to perform optimization with regard to a general function $\delta : \mathbb{R}^h \to \mathbb{R}_{\geq 0}$, which is usually non-linear. The crucial use of gradient information to achieve this goal is the reason for the designation of these methods (cf. Hackbusch, 1994, pp. 248 ff; Nesterov, 2004, pp. 25 ff). Note that throughout the subsequent discussion, optimization techniques are presented for minimization problems. Maximization can nevertheless be achieved in a corresponding manner, e.g. by minimizing the negative of a function. Relying on the linear solving techniques introduced in the previous section, a summary of unconstrained (section 4.2.1) as well as constrained non-linear optimization (section 4.2.2) is provided in the following discussion. Simplifying approximations for the Hessian matrix of δ that are frequently used for this purpose are considered in section 4.2.3. These optimization methods are fundamental for fitting most of the prediction and weighting models for non-probability samples discussed in the following chapter 5.

4.2.1 Unconstrained Non-linear Optimization

The objective in unconstrained optimization is to find a minimum

$$\boldsymbol{\Theta}^* = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left(\delta \left(\boldsymbol{\Theta} \right) \right) \quad . \tag{4.5}$$

for a function $\delta : \mathbb{R}^h \to \mathbb{R}_{\geq 0}$ with respect to a vector of parameters $\boldsymbol{\Theta} \in \mathbb{R}^h$ of given size $h \in \mathbb{N}$. As outlined above, maximization is not considered separately because it can be achieved analogously. The optimality criteria require the Jacobian matrix $\mathbf{J}_\delta \left(\boldsymbol{\Theta} \right) \in \mathbb{R}^{1 \times h}$ of $\delta \left(\boldsymbol{\Theta} \right)$ to be zero at this optimal solution (cf. Geiger and Kanzow, 2002, pp. 46 f):

$$\mathbf{J}_\delta \left(\boldsymbol{\Theta}^* \right) = \mathbf{0}_{1 \times h} \quad . \tag{4.6}$$

One of the most influential methods to find $\boldsymbol{\Theta}^*$ in this context is the Newton-Raphson algorithm, for which a historical overview is provided by Cajori (1911). He states that the idea was first published in Wallis (1685, pp. 338 ff) and later refined by Raphson (1690) to its present form shown in algorithm 5 (cf. Abramowitz and Stegun, 1970, p. 18). Applying the Newton-Raphson method to find $\boldsymbol{\Theta}^*$ is based on a first order Taylor approximation of the Jacobian matrix (or equivalently, a second order approximation of δ itself; cf. Nesterov, 2004, pp. 37 ff), which yields

$$\mathbf{J}_\delta \left(\boldsymbol{\Theta} + \boldsymbol{\Delta}_{\boldsymbol{\Theta}} \right) \approx \mathbf{J}_\delta \left(\boldsymbol{\Theta} \right) + \left(\mathbf{H}_\delta \left(\boldsymbol{\Theta} \right) \boldsymbol{\Delta}_{\boldsymbol{\Theta}} \right)^\mathsf{T} \quad , \tag{4.7}$$

where $\mathbf{H}_\delta \left(\boldsymbol{\Theta} \right) \in \mathbb{R}^{h \times h}$ is the Hessian matrix of $\delta \left(\boldsymbol{\Theta} \right)$, and $\boldsymbol{\Delta}_{\boldsymbol{\Theta}} \in \mathbb{R}^h$ defines some distance from $\boldsymbol{\Theta}$ for which the change in $\mathbf{J}_\delta \left(\boldsymbol{\Theta} \right)$ is approximated. It is referred to as the *step direction* (cf. algorithm 5). By using this approximation 4.7 for optimality condition 4.6, it is evident that $\mathbf{J}_\delta \left(\boldsymbol{\Theta} + \boldsymbol{\Delta}_{\boldsymbol{\Theta}} \right) \approx \mathbf{0}_{1 \times h}$ if

$$\begin{aligned} \mathbf{H}_\delta \left(\boldsymbol{\Theta} \right) \boldsymbol{\Delta}_{\boldsymbol{\Theta}} &= - \left(\mathbf{J}_\delta \left(\boldsymbol{\Theta} \right) \right)^\mathsf{T} \\ \Leftrightarrow \qquad \boldsymbol{\Delta}_{\boldsymbol{\Theta}} &= - \left(\mathbf{H}_\delta \left(\boldsymbol{\Theta} \right) \right)^{-1} \left(\mathbf{J}_\delta \left(\boldsymbol{\Theta} \right) \right)^\mathsf{T} \quad . \end{aligned} \tag{4.8}$$

Equations 4.8 constitute a system of linear equations that can be solved by the methods discussed in section 4.1 because the Hessian is always a square matrix (cf. Lawson and Hanson, 1995, pp. 36 ff). The Newton-Raphson method is based on repeatedly updating $\boldsymbol{\Theta}^*$ by solving this approximation to condition 4.6 until the optimality requirement is actually fulfilled (cf. Jarre and Stoer, 2004, pp. 68 ff; Nesterov, 2004, pp. 37 ff). Algorithm 5 formally represents this approach.

In this algorithm, v denotes an arbitrary boundary for numerical tolerance, depending on the required precision, and $\boldsymbol{\Theta}^{(a)}$ is the value of $\boldsymbol{\Theta}$ at iteration a. Equations 4.8 are solved repeatedly until optimality condition 4.6 is met, or until all changes in parameters $\boldsymbol{\Theta}$ are smaller than v in magnitude. However, to guarantee that this procedure actually reduces $\mathbf{J}_\delta \left(\boldsymbol{\Theta}^{(a-1)} \right)$ in each iteration, it needs to hold that

$$\mathbf{J}_\delta \left(\boldsymbol{\Theta}^{(a-1)} \right) \boldsymbol{\Delta}_{\boldsymbol{\Theta}} < 0 \tag{4.9}$$

(cf. Nocedal and Wright, 1999, p. 36). As can be seen by using equalities 4.8, this requirement 4.9 is readily fulfilled in cases where $\mathbf{H}_\delta \left(\boldsymbol{\Theta}^{(a-1)} \right)$ is positive definite. To achieve a decrease such that the new value $\delta \left(\boldsymbol{\Theta}^{(a)} \right) = \delta \left(\boldsymbol{\Theta}^{(a-1)} + t \cdot \boldsymbol{\Delta}_{\boldsymbol{\Theta}} \right)$ is sufficiently

Algorithm 5: Newton-Raphson algorithm

1: **Input:** $\boldsymbol{\Theta}^{(0)} \in \mathbb{R}^h$; $\delta : \mathbb{R}^h \to \mathbb{R}_{\geq 0}$; $v > 0$
2: Set $a = 1$
3: Use algorithms 2 and 3 to compute the step direction $\boldsymbol{\Delta}_{\boldsymbol{\Theta}}$ as the solution of

$$\mathbf{H}_\delta\left(\boldsymbol{\Theta}^{(a-1)}\right)\boldsymbol{\Delta}_{\boldsymbol{\Theta}} \;=\; -\left(\mathbf{J}_\delta\left(\boldsymbol{\Theta}^{(a-1)}\right)\right)^{\mathsf{T}}$$

4: **Optional:** Compute a step size t through algorithm 6, otherwise use $t = 1$
5: Set $\boldsymbol{\Theta}^{(a)} \;:=\; \boldsymbol{\Theta}^{(a-1)} + t \cdot \boldsymbol{\Delta}_{\boldsymbol{\Theta}}$
6: **if** $\left(\mathbf{J}_\delta\left(\boldsymbol{\Theta}^{(a-1)}\right) = \mathbf{0}_{1\times h}$ or $\operatorname{Max}\left(\operatorname{Abs}\left(\boldsymbol{\Delta}_{\boldsymbol{\Theta}} \oslash \boldsymbol{\Theta}^{(a)} - 1\right)\right) < v\right)$ **then**
7: **Return:** $\boldsymbol{\Theta}^{(a)}$
8: **else**
9: Update $a \leftarrow a + 1$ and **go to** step 3
10: **end if**

lower than the previous one regardless of the function δ and starting point $\boldsymbol{\Theta}^{(0)}$, the step direction $\boldsymbol{\Delta}_{\boldsymbol{\Theta}}$ can be multiplied by a step-size factor $t \in (0, 1]$ determined by an Armijo-type line search to achieve better convergence properties (cf. Armijo, 1966, p. 2; Jarre and Stoer, 2004, pp. 140 ff; Nocedal and Wright, 1999, pp. 55 ff). For this reason, the following algorithm 6 can be used in step 4 of algorithm 5.

Algorithm 6: Armijo step-size rule for unconstrained optimization

1: **Input:** $\boldsymbol{\Theta}, \boldsymbol{\Delta}_{\boldsymbol{\Theta}} \in \mathbb{R}^h$; $b, c \in (0, 1)$
2: Set $v = 0$
3: **if** $\left(\delta\left(\boldsymbol{\Theta} + c^v \cdot \boldsymbol{\Delta}_{\boldsymbol{\Theta}}\right) \leq \delta\left(\boldsymbol{\Theta}\right) + b \cdot c^v \cdot \left(\mathbf{J}_\delta\left(\boldsymbol{\Theta}\right)\right)^{\mathsf{T}} \boldsymbol{\Delta}_{\boldsymbol{\Theta}}\right)$ **then**
4: **Return:** $t = c^v$
5: **else**
6: Set $v \leftarrow v + 1$ and **go to** step 3
7: **end if**

In this algorithm, b is a quantifier for a decrease in δ that is sufficient in relation to the change projected by the Jacobian, and c determines the step size for searching along this projection. Such a line search enforces the Armijo (1966) condition, which corresponds to the first condition introduced by Wolfe (1969). In comparison to using both (or even the strong) Wolfe conditions, this approach reduces the computational effort, resulting usually in more but cheaper iterations (cf. Geiger and Kanzow, 2002, pp. 273 f; Nocedal and Wright, 1999, pp. 35 ff).

There are different strategies that attempt to reduce the computational burden that is imposed by using the Hessian matrix in step 3 of algorithm 5. For example, the `BFGS`-method approximates the Hessian matrix using only parameter and gradient information. As an additional benefit, it ensures that all sub-problems of algorithm 5 are convex (cf. Nocedal and Wright, 1999, pp. 194 ff). An overview of this and selected other quasi-Newton methods is provided jointly for unconstrained and constrained optimization in section 4.2.3, after considering constrained optimization in the following section 4.2.2.

4.2.2 Constrained Non-linear Optimization

In constrained optimization, the goal is again to find a vector of parameters $\boldsymbol{\Theta}^* \in \mathbb{R}^h$ of given size $h \in \mathbb{N}$ that minimizes the function δ, but with respect to some general equality and/or inequality constraints. Problem 4.5 is thus extended to

$$\begin{aligned} \boldsymbol{\Theta}^* &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left(\delta\left(\boldsymbol{\Theta}\right)\right) \\ \text{s.t.} \quad & \bar{\mathbf{g}}\left(\boldsymbol{\Theta}^*\right) = \mathbf{0} \\ & \widetilde{\mathbf{g}}\left(\boldsymbol{\Theta}^*\right) \leq \mathbf{0} \quad . \end{aligned} \tag{4.10}$$

As before, $\delta : \mathbb{R}^h \to \mathbb{R}_{\geq 0}$ is the function to be minimized, while $\bar{\mathbf{g}} : \mathbb{R}^h \to \mathbb{R}^s$ and $\widetilde{\mathbf{g}} : \mathbb{R}^h \to \mathbb{R}^u$ respectively are functions expressing the equality and inequality constraints of arbitrary dimensions $s, u \in \mathbb{N}$ imposed to the problem. Here and in the following, inequalities are applied element-wise, just as equalities. Denoting the corresponding Lagrange multipliers for equality and inequality constraints by $\boldsymbol{\alpha} \in \mathbb{R}^s$ and $\boldsymbol{\lambda} \in \mathbb{R}^u$, respectively, the Lagrange function of this optimization problem is defined by

$$\mathrm{L}\left(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}\right) = \delta\left(\boldsymbol{\Theta}\right) + \boldsymbol{\alpha}^{\mathsf{T}} \bar{\mathbf{g}}\left(\boldsymbol{\Theta}\right) + \boldsymbol{\lambda}^{\mathsf{T}} \widetilde{\mathbf{g}}\left(\boldsymbol{\Theta}\right) \tag{4.11}$$

(cf. Geiger and Kanzow, 2002, p. 242; Nocedal and Wright, 1999, p. 327).

The Karush-Kuhn-Tucker optimality criteria (KKT-conditions; cf. Karush, 1939, quoted in Kjeldsen, 2000; Kuhn and Tucker, 1951) that parameters $\boldsymbol{\Theta}^*$ must fulfill in order to be an optimal solution for the non-linear problem 4.10 are then given by

$$\begin{aligned} \mathbf{J}_{\delta}\left(\boldsymbol{\Theta}^*\right) + \boldsymbol{\alpha}^{\mathsf{T}} \mathbf{J}_{\bar{\mathbf{g}}}\left(\boldsymbol{\Theta}^*\right) + \boldsymbol{\lambda}^{\mathsf{T}} \mathbf{J}_{\widetilde{\mathbf{g}}}\left(\boldsymbol{\Theta}^*\right) &= \mathbf{0} \\ \bar{\mathbf{g}}\left(\boldsymbol{\Theta}^*\right) &= \mathbf{0} \\ \widetilde{\mathbf{g}}\left(\boldsymbol{\Theta}^*\right) &\leq \mathbf{0} \\ \boldsymbol{\lambda} &\geq \mathbf{0} \\ \boldsymbol{\lambda}^{\mathsf{T}} \widetilde{\mathbf{g}}\left(\boldsymbol{\Theta}^*\right) &= \mathbf{0} \quad . \end{aligned} \tag{4.12}$$

In conditions 4.12, $\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}\right) \in \mathbb{R}^{1 \times h}$, $\mathbf{J}_{\bar{\mathbf{g}}}\left(\boldsymbol{\Theta}\right) \in \mathbb{R}^{s \times h}$ and $\mathbf{J}_{\widetilde{\mathbf{g}}}\left(\boldsymbol{\Theta}\right) \in \mathbb{R}^{u \times h}$ respectively denote the Jacobian matrices of the distance, equality and inequality constraint function (cf. Geiger and Kanzow, 2002, p. 46; Gill, Murray and Wright, 1981, pp. 77 ff; Nocedal and Wright, 1999, p. 328). In the course of the following sections, the foundation for unconstrained minimization of potentially non-linear functions through iterative approximation of the optimality condition presented in section 4.2.1 is extended to account for constraints in the optimization process.

4.2.2.1 Quadratic Programming

First, the special case of problem 4.10 where the loss function is quadratic and the constraints are linear is considered. It is defined by

$$\begin{aligned} \delta\left(\boldsymbol{\Theta}\right) &:= \boldsymbol{\Theta}^{\mathsf{T}} \boldsymbol{Q} \boldsymbol{\Theta} + \boldsymbol{\Theta}^{\mathsf{T}} \boldsymbol{c} \\ \bar{\mathbf{g}}\left(\boldsymbol{\Theta}\right) &:= \bar{\boldsymbol{G}} \boldsymbol{\Theta} - \bar{\boldsymbol{t}} \\ \widetilde{\mathbf{g}}\left(\boldsymbol{\Theta}\right) &:= \widetilde{\boldsymbol{G}} \boldsymbol{\Theta} - \widetilde{\boldsymbol{t}} \quad , \end{aligned} \tag{4.13}$$

where $\boldsymbol{Q} \in \mathbb{R}^{h \times h}$ and $\boldsymbol{c} \in \mathbb{R}^h$ constitute multipliers for the quadratic and linear part of the distance, respectively. Furthermore are $\bar{\boldsymbol{G}} \in \mathbb{R}^{s \times h}$, $\widetilde{\boldsymbol{G}} \in \mathbb{R}^{u \times h}$, $\bar{\boldsymbol{t}} \in \mathbb{R}^s$ and $\widetilde{\boldsymbol{t}} \in \mathbb{R}^u$ the linear multipliers, targets and upper bounds for equality and inequality constraints. It is easy to see that KKT-conditions 4.12 for problem 4.13 are

$$\begin{aligned} \boldsymbol{Q}\boldsymbol{\Theta}^* + \boldsymbol{\alpha}^\mathsf{T}\bar{\boldsymbol{G}} + \boldsymbol{\lambda}^\mathsf{T}\widetilde{\boldsymbol{G}} + \boldsymbol{c} &= \boldsymbol{0} \\ \bar{\boldsymbol{G}}\boldsymbol{\Theta}^* &= \bar{\boldsymbol{t}} \\ \widetilde{\boldsymbol{G}}\boldsymbol{\Theta}^* &\leq \widetilde{\boldsymbol{t}} \\ \boldsymbol{\lambda} &\geq \boldsymbol{0} \\ \boldsymbol{\lambda}^\mathsf{T}\widetilde{\boldsymbol{G}}\boldsymbol{\Theta}^* &= \boldsymbol{0} \end{aligned} \tag{4.14}$$

and in this case all linear in $\boldsymbol{\Theta}^*$, but still include inequality constraints (cf. Geiger and Kanzow, 2002, pp. 197 ff). If there were only equality constraints, the solution could simply be found through solving

$$\begin{bmatrix} \boldsymbol{Q} & \bar{\boldsymbol{G}}^\mathsf{T} \\ \bar{\boldsymbol{G}} & \boldsymbol{0}_{s \times s} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Theta} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{c} \\ \bar{\boldsymbol{t}} \end{bmatrix} \tag{4.15}$$

by means of algorithms 2 and 3. Thus, only the inequality constraints prevent the use of linear solvers for this problem directly. Yet, these inequality restrictions can be either exactly binding or negligible for the optimization problem. If it holds that $\widetilde{\boldsymbol{g}}_{i\cdot}\boldsymbol{\Theta}^* = \widetilde{t}_i$ at a point $\boldsymbol{\Theta}^*$ that is feasible for conditions 4.14, then the restriction imposed by the i-th row of $\widetilde{\boldsymbol{G}}$ is said to be *active* (or *binding*) and, hence, constitutes an equality constraint. If $\widetilde{\boldsymbol{g}}_{i\cdot}\boldsymbol{\Theta}^* < \widetilde{t}_i$, then the constraint is *inactive* and does not restrict the feasibility at $\boldsymbol{\Theta}^*$. Following this reasoning, treating a subset of the inequalities – the *active set* – as equalities while all other inequality constraints are disregarded allows bringing these conditions into the form of equation 4.15. Such an *active set strategy* is formalized by repeatedly selecting a working set (rows of $\widetilde{\boldsymbol{G}}$) and solving the system of linear equations while expanding or reducing the working set. This can again be done by means of the techniques described in section 4.1 (cf. Fletcher, 1971, pp. 80 ff; Geiger and Kanzow, 2002, pp. 197 ff; Gill, Murray and Wright, 1981, pp. 71 ff, 167 ff; Nocedal and Wright, 1999, pp. 444 f).

The outlined procedure is described in algorithm 7. By starting from a feasible point (cf. step 1) and iteratively updating the active constraints and parameters, the algorithm optimizes $\boldsymbol{\Theta}$ without violating the feasible region. Steps 16 and 19 lead to the largest possible update that keeps $\boldsymbol{\Theta}$ in this region while potentially activating relevant constraints. Step 12 discards inequality constraints that are no longer binding. Through those steps, inequality constraints enter or leave the working set until a feasible minimum of the distance function is found, such that problem 4.13 can be solved by linear techniques (cf. Fletcher, 1971; Geiger and Kanzow, 2002, pp. 199 ff; Gill and Murray, 1978, p. 351; Gill, Murray and Wright, 1981, pp. 167 ff; Lenard, 1979; Nocedal and Wright, 1999, pp. 444 f).

Note that quadratic problems without any or with exclusively linear equality constraints are special cases of problem 4.13 that require only one iteration in algorithm 7 because the active set is fixed in these cases. Based on this algorithm, optimization of general non-linear loss and constraint functions can be implemented as described in the following section 4.2.2.2.

Algorithm 7: Quadratic programming using an active set strategy (QP)

1: **Input:** $\boldsymbol{\Theta}^{(0)} \in \mathbb{R}^h$ feasible for problem 4.13; $\boldsymbol{Q} \in \mathbb{R}^{h\times h}$; $\bar{\boldsymbol{G}} \in \mathbb{R}^{s\times h}$; $\widetilde{\boldsymbol{G}} \in \mathbb{R}^{u\times h}$; $\tilde{\boldsymbol{t}} \in \mathbb{R}^u$

2: Initialize $\boldsymbol{\alpha}^{(0)} \coloneqq \mathbf{0}_{s\times 1}$, $\boldsymbol{\lambda}^{(0)} \coloneqq \mathbf{0}_{u\times 1}$, $a \coloneqq 0$ and find the current working set

$$\mathscr{A}^{(0)} \coloneqq \left\{ i : \tilde{\boldsymbol{g}}_{i\cdot}\boldsymbol{\Theta}^{(0)} = \tilde{t}_i \right\}$$

3: **if** $\boldsymbol{\Theta}^{(a)}$, $\boldsymbol{\alpha}^{(a)}$ and $\boldsymbol{\lambda}^{(a)}$ fulfill KKT-conditions 4.14 **then**

4: **Return:** $\left[\left(\boldsymbol{\Theta}^{(a)}\right)^{\mathsf{T}} \quad \left(\boldsymbol{\alpha}^{(a)}\right)^{\mathsf{T}} \quad \left(\boldsymbol{\lambda}^{(a)}\right)^{\mathsf{T}} \right]^{\mathsf{T}}$

5: **end if**

6: Initialize $\boldsymbol{\lambda}^{(a+1)} \coloneqq \mathbf{0}_{s\times 1}$ and let $\widetilde{\boldsymbol{G}}_{\mathscr{A}^{(a)}\cdot}$ be the sub-matrix of $\widetilde{\boldsymbol{G}}$ containing the active rows with row-indices given by $\mathscr{A}^{(a)}$ and $\boldsymbol{\lambda}^{(a+1)}_{\mathscr{A}^{(a)}}$ the corresponding sub-vector of $\boldsymbol{\lambda}^{(a+1)}$. Use algorithms 2 and 3 to solve

$$\begin{bmatrix} \boldsymbol{Q} & \bar{\boldsymbol{G}}^{\mathsf{T}} & \widetilde{\boldsymbol{G}}^{\mathsf{T}}_{\mathscr{A}^{(a)}\cdot} \\ \bar{\boldsymbol{G}} & \mathbf{0} & \mathbf{0} \\ \widetilde{\boldsymbol{G}}_{\mathscr{A}^{(a)}\cdot} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{\Theta}} \\ \boldsymbol{\alpha}^{(a+1)} \\ \boldsymbol{\lambda}^{(a+1)}_{\mathscr{A}^{(a)}} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{c} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

7: **if** $\boldsymbol{\Delta}_{\boldsymbol{\Theta}} = \mathbf{0}$ **then**

8: **if** ${\lambda_i}^{a+1} \geq 0$ for all i **then**

9: **Return:** $\left[\left(\boldsymbol{\Theta}^{(a)}\right)^{\mathsf{T}} \quad \left(\boldsymbol{\alpha}^{(a)}\right)^{\mathsf{T}} \quad \left(\boldsymbol{\lambda}^{(a)}\right)^{\mathsf{T}} \right]^{\mathsf{T}}$

10: **else**

11: Select one index $j \in \left\{ k : {\lambda_k}^{a+1} = \mathrm{Min}\left(\boldsymbol{\lambda}^{(a+1)}\right) \right\}$,

12: Set $\boldsymbol{\Theta}^{(a+1)} \coloneqq \boldsymbol{\Theta}^{(a)}$, $\mathscr{A}^{(a+1)} \coloneqq \mathscr{A}^{(a)} \setminus \{j\}$ and **go to** step 22

13: **end if**

14: **else**

15: **if** $\boldsymbol{\Theta}^{(a)} + \boldsymbol{\Delta}_{\boldsymbol{\Theta}}$ is feasible for problem 4.13 **then**

16: Set $\boldsymbol{\Theta}^{(a+1)} \coloneqq \boldsymbol{\Theta}^{(a)} + \boldsymbol{\Delta}_{\boldsymbol{\Theta}}$, $\mathscr{A}^{(a+1)} \coloneqq \mathscr{A}^{(a)}$ and **go to** step 22

17: **else**

18: Select one index

$$j \in \left\{ k : \left(\frac{\tilde{t}_k - \tilde{\boldsymbol{g}}_{k\cdot}\boldsymbol{\Theta}^{(a)}}{\tilde{\boldsymbol{g}}_{k\cdot}\boldsymbol{\Delta}_{\boldsymbol{\Theta}}} = \mathrm{Min}\left(\left. \frac{\tilde{t}_l - \tilde{\boldsymbol{g}}_{l\cdot}\boldsymbol{\Theta}^{(a)}}{\tilde{\boldsymbol{g}}_{l\cdot}\boldsymbol{\Delta}_{\boldsymbol{\Theta}}} \right| l \notin \mathscr{A}^{(a)} \text{ and } \tilde{\boldsymbol{g}}_{l\cdot}\boldsymbol{\Delta}_{\boldsymbol{\Theta}} > 0 \right) \right) \right\}$$

19: Set $\boldsymbol{\Theta}^{(a+1)} \coloneqq \boldsymbol{\Theta}^{(a)} + \dfrac{\tilde{t}_j - \tilde{\boldsymbol{g}}_{j\cdot}\boldsymbol{\Theta}^{(a)}}{\tilde{\boldsymbol{g}}_{j\cdot}\boldsymbol{\Delta}_{\boldsymbol{\Theta}}} \cdot \boldsymbol{\Delta}_{\boldsymbol{\Theta}}$, $\mathscr{A}^{(a+1)} \coloneqq \mathscr{A}^{(a)} \cup \{j\}$ and **go to** step 22

20: **end if**

21: **end if**

22: Set $a \leftarrow a + 1$ and **go to** step 3

4.2.2.2 Sequential Quadratic Programming

Based on the solution for quadratic optimization problems under linear constraints presented in the previous section 4.2.2.1, the general non-linear problem 4.10 can be tackled. Building on ideas of the Newton-Raphson method presented in section 4.2.1, sequential quadratic programming (SQP) solves this non-linearly constrained non-linear problem by iterative linear approximations of its optimality conditions 4.12 (cf. Geiger and Kanzow, 2002, p. 239). Similar to algorithm 5, the distance function is approximated by its Jacobian $\mathbf{J}_{\delta}(\boldsymbol{\Theta})$ and a matrix $\widetilde{\mathbf{H}} \in \mathbb{R}^{h \times h}$, which usually is an approximated Hessian of the Lagrange function 4.11 (cf. section 4.2.3). The reason for choosing the Hessian matrix of the Lagrange function L rather than that of the distance function δ is that faster convergence can be achieved by additionally incorporating information about the constraints (cf. Powell, 1978, pp. 146 ff). Linear approximations of the constraint functions are determined by their respective Jacobian matrices, $\mathbf{J}_{\bar{\mathbf{g}}}(\boldsymbol{\Theta})$ and $\mathbf{J}_{\widetilde{\mathbf{g}}}(\boldsymbol{\Theta})$. These approximations were first introduced for constrained non-linear optimization by Wilson (1963, p. 41) and Fletcher (1972, p. 136). With respect to a step direction $\boldsymbol{\Delta}_{\boldsymbol{\Theta}}$ as before, they constitute a quadratic distance function under linear constraint functions and, thus, can be solved by means of algorithm 7 (cf. Geiger and Kanzow, 2002, p. 243; Kraft, 1988). The following algorithm 8 represents such an iterative updating procedure for $\boldsymbol{\Theta}$ to find an optimal solution $\boldsymbol{\Theta}^{*}$.

As before, v denotes an arbitrary boundary for numerical tolerance, depending on the required precision. Further, $\boldsymbol{\alpha}^{+}, \boldsymbol{\alpha}^{-} \in \mathbb{R}^{s}$ and $\boldsymbol{\lambda}^{*} \in \mathbb{R}^{u}$ are Lagrange multipliers to restrict the slack variables $\boldsymbol{\xi}^{+}, \boldsymbol{\xi}^{-} \in \mathbb{R}^{s}_{\geq 0}$ and $\boldsymbol{\xi}^{*} \in \mathbb{R}^{u}_{\geq 0}$ to be non-negative. These slack variables are required since some of the sub-problems in step 6 may have no feasible solution otherwise. They allow for violations of the constraints and are penalized by the parameter $\varsigma^{(a)}$ that is updated in every iteration using a prespecified constant $\bar{\varsigma}$. Updating $\widetilde{\mathbf{H}}$ in step 11 can, for example, be done by means of algorithm 10 (cf. Geiger and Kanzow, 2002, pp. 234 ff; Jarre and Stoer, 2004, pp. 327 ff; Kraft, 1988; Nocedal and Wright, 1999, pp. 528 ff).

Just like in the unconstrained scenario, step size determination (step 10) can be used to achieve global convergence of the SQP method. In the constrained case, this is achieved by using a merit function $\varphi(t, \varsigma)$ for a line search, which was proposed by Armijo (1966, p. 2). Han (1977, p. 299) and Schittkowski (1981, p. 87) introduced further simplifications that lead to algorithm 9.

The merit function used in that algorithm throughout the following chapters is given by

$$\begin{aligned} \varphi(t, \varsigma) \; &= \delta(\boldsymbol{\Theta} + t \cdot \boldsymbol{\Delta}_{\boldsymbol{\Theta}}) \\ &+ \; \varsigma \cdot \Big(\mathrm{Abs}(\bar{\mathbf{g}}(\boldsymbol{\Theta} + t \cdot \boldsymbol{\Delta}_{\boldsymbol{\Theta}})) \\ &\qquad + \; \mathbf{Rowmax}([\widetilde{\mathbf{g}}(\boldsymbol{\Theta} + t \cdot \boldsymbol{\Delta}_{\boldsymbol{\Theta}}), \mathbf{0}_{u \times 1}]) \Big) \end{aligned} \quad , \tag{4.16}$$

where ς is a penalty parameter, for which an updating rule proposed by Powell (1978, p. 151) is applied in algorithm 8. Note that the approximation $\dfrac{\partial(\varphi(t, \varsigma))}{\partial(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})} \approx \boldsymbol{\Delta}_{\boldsymbol{\Theta}}{}^{\mathsf{T}} \widetilde{\mathbf{H}}$ is used in this line search algorithm, which is based on $\widetilde{\mathbf{H}}$ being the (approximate) Hessian matrix of the Lagrange function (cf. section 4.2.3). Theoretical justification for this line search approach is closely related to that in the unconstrained case (cf. algorithm 6).

Algorithm 8: Sequential quadratic programming (SQP)

1: **Input:** $\boldsymbol{\Theta}^{(0)} \in \mathbb{R}^h$; $\delta : \mathbb{R}^h \to \mathbb{R}_{\geq 0}$; $\bar{\mathbf{g}} : \mathbb{R}^h \to \mathbb{R}^s$; $\tilde{\mathbf{g}} : \mathbb{R}^h \to \mathbb{R}^u$; symmetric $\widetilde{\mathbf{H}}^{(0)} \in \mathbb{R}^{h \times h}$; $b, c \in (0,1)$; $\varsigma^{(0)}, \bar{\varsigma} \in \mathbb{R}_+$; $v > 0$ depending on the required precision

2: Initialize $\boldsymbol{\alpha}^{(0)} \coloneqq \mathbf{0}_{s \times 1}$, $\boldsymbol{\lambda}^{(0)} \coloneqq \mathbf{0}_{u \times 1}$ and $a \coloneqq 0$

3: **if** $\boldsymbol{\Theta}^{(a)}$, $\boldsymbol{\alpha}^{(a)}$ and $\boldsymbol{\lambda}^{(a)}$ fulfill KKT-conditions 4.12 **then**

4: **Return:** $\boldsymbol{\Theta}^{(a)}$

5: **end if**

6: By using algorithm 7, compute parameters $\Psi^{(a)} \coloneqq \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{\Theta}}^{\mathsf{T}} & \left(\boldsymbol{\xi}^{+}\right)^{\mathsf{T}} & \left(\boldsymbol{\xi}^{-}\right)^{\mathsf{T}} & (\boldsymbol{\xi}^{*})^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$ and Lagrange multipliers $\boldsymbol{\Lambda}^{(a+1)} \coloneqq \begin{bmatrix} \left(\boldsymbol{\alpha}^{(a+1)}\right)^{\mathsf{T}} & \left(\boldsymbol{\lambda}^{(a+1)}\right)^{\mathsf{T}} & (\boldsymbol{\alpha}^{+})^{\mathsf{T}} & (\boldsymbol{\alpha}^{-})^{\mathsf{T}} & (\boldsymbol{\lambda}^{*})^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$ from

$$\begin{bmatrix} \Psi^{(a)} \\ \boldsymbol{\Lambda}^{(a+1)} \end{bmatrix} = \underset{(\Psi, \boldsymbol{\Lambda})}{\operatorname{argmin}} \left(\frac{1}{2} \cdot \Psi \begin{bmatrix} \widetilde{\mathbf{H}}^{(a)} & \mathbf{0}_{h \times (s+u)} \\ \mathbf{0}_{(s+u) \times h} & \mathbf{0}_{(s+u) \times (s+u)} \end{bmatrix} \Psi + \begin{bmatrix} \mathbf{J}_{\delta}\left(\boldsymbol{\Theta}^{(a)}\right) & \varsigma^{(a)} \cdot \mathbf{1}_{1 \times (s+u)} \end{bmatrix} \Psi \right)$$

$$\text{s. t.} \quad \begin{bmatrix} \mathbf{J}_{\bar{\mathbf{g}}}\left(\boldsymbol{\Theta}^{(a)}\right) & \mathbf{I}_s & -\mathbf{I}_s & \mathbf{0}_{s \times u} \end{bmatrix} \Psi = -\bar{\mathbf{g}}\left(\boldsymbol{\Theta}^{(a)}\right)$$

$$\begin{bmatrix} \mathbf{J}_{\tilde{\mathbf{g}}}\left(\boldsymbol{\Theta}^{(a)}\right) & \mathbf{0}_{u \times (2s)} & -\mathbf{I}_u \\ \mathbf{0}_{(2s+u) \times h} & & -\mathbf{I}_{(2s+u)} \end{bmatrix} \Psi \leq -\begin{bmatrix} \tilde{\mathbf{g}}\left(\boldsymbol{\Theta}^{(a)}\right) \\ \mathbf{0}_{(2s+u) \times 1} \end{bmatrix}$$

7: **if** $\left(\operatorname{Max}\left(\operatorname{Abs}\left(\boldsymbol{\Delta}_{\boldsymbol{\Theta}} \oslash \boldsymbol{\Theta}^{(a)} - 1\right)\right) \leq v\right)$ **then**

8: **Return:** $\boldsymbol{\Theta}^{(a)}$

9: **end if**

10: **Optional:** Compute a step size t through algorithm 9, otherwise use $t = 1$

11: Update $\boldsymbol{\Theta}^{(a+1)} \coloneqq \boldsymbol{\Theta}^{(a)} + t \cdot \boldsymbol{\Delta}_{\boldsymbol{\Theta}}$ and $\varsigma^{(a+1)} \coloneqq \operatorname{Max}\left(\varsigma^{(a)}, \operatorname{Max}\left(\boldsymbol{\lambda}^{(a+1)}, \operatorname{Abs}\left(\boldsymbol{\alpha}^{(a+1)}\right)\right) + \bar{\varsigma}\right)$, select symmetric $\widetilde{\mathbf{H}}^{(a+1)} \in \mathbb{R}^{h \times h}$, set $a \leftarrow a + 1$ and go to 3

Algorithm 9: Armijo step-size rule for constrained optimization

1: **Input:** $\boldsymbol{\Theta}, \boldsymbol{\Delta}_{\boldsymbol{\Theta}} \in \mathbb{R}^h$; $\widetilde{\mathbf{H}} \in \mathbb{R}^{h \times h}$; $\varphi : (0,1] \times \mathbb{R}_+ \to \mathbb{R}_+$; $b, c \in (0,1)$

2: Set $v = 0$

3: **if** $\left(\varphi\left(c^v, \varsigma\right) \leq \varphi\left(0, \varsigma\right) - b \cdot c^v \cdot \boldsymbol{\Delta}_{\boldsymbol{\Theta}}{}^{\mathsf{T}} \widetilde{\mathbf{H}} \boldsymbol{\Delta}_{\boldsymbol{\Theta}}\right)$ **then**

4: **Return:** $t = c^v$

5: **else**

6: Set $v \leftarrow v + 1$ and **go to** step 3

7: **end if**

Details and theoretical evaluations of the resulting convergence properties can be found in Geiger and Kanzow (2002, pp. 274 ff), Jarre and Stoer (2004, pp. 333 ff) as well as Nocedal and Wright (1999, pp. 544 ff).

SQP-methods as presented in algorithm 8 are widespread and successfully applied in various areas. In comparison to other methods for solving problem 4.10, their subproblems are often more complex. Yet, they are considered favorable when second

derivatives are approximated rather than exact (cf. Boggs and Tolle, 1995; Geiger and Kanzow, 2002, pp. 256 ff; Jarre and Stoer, 2004, p. 337). This aspect is important in the subsequent chapters, especially for implementing calibrated semi-parametric artificial neural networks (cf. section 5.2.3). In the following section 4.2.3, an overview of selected strategies for approximating a function's Hessian matrix is provided.

4.2.3 Substitutes for the Hessian Matrix

Computing the Hessian matrix in each iteration of the Newton-Raphson or `SQP` algorithm imposes considerable computational burden, especially in case of a large number of optimization parameters $\boldsymbol{\Theta}$. In addition, positive definiteness of the Hessian is an important property for achieving global convergence that is, especially in the constrained case, not generally guaranteed for arbitrary values of $\boldsymbol{\Theta}$ (cf. also inequality 4.9 and the related discussion). To overcome these issues, many computational implementations, such as *quasi-Newton* (or *variable metric*; cf. Nesterov, 2004, p. 38) methods, rely on an approximation $\widetilde{\mathbf{H}}$ of the actual Hessian matrix. Typically, $\widetilde{\mathbf{H}}$ is positive definite, iteratively updated in the optimization algorithm and computationally simpler than the Hessian itself (cf. Geiger and Kanzow, 2002, p. 256; Nocedal and Wright, 1999, pp. 128 ff, 540; Powell, 1978, p. 145). For example, Fisher (1925) proposes replacing the Hessian by its expected value (called the *Fisher information*), which results in the *Fisher scoring algorithm* (cf. Osborne, 1992, p. 105). The *steepest descent* (or gradient descent) method simply replaces the Hessian by an identity matrix (cf. Gill, Murray and Wright, 1981, p. 103; Nocedal and Wright, 1999, p. 35).

These examples primarily aim at unconstrained optimization problems. A quasi-Newton approach applicable to both unconstrained and constrained optimization is the (damped) Broyden-Fletcher-Goldfarb-Shanno-algorithm (`BFGS`-algorithm) introduced by Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970), which is commonly considered to be the most relevant quasi-Newton method (cf. e.g. Jarre and Stoer, 2004, p. 180 Nocedal and Wright, 1999, p. 197). The rationale behind the original `BFGS`-approach is to approximate the *inverse* Hessian matrix, in order to skip the common inversion step when using it (e.g. in algorithm 5):

$$\widetilde{\mathbf{B}} \coloneqq \left(\widetilde{\mathbf{H}}\right)^{-1} \approx \left(\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}\right)\right)^{-1} \quad . \tag{4.17}$$

Approximating the inverse rather than the Hessian itself is the sole difference between the `BFGS`-approximation and its predecessor, the Davidon-Fletcher-Powell (`DFP`) method proposed by Davidon (1959), Fletcher and Powell (1963), where the Hessian itself is approximated in an equivalent way.

By definition, the Hessian matrix has two important properties for optimization: it is symmetric and fulfills the Taylor approximation given in equation 4.7, which can be reformulated as

$$\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}\right)\boldsymbol{\Delta}_{\boldsymbol{\Theta}} \approx \left(\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}+\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\right)-\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}\right)\right)^{\mathsf{T}} \tag{4.18}$$

(cf. Jarre and Stoer, 2004, pp. 127 f; Nesterov, 2004, p. 19; Nocedal and Wright, 1999, pp. 24, 194 f). The idea behind the `BFGS`-method is to enforce these properties when updating the current approximation $\widetilde{\mathbf{B}}^{(a)}$ for iteration a while modifying it as little as possible. Denoting the changes in parameter and gradient vector by

$$\boldsymbol{s} \coloneqq \boldsymbol{\Theta}^{(a+1)} - \boldsymbol{\Theta}^{(a)} \tag{4.19}$$

and

$$\boldsymbol{y} \coloneqq \mathbf{J}_{\mathrm{L}}\left(\boldsymbol{\Theta}^{(a+1)}\right) - \mathbf{J}_{\mathrm{L}}\left(\boldsymbol{\Theta}^{(a)}\right) \quad , \tag{4.20}$$

respectively, approximation 4.18 leads to the secant (or quasi-Newton) condition

$$\widetilde{\mathbf{B}}^{(a+1)}\boldsymbol{y} = \boldsymbol{s} \quad , \tag{4.21}$$

while the symmetry condition is given by

$$\widetilde{\mathbf{B}}^{(a+1)} = \left(\widetilde{\mathbf{B}}^{(a+1)}\right)^{\mathsf{T}} \tag{4.22}$$

(cf. Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970; Jarre and Stoer, 2004, pp. 176 ff; Nocedal and Wright, 1999, pp. 194 ff).[3] The distance between the current and the updated approximation is measured in terms of the squared weighted Frobenius norm, i.e. by

$$\left\|\widetilde{\mathbf{B}}^{(a+1)} - \widetilde{\mathbf{B}}^{(a)}\right\|^2_{\mathrm{F}_{\boldsymbol{W}}} = \mathrm{tr}\left(\left(\widetilde{\mathbf{B}}^{(a+1)} - \widetilde{\mathbf{B}}^{(a)}\right)^{\mathsf{T}} \boldsymbol{W}^{\mathsf{T}} \left(\widetilde{\mathbf{B}}^{(a+1)} - \widetilde{\mathbf{B}}^{(a)}\right) \boldsymbol{W}\right) \quad , \tag{4.23}$$

where $\boldsymbol{W}$ can be *any* symmetric matrix with property

$$\boldsymbol{W}\boldsymbol{s} \stackrel{!}{=} \boldsymbol{y} \quad , \tag{4.24}$$

which is used to make the solution independent of the units of $\boldsymbol{s}$ (cf. Greenstadt, 1970, p. 4; Goldfarb, 1970, p. 23; Trefethen and Bau, 1997, pp. 24 ff; Gill, Murray and Wright, 1981, p. 29; Nocedal and Wright, 1999, p. 196). This leads to the update rule

$$\widetilde{\mathbf{B}}^{(a+1)} = \left(\mathbf{I}_h - \frac{\boldsymbol{s}\boldsymbol{y}^{\mathsf{T}}}{\boldsymbol{y}^{\mathsf{T}}\boldsymbol{s}}\right) \widetilde{\mathbf{B}}^{(a)} \left(\mathbf{I}_h - \frac{\boldsymbol{y}\boldsymbol{s}^{\mathsf{T}}}{\boldsymbol{y}^{\mathsf{T}}\boldsymbol{s}}\right) + \frac{\boldsymbol{s}\boldsymbol{s}^{\mathsf{T}}}{\boldsymbol{y}^{\mathsf{T}}\boldsymbol{s}} \tag{4.25}$$

for $\widetilde{\mathbf{B}}$. Due to the Sherman-Morrison-Woodbury formula (cf. Nocedal and Wright, 1999, p. 605; Hager, 1989; Sherman and Morrison, 1950), it is equivalent to update $\widetilde{\mathbf{H}}$ by

$$\widetilde{\mathbf{H}}^{(a+1)} = \widetilde{\mathbf{H}}^{(a)} + \frac{\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}}{\boldsymbol{y}^{\mathsf{T}}\boldsymbol{s}} - \frac{\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)}}{\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}} \quad . \tag{4.26}$$

Details for obtaining equations 4.25 and 4.26 are provided in appendix A.

The starting point for the `BFGS` approximation is an initial guess $\widetilde{\mathbf{B}}^{(0)}$, which is a symmetric positive definite matrix chosen by the user. A scalar multiple of the identity matrix or an approximation based on finite differences are common choices if no additional information is available (cf. Nocedal and Wright, 1999, p. 198). Using this approximation, $\widetilde{\mathbf{B}}^{(a)}$ or $\widetilde{\mathbf{H}}^{(a)}$ can be updated iteratively while using only first order information that is anyway required for the remaining parts of the optimization algorithms 5 and 8.

[3] Note that to achieve notational compliance with the referred literature for the `BFGS`-update, $\boldsymbol{y}$ is used here in a manner deviating from the use of $\boldsymbol{Y}$ and its sub-matrices throughout the other chapters of this thesis. For the same reason, $\boldsymbol{s}$ is used here, which is not related to the symbol s used in other contexts.

As described in section 4.2.1, Armijo condition 4.9 holds if $\widetilde{\mathbf{H}}^{(a)}$ and thus $\widetilde{\mathbf{B}}^{(a)}$ are positive definite. In order to fulfill this, the BFGS-update (due to equality 4.21) requires that

$$\boldsymbol{s}^\mathsf{T}\boldsymbol{y} > 0 \quad , \tag{4.27}$$

which at the same time is a necessary condition for fulfilling the curvature (or second Wolfe) condition (cf. Nocedal and Wright, 1999, p. 195; Jarre and Stoer, 2004, p. 181; Wolfe, 1969). However, inequality 4.27 does not always hold automatically, especially in case of constrained optimization problems (cf. Nocedal and Wright, 1999, pp. 540 f). Consequently, Powell (1978, pp. 147 ff) proposes a modified updating rule, which is called the damped BFGS-method. Where possible while respecting condition 4.27, the original update is used, but if equalities 4.25 and 4.26 result in matrices violating this inequality, a convex combination of left- and right-hand side of the quasi-Newton condition 4.21 is used to assure positive definiteness. This updating rule is summarized in algorithm 10.

Algorithm 10: (Damped) BFGS-update rule

1: **Input:** $\boldsymbol{\Theta}^{(a)}, \boldsymbol{\Theta}^{(a+1)} \in \mathbb{R}^h$; $\mathbf{J}_\delta\left(\boldsymbol{\Theta}^{(a)}\right), \mathbf{J}_\delta\left(\boldsymbol{\Theta}^{(a+1)}\right) \in \mathbb{R}^{1\times h}$; symmetric $\widetilde{\mathbf{H}}^{(a)} \in \mathbb{R}^{h\times h}$

2: Calculate the changes in parameters and Jacobian matrix as
$\boldsymbol{s} \coloneqq \boldsymbol{\Theta}^{(a+1)} - \boldsymbol{\Theta}^{(a)}$ and $\boldsymbol{y} \coloneqq \mathbf{J}_\mathrm{L}\left(\boldsymbol{\Theta}^{(a+1)}\right) - \mathbf{J}_\mathrm{L}\left(\boldsymbol{\Theta}^{(a)}\right)$

3: **Optional:** Determine

$$\theta^{(a)} \coloneqq \begin{cases} 1 & \text{, if} \quad \boldsymbol{s}^\mathsf{T}\boldsymbol{y} \geq 0.2 \cdot \boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} \\ \dfrac{0.8 \cdot \boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}}{\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} - \boldsymbol{s}^\mathsf{T}\boldsymbol{y}} & \text{, else} \end{cases} ,$$

otherwise use $\theta^{(a)} = 1$

4: Update $\boldsymbol{y} \leftarrow \theta^{(a)} \cdot \boldsymbol{y} + \left(1 - \theta^{(a)}\right) \cdot \widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}$

5: **Return:** $\widetilde{\mathbf{H}}^{(a+1)} \coloneqq \widetilde{\mathbf{H}}^{(a)} + \dfrac{\boldsymbol{y}\boldsymbol{y}^\mathsf{T}}{\boldsymbol{s}^\mathsf{T}\boldsymbol{y}} - \dfrac{\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}}{\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}}$

The modification solely differs from the original one in potentially setting $\theta^{(a)} < 1$ in step 3 of the algorithm. It therefore reduces to the original update rule 4.26 in cases where $\theta^{(a)} = 1$, while $\widetilde{\mathbf{H}}^{(a+1)} \approx \widetilde{\mathbf{H}}^{(a)}$ for $\theta^{(a)} = 0$. This approach is, thus, called a damped BFGS method. The convex combination that is determined by $\theta^{(a)}$ is used to assure that curvature condition 4.27 holds for the approximated Hessian matrix, especially in case of the constrained quadratic sub-problems that occur in step 6 of algorithm 8.

Deeper theoretical justification for the BFGS-method, the dampening procedure and especially its global convergence properties are e.g. given by Geiger and Kanzow (2002, pp. 256 ff), Gill, Murray and Wright (1981, pp. 116 ff), Jarre and Stoer (2004, pp. 330 ff), Nocedal and Wright (1999, pp. 193 ff, 540 ff) and Saad (2003, pp. 32 ff).

5 Approaches for Estimation from Non-probability Samples

In the previous chapters, the challenges in dealing with non-probability samples and their potential selectivity are discussed, together with approaches for assessing these issues. Based on that discussion (and the references cited therein), it becomes evident that the limitations and pitfalls of such data have to be considered and as far as possible compensated when it comes to estimation from non-probability samples. Consequently, various methods to deal with potential selectivity are proposed in the scientific literature. In the following, a summarizing overview of these strategies is given. In addition, new methods for estimation from non-probability samples are proposed. These are constituted by semi-parametric artificial neural networks as well as a calibrated version thereof, which extend and integrate the ideas underlying some of the pre-existing estimation approaches.

As described in section 2.3, the challenges of non-probability samples can be characterized as potential coverage errors and missingness of information that would be required for classical design-based estimation. This missing information corresponds to

a) (parts of the) variables of interest $\boldsymbol{Y}$ that are not observed in an (adequate) probability sample, e.g. the auxiliary data introduced in chapter 3, and

b) the design weights $\boldsymbol{w}^{\mathsf{nps}}$ (or equivalently inclusion probabilities $\boldsymbol{\pi}^{\mathsf{nps}}$) for the non-probability sample

(cf. also Yang and Kim, 2018, p. 3), which in principle constitute two separable issues. Correspondingly, methods proposed for estimation from non-probability samples can be divided into two broader paradigms (cf. e.g. Baker et al., 2013b, pp. 96 f; Buelens, Burger and van den Brakel, 2018, pp. 329 ff; Valliant and Dever, 2011, p. 109). Each of these paradigms tackles one the of issues ***a)*** and ***b)*** in resemblance to established methods that are used for non-response adjustment in a probability sampling context (cf. e.g. van Buuren, 2018; Little and Rubin, 2019; Särndal and Lundström, 2005; Särndal, 2011; Kott, 2006) and observational studies (cf. e.g. Cochran, Moses and Mosteller, 1983; Rosenbaum, 2010; Rubin, 2006). The common ground underlying both paradigms is that they aim at accounting for selection bias by using some auxiliary variables $\boldsymbol{X}$ or $\boldsymbol{Z}$ which (ideally) assure conditional independence of $\boldsymbol{Y}$ and the inclusion indicator $\boldsymbol{r}^{\mathsf{nps}}$, e.g.

$$(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{r}^{\mathsf{nps}}) \,|\, \boldsymbol{X} \quad , \tag{5.1}$$

such that selectivity is MAR (cf. sections 2.3 and 3.5). This is typically referred to as the *conditional independence assumption* (cf. e.g. Bethlehem and Biffignandi, 2012, p. 394; Schonlau et al., 2009, p. 299). The methods discussed in chapter 3 can provide an indication on variables that are useful for fulfilling this assumption. To make use of such variables, it is typically required for any of the following methods that some information about $\boldsymbol{X}$ or $\boldsymbol{Z}$ outside the non-probability sample is available (cf. Buelens, Burger and van den Brakel, 2018, p. 340; Kim et al., 2018, p. 18). To illustrate how and which auxiliary information is used, a schematic representation based on the discussion in chapter 2 is given in figure 5.1 (cf. also Yang and Kim, 2018, p. 3). As before, the sets of variables $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ do not necessarily need to be mutually exclusive. Furthermore, the probability sample can by definition coincide with the whole population (cf. sections 2.2 and 2.3; Pfeffermann, 2011, p. 117).

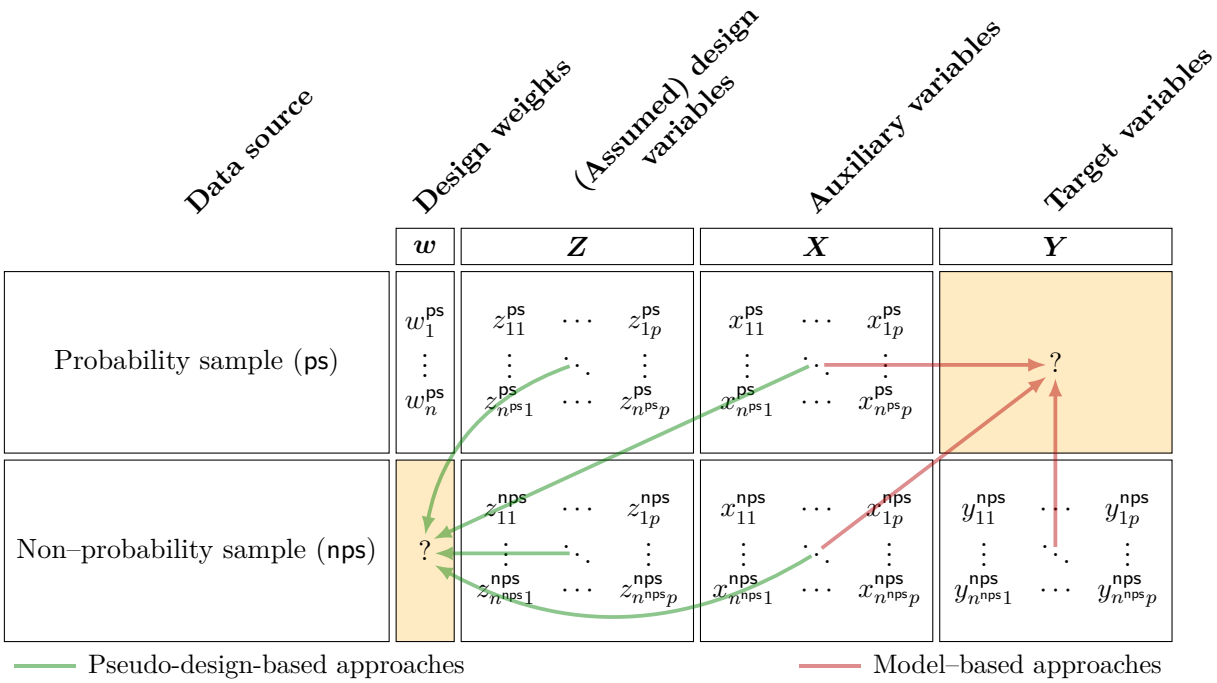


Figure 5.1: Schematic representation of estimation approaches for non-probability samples

The *model-based* paradigm attempts to predict the variable(s) of interest for an auxiliary data set to solve the outlined issue ***a)***. It relies on models for expressing certain aspects of the target variables $\boldsymbol{Y}$ in dependency of the auxiliaries $\boldsymbol{X}$, assuming an underlying structural relationship between both groups of variables. These models can be fit in the non-probability sample, while the distribution of $\boldsymbol{X}$ or at least some of its properties are considered to be known also for the probability reference sample. Estimation and inference for statistics of $\boldsymbol{Y}$ is then based on this distribution of $\boldsymbol{X}$ by using the structural relationship expressed by the model. This is typically done by predicting $\boldsymbol{Y}$ for the probability sample, which leads to imputation of entire variables and is thus also called *'mass imputation'* (e.g. Beręsewicz, 2016, p. 79; Kim et al., 2018, p. 1). A difficulty with this paradigm is that model-based methods usually require a separate prediction model for every single target variable (cf. Baker et al., 2013a, p. 10; Boonstra and Buelens, 2011, p. 5; Buelens, Burger and van den Brakel, 2018, pp. 329 f; Japec et al., 2015, p. 867).

As an alternative, the *pseudo-design-based* paradigm tackles issue ***b)*** by interpreting non-probability samples in the classical design-based context of random sampling. As there is no known probability sampling design in this case, the aim is to obtain weights based on auxiliary and/or assumed design variables $\boldsymbol{X}$ and $\boldsymbol{Z}$. These *pseudo-design weights* are intended to mimic classical design weights in probability sampling (cf. definition 2.13). Estimation is then performed as if the non-probability sample was generated by a probability sampling design yielding the pseudo-design weights. In contrast to the model-based paradigm, one set of weights is usually intended for use with all variables of interest (cf. Buelens et al., 2012, p. 10; Breidt and Opsomer, 2017, p. 196).

The model- and the pseudo-design-based paradigm for non-probability samples are often considered as mutually exclusive (cf. e.g. Buelens, Burger and van den Brakel, 2018; Elliott and Valliant, 2017; Kim et al., 2018). Nevertheless, there are some inclusive proposals to combine methods from both frameworks, attempting to integrate desirable properties from both (cf. e.g. Beaumont, 2000; Gelman et al., 2016b; Pfeffermann and Sikov, 2011; Valliant and Dever, 2011, p. 109; Wang et al., 2015).

These two paradigms and their syntheses partition the variety of methods for handling non-probability samples and shape the scientific discussion (cf. e.g. Baker et al., 2013b; Buelens, Burger and van den Brakel, 2018; Chen, Valliant and Elliott, 2019; Elliott and Valliant, 2017; Kim et al., 2018; Yang and Kim, 2018). Since one purpose of this thesis is to summarize and compare methods proposed for non-probability samples, an overview of important realizations from both paradigms is given in the following section. Model-based methods are discussed first in section 5.1. Some of these are required for pseudo-design-based approaches as well, which are considered in section 5.2. Syntheses between both lines of thought are then presented in section 5.3. Inferential approaches, which typically refer to one of these paradigms as well, are discussed in section 5.4.

5.1 Model-based Methods: Prediction

As outlined above, the model-based paradigm aims at modeling the target variables in different ways. These variables of interest are observed in the non-probability sample, and their unobserved values or distributions in a probability reference sample, which may be the whole population, are predicted (cf. figure 5.1). Design-based estimation can then be applied to the imputed reference data set. Prediction errors of the imputed variables, however, have to be taken into account in this case as well (cf. e.g. section 5.4).

To obtain predictions, a general statistical or machine learning model for an arbitrary data set denoted by s is defined as a function $\mathbf{m} : \mathbb{R}^{n^{\mathsf{s}} \times p} \times \mathbb{R}^{h \times u} \to \mathbb{R}^{n^{\mathsf{s}} \times o}$. This is usually a prespecified function that maps the matrix of *independent variables* $\boldsymbol{X}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}} \times p}$ to the matrix of predicted *dependent variables* $\widehat{\boldsymbol{Y}}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}} \times o}$. A general matrix of parameters $\boldsymbol{\Theta} \in \mathbb{R}^{h \times u}$ is used for this purpose. Structure and impact of $\boldsymbol{\Theta}$ as well as its dimensions $h, u \in \mathbb{N}$ depend on the applied model (cf. Breiman, 2001b, p. 205; Buelens et al., 2012, p. 9). The model's output

$$\widehat{\boldsymbol{Y}}^{\mathsf{s}} = \begin{bmatrix} \widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot 1} & \dots & \widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot o} \end{bmatrix} \coloneqq \mathbf{m}\left(\boldsymbol{X}^{\mathsf{s}}, \boldsymbol{\Theta}\right) \tag{5.2}$$

is a prediction for

$$\boldsymbol{Y}^{\mathsf{s}} = \widehat{\boldsymbol{Y}}^{\mathsf{s}} + \boldsymbol{E}^{\mathsf{s}} \quad . \tag{5.3}$$

In this context $\boldsymbol{E}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}} \times o}$ is a matrix of *prediction errors (residuals)* of the same dimensions as $\boldsymbol{Y}^{\mathsf{s}}$ (cf. Hastie, Tibshirani and Friedman, 2008, pp. 9 ff).

The model parameters $\boldsymbol{\Theta}$ are usually determined by minimizing a loss-function $\delta : \mathbb{R}^{h \times u} \to \mathbb{R}_{\geq 0}$, which is used to quantify the error when predicting observed values $\boldsymbol{Y}^{\mathsf{s}}$ by $\widehat{\boldsymbol{Y}}^{\mathsf{s}}$. This is referred to as *model fitting* and commonly done by means of the optimization framework presented in chapter 4. In the following discussion and subsequent sections, s generally denotes the data set in which the model is fit to find parameters $\boldsymbol{\Theta}$. The distance function δ to be minimized is denoted with regard to $\boldsymbol{\Theta}$ as its only argument because $\boldsymbol{X}^{\mathsf{s}}$ and $\boldsymbol{Y}^{\mathsf{s}}$ are considered fix in the present finite population sampling context (cf. section 2.2). For example, the quadratic loss functions defined by

$$\begin{aligned} \delta\left(\boldsymbol{\Theta}\right) &= \mathrm{E}\left(\left(\boldsymbol{Y}^{\mathsf{s}} - \mathbf{m}\left(\boldsymbol{X}^{\mathsf{s}}, \boldsymbol{\Theta}\right)\right)^2\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(\left(\boldsymbol{Y}^{\mathsf{s}} - \mathbf{m}\left(\boldsymbol{X}^{\mathsf{s}}, \boldsymbol{\Theta}\right)\right)^2 \,\middle|\, \boldsymbol{X}^{\mathsf{s}}\right)\right) \end{aligned} \tag{5.4}$$

is frequently used for this purpose. Due to the fact that

$$\mathbf{m}\left(\boldsymbol{X}^{\mathsf{s}}, \boldsymbol{\Theta}\right) \;=\; \mathrm{E}\left(\boldsymbol{Y}^{\mathsf{s}} \,|\, \boldsymbol{X}^{\mathsf{s}}\right) \tag{5.5}$$

is optimal in terms of equations 5.4, this conditional expectation is explicitly represented by many models. When using such models in form of equation 5.5, unbiasedness of the overall mean $\mathrm{E}\left(\boldsymbol{Y}^{\mathsf{s}}\right) = \mathrm{E}\left(\mathrm{E}\left(\boldsymbol{Y}^{\mathsf{s}} \,|\, \boldsymbol{X}^{\mathsf{s}}\right)\right)$ is commonly facilitated by including a constant column in $\boldsymbol{X}$. This *intercept column* is typically specified to be the first in $\boldsymbol{X}$ and filled with ones, i.e. $\boldsymbol{x}_{\cdot 1} = \mathbf{1}_{\mathrm{N}\times 1}$, (cf. Berk, 2008, pp. 35 ff; Hastie, Tibshirani and Friedman, 2008, pp. 11 ff). This specification is commonly used in the subsequent discussion.

The concept of minimizing prediction errors for fitting the model requires values of $\boldsymbol{Y}$ that are actually observed and is therefore referred to as 'supervised learning' (cf. e.g. Hastie, Tibshirani and Friedman, 2008, p. 29). As indicated in figures 2.1 and 5.1 as well as the related discussion, a partial overlap of variables between non-probability and reference sample is assumed, such that $\boldsymbol{Y}$ is observed only in the non-probability sample. Therefore, prediction models are typically fit to the non-probability sample ($\mathsf{s} = \mathsf{nps}$), for which $\boldsymbol{X}^{\mathsf{nps}}$ and $\boldsymbol{Y}^{\mathsf{nps}}$ are both available (cf. section 2.3). In line with the referred literature, the following discussion is hence focused on supervised learning. 'Unsupervised learning' (cf. e.g. Hastie, Tibshirani and Friedman, 2008, pp. 485 ff; Ripley, 1996, pp. 287 ff) is usually not considered to compensate for non-probability sample selection because it would use only the overlapping variables and, thus, neglect observed values and predictions for target variables $\boldsymbol{Y}$. As indicated in equation 5.5, supervised learning is generally based on properties of the conditional distribution $\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{x}_{i\cdot}\right)$. This may, but does not necessarily, imply to explicitly model this conditional distribution as a whole (cf. Hastie, Tibshirani and Friedman, 2008, p. 485).

Once parameters $\boldsymbol{\Theta}$ are determined by minimizing δ in data set s as described above, the model can be used for prediction in any data set t for which independent variables $\boldsymbol{X}^{\mathsf{t}}$ are observed. Definition 5.2 allows for such predictions even if $\boldsymbol{Y}^{\mathsf{t}}$ is unknown since

$$\widehat{\boldsymbol{Y}}^{\mathsf{t}} = \begin{bmatrix} \widehat{\boldsymbol{y}}^{\mathsf{t}}_{\cdot 1} & \dots & \widehat{\boldsymbol{y}}^{\mathsf{t}}_{\cdot o} \end{bmatrix} \;\coloneqq\; \mathbf{m}\left(\boldsymbol{X}^{\mathsf{t}}, \boldsymbol{\Theta}\right) \tag{5.6}$$

does not depend on $\boldsymbol{Y}^{\mathsf{t}}$. A data set for which predictions are obtained from a model that is fit on a different data set is generally denoted by t in the following discussion and subsequent sections. The main purpose for fitting models in the context of the current section 5.1 is to obtain predictions as in equality 5.6 since they constitute substitutes for unobserved values of the target variables $\boldsymbol{Y}$ (cf. figure 5.1).

The rationale behind this approach is that when assuming conditional independence of sample inclusion and target variables given $\boldsymbol{X}$ (cf. assumption 5.1), the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ in sample s can be used for estimation and inference for statistics of $\boldsymbol{Y}$. By Bayes' theorem and in analogy to equation 2.24, the conditional distribution in the population is defined by

$$\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{x}_{i\cdot}\right) \;=\; \frac{\mathrm{P}\left(r_i^{\mathsf{s}} = 1 \,|\, \boldsymbol{x}_{i\cdot}\right)}{\mathrm{P}\left(r_i^{\mathsf{s}} = 1 \,|\, \boldsymbol{x}_{i\cdot}, \boldsymbol{y}_{i\cdot}\right)} \cdot \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{x}_{i\cdot}, r_i^{\mathsf{s}} = 1\right) \qquad , \tag{5.7}$$

where $\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{x}_{i\cdot}, r_i^{\mathsf{s}} = 1\right) = \mathrm{f}_{\boldsymbol{Y}^{\mathsf{s}}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{x}_{j\cdot}\right)$ is the conditional distribution in sample s, and $\boldsymbol{r}^{\mathsf{s}}$ is the inclusion indicator for this sample (cf. equation 2.2). When the conditional independence assumption holds, it follows that $\mathrm{P}\left(r_i^{\mathsf{s}} = 1 \,|\, \boldsymbol{x}_{i\cdot}\right) = \mathrm{P}\left(r_i^{\mathsf{s}} = 1 \,|\, \boldsymbol{x}_{i\cdot}, \boldsymbol{y}_{i\cdot}\right)$ (cf.

Dawid, 1979, p. 3), and if this conditional probability is positive for all $i \in \mathcal{S}^{\mathrm{P}}$, an unbiased estimate for the conditional distribution $\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \mid \boldsymbol{x}_{j\cdot}\right)$ can be obtained from sample s. If, furthermore, $\mathrm{f}_{\boldsymbol{X}}\left(\boldsymbol{x}_{j\cdot}\right)$ is considered to be known for the population, the distribution of $\boldsymbol{Y}$ can be computed by integrating out all p variables in $\boldsymbol{x}_{j\cdot}$:

$$\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot}\right) \;=\; \int \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \mid \boldsymbol{x}_{j\cdot}\right) \cdot \mathrm{f}_{\boldsymbol{X}}\left(\boldsymbol{x}_{j\cdot}\right) \, d\,\boldsymbol{x}_{j\cdot} \qquad . \tag{5.8}$$

The same reasoning holds when using an unbiased estimate for $\mathrm{f}_{\boldsymbol{X}}\left(\boldsymbol{x}_{j\cdot}\right)$ that is obtained from data set t, but one has to consider its uncertainty when it comes to inference. In special cases, some marginal information for $\boldsymbol{X}$ instead of the full distribution may be sufficient as well (cf. appendix B; Pfeffermann, 2011; Smith, 1983).

The typical case for the model-based paradigm in the context of non-probability samples is to fit a model in the non-probability sample ($\mathsf{s} = \mathsf{nps}$), using the observed values of $\boldsymbol{Y}^{\mathsf{nps}}$ and $\boldsymbol{X}^{\mathsf{nps}}$ for supervised learning. This model represents the conditional distribution $\mathrm{f}_{\boldsymbol{Y}^{\mathsf{nps}}}\left(\boldsymbol{y}_{i\cdot} \mid \boldsymbol{x}_{j\cdot}\right)$ or its relevant properties, e.g. the conditional mean (cf. equalities 5.4 and 5.5). As a consequence of equality 5.7, this estimate is unbiased for the population's conditional distribution if conditional independence assumption 5.1 and full coverage of the population are fulfilled (cf. appendix B; Pfeffermann, 2011; Smith, 1983). The strategy represented by equality 5.8 is then commonly implemented by explicitly imputing (predicting) the unknown values of $\boldsymbol{Y}^{\mathsf{ps}}$ by $\widehat{\boldsymbol{Y}}^{\mathsf{ps}}$ for the probability reference sample ($\mathsf{t} = \mathsf{ps}$; cf. equation 5.6). In that way, the distribution $\mathrm{f}_{\boldsymbol{X}^{\mathsf{ps}}}\left(\boldsymbol{x}_{j\cdot}\right)$ of observed values $\boldsymbol{X}^{\mathsf{ps}}$ is used to obtain a substitute $\mathrm{f}_{\widehat{\boldsymbol{Y}}^{\mathsf{ps}}}\left(\widehat{\boldsymbol{y}}_{i\cdot}\right) \approx \mathrm{f}_{\boldsymbol{Y}^{\mathsf{ps}}}\left(\boldsymbol{y}_{i\cdot}\right)$ from the model. Estimation is then based on this modeled distribution $\mathrm{f}_{\widehat{\boldsymbol{Y}}^{\mathsf{ps}}}\left(\widehat{\boldsymbol{y}}_{i\cdot}\right)$ since the actual values $\boldsymbol{Y}^{\mathsf{ps}}$ are not observed (cf. figure 5.1). A special case of this approach occurs if information about the independent variables is known on population level, such that predictions can be made for the whole population ($\mathsf{t} = \mathsf{ps} = \mathrm{P}$). In a finite population, inference is then solely based on the model's prediction error (cf. Buelens, Burger and van den Brakel, 2018, pp. 329 f; Kim, Kwon and Paik, 2016; Särndal, 1978, p. 34; Sverchkov and Pfeffermann, 2004, p. 81).

In case of *classical statistical models*, $\mathbf{m}$ is typically used to express a functional relationship between $\boldsymbol{X}^{\mathsf{s}}$ and $\boldsymbol{Y}^{\mathsf{s}}$ by explicitly representing the underlying conditional distribution $\mathrm{f}_{\boldsymbol{Y}^{\mathsf{s}}}\left(\boldsymbol{y}_{i\cdot} \mid \boldsymbol{x}_{j\cdot}\right)$. An extensive theoretical framework is built around different assumptions about how this distribution is modeled (cf. e.g. Wood, 2017; Hastie and Tibshirani, 1990; McCullagh and Nelder, 1989), with linear or generalized linear models described in sections 5.1.2 and 5.1.3 as the most popular examples (cf. Buelens, Burger and van den Brakel, 2018, p. 323). An important aspect for most of these models is the trade-off between interpretability and predictive power that results from their explicit distributional assumptions (cf. Breiman, 2001b, pp. 209 f; Hastie, Tibshirani and Friedman, 2008, p. 304). For example, modeling $\mathrm{f}_{\boldsymbol{Y}^{\mathsf{s}}}\left(\boldsymbol{y}_{i\cdot} \mid \boldsymbol{x}_{j\cdot}\right)$ implies that parameters $\boldsymbol{\Theta}$ can not only represent the dependency between $\boldsymbol{Y}^{\mathsf{s}}$ and $\boldsymbol{X}^{\mathsf{s}}$ in form of equalities 5.2 and 5.3 but also on a distributional level. Model selection can therefore be based on the distribution of $\boldsymbol{Y}$, e.g. when using likelihood-ratio tests or information criteria. This is frequently used to choose a subset of potential variables with regard to their predictive accuracy for modeling $\boldsymbol{Y}$, or to determine shrinkage or other parameters that are not part of the original optimization parameters $\boldsymbol{\Theta}$ (cf. e.g. section 5.1.11). An important further benefit of statistical models relying on distributional assumptions is that the actual or asymptotic distribution $\mathrm{f}_{\boldsymbol{\Theta}}\left(\boldsymbol{\Theta}\right)$ of $\boldsymbol{\Theta}$ can be derived from that of $\boldsymbol{Y}^{\mathsf{s}}$. Confidence intervals or statistical tests can then be based on $\mathrm{f}_{\boldsymbol{\Theta}}\left(\boldsymbol{\Theta}\right)$ to facilitate inference for the parameters (cf. e.g. Akaike, 1973; Green

and Silverman, 1994, pp. 95 ff; Hastie and Tibshirani, 1990, pp. 65 ff, 155 ff; Hastie, Tibshirani and Friedman, 2008, pp. 219 ff; Lee and Nelder, 1996, pp. 635 f; Lee, Nelder and Pawitan, 2006, pp. 97 ff, 183 ff; Wood, 2017). Alternative methods for model selection and inference are discussed in sections 5.1.11 and 5.4 as well as the references cited therein. Conditional independence assumption 5.1 is nevertheless a core element when it comes to model-based estimation from non-probability samples. Variable selection should therefore always consider variables that allow fulfilling this assumption as far as possible, which can e.g. be identified using the methods discussed in chapter 3. Similarly as for the tests discussed in section 3.4, inference for models and parameters has to be applied cautiously in the context of non-probability samples since the conditional distribution expressed by the model is not unbiased if conditional independence is violated (cf. equality 5.7).

Extending the view from such classical statistical models, Buelens, Burger and van den Brakel (2018), Lee, Lessler and Stuart (2010, p. 338), Pfeffermann (2015, p. 431) as well as Rafei, Flannagan and Elliott (2020, p. 175) emphasize the importance and potential of *machine learning models* for dealing with non-probability samples and their possible selectivity. In this context, simplicity and interpretability regarding the relation between independent and dependent variables is usually of less importance than for the classical statistical models. However, despite originating and being motivated in somewhat different contexts, various similarities and overlaps between both types of models exist. In many cases, classical statistical models can be interpreted as machine learning methods since the common aim of both is to achieve good predictions for $\boldsymbol{Y}$ based on $\boldsymbol{X}$, and similar strategies are applied for this purpose. For example, coinciding loss functions and model specifications are commonly used in either setting, leading to similar distributional characteristics and predictions (cf. e.g. equations 5.4 and 5.5). Therefore, classical statistical and machine learning models are closely related to and often not uniquely differentiated from each other (cf. Breiman, 2001a, p. 23; Buelens, Burger and van den Brakel, 2018, p. 323). Machine learning methods are also labeled "algorithmic models" (Breiman, 2001b, p. 199) or "learning algorithms" (Hastie, Tibshirani and Friedman, 2008, p. 29) because their typical focus is to find algorithms that are capable of producing good predictions. Unlike classical statistical models, machine learning methods do not generally assume a prespecified type of underlying conditional distribution to be modeled for the purpose of prediction, but merely rely on a functional relationship between $\boldsymbol{X}^{\mathsf{s}}$ and $\boldsymbol{Y}^{\mathsf{s}}$ as in equations 5.2 and 5.3. The sole distributional assumption that is commonly made in this context concerns the requirement that observations are independently and identically distributed (i.i.d.), which is necessary for many of the model fitting techniques (cf. e.g. Breidt and Opsomer, 2017; Breiman, 2001b; Hastie, Tibshirani and Friedman, 2008; Lee, Lessler and Stuart, 2010, p. 2; Sra, Nowozin and Wright, 2012).

Since all model-based methods rely more or less explicitly on the common distribution of $\boldsymbol{Y}^{\mathsf{s}}$ and $\boldsymbol{X}^{\mathsf{s}}$, potential difficulties in this context lie on the one hand in identifying proper prediction models for all variables of interest (cf. Baker et al., 2013a, p. 76; Buelens, Burger and van den Brakel, 2015, p. 6). On the other hand, even a good model rarely provides perfect predictions for real values of $\boldsymbol{Y}$, such that some uncertainty is usually left after applying these methods. The same holds for potentially remaining selection biases since perfect conditional independence of sample inclusion and target variables is often not realistic (cf. Buelens et al., 2012, pp. 9 f; Magnussen, 2015, p. 317).

The scientific discourse revolves around a bandwidth of specific models for handling non-probability samples (cf. e.g. Buelens, Burger and van den Brakel, 2018; Elliott and

Valliant, 2017; Kim et al., 2018; Rafei, Flannagan and Elliott, 2020; Yang and Kim, 2018). Since one aim of this thesis is to summarize and compare methods for non-probability samples, an overview is given throughout the following sections. Note that while some of these models are straightforwardly representable for a general matrix of predictions $\widehat{\boldsymbol{Y}}^{\mathsf{s}}$ as in definition 5.2, others are written for single output variables $\boldsymbol{y}^{\mathsf{s}}_{\cdot l}$ to not overcomplicate notation. However, the latter can be applied to each variable $\boldsymbol{y}^{\mathsf{s}}_{\cdot l}$ for $l = 1, \ldots, o$ separately to constitute a matrix of predictions $\widehat{\boldsymbol{Y}}^{\mathsf{s}}$ if the residuals' covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{E}^{\mathsf{s}}}$ is assumed to be diagonal. If this assumption is not reasonable, a model representing all required interactions of dependent variables (cf. McCullagh and Nelder, 1989, pp. 219 ff) or optimization based on a decorrelating (e.g. Mahalanobis) transformation can be used (cf. Hastie, Tibshirani and Friedman, 2008, pp. 84 ff; Kessy, Lewin and Strimmer, 2018; Schaid et al., 2019, pp. 113 f; Wood, Pya and Säfken, 2016, pp. 15 f).

The following discussion is structured as follows. The very basic ideas of modeling are closely related to those of matching (cf. section 3.5), which is adapted for prediction in section 5.1.1. However, models that apply stronger structural assumptions are more common. Starting with the presumably most popular ones of these (cf. Berk, 2008, p. 8), (generalized) linear models are discussed in sections 5.1.2 and 5.1.3, extending the view to (generalized) additive models in section 5.1.4 and the corresponding mixed models in section 5.1.5. Based on these methods, (multivariate adaptive) regression splines are reviewed in sections 5.1.6 and 5.1.7. Artificial neural networks are introduced in section 5.1.8, followed by a proposal for extending and integrating these with regression splines in form of semi-parametric artificial neural networks that is presented in section 5.1.9. As an alternative non-linear prediction model, support vector machines are discussed in section 5.1.10. Shrinkage methods are an important part or extension applicable to the fitting techniques used for all of these models, and therefore summarized in section 5.1.11.

5.1.1 Matching

In section 3.5, matching is introduced as a way to compare the conditional distribution $\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \mid \boldsymbol{x}_{i\cdot}\right)$ in non-probability and reference sample. As discussed with regard to equation 5.5, predictions in a supervised learning context are typically based on this conditional distribution. Therefore, an apparent way to obtain predictions is to rely on matching again. To predict $\boldsymbol{Y}^{\mathsf{t}}$ for data set t based on observed values in data set s, a set $\mathcal{J}^{(i)} \subseteq \mathcal{S}^{\mathsf{s}}$ of units which are matched to each unit $i = 1, \ldots, n^{\mathsf{t}}$ observed in t is defined by

$$\mathcal{J}^{(i)} \coloneqq \left\{ j : \delta\left(\boldsymbol{x}^{\mathsf{t}}_{i\cdot}, \boldsymbol{x}^{\mathsf{s}}_{j\cdot}\right) \leq a_i \right\} \quad . \tag{5.9}$$

Similar as for equation 3.23, $\boldsymbol{a} \in \mathbb{R}^{n^{\mathsf{t}}}_{\geq 0}$ is a vector of cut-off constants. Due to definition 5.9, the rows $\boldsymbol{X}^{\mathsf{s}}_{\mathcal{J}^{(i)}\cdot}$ for the matched units each are (at least) similar to $\boldsymbol{x}^{\mathsf{t}}_{i\cdot}$. Predictions that are based on the conditional expectation $\mathrm{E}\left(\boldsymbol{Y}^{\mathsf{s}} \mid \boldsymbol{X}^{\mathsf{s}}\right)$ as in equality 5.5 can, therefore, be approximated by the expected values over these observations, i.e.

$$\widehat{\boldsymbol{y}}^{\mathsf{t}}_{i\cdot} \coloneqq \frac{1}{\left|\mathcal{J}^{(i)}\right|} \sum_{j \in \mathcal{J}^{(i)}} \boldsymbol{y}^{\mathsf{s}}_{j\cdot} \quad \text{for all} \quad i \in \mathcal{S}^{\mathsf{t}} \quad , \tag{5.10}$$

where the number of elements in $\mathcal{J}^{(i)}$ is denoted by $\left|\mathcal{J}^{(i)}\right|$. In the general model formulation 5.2, parameters $\boldsymbol{\Theta} = \left[\mathcal{J}^{(1)} \ \ldots \ \mathcal{J}^{(n^{\mathsf{t}})}\right]^{\mathsf{T}} \in \mathbb{N}^h$ represent the concatenation of all $h = \sum_{i \in \mathcal{S}^{\mathsf{t}}} \left|\mathcal{J}^{(i)}\right|$

sets of matching observations for all units $i \in \mathcal{S}^{\mathsf{t}}$. These parameters can be found by using the framework of matching techniques described in section 3.5, in which different distance measures as well as cut-off values can be applied. As before, exact matching is often infeasible since there are usually no two exactly identical rows in $\boldsymbol{X}$ if it contains more than a few categorical variables. In that case, conditioning is relaxed to the neighborhood-region in data set s instead of a single point (cf. equation 5.5; Hastie, Tibshirani and Friedman, 2008, p. 18). These ideas are often used for imputation methods (cf. e.g. Andridge and Little, 2010; Okner, 1972).

However, there are several problems with such applications of matching. On the one hand, the approximation via neighborhood-regions gets worse with an increasing number of matching variables because similarity is less easily achievable in higher dimensions. On the other hand, the predictions are rather instable since the matching model is constituted by a locally constant function. To overcome these limitations, models that rely on stronger structural assumptions for the dependency between $\boldsymbol{X}$ and $\boldsymbol{Y}$ are therefore more common (cf. Hastie, Tibshirani and Friedman, 2008, pp. 16 ff). The following sections introduce a number of such models, starting with linear regression in section 5.1.2.

5.1.2 Linear Models

Dating back to the beginning of the 19th century (cf. Gauss, 1809; Legendre, 1805, both cited in Farebrother, 1999, p. 165), linear regression models are still the presumably most prevalent statistical models. Beyond their canonical value, they constitute a valuable foundation for the more complex models that are introduced subsequently and can often be interpreted as extensions and generalizations of linear models (cf. e.g. Berk, 2008, p. 8; Hastie, Tibshirani and Friedman, 2008, p. 35; James et al., 2013, p. 59; Royall, 1970).

A linear regression model with potentially multiple independent and dependent variables for data set t is defined by

$$\widehat{\boldsymbol{Y}}^{\mathsf{t}} \;=\; \mathbf{m}\left(\boldsymbol{X}^{\mathsf{t}}, \boldsymbol{\Theta}\right) \coloneqq \boldsymbol{X}^{\mathsf{t}}\boldsymbol{\beta} \qquad . \tag{5.11}$$

It is called a linear regression due to the fact that $\mathbf{m}$ represents a linear function of $\boldsymbol{X}^{\mathsf{t}}$. Note that in general, optimization is denoted with respect to $\boldsymbol{\Theta}$ to avoid ambiguities when it comes to the iterative updating procedures and multiple types of parameters in the following sections. Therefore, the optimization parameters are defined by $\boldsymbol{\Theta} \coloneqq \boldsymbol{\beta} \in \mathbb{R}^{p \times o}$ in this context, using $\boldsymbol{\beta}$ in definition 5.11 because it is the common symbol for regression coefficients. Assuming that an intercept column $\boldsymbol{x}^{\mathsf{t}}_{\cdot 1} = \mathbf{1}_{n^{\mathsf{s}} \times 1}$ is included in $\boldsymbol{X}^{\mathsf{t}}$, β_1 is the *intercept* of the linear model.

As described in the introduction to section 5.1, the data set in which the model is fit is generally denoted by s. For fitting a linear regression model, the *least squares approach* is commonly used (cf. e.g. equation 5.4; Hastie, Tibshirani and Friedman, 2008, pp. 44 ff). It corresponds to applying the weighted residual sum of squares as distance function for estimating $\boldsymbol{\Theta}$, which is defined by

$$\begin{aligned} \delta\left(\boldsymbol{\Theta}\right) \;&=\; \left(\boldsymbol{E}^{\mathsf{s}}\right)^{\mathsf{T}} \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{s}}\right) \boldsymbol{E}^{\mathsf{s}} \\ &=\; \left(\boldsymbol{Y}^{\mathsf{s}} - \boldsymbol{X}^{\mathsf{s}}\boldsymbol{\beta}\right)^{\mathsf{T}} \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{s}}\right) \left(\boldsymbol{Y}^{\mathsf{s}} - \boldsymbol{X}^{\mathsf{s}}\boldsymbol{\beta}\right) \qquad . \end{aligned} \tag{5.12}$$

The residuals $\boldsymbol{E}^{\mathsf{s}}$ represent the observed prediction errors, and design weights $\boldsymbol{w}^{\mathsf{s}}$ are applied to obtain a HT-estimator for the sum of these squared errors in the population (cf. equations 2.15 and 5.3; Binder, 1983, p. 282; Pfeffermann, 2011, p. 122). Since the

conditional expectation is the best prediction in terms of squared loss, it is equivalent to assume that $\mathrm{E}\left(\boldsymbol{Y}^{\mathsf{s}} \mid \boldsymbol{X}^{\mathsf{s}}\right) = \boldsymbol{X}^{\mathsf{s}}\boldsymbol{\beta}$ (cf. equation 5.5; Hastie, Tibshirani and Friedman, 2008, p. 18).

The parameters are found using the methods presented in section 4.2.1. Setting the Jacobian matrix

$$\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}\right) = 2 \cdot \left(\boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta} - \boldsymbol{Y}^{\mathsf{s}}\right)^{\mathsf{T}} \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{s}}\right) \boldsymbol{X}^{\mathsf{s}} \tag{5.13}$$

to zero yields

$$\boldsymbol{\Theta} = \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{s}}\right) \boldsymbol{X}^{\mathsf{s}}\right)^{-1} \left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{s}}\right) \boldsymbol{Y}^{\mathsf{s}} \tag{5.14}$$

as a closed form solution for the minimum of equation 5.12. This is equivalent to the result of a single iteration of the Newton-Raphson algorithm 5 since the Hessian matrix in this case is given by

$$\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}\right) = \left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{s}}\right) \boldsymbol{X}^{\mathsf{s}} \tag{5.15}$$

(cf. appendix B.4.1). Furthermore, this least squares estimate 5.14 for the linear regression's coefficients coincides with the generalized method of moments estimator and the maximum likelihood estimator when assuming that the residuals $\boldsymbol{e}^{\mathsf{s}}_{\cdot l}$ all follow a normal distribution. In the linear regression model, this residual distribution is equivalent to the conditional distributions of $\boldsymbol{Y}$ given $\boldsymbol{X}$ (cf. Amemiya, 1985, pp. 4 ff; Greene, 2008, pp. 168, 456). To obtain models for other kinds of conditional distributions, generalized linear models are introduced in the following section 5.1.3.

5.1.3 Generalized Linear Models

As described in section 5.1.2, linear regression is designed for predicting continuous variables $\boldsymbol{Y}$ and can be interpreted as assuming a conditional normal distribution for $\boldsymbol{Y}$ given $\boldsymbol{X}$. For many variables, however, it is not reasonable to assume such a conditional distribution, e.g. in case of count or categorical data. For such variables, linear regression is not an adequate model and can produce invalid results due to violated assumptions. To facilitate models and predictions for other types of variables and conditional distributions, Nelder and Wedderburn (1972) propose a unified framework for generalized linear models (GLMs).

Three assumptions are used as foundation for this generalization (cf. Nelder and Wedderburn, 1972, pp. 371 f; McCullagh and Nelder, 1989, pp. 21 ff; Venables and Ripley, 2002, pp. 183 ff):

a) The variable of interest $\boldsymbol{y}^{\mathsf{t}}_{\cdot l} \in \mathbb{R}^{n^{\mathsf{t}} \times 1}$ in an arbitrary data set t is a realization from a n^{t}-dimensional probability density function belonging to the exponential family, with mean $\boldsymbol{\mu}^{(\mathsf{t})}_{\boldsymbol{y}_{\cdot l}} \in \mathbb{R}^{n^{\mathsf{t}}}$ and covariance matrix $\boldsymbol{\Sigma}^{(\mathsf{t})}_{\boldsymbol{y}_{\cdot l}} \in \mathbb{R}^{n^{\mathsf{t}} \times n^{\mathsf{t}}}$. The general form for its density $\mathrm{f}_{\boldsymbol{y}^{\mathsf{t}}_{\cdot 1}} : \mathbb{R}^{n^{\mathsf{t}}} \to [0;1]^{n^{\mathsf{t}}}$ is

$$\mathrm{f}_{\boldsymbol{y}^{\mathsf{t}}_{\cdot 1}}\left(\boldsymbol{y}^{\mathsf{t}}_{\cdot l} \,|\boldsymbol{\theta}, \phi\right) = \exp\left(\left(\mathbf{a}\left(\phi\right)\right)^{-1} \left(\boldsymbol{y}^{\mathsf{t}}_{\cdot l} \circ \boldsymbol{\theta} - \mathbf{b}\left(\boldsymbol{\theta}\right)\right) + \mathbf{c}\left(\boldsymbol{y}^{\mathsf{t}}_{\cdot l}, \phi\right)\right) \tag{5.16}$$

in vector-notation. Here, $\mathbf{a} : \mathbb{R}_{\geq 0} \to \mathbb{R}^{n^{\mathsf{t}} \times n^{\mathsf{t}}}$, $\mathbf{b} : \mathbb{R}^{n^{\mathsf{t}}} \to \mathbb{R}^{n^{\mathsf{t}}}$ and $\mathbf{c} : \mathbb{R}^{n^{\mathsf{t}}} \times \mathbb{R}_{\geq 0} \to \mathbb{R}^{n^{\mathsf{t}}}$ denote component functions of the exponential family. Further, $\boldsymbol{\theta} \in \mathbb{R}^{n^{\mathsf{t}}}$ is the *parameter of the exponential family*, and $\phi \in \mathbb{R}_{\geq 0}$ is a *dispersion parameter*. For the generalized linear models, it must hold that $\mathbf{a}\left(\phi\right)$ is positive definite, $\mathbf{b}\left(\boldsymbol{\theta}\right)$ is twice differentiable and normalization of $\mathrm{f}_{\boldsymbol{y}^{\mathsf{t}}_{\cdot 1}}\left(\boldsymbol{y}^{\mathsf{t}}_{\cdot l} \,|\boldsymbol{\theta}, \phi\right)$ is possible.

b) The distribution of $\boldsymbol{y}_{\cdot l}^{\mathsf{t}}$ is influenced by the independent variables $\boldsymbol{X}^{\mathsf{t}}$ through a linear combination $\boldsymbol{\eta}^{\mathsf{t}} \in \mathbb{R}^{n^{\mathsf{t}}}$, called the *systematic component*.

c) The predictor $\boldsymbol{\eta}^{\mathsf{t}}$ is related to the expectation by a smooth and invertible linking function $\boldsymbol{l} : \mathbb{R}^{n^{\mathsf{t}}} \to \mathbb{R}^{n^{\mathsf{t}}}$, such that $\boldsymbol{l}\left(\boldsymbol{\mu}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})}\right) = \boldsymbol{\eta}^{\mathsf{t}}$. If $\boldsymbol{l}\left(\boldsymbol{\mu}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})}\right) = \boldsymbol{\theta}$ holds, $\boldsymbol{l}$ is called a *canonical* link function.

For distributions belonging to the exponential family, it holds that the expected value and covariance for observations $\boldsymbol{y}_{\cdot l}^{\mathsf{t}}$ are determined by

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})} &= \frac{\partial\left(\mathbf{b}\left(\boldsymbol{\theta}\right)\right)}{\partial\left(\boldsymbol{\theta}\right)} \\ \boldsymbol{\Sigma}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})} = \boldsymbol{\Sigma}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})}\left(\phi\right) = \mathbf{a}\left(\phi\right) \frac{\partial^2\left(\mathbf{b}\left(\boldsymbol{\theta}\right)\right)}{\partial\left(\boldsymbol{\theta}\right)^2} &= \mathbf{a}\left(\phi\right) \mathrm{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})}\right) \end{aligned} \tag{5.17}$$

for $\boldsymbol{\Sigma}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})} : \mathbb{R}_{\geq 0} \to \mathbb{R}^{n^{\mathsf{t}} \times n^{\mathsf{t}}}$. These equalities are important for some of the following derivations in relation to this family of distributions (cf. McCullagh and Nelder, 1989, pp. 29, 42; Simonoff, 2003, p. 126; Wood, 2017, pp. 62 f).

In the generalized linear models, different choices of $\mathbf{a}, \mathbf{b}, \mathbf{c}, \boldsymbol{\theta}$ and ϕ lead to a variety of distributions. For example, Nelder and Wedderburn (1972, p. 375) as well as McCullagh and Nelder (1989, p. 28) show that the linear regression introduced in section 5.1.2 results from assuming

a) a normal distribution for dependent variable, i.e. $\boldsymbol{y}_{\cdot l}^{\mathsf{t}} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})}, \sigma^2\right)$, by choosing

$$\begin{aligned} \boldsymbol{\theta} &= \boldsymbol{\mu}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})} \qquad \phi = \sigma^2 \qquad \mathbf{a}\left(\phi\right) = \phi \cdot \mathbf{I}_{n^{\mathsf{t}}} \qquad \mathbf{b}\left(\boldsymbol{\theta}\right) = \frac{1}{2} \cdot \boldsymbol{\theta}^{\circ 2} \\ \mathbf{c}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{t}}, \phi\right) &= -\frac{1}{2} \cdot \left(\mathbf{diag}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{t}}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{t}}\right)^{\mathsf{T}}\left(\mathbf{a}\left(\phi\right)\right)^{-1}\right) + \log\left(\mathbf{a}\left(\phi\right)\right) + \log\left(2\pi\right)\right) \ , \end{aligned} \tag{5.18}$$

b) the systematic component $\boldsymbol{\eta}^{\mathsf{t}} \coloneqq \boldsymbol{X}^{\mathsf{t}}\boldsymbol{\beta}$ for $\boldsymbol{\beta} \in \mathbb{R}^p$, and

c) the identity as canonical link function, such that

$$\boldsymbol{l}\left(\boldsymbol{\mu}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})}\right) = \boldsymbol{\theta} = \boldsymbol{\mu}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})} \qquad . \tag{5.19}$$

In case of a dichotomous variable of interest $\boldsymbol{y}_{\cdot l}^{\mathsf{t}}$, the *logit model* can be applied. It is based on the following assumptions (cf. Nelder and Wedderburn, 1972, p. 375; McCullagh and Nelder, 1989, p. 30):

a) $\boldsymbol{y}_{\cdot l}^{\mathsf{t}}$ is a realization from a binomial distribution with h trials and success probability $\boldsymbol{p} = \boldsymbol{\mu}_{\boldsymbol{y}\cdot l}^{(\mathsf{t})} \cdot h^{-1}$, i.e. $\boldsymbol{y}_{\cdot l}^{\mathsf{t}} \sim \mathrm{B}\left(h, \boldsymbol{p}\right)$. This is expressed by

$$\begin{aligned} \boldsymbol{\theta} &= \log\left(\boldsymbol{p} \oslash \left(1 - \boldsymbol{p}\right)\right) \qquad \phi = 1 \qquad \mathbf{a}\left(\phi\right) = \mathbf{I}_{n^{\mathsf{t}}} \\ \mathbf{b}\left(\boldsymbol{\theta}\right) &= h \cdot \log\left(1 + \exp\left(\boldsymbol{\theta}\right)\right) \qquad \mathbf{c}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{t}}, \phi\right) = \log\left(\binom{h}{\boldsymbol{y}_{\cdot l}^{\mathsf{t}}}\right) \ , \end{aligned} \tag{5.20}$$

where the binomial coefficient $\binom{h}{\boldsymbol{y}_{\cdot l}^{\mathsf{t}}}$ is applied element-wise.

b) The systematic component is $\boldsymbol{\eta}^{\mathsf{t}} \coloneqq \boldsymbol{X}^{\mathsf{t}}\boldsymbol{\beta}$ for $\boldsymbol{\beta} \in \mathbb{R}^p$, and

c) the link function is canonical, such that

$$\boldsymbol{l}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{t})}\right) = \boldsymbol{\theta} = \log\left(\boldsymbol{p} \oslash (1 - \boldsymbol{p})\right) \qquad . \tag{5.21}$$

As before (cf. section 5.1.2), the data set used for fitting a model is generally denoted by s. The regression parameters $\boldsymbol{\beta}$ are typically obtained by maximizing the log-likelihood resulting from equation 5.16, which is referred to as *maximum likelihood (ML) estimation* (cf. Nelder and Wedderburn, 1972, pp. 372 ff; McCullagh and Nelder, 1989, pp. 40 ff). Unlike for the linear regression itself (cf. equation 5.14), there is usually no closed form solution in this generalized case. Consequently, the iterative methods discussed in section 4.2 are commonly used to find the optimization parameters $\boldsymbol{\Theta} \coloneqq \boldsymbol{\beta}$, in particular the Fisher scoring algorithm (cf. Green, 1984; Jørgensen, 1984). A general assumption made for fitting (generalized) linear models is that all observations are i.i.d. given the independent variables (cf. Hastie, Tibshirani and Friedman, 2008, p. 28; Nelder and Wedderburn, 1972, p. 372). In this case, the covariance matrix of all observations in $\boldsymbol{y}_{\cdot l}^{\mathsf{s}}$ (cf. equation 5.17), is a scalar multiple of the identity matrix. Therefore, the weighted log-likelihood of the generalized linear model in data set s can be written component-wise as

$$\mathscr{L}\left(\boldsymbol{\Theta}\right) \;=\; \sum_{i=1}^{n^{\mathsf{s}}} w_i^{\mathsf{s}} \cdot \mathscr{L}_{\mathrm{i}}\left(\boldsymbol{\Theta}\right) = \sum_{i=1}^{n^{\mathsf{s}}} w_i^{\mathsf{s}} \cdot \frac{y_{il}^{\mathsf{s}} \cdot \theta_i - \mathbf{b}\left(\theta_i\right)}{\left[\mathbf{a}\left(\phi\right)\right]_{ii}} + \left[\mathbf{c}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}, \phi\right)\right]_i \quad , \tag{5.22}$$

where $\left[\mathbf{c}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}, \phi\right)\right]_i$ and $\left[\mathbf{a}\left(\phi\right)\right]_{ii}$ respectively denote the i-th (diagonal) element of $\mathbf{c}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}, \phi\right)$ and $\mathbf{a}\left(\phi\right)$. Due to the i.i.d. assumption, the dispersion parameter ϕ can be separated from $\boldsymbol{\beta}$ for optimization and can be obtained from the (intermediate) values of $\boldsymbol{\Theta}$ based on equation 5.17 if it is not fixed in advance (cf. McCullagh and Nelder, 1989, p. 295; Ruppert, Wand and Carroll, 2003, pp. 197 ff). Incorporating design weights w_i^{s}, equation 5.22 is again a weighted total (HT-)estimator for the population log-likelihood, often referred to as 'pseudo log-likelihood' (cf. equation 2.15; Binder, 1983, p. 282; Fuller, 2009, p. 378; Lumley and Scott, 2017, p. 268; Pfeffermann, 2011, p. 122).

The distance function that is used for minimization is typically minus twice the weighted log-likelihood, i.e.

$$\delta\left(\boldsymbol{\Theta}\right) \;=\; -\,2 \cdot \mathscr{L}\left(\boldsymbol{\Theta}\right) \qquad . \tag{5.23}$$

The negative log-likelihood is also referred to as the *deviance* or *cross-entropy* and is a sample estimate for the prediction error measured by the Kullback-Leibler distance (cf. Hastie and Tibshirani, 1986, p. 300; Hastie, Tibshirani and Friedman, 2008, p. 32; Kullback and Leibler, 1951). To minimize this distance, the Jacobian and negative Fisher information matrix of δ are required. These are determined by

$$\begin{aligned} \mathbf{J}_{\delta}\left(\boldsymbol{\Theta}\right) \;&=\; -\,2 \cdot \sum_{i=1}^{n^{\mathsf{s}}} w_i^{\mathsf{s}} \cdot \boldsymbol{x}_{i\cdot}^{\mathsf{s}} \cdot \left[\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}\left(\phi\right)\right]_{ii}^{-1} \cdot \left(y_{il}^{\mathsf{s}} - \left[\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}\right]_i\right) \cdot \left[\mathbf{J}_{l^{-1}}\left(\boldsymbol{\eta}^{\mathsf{s}}\right)\right]_{ii} \\ &=\; -\,2 \cdot \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W} \left(\mathbf{J}_l\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}\right)\right) \left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}} - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}\right)\right)^{\mathsf{T}} \end{aligned} \tag{5.24}$$

and

$$\begin{aligned} \mathrm{E}\left(\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}\right)\right) \;&=\; 2 \cdot \sum_{i=1}^{n^{\mathsf{s}}} \left[\mathbf{J}_{l^{-1}}\left(\boldsymbol{\eta}^{\mathsf{s}}\right)\right]_{ii}^{\circ 2} \cdot \left[\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}\left(\phi\right)\right]_{ii}^{-1} \cdot \mathrm{E}\left(\left(\boldsymbol{x}_{i\cdot}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{x}_{i\cdot}^{\mathsf{s}}\right) \\ &=\; 2 \cdot \left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W} \boldsymbol{X}^{\mathsf{s}} \qquad , \end{aligned} \tag{5.25}$$

using

$$\boldsymbol{W} \coloneqq \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{s}}\right)\left(\mathbf{J}_{\boldsymbol{l}^{-1}}\left(\boldsymbol{\eta}^{\mathsf{s}}\right)\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\left(\phi\right)\right)^{-1}\mathbf{J}_{\boldsymbol{l}^{-1}}\left(\boldsymbol{\eta}^{\mathsf{s}}\right) \quad . \tag{5.26}$$

The matrix $\boldsymbol{W}$ is diagonal due to the i.i.d. assumption, and Jacobian matrices of the link and inverse link function are respectively denoted by $\mathbf{J}_{\boldsymbol{l}}\left(\boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right)$ and $\mathbf{J}_{\boldsymbol{l}^{-1}}\left(\boldsymbol{\eta}^{\mathsf{s}}\right)$. When using a canonical link function, Fisher scoring and Newton-Raphson algorithm coincide due to the fact that expected and actual value of the Hessian matrix are the same in this case (cf. appendix B.4.2; Kale, 1961, p. 453; McCullagh and Nelder, 1989, pp. 29–43; Nelder and Wedderburn, 1972, pp. 372 ff; Rao, 1952, p. 165).

The Fisher scoring algorithm can be implemented as an iteratively reweighted least squares (IRWLS) procedure. The updating rule for $\boldsymbol{\Theta}$ introduced in algorithm 5 can be written as

$$\begin{aligned}
\boldsymbol{\Theta}^{(a)} &\coloneqq \boldsymbol{\Theta}^{(a-1)} - \left(\mathrm{E}\left(\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}\right)\right)\right)^{-1}\left(\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}\right)\right)^{\mathsf{T}} \\
&= \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}}\boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\right)^{-1}\left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}}\boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\right)\boldsymbol{\Theta}^{(a-1)} \\
&\quad + \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}}\boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\right)^{-1}\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}}\boldsymbol{W}\mathbf{J}_{\boldsymbol{l}}\left(\boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right)\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l} - \boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right) \\
&= \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}}\boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\right)^{-1}\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}}\boldsymbol{W}\left(\boldsymbol{\eta}^{\mathsf{s}} + \mathbf{J}_{\boldsymbol{l}}\left(\boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right)\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l} - \boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right)\right) \quad ,
\end{aligned} \tag{5.27}$$

which is exactly the form of the weighted least squares coefficients defined in equation 5.14. Therefore, equalities 5.27 can be interpreted as the result of a linear regression of the *adjusted dependent variable*

$$\widetilde{\boldsymbol{y}}^{\mathsf{s}}_{\cdot l} \coloneqq \boldsymbol{\eta}^{\mathsf{s}} + \mathbf{J}_{\boldsymbol{l}}\left(\boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right)\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l} - \boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right) \tag{5.28}$$

on $\boldsymbol{X}^{\mathsf{s}}$, with weights defined by the diagonal entries of $\boldsymbol{W}$ in equation 5.26. This property is handy due to the importance and implementation of linear regression in most software, which can be used to iteratively solve GLMs (cf. Green, 1984; Jørgensen, 1984, p. 287; McCullagh and Nelder, 1989, p. 40; Simonoff, 2003, pp. 127 f).

The (generalized) linear regression models belong to the most popular models for data analysis and offer a straightforward interpretability. This particularly concerns the distribution of parameters $\boldsymbol{\beta} = \boldsymbol{\Theta}$ as well as the way in which they describe the influence of $\boldsymbol{X}$ on $\boldsymbol{Y}$ (cf. e.g. Green and Silverman, 1994, p. 1; Hastie and Tibshirani, 1990, pp. 1, 86; Wood, 2017, p. 107). Once parameters $\boldsymbol{\beta}$ are estimated based on equalities 5.22 to 5.28, GLMs can readily solve prediction problems when the underlying relationships between independent and dependent variables (approximately) follow their respective assumptions. As outlined in equation 5.5, predictions for data set t are then determined by $\widehat{\boldsymbol{y}}^{\mathsf{t}}_{\cdot l} = \mathrm{E}\left(\boldsymbol{y}^{\mathsf{t}}_{\cdot l} \,|\, \boldsymbol{X}^{\mathsf{t}}\right) = \boldsymbol{l}^{-1}\left(\boldsymbol{\eta}^{\mathsf{t}}\right) = \boldsymbol{l}^{-1}\left(\boldsymbol{X}^{\mathsf{t}}\boldsymbol{\beta}\right)$. One assumption of GLMs that needs to hold for this purpose is that the functional relationship between auxiliary variables $\boldsymbol{X}$ and systematic components $\boldsymbol{\eta}$ is a fixed linear function. For many dependencies between $\boldsymbol{X}$ and $\boldsymbol{Y}$, however, this requirement of linearity is not fulfilled. For example, it is often more reasonable to assume a non-linear influence of a person's age on $\boldsymbol{\eta}$, e.g. when modeling wages, the risk of certain diseases, or the willingness to respond in a survey. Such non-linear relationships can be incorporated in (generalized) linear models through adjustments, e.g. by splitting the respective variable into groups to allow for different slopes or by applying transformations like logarithms or exponentials to $\boldsymbol{X}$. However, such adjustments reduce interpretability and require a decent amount of time and expertise

to find transformations that suit the specific variables (cf. Greene, 2008, pp. 12, 111 ff; Hastie and Tibshirani, 1990, pp. 1 ff; James et al., 2013, pp. 265 ff; Ruppert, Wand and Carroll, 2003, pp. 2 ff; Weisberg, 2005, pp. 172 ff). Relaxations of the linearity assumptions with lower demand in selecting appropriate adjustments can be achieved by using non- or semi-parametric models, which do not assume a fixed number of parameters to model the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$. An important realization thereof is constituted by generalized additive models, which are introduced in the following section 5.1.4.

5.1.4 Generalized Additive Models

Generalized linear regression models consider the systematic component as a fixed linear function of the independent variables. Since the conditional distribution of $\boldsymbol{Y}$ is fully described by a fixed number (p) of model parameters, these models are typically referred to as *parametric models*. They require refining adjustments when their respective assumptions about the shape of the functional relationship do not hold. *Non-parametric models* achieve relaxation of the linearity assumption with lower demands in selecting appropriate adjustments while still preserving many desirable properties, such as continuity and differentiability of the function $\mathbf{m}$. Rather than assuming the systematic component to be defined by $\boldsymbol{\eta}^{\mathrm{t}} = \boldsymbol{X}^{\mathrm{t}}\boldsymbol{\beta}$ (cf. section 5.1.3), non-parametric models postulate a more general functional relationship

$$\boldsymbol{\eta}^{\mathrm{t}} \coloneqq \mathrm{F}\left(\boldsymbol{X}^{\mathrm{t}}\right) \quad , \tag{5.29}$$

representing by $\mathrm{F} : \mathbb{R}^{n^{\mathrm{t}} \times p} \to \mathbb{R}^{n^{\mathrm{t}}}$ a smooth function of $\boldsymbol{X}^{\mathrm{t}}$ to be estimated. Models in form of equality 5.29 are called non-parametric because F can depend on an adaptable number of (effective) parameters (cf. also section 5.1.11). The case of mixing non-parametric and parametric ideas in the form of $\boldsymbol{\eta}^{\mathrm{t}} \coloneqq \boldsymbol{X}^{\mathrm{t}}\boldsymbol{\beta} + \mathrm{F}\left(\boldsymbol{X}^{\mathrm{t}}\right)$ is often referred to as *semi-parametric models*, but can as well be expressed as a special case of the non-parametric formulation (cf. Green and Silverman, 1994, pp. 8, 64; Hastie and Tibshirani, 1990, pp. 3 ff, 118; Ruppert, Wand and Carroll, 2003, pp. 57 ff, 161 ff).

Generalized additive models (GAMs) introduced by Hastie and Tibshirani (1984; 1986; 1990) are the presumably most common approach for implementing non-parametric models (cf. Green and Silverman, 1994, pp. 83 ff; Ruppert, Wand and Carroll, 2003, p. 36). Just like for generalized linear models, the dependent variable $\boldsymbol{y}_{\cdot l}^{\mathrm{t}}$ is assumed to be a realization from a distribution belonging to the exponential family that is influenced by the systematic component $\boldsymbol{\eta}^{\mathrm{t}}$. The difference between GAMs to GLMs lies in defining $\boldsymbol{\eta}^{\mathrm{t}}$ by

$$\boldsymbol{\eta}^{\mathrm{t}} \coloneqq \sum_{j=1}^{h} \boldsymbol{t}_{\mathrm{j}}\left(\boldsymbol{X}^{\mathrm{t}}\right) = \boldsymbol{t}\left(\boldsymbol{X}^{\mathrm{t}}\right) \quad . \tag{5.30}$$

The overall transformation $\boldsymbol{t} : \mathbb{R}^{n^{\mathrm{t}} \times p} \to \mathbb{R}^{n^{\mathrm{t}}}$ is a sum of h smooth component transformations $\boldsymbol{t}_{\mathrm{j}} : \mathbb{R}^{n^{\mathrm{t}} \times p} \to \mathbb{R}^{n^{\mathrm{t}}}$ for $j = 1, \ldots, h$, where $h \in \mathbb{N}$ is an arbitrary given number of transformations. One of these transformations can constitute an intercept. To achieve relatively simple models, it is common to consider each component $\boldsymbol{t}_{\mathrm{j}}$ to be a univariate function that transforms only a single column of $\boldsymbol{X}^{\mathrm{t}}$. However, the more complex case where multivariate transformations are applied is also possible (cf. e.g. section 5.1.7). Besides definition 5.30 not requiring linearity of $\boldsymbol{t}_{\mathrm{j}}$ with respect to $\boldsymbol{X}^{\mathrm{t}}$, the remaining properties and assumptions of generalized linear models are kept (cf. Coull, Ruppert and Wand, 2001; Hastie and Tibshirani, 1986, p. 300; 1990, pp. 264 ff; Wood, 2017, pp. 119 ff).

Although this is not required for definition 5.30, the transformation for GAMs is typically chosen to be fully parametrizable. In this case, each of the h component transformations $\boldsymbol{t}_{\mathsf{j}} : \mathbb{R}^{n^{\mathsf{t}} \times p} \times \mathbb{R}^{c_j} \to \mathbb{R}^{n^{\mathsf{t}}}$ is completely described by a vector of *transformation parameters* $\boldsymbol{\kappa}^{(j)} \in \mathbb{R}^{c_j}$. The numbers of parameters used for all transformations constitute a vector $\boldsymbol{c} = \begin{bmatrix} c_1 & \dots & c_h \end{bmatrix}^{\mathsf{T}} \in \mathbb{N}^h$. Equality 5.30 is then written as

$$\boldsymbol{\eta}^{\mathsf{t}} \; = \boldsymbol{\eta}^{\mathsf{t}}\left(\boldsymbol{\kappa}\right) \coloneqq \sum_{j=1}^{h} \boldsymbol{t}_{\mathsf{j}}\left(\boldsymbol{X}^{\mathsf{t}}, \boldsymbol{\kappa}^{(j)}\right) \; = \; \boldsymbol{t}\left(\boldsymbol{X}^{\mathsf{t}}, \boldsymbol{\kappa}\right) \qquad , \tag{5.31}$$

defining the systematic component as a function $\boldsymbol{\eta}^{\mathsf{t}} : \mathbb{R}^d \to \mathbb{R}^{n^{\mathsf{t}}}$ of parameters $\boldsymbol{\kappa}$ to be determined. Here, $\boldsymbol{\kappa} \coloneqq \left[\left(\boldsymbol{\kappa}^{(1)}\right)^{\mathsf{T}}, \dots, \left(\boldsymbol{\kappa}^{(h)}\right)^{\mathsf{T}}\right]^{\mathsf{T}} \in \mathbb{R}^d$ for $d = ||\boldsymbol{c}||_1$. is the concatenation of all parameters defining the parametrized overall transformation $\boldsymbol{t} : \mathbb{R}^{n^{\mathsf{t}} \times p} \times \mathbb{R}^d \to \mathbb{R}^{n^{\mathsf{t}}}$. Similarly as for GLMs, an intercept is typically included by defining $\boldsymbol{t}_1\left(\boldsymbol{X}^{\mathsf{t}}, \boldsymbol{\kappa}^{(1)}\right) \coloneqq \kappa_1^{(1)} \cdot \mathbf{1}_{n^{\mathsf{t}} \times 1}$. Frequently used transformations for GAMs are spline functions, which are discussed in section 5.1.6 (cf. Coull, Ruppert and Wand, 2001; Hastie and Tibshirani, 1984; 1986; 1990; Wood, 2017).

Just as in section 5.1.3, parameters $\boldsymbol{\kappa}$ are commonly obtained by ML estimation in data set s, using the optimization techniques discussed in chapter 4. As before, it is assumed that observations of $\boldsymbol{y}_{\cdot l}^{\mathsf{s}}$ given $\boldsymbol{X}^{\mathsf{s}}$ in data set s are i.i.d. The distance function to be minimized for estimating parameters $\boldsymbol{\Theta} \coloneqq \boldsymbol{\kappa}$ is again twice the weighted deviance (negative weighted log-likelihood):

$$\begin{aligned} \delta\left(\boldsymbol{\Theta}\right) \; &= -2 \cdot \mathscr{L}\left(\boldsymbol{\Theta}\right) = -2 \cdot \sum_{i=1}^{n} w_i^{\mathsf{s}} \cdot \mathscr{L}_{\mathrm{i}}\left(\boldsymbol{t}\left(\boldsymbol{x}_{i\cdot}^{\mathsf{s}}, \boldsymbol{\kappa}\right)\right) \\ &= \; -2 \cdot \sum_{i=1}^{n} w_i^{\mathsf{s}} \cdot \frac{y_{il}^{\mathsf{s}} \cdot \theta_i - \mathbf{b}\left(\theta_i\right)}{\left[\mathbf{a}\left(\phi\right)\right]_{ii}} + \left[\mathbf{c}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}, \phi\right)\right]_i \qquad . \end{aligned} \tag{5.32}$$

This is basically the same that is used for GLMs (cf. equations 5.22 and 5.23), but the parametrization is different (cf. Hastie and Tibshirani, 1986, p. 300). As before, the dispersion parameter ϕ can be separated from $\boldsymbol{\Theta}$ for optimization. If it is not fixed in advance, ϕ can be obtained from the (intermediate) values of $\boldsymbol{\Theta}$ by using equation 5.17 where necessary (cf. McCullagh and Nelder, 1989, p. 295; Ruppert, Wand and Carroll, 2003, pp. 197 ff). Assuming that all required derivatives exist, the corresponding Jacobian and expected Hessian matrix of δ are determined by

$$\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}\right) \; = \; -2 \cdot \left(\left(\mathbf{J}_{\boldsymbol{\eta}^{\mathsf{s}}}\left(\boldsymbol{\Theta}\right)\right)^{\mathsf{T}} \boldsymbol{W} \left(\mathbf{J}_{\boldsymbol{l}}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}\right)\right)\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}} - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}\right)\right)^{\mathsf{T}} \tag{5.33}$$

and

$$\mathrm{E}\left(\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}\right)\right) \; = 2 \cdot \left(\mathbf{J}_{\boldsymbol{\eta}^{\mathsf{s}}}\left(\boldsymbol{\Theta}\right)\right)^{\mathsf{T}} \boldsymbol{W} \left(\mathbf{J}_{\boldsymbol{\eta}^{\mathsf{s}}}\left(\boldsymbol{\Theta}\right)\right) \qquad . \tag{5.34}$$

The weights

$$\boldsymbol{W} \; \coloneqq \; \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{s}}\right)\left(\mathbf{J}_{\boldsymbol{l}^{-1}}\left(\boldsymbol{\eta}^{\mathsf{s}}\right)\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}\left(\phi\right)\right)^{-1} \mathbf{J}_{\boldsymbol{l}^{-1}}\left(\boldsymbol{\eta}^{\mathsf{s}}\right) \tag{5.35}$$

are equal to those used for GLMs (cf. equation 5.26). Therefore, this Jacobian and Fisher information matrix resemble the ones that are used for GLMs, but $\boldsymbol{X}^{\mathsf{s}}$ is replaced by

$\mathbf{J}_{\boldsymbol{\eta}^{\mathsf{s}}}(\boldsymbol{\Theta})$ (cf. equations 5.24 and 5.25). Consequently, GLMs are a special case of GAMs, where $\boldsymbol{\kappa} = \boldsymbol{\beta}$ and $\boldsymbol{t}(\boldsymbol{X}^{\mathsf{s}}, \boldsymbol{\kappa}) = \boldsymbol{X}^{\mathsf{s}}\boldsymbol{\beta}$. As before, the Fisher scoring and Newton-Raphson algorithm coincide in case of a canonical link function if $\boldsymbol{t}$ is linear in $\boldsymbol{\kappa}$ since expected and actual value of the Hessian matrix are again the same in this case (cf. appendix B.4.2; Hastie and Tibshirani, 1986, p. 302; McCullagh and Nelder, 1989, pp. 29 ff; Nelder and Wedderburn, 1972, pp. 372 ff). The resulting update rule for Fisher scoring (cf. algorithm 5) is

$$\begin{aligned}\boldsymbol{\Theta}^{(a)} &\coloneqq \boldsymbol{\Theta}^{(a-1)} - \left(\mathrm{E}\left(\mathbf{H}_{\delta}(\boldsymbol{\Theta})\right)\right)^{-1}\left(\mathbf{J}_{\delta}(\boldsymbol{\Theta})\right)^{\mathsf{T}} \\ &= \boldsymbol{\Theta}^{(a-1)} + \left(\left(\mathbf{J}_{\boldsymbol{\eta}^{\mathsf{s}}}(\boldsymbol{\Theta})\right)^{\mathsf{T}}\boldsymbol{W}\left(\mathbf{J}_{\boldsymbol{\eta}^{\mathsf{s}}}(\boldsymbol{\Theta})\right)\right)^{-1}\left(\mathbf{J}_{\boldsymbol{\eta}^{\mathsf{s}}}(\boldsymbol{\Theta})\right)^{\mathsf{T}}\boldsymbol{W}\left(\mathbf{J}_{\boldsymbol{l}}\left(\boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right)\right)\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l} - \boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right) .\end{aligned} \tag{5.36}$$

This solution based on the methods discussed in section 4.2.1 is also called local scoring for GAMs (cf. Hastie and Tibshirani, 1986; Nelder and Wedderburn, 1972; Yee and Wild, 1996). It can once again be implemented by means of iteratively reweighted least squares. The update $\boldsymbol{\Delta}_{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(a)} - \boldsymbol{\Theta}^{(a-1)}$ is the resulting vector of coefficients when regressing the adjusted dependent variable

$$\widetilde{\boldsymbol{y}}^{\mathsf{s}}_{\cdot l} \coloneqq \left(\mathbf{J}_{\boldsymbol{l}}\left(\boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right)\right)\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l} - \boldsymbol{\mu}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\right) \quad \in \quad \mathbb{R}^{n^{\mathsf{s}}} \tag{5.37}$$

on the adjusted auxiliaries

$$\widetilde{\boldsymbol{X}}^{\mathsf{s}} \coloneqq \mathbf{J}_{\boldsymbol{\eta}^{\mathsf{s}}}(\boldsymbol{\Theta}) \quad \in \quad \mathbb{R}^{n^{\mathsf{s}} \times d} \qquad . \tag{5.38}$$

Again, a vector of weights containing the diagonal elements of $\boldsymbol{W}$ defined in equation 5.35 is used, similarly as for the GLMs (cf. Wood, 2017, p. 165; Yee and Wild, 1996, p. 484).

For optimization by means of Fisher scoring as implied in equalities 5.33 to 5.36, it is required that the transformation $\boldsymbol{t}$ is fully described by and differentiable with respect to the parameters $\boldsymbol{\kappa}$. Nevertheless, transformations for GAMs can be chosen differently and do not necessarily fulfill these requirements. For example, this is the case when one of the additive components in equality 5.30 incorporates recursive partitioning as in regression trees (cf. section 5.1.7). In such settings, the optimization parameters are commonly defined as the systematic component itself, i.e. $\boldsymbol{\Theta} = \boldsymbol{\eta}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}}}$. Updating the actual $j = 1, \ldots, h$ transformations $\boldsymbol{t}_{\mathrm{j}}$ is, hence, not part of equality 5.36 (cf. definition 5.30) but rather done in a separate step that is additionally incorporated in each iteration. This update of the transformations can be done using the backfitting algorithm proposed by Friedman and Stuetzle (1981) and leads to the fitting procedure originally proposed for GAMs by Hastie and Tibshirani (1984, p. 16; 1986, pp. 305 f; 1990, pp. 140 ff). To that end, the desired transformations of $\boldsymbol{X}^{\mathsf{s}}$ are fit to the adjusted dependent variable as defined in equations 5.37 by means of the following algorithm 11. This adaptation results in iteratively reweighted backfitting in place of iteratively reweighted least squares (cf. also Friedman, 1991b; Wood, 2017, pp. 208 ff).

Starting from an intercept-only model, this algorithm proceeds to fit each transformation $\boldsymbol{t}_{\mathrm{k}}$ to the residuals remaining from all other transformations $\boldsymbol{t}_{\mathrm{j}}$ ($j \in \{1, \ldots, h\} \setminus k$). This process is repeated until the transformations converge, which is measured by the change in their output being smaller than a given tolerance v. As long as the predictions are linear in $\boldsymbol{Y}^{\mathsf{s}}$ (i.e. application of a linear smoother / hat-matrix), which is almost always the case, backfitting is equivalent to the Gauss-Seidel algorithm exploiting the special

Algorithm 11: Backfitting algorithm for additive models

1: **Input:** $\boldsymbol{X}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}} \times p}$; $\boldsymbol{y}^{\mathsf{s}}_{\cdot l} \in \mathbb{R}^{n^{\mathsf{s}}}$; $\boldsymbol{w}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}}}$; $\delta : n^{\mathsf{s}} \to \mathbb{R}_{\geq 0}$ $v \in \mathbb{R}_{\geq 0}$; $h \in \mathbb{N}$

2: Initialize intercept $\boldsymbol{t}_1\left(\boldsymbol{X}^{\mathsf{s}}\right) \coloneqq \widehat{\boldsymbol{\mu}}_{\boldsymbol{y}^{\mathsf{s}}_{\cdot 1}}\left(\boldsymbol{w}^{\mathsf{s}}\right)$ and transformations $\boldsymbol{t}_{\mathrm{j}}\left(\boldsymbol{X}^{\mathsf{s}}\right) \coloneqq \boldsymbol{0}_{n^{\mathsf{s}} \times 1}$ for all $j > 1$

3: Set $\boldsymbol{B} \leftarrow \boldsymbol{t}\left(\boldsymbol{X}^{\mathsf{s}}\right)$

4: **for** $k = 1$ to h **do**

5: Keeping all $h-1$ other transformations fixed, update $\boldsymbol{t}_{\mathrm{k}}\left(\boldsymbol{X}^{\mathsf{s}}\right)$ such that it is optimal w.r.t. δ:

$$\begin{aligned} \boldsymbol{t}^{*}_{k}\left(\boldsymbol{X}^{\mathsf{s}}\right) &\leftarrow \underset{\boldsymbol{t}_k(\boldsymbol{X}^{\mathsf{s}})}{\operatorname{argmin}}\left(\delta\left(\boldsymbol{t}\left(\boldsymbol{X}^{\mathsf{s}}\right)\right)\right) \\ \boldsymbol{t}_{k}\left(\boldsymbol{X}^{\mathsf{s}}\right) &\leftarrow \boldsymbol{t}^{*}_{k}\left(\boldsymbol{X}^{\mathsf{s}}\right) - \widehat{\boldsymbol{\mu}}_{\boldsymbol{t}^{*}_{\mathrm{k}}(\boldsymbol{X}^{\mathsf{s}})}\left(\boldsymbol{w}^{\mathsf{s}}\right) \end{aligned}$$

6: **end for**

7: **if** $\operatorname{Max}\left(\operatorname{Abs}\left(\boldsymbol{t}\left(\boldsymbol{X}^{\mathsf{s}}\right) - \boldsymbol{B}\right)\right) \leq v$ **then**

8: **Return:** $\boldsymbol{t}\left(\boldsymbol{X}^{\mathsf{s}}\right)$

9: **else**

10: go to step 3

11: **end if**

structure of GAMs. It is not equivalent to the Newton-Raphson or Fisher scoring method presented above (cf. e.g. equation 5.36) and does not require a specific type of optimization in step 5. A weighted distance function can be applied, and the technique is usable for transformations that do not meet the requirements of quasi-Newton methods. Algorithm 11 can nevertheless be used as an alternative for parametrizable and differentiable transformations as well (cf. Buja, Hastie and Tibshirani, 1989; Friedman and Stuetzle, 1981, p. 818; Hackbusch, 1994, p. 70; Hastie and Tibshirani, 1984; 1986; 1990, pp. 90 ff; Hastie, Tibshirani and Friedman, 2008, p. 298; Wood, 2017, pp. 209 ff).

Once all $j = 1, \ldots, h$ component transformations $\boldsymbol{t}_{\mathrm{j}}$ are fit by using IRWLS or iteratively reweighted backfitting, the predictions for data set t are $\widehat{\boldsymbol{y}}^{\mathsf{t}}_{\cdot l} = \mathrm{E}\left(\boldsymbol{y}^{\mathsf{t}}_{\cdot l} \,|\, \boldsymbol{X}^{\mathsf{t}}\right) = \boldsymbol{l}^{-1}\left(\boldsymbol{\eta}^{\mathsf{t}}\right)$. By relaxing the linearity assumption of generalized linear models for computing $\boldsymbol{\eta}^{\mathsf{t}}$, GAMs provide more flexibility regarding the dependencies between $\boldsymbol{X}$ and $\boldsymbol{Y}$ that can be considered in the model. However, one limitation still lies in the i.i.d. assumption, especially since observations in sampling are not always independent. In the following section 5.1.5, generalized additive mixed models are introduced, which help to overcome this restriction.

5.1.5 Generalized Additive Mixed Models

In sections 5.1.3 and 5.1.4, the fitting methods for generalized linear and additive models generally require that values of the dependent variable $\boldsymbol{y}_{\cdot l}^{\mathsf{s}}$ are assumed to be i.i.d. given the auxiliary variables $\boldsymbol{X}^{\mathsf{s}}$. Under this assumption, the unknown matrix of covariances between values of $\boldsymbol{y}_{\cdot l}^{\mathsf{s}}$ given $\boldsymbol{X}^{\mathsf{s}}$ is a scalar multiple of the identity matrix that depends on a single dispersion parameter ϕ, such that

$$\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{t})}(\phi) \;=\; \mathbf{a}(\phi)\,\mathrm{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{t})}\right) \tag{5.39}$$

(cf. e.g. equalities 5.18 and 5.20). As a consequence, the log-likelihood can be written as a sum over all independent observations, and the variance component can be factored out when updating the parameters (cf. e.g. McCullagh and Nelder, 1989, p. 295). However, the i.i.d. assumption does not always hold. For example, this can be the case for repeated measurements in observational studies or non-probability web-panels. Similarly, respondent-driven sampling or recruitment via website or newspaper advertisements can lead to clusters in which observations are not independent of each other. The only alternative for fitting the models considered so far is to assume that $\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{t})}(\phi)$ is known (cf. Elliott and Valliant, 2017, p. 257; Heckathorn, 2002, p. 16; Henderson et al., 1959, p. 196; Lee, Nelder and Pawitan, 2006, pp. 65 ff; Wood, 2017, pp. 274 ff).

To allow for dependencies between observations when estimating regression models, generalized additive mixed models (GAMMs) are used to relax the i.i.d. assumption. For this purpose, the total covariance matrix of observations $\boldsymbol{y}_{\cdot l}^{\mathsf{t}}$ given $\boldsymbol{X}^{\mathsf{t}}$ is expressed as a function $\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{t})} : \mathbb{R}^s \to \mathbb{R}^{n^{\mathsf{t}} \times n^{\mathsf{t}}}$ of a *vector of dispersion (or variance) parameters*

$$\boldsymbol{\phi} \;\coloneqq\; \left[\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)^{\mathsf{T}} \quad \left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)^{\mathsf{T}}\right]^{\mathsf{T}} \quad \in \quad \mathbb{R}^s \qquad , \tag{5.40}$$

where $\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})} \in \mathbb{R}^b$ and $\boldsymbol{\phi}^{(\boldsymbol{u})} \in \mathbb{R}^c$, such that $s = b + c$. This vector $\boldsymbol{\phi}$ is usually considered unknown and estimated within the fitting procedure of the model. A common choice for $\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{t})}$ is

$$\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{t})}(\boldsymbol{\phi}) \;\coloneqq \boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot l}}^{(\mathsf{t})}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right) + \boldsymbol{D}^{\mathsf{t}}\boldsymbol{\Sigma}_{\boldsymbol{u}}^{(\mathsf{t})}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\left(\boldsymbol{D}^{\mathsf{t}}\right)^{\mathsf{T}} \qquad , \tag{5.41}$$

where $\boldsymbol{D}^{\mathsf{t}} \in \mathbb{R}^{n^{\mathsf{t}} \times v}$ is a known design matrix for some *random effects* $\boldsymbol{u} \in \mathbb{R}^v$. Note that this is basically a form of decomposing the conditional covariance of observations $\boldsymbol{y}_{\cdot l}^{\mathsf{s}}$ into a within and a between component denoted by $\boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot l}}^{(\mathsf{t})} : \mathbb{R}^b \to \mathbb{R}^{n^{\mathsf{t}} \times n^{\mathsf{t}}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{(\mathsf{t})} : \mathbb{R}^c \to \mathbb{R}^{v \times v}$, respectively (cf. Pinheiro and Bates, 2000, pp. 57 ff; Harville, 1977, p. 321; Henderson, 1953). Here and in the following discussion, indices are used to denote whether a mean, covariance or component function of the exponential family (cf. equation 5.16) are meant with respect to $\boldsymbol{y}_{\cdot l}$ as a whole or its components $\boldsymbol{e}_{\cdot l}$ or $\boldsymbol{u}$. In the context of GAMMs, usually only $\boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot l}}^{(\mathsf{t})}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)$ is assumed to be a scalar multiple of the identity matrix, such that it holds due to equations 5.17 that

$$\mathbf{a}_{\boldsymbol{y}_{\cdot l}}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right) \;\coloneqq\; \left[\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right]_1 \cdot \mathbf{I}_{n^{\mathsf{t}}} \tag{5.42}$$

and

$$\mathrm{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{t})}\right) \;\coloneqq\; \left(\mathbf{I}_{n^{\mathsf{t}}} + \frac{1}{\left[\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right]_1} \cdot \boldsymbol{D}^{\mathsf{t}}\boldsymbol{\Sigma}_{\boldsymbol{u}}^{(\mathsf{t})}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\left(\boldsymbol{D}^{\mathsf{t}}\right)^{\mathsf{T}}\right) \qquad , \tag{5.43}$$

which means that observations need only to be independent and identically distributed given the random effects (cf. Henderson, 1950; Henderson et al., 1959, p. 204; Pinheiro and Bates, 2000, pp. 202 f; Wood, 2017, pp. 287 f). To that end, the coefficients to be estimated are extended to include the random effects $\boldsymbol{u}$ and dispersion parameters $\boldsymbol{\phi}$:

$$\boldsymbol{\Theta} \coloneqq \begin{bmatrix} \boldsymbol{\kappa}^\mathsf{T} & \boldsymbol{u}^\mathsf{T} & \boldsymbol{\phi}^\mathsf{T} \end{bmatrix}^\mathsf{T} \quad \in \quad \mathbb{R}^{d+v+s} \qquad , \tag{5.44}$$

where $\boldsymbol{\kappa} \coloneqq \left[\left(\boldsymbol{\kappa}^{(1)} \right)^\mathsf{T}, \ldots, \left(\boldsymbol{\kappa}^{(h)} \right)^\mathsf{T} \right]^\mathsf{T} \in \mathbb{R}^d$ denotes a vector of transformation parameters of arbitrary given size, as in the previous section 5.1.4. To model the covariance structure that is implied by equation 5.41, the random effects are included in the systematic component. It is now defined as a function $\boldsymbol{\eta}^\mathsf{t} : \mathbb{R}^{d+v+s} \to \mathbb{R}^{n^\mathsf{t}}$ of the form

$$\begin{aligned} \boldsymbol{\eta}^\mathsf{t} \;=\; \boldsymbol{\eta}^\mathsf{t}(\boldsymbol{\Theta}) \;&\coloneqq\; \sum_{j=1}^{h} \boldsymbol{t}_\mathrm{j}\left(\boldsymbol{X}^\mathsf{t}, \boldsymbol{\kappa}^{(j)} \right) + \boldsymbol{D}^\mathsf{t}\boldsymbol{u} \\ &=\; \boldsymbol{t}\left(\boldsymbol{X}^\mathsf{t}, \boldsymbol{\kappa} \right) + \boldsymbol{D}^\mathsf{t}\boldsymbol{u} \qquad , \end{aligned} \tag{5.45}$$

using (component) transformations $\boldsymbol{t}_\mathrm{j}$ and $\boldsymbol{t}$ as in section 5.1.4. Similar as before, *generalized linear mixed model (GLMMs)* are constituted by the special case when $\boldsymbol{\kappa} \coloneqq \boldsymbol{\beta}$ and $\boldsymbol{t}\left(\boldsymbol{X}^\mathsf{t}, \boldsymbol{\kappa} \right) \coloneqq \boldsymbol{X}^\mathsf{t}\boldsymbol{\beta}$ (cf. section 5.1.4; Hastie and Tibshirani, 1984, p. 9; Wood, 2017, p. 309).

The remaining assumptions for the GAMs are kept for fitting the model, which is again denoted for a data set s as in the previous sections. Consequently, it is assumed that the conditional distribution for the dependent variable given the random effects as well as the distribution of the random effects both belong to the exponential family. In this case, the log-likelihood

$$\begin{aligned} \mathscr{L}\left(\boldsymbol{y}^\mathsf{s}_{\cdot l} \,|\, \boldsymbol{u} \right) \;&=\; \sum_{i \in \mathcal{S}^\mathsf{s}} \log\left(\mathrm{f}_{\boldsymbol{y}^\mathsf{s}_{\cdot 1}}\left(y^\mathsf{s}_{il} \,|\, \boldsymbol{u} \right) \right) \\ &=\; \left\| \left(\mathbf{a}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})} \right) \right)^{-1} \left(\boldsymbol{y}^\mathsf{s}_{\cdot l} \circ \boldsymbol{\theta}_{\boldsymbol{y}^\mathsf{s}_{\cdot l}} - \mathbf{b}_{\boldsymbol{y}^\mathsf{s}_{\cdot 1}}\left(\boldsymbol{\theta}_{\boldsymbol{y}^\mathsf{s}_{\cdot l}} \right) \right) + \mathbf{c}_{\boldsymbol{y}^\mathsf{s}_{\cdot 1}}\left(\boldsymbol{y}^\mathsf{s}_{\cdot l}, \boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})} \right) \right\|_1 \end{aligned} \tag{5.46}$$

as well as the logarithm of the density function

$$\ell\left(\boldsymbol{u} \right) \;=\; \sum_{i=1}^{v} \log\left(\mathrm{f}_{\boldsymbol{u}}\left(u_i \right) \right) = \left\| \left(\boldsymbol{u} \circ \boldsymbol{\theta}_{\boldsymbol{u}} - \mathbf{b}_{\boldsymbol{u}}\left(\boldsymbol{\theta}_{\boldsymbol{u}} \right) \right) \left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(\boldsymbol{u})} \right) \right)^{-1} + \mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u}, \boldsymbol{\phi}^{(\boldsymbol{u})} \right) \right\|_1 \tag{5.47}$$

emerge. The function to be maximized for finding the optimization parameters $\boldsymbol{\Theta}$ is denoted by

$$\mathscr{L}\left(\boldsymbol{\Theta} \right) \;=\; \mathscr{L}\left(\boldsymbol{y}^\mathsf{s}_{\cdot l} \,|\, \boldsymbol{u} \right) + \ell\left(\boldsymbol{u} \right) \tag{5.48}$$

(cf. Bates, 2018, p. 5; Lee and Nelder, 1996, pp. 620 f). Note that equation 5.47 and consequently 5.48 are called a log-likelihood in early publications of Henderson (e.g. 1950). Out of equalities 5.46 to 5.48, however, only equation 5.46 is the logarithm of a function that is usually considered a classical likelihood in the sense of Fisher (1922b) because $\boldsymbol{u}$ is a random variable (cf. Robinson et al., 1991, pp. 18, 29). For this reason, $\mathscr{L}\left(\boldsymbol{\Theta} \right)$ as defined in equation 5.48 is called a *hierarchical* log-likelihood and thus denoted by a different symbol in some publications (cf. e.g. Lee and Nelder, 1996; Lee, Nelder and Pawitan, 2006; Noh, Wu and Lee, 2012). Other authors still call it a log-likelihood and

denote it by the corresponding symbol (cf. e.g. Bates, 2018, p. 16; Burgard, 2013, pp. 22 f; Pinheiro and Bates, 2000, p. 62). As there is no risk of ambiguity in the context of this thesis, the latter convention is used here for the sake of linguistic and notational brevity.

As for GLMs and GAMs, the parameters for GAMMs are typically determined by ML estimation using the Fisher scoring algorithm (cf. section 4.2). The model is fit by minimizing twice the weighted deviance (negative weighted log-likelihood), such that

$$\delta\left(\boldsymbol{\Theta}\right) \;=\; -\,2\cdot\mathscr{L}\left(\boldsymbol{\Theta}\right) \tag{5.49}$$

is the distance function. The Jacobian matrix of δ is

$$\begin{aligned}&\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}\right) \;=\\ &-\,2\cdot\left[\left(\frac{\partial(\mathscr{L}\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\kappa}\right)}\right)^{\mathsf{T}}\left(\frac{\partial(\ell\left(\boldsymbol{u}\right)+\mathscr{L}\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{u}\right)}\right)^{\mathsf{T}}\left(\frac{\partial(\mathscr{L}\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)}\right)^{\mathsf{T}}\left(\frac{\partial(\ell\left(\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\phi}^{(u)}\right)}\right)^{\mathsf{T}}\right]\quad,\end{aligned} \tag{5.50}$$

while the expected Hessian matrix used for Fisher scoring is

$$\begin{aligned}&\mathrm{E}\left(\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}\right)\right) =\\ &-\,2\cdot\mathrm{E}\left(\left[\begin{array}{cccc}\dfrac{\partial^2\left(\mathscr{L}\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}\,|\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\kappa}\right)\partial\left(\boldsymbol{\kappa}\right)} & \dfrac{\partial^2\left(\mathscr{L}\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}\,|\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\kappa}\right)\partial\left(\boldsymbol{u}\right)} & \mathbf{0} & \mathbf{0}\\ \dfrac{\partial^2\left(\mathscr{L}\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}\,|\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{\kappa}\right)} & \dfrac{\partial^2\left(\mathscr{L}\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}\,|\boldsymbol{u}\right)+\ell\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{u}\right)} & \mathbf{0} & \dfrac{\partial^2\left(\ell\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{\phi}^{(u)}\right)}\\ \mathbf{0} & \mathbf{0} & \dfrac{\partial^2\left(\mathscr{L}\left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}\,|\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)} & \mathbf{0}\\ \mathbf{0} & \dfrac{\partial^2\left(\ell\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(u)}\right)\partial\left(\boldsymbol{u}\right)} & \mathbf{0} & \dfrac{\partial^2\left(\ell\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(u)}\right)\partial\left(\boldsymbol{\phi}^{(u)}\right)}\end{array}\right]\right)\quad.\end{aligned} \tag{5.51}$$

Elements of these matrices are rather extensive in derivation and representation (cf. e.g. Bates, 2018; Lee and Nelder, 1996; Rao, 2003, pp. 100 ff; Searle, Casella and McCulloch, 2006, pp. 235 ff, 456). They are therefore deferred to appendix B.4.3, and only a few selected results are presented below. The weights to maximize the pseudo log-likelihood as a HT-type estimator (cf. equation 2.15; Binder, 1983, p. 282; Fuller, 2009, p. 378; Lumley and Scott, 2017, p. 268; Pfeffermann, 2011, p. 122) are defined as

$$\boldsymbol{W} \;\coloneqq\; \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{s}}\right)\mathbf{J}_{l^{-1}}\left(\boldsymbol{\eta}^{\mathsf{s}}\right)\left(\boldsymbol{\Sigma}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}\left(\boldsymbol{\phi}\right)\right)^{-1}\left(\mathbf{J}_{l^{-1}}\left(\boldsymbol{\eta}^{\mathsf{s}}\right)\right)^{\mathsf{T}}\quad. \tag{5.52}$$

Apart from $\boldsymbol{\phi}$ being part of $\boldsymbol{\Theta}$ and a more complex structure for $\boldsymbol{\Sigma}^{(\mathsf{s})}_{\boldsymbol{y}_{\cdot l}}$, definition 5.52 coincides with those for GLMs and GAMs (cf. equations 5.26 and 5.35). Furthermore, it becomes evident from

$$\mathrm{E}\left(\frac{\partial^2\left(\ell\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(u)}\right)\partial\left(\boldsymbol{u}\right)}\right) \;=\; \mathrm{E}\left(\frac{\partial\left(\left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(u)}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}^{(u)}\right)}\boldsymbol{\theta}_{\boldsymbol{u}}\right)+\mathrm{E}\left(\frac{\partial^2\left(\mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u},\boldsymbol{\phi}^{(u)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(u)}\right)\partial\left(\boldsymbol{u}\right)}\right) \tag{5.53}$$

why most commonly, $\boldsymbol{u}$ is assumed to be normally distributed (cf. e.g. Bates et al., 2015; Breslow and Clayton, 1993; Ghosh et al., 1998; Harville, 1977; Henderson, 1950; Schall,

1991), although other distributions may be used as well (cf. Lee and Nelder, 1996; Rao, 2003, p. 206; Searle, Casella and McCulloch, 2006, p. 341). If it holds that

$$\boldsymbol{u} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}^{(\mathsf{s})}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right) \quad , \tag{5.54}$$

it can be shown (by using equalities 5.18) that equation 5.53 results in a matrix of zeros (cf. appendix B.4.3). This is a convenient characteristic because it reduces the expected Hessian matrix defined in equation 5.51 to a block-diagonal form. In that case, the updating rules can be split in a block-wise manner. This means that fitting GAMMs can be done by updating the parameters for modeling the conditional mean by

$$\begin{bmatrix} \boldsymbol{\kappa}^{(a)} \\ \boldsymbol{u}^{(a)} \end{bmatrix} \coloneqq \begin{bmatrix} \boldsymbol{\kappa}^{(a-1)} \\ \boldsymbol{u}^{(a-1)} \end{bmatrix} - \left[\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}^{(a-1)}\right)\right]_{\cdot\mathcal{J}} \left(\left[\mathrm{E}\left(\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}^{(a-1)}\right)\right)\right]_{\mathcal{J}\mathcal{J}}\right)^{-1} \quad , \tag{5.55}$$

where using $\mathcal{J} = \{1, \ldots, d+v\}$ results in the submatrices of $\mathbf{J}_{\mathscr{L}}(\boldsymbol{\Theta})$ and $\mathrm{E}(\mathbf{H}_{\mathscr{L}}(\boldsymbol{\Theta}))$ that correspond to $\boldsymbol{\kappa}$ and $\boldsymbol{u}$. If the model is a linear mixed model, the update rule for all other parameters is furthermore independent of the current values $\boldsymbol{u}^{(a-1)}$. In this case, the random effects do not need to be updated in every Fisher scoring iteration and can be determined once optimization is completed. When using the marginal log-likelihood that results from integrating out the random effects $\boldsymbol{u}$ for optimization, this advantage is preserved for fitting generalized linear mixed models as well (cf. Lee and Nelder, 1996, pp. 631 f; Pinheiro and Bates, 2000, p. 71; Wood, 2017, pp. 292 f). In correspondence to equality 5.55, variance components are updated by

$$\begin{aligned} \left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)^{(a)} &\coloneqq \left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)^{(a-1)} - \left[\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}^{(a-1)}\right)\right]_{\cdot\mathcal{J}} \left[\mathrm{E}\left(\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}^{(a-1)}\right)\right)\right]^{-1}_{\mathcal{J}\mathcal{J}} \\ \left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)^{(a)} &\coloneqq \left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)^{(a-1)} - \left[\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}^{(a-1)}\right)\right]_{\cdot\mathcal{M}} \left(\left[\mathrm{E}\left(\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}^{(a-1)}\right)\right)\right]_{\mathcal{M}\mathcal{M}}\right)^{-1} \end{aligned} \tag{5.56}$$

for $\mathcal{J} = \{d+v+1, \ldots, d+v+b\}$ and $\mathcal{M} = \{d+v+b+1, \ldots, d+v+s\}$ (cf. e.g. Harville, 1977, pp. 322 ff; Henderson, 1950; Rao and Molina, 2015, pp. 102 ff; Searle, Casella and McCulloch, 2006, pp. 235 ff). When comparing this case to the GAMs presented in section 5.1.4, extending update 5.55 to include the random effects and introducing update 5.56 as an additional step in the optimization procedure are the extensions made to account for dependency between observations.

An advantage of this strategy is that it facilitates the use of residual (or restricted) maximum likelihood (REML) estimation for variance parameters. In general, ML estimation of variance components $\boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot l}}^{(\mathsf{s})}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)$ and $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{(\mathsf{s})}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)$ as described above is not unbiased because the degrees of freedom that are used for estimating parameters $\boldsymbol{\kappa}$ are not accounted for. In case of mixed models, this can even result in negative variance estimates, although the true variances are always non-negative by definition. If variance components can be updated separately from the remaining parameters as in equalities 5.56, this bias can be compensated by using the REML criterion

$$\mathscr{L}_{\mathrm{R}}(\boldsymbol{\Theta}) \coloneqq \int \mathscr{L}(\boldsymbol{\Theta})\, d\boldsymbol{\kappa} \tag{5.57}$$

instead of the joint log-likelihood $\mathscr{L}(\boldsymbol{\Theta})$ for updating $\boldsymbol{\phi}$. Criterion 5.57 corresponds to the scaled average of $\mathscr{L}(\boldsymbol{\Theta})$ over all possible values of $\boldsymbol{\kappa}$. This renders variance parameter updates independent from $\boldsymbol{\kappa}$ and results in unbiased variance estimates (cf. Bates, 2018,

pp. 9 ff; Patterson and Thompson, 1971; Pinheiro and Bates, 2000, pp. 75 f; Searle, Casella and McCulloch, 2006, pp. 249 ff; Thompson et al., 1962; Wood, 2017, pp. 293 f).

For computing updates 5.55 and 5.56, one can either use a mixture of parameters already updated in the current iteration together with non-updated ones from the previous iteration (Gauss-Seidel iteration) or the full set of parameters from the previous iteration (Jacobi iteration). In particular, one can calculate $\boldsymbol{\kappa}^{(a)}$ and $\boldsymbol{u}^{(a)}$ based on $\boldsymbol{\phi}^{(a-1)}$ and then use these new values to obtain $\boldsymbol{\phi}^{(a)}$, or still use $\boldsymbol{\kappa}^{(a-1)}$ and $\boldsymbol{u}^{(a-1)}$ to determine the dispersion parameters $\boldsymbol{\phi}^{(a)}$ (cf. Hackbusch, 1994, pp. 68 f). For example, Burgard (2013, p. 11) as well as Rao and Molina (2015, pp. 102 f) use the Jacobi variant, while Schall (1991, p. 722), Breslow and Clayton (1993, p. 12) and Wolfinger and O'connell (1993, pp. 238 f) tend to use the Gauss-Seidel variant.

In the general case where no explicit distribution is assumed for $\boldsymbol{u}$ to simplify equation 5.53 (e.g. as in assumption 5.54), the parameter updates for mean and covariance structure can not necessarily be separated. Yet, it is still possible to use the Fisher scoring algorithm with Jacobian and expected Hessian matrix defined in equations 5.50 and 5.51 for ML estimation. As the updates in general form a system of linear equations, they can in principle be calculated using a linear regression (cf. Bates et al., 2015; Henderson, 1950; 1953; 1963; Wood, 2017, pp. 309 ff; Rao and Molina, 2015, p. 99; Wolfinger and O'connell, 1993, p. 239). However, $\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}^{(\mathsf{s})}(\boldsymbol{\phi})$ and hence $\boldsymbol{W}$ are no longer diagonal matrices, but the input for linear regression in most common software packages does only allow for a vector of diagonal elements (cf. e.g. Faraway, 2002, p. 62; Ruppert, Wand and Carroll, 2003, p. 85). Therefore, the handy implementation of Fisher scoring as an IRWLS scheme that is used for GLMs and GAMs (cf. equations 5.27 and 5.36) is not generally applicable for GAMMs. As before (cf. section 5.1.4), a strategy based on algorithm 11 can be used for transformations $\boldsymbol{t}$ that are not (fully) described by a parametrization $\boldsymbol{\kappa}$ (cf. section 5.1.4).

Predictions for data set t can be obtained by $\widehat{\boldsymbol{y}}_{\cdot l}^{\mathsf{t}} = \mathrm{E}(\boldsymbol{y}_{\cdot l}^{\mathsf{t}} \mid \boldsymbol{X}^{\mathsf{t}}, \boldsymbol{D}^{\mathsf{t}}) = \boldsymbol{l}^{-1}(\boldsymbol{\eta}^{\mathsf{t}})$ once optimization of parameters $\boldsymbol{\Theta}$ is completed, similarly as in the previous section 5.1.4. For commonly used models where $\mathrm{E}(\boldsymbol{u}) = 0$ (cf. assumption 5.54), prediction is even possible when $\boldsymbol{D}^{\mathsf{t}}$ is not observed because the marginal expectation can be written as $\mathrm{E}(\boldsymbol{y}_{\cdot l}^{\mathsf{t}} \mid \boldsymbol{X}^{\mathsf{t}}) = \boldsymbol{l}^{-1}(\boldsymbol{t}(\boldsymbol{X}^{\mathsf{t}}, \boldsymbol{\kappa}))$ (cf. Wood, 2017, p. 292; Pinheiro and Bates, 2000, p. 94).

To introduce generalized additive (mixed) models, general transformation functions $\boldsymbol{t}_{\mathrm{j}}$ are assumed throughout the current and previous section. B-spline transformations constitute an important example of such functions and are very common for GAMs and GAMMs (cf. Hastie and Tibshirani, 1986; 1990; Wood, 2017, pp. 142 ff). These transformations are introduced in the following section 5.1.6.

5.1.6 Regression Splines

In the previous sections 5.1.4 and 5.1.5, generalized additive (mixed) models for expressing dependencies between auxiliary and target variables are discussed with regard to general transformation functions $\boldsymbol{t}$. One specific type of such transformations that is of particular relevance for GAMs and GAMMs (cf. Hastie and Tibshirani, 1986; 1990; Wood, 2017, pp. 142 ff) as well as in various other applications (cf. e.g. Böhm, Farin and Kahmann, 1984; Ward and Ronald, 2008, pp. 404 ff) are *basis splines (B-splines)*. Favorable properties of these transformations include high numerical stability (cf. Hastie and Tibshirani, 1990, p. 25; Reinsch, 1967; Ruppert, Wand and Carroll, 2003, p. 70), continuity and smoothness of the resulting functions (cf. Wood, 2017, pp. 142 ff) as well as analytically solvable derivatives and integrals for linear combinations (cf. de Boor, 1978, p. 138; Ward and Ronald, 2008, pp. 408 f).

B-spline base functions for a single input variable $\boldsymbol{x}_{\cdot j}$ are recursively defined by

$$\mathbf{B}_k^0\left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \coloneqq \mathbb{I}\left(K_k^{\boldsymbol{x}_{\cdot j}} \leq \boldsymbol{x}_{\cdot j} < K_{k+1}^{\boldsymbol{x}_{\cdot j}}\right) \tag{5.58a}$$

$$\begin{aligned}\mathbf{B}_k^l\left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \coloneqq\ & \frac{\boldsymbol{x}_{\cdot j} - K_k^{\boldsymbol{x}_{\cdot j}}}{K_{k+l}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^{l-1}\left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \\ & + \frac{K_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - x_{ij}}{K_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^{l-1}\left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \quad ,\end{aligned} \tag{5.58b}$$

where k and l respectively denote the interval and order of the splines. The points $K_k^{\boldsymbol{x}_{\cdot j}}$ splitting the range of a variable $\boldsymbol{x}_{\cdot j}$ into intervals are called *knots* and denoted by the vector

$$\boldsymbol{K}^{\boldsymbol{x}_{\cdot j}} \coloneqq \left[K_1^{\boldsymbol{x}_{\cdot j}} \ \cdots \ K_{a_j}^{\boldsymbol{x}_{\cdot j}}\right]^{\mathsf{T}} \quad \in \quad \mathbb{R}^{a_j} \qquad . \tag{5.59}$$

The number of resulting base functions for variable $\boldsymbol{x}_{\cdot j}$ is determined by $b_j \coloneqq a_j + l - 1$, where $a_j \coloneqq |\boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}|$ is the number of knots that is used for this variable (cf. Curry and Schoenberg, 1947; 1966; de Boor, 1978, pp. 109, 131).

Depending on the order of the splines, computation of these base functions requires l additional outer knots each before the first and after the last knot. As their definition is arbitrary, these outer knots are often defined by repeating the first and last knot l times, i.e. in form of

$$K_k^{\boldsymbol{x}_{\cdot j}} \coloneqq \begin{cases} K_1^{\boldsymbol{x}_{\cdot j}} & \text{if} \quad k < 1 \\ K_{a_j}^{\boldsymbol{x}_{\cdot j}} & \text{if} \quad k > h \end{cases} \qquad . \tag{5.60}$$

To avoid division by zero, one then has to additionally define $0/0 \coloneqq 0$ because denominator(s) and base function of degree $(l-1)$ in definition 5.58b can both be zero, but only simultaneously (cf. Boor, 2001, pp. 110 ff; Curry and Schoenberg, 1966, p. 79; Hastie, Tibshirani and Friedman, 2008, pp. 186 ff).

For multiple $\boldsymbol{X}$-variables, B-spline transformations can be applied in a column-wise manner, i.e.

$$\boldsymbol{t}_{\mathrm{j}}\left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \coloneqq \left[\mathbf{B}_1^l\left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \ \cdots \ \mathbf{B}_s^l\left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right] \qquad , \tag{5.61}$$

where $\boldsymbol{t}_{\mathrm{j}} : \mathbb{R}^{n^{\mathrm{t}}} \times \mathbb{R}^{a_j} \to [0;1]^{n^{\mathrm{t}} \times b_j}$ is a component transformation used for input variable $\boldsymbol{x}_{\cdot j}$ as in the previous sections. When applying such a B-spline transformation to all p columns in $\boldsymbol{X}$, the concatenation of the knots for all $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{\mathrm{p}}$ resulting transformations

is denoted by

$$\boldsymbol{K}^{\boldsymbol{X}} \coloneqq \left[\left(\boldsymbol{K}^{\boldsymbol{x}_{\cdot 1}}\right)^{\mathsf{T}} \cdots \left(\boldsymbol{K}^{\boldsymbol{x}_{\cdot p}}\right)^{\mathsf{T}} \right]^{\mathsf{T}} \quad \in \quad \mathbb{R}^{||\boldsymbol{a}||_1} \tag{5.62}$$

for $\boldsymbol{a} = \left[a_1, \ldots, a_p\right]^{\mathsf{T}} \in \mathbb{N}^p$ containing the number of knots for all variables. The output of these p transformations is the transformed predictor matrix

$$\widetilde{\boldsymbol{X}} \coloneqq \boldsymbol{t}\left(\boldsymbol{X}, \boldsymbol{K}^{\boldsymbol{X}}\right) \coloneqq \left[\boldsymbol{t}_1\left(\boldsymbol{x}_{\cdot 1}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot 1}}\right) \cdots \boldsymbol{t}_{\mathrm{p}}\left(\boldsymbol{x}_{\cdot p}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot p}}\right) \right] \quad \in \quad [0;1]^{n \times ||\boldsymbol{b}||_1} \quad . \tag{5.63}$$

In equation 5.63, $\boldsymbol{b} = \left[b_1, \ldots, b_p\right]^{\mathsf{T}} \in \mathbb{N}^p$ with elements defined above is the vector determining the number of output columns resulting from each of the p B-spline transformations. Combining the component transformations $\boldsymbol{t}_{\mathrm{j}}$ for each variable in form of a general transformation function $\boldsymbol{t} : \mathbb{R}^{n^{\mathsf{t}} \times p} \times \mathbb{R}^{||\boldsymbol{a}||_1} \to \mathbb{R}^{n^{\mathsf{t}} \times ||\boldsymbol{b}||_1}$ achieves notational coherence with the previous and following sections.

As outlined above, B-splines are an attractive and widespread approach for GAMs. When using transformations $\widetilde{\boldsymbol{X}}$ in place of $\boldsymbol{X}$ as independent variables for GLMs, non-parametric models emerge. Mixtures of both that use a subset of columns $\boldsymbol{X}_{\cdot \mathcal{J}}$ from $\boldsymbol{X}$ together with a subset of columns $\widetilde{\boldsymbol{X}}_{\cdot \mathcal{J}}$ from $\widetilde{\boldsymbol{X}}$ result in semi-parametric models. In any case, these models can be fit by means of the techniques discussed in the previous sections 5.1.2 to 5.1.5 (cf. Ruppert, Wand and Carroll, 2003, p. 161; Hastie and Tibshirani, 1990, p. 118). For example, the model $\widehat{\boldsymbol{Y}}^{\mathsf{s}} = \widetilde{\boldsymbol{X}}^{\mathsf{s}}\boldsymbol{\beta}$ specifies a *regression spline* that is equivalent to an additive model with B-spline transformation. Since this model is linear in $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$, $\boldsymbol{\beta} \in \mathbb{R}^{||\boldsymbol{b}||_1 \times o}$ may in principle be estimated by a linear regression of $\boldsymbol{Y}^{\mathsf{s}}$ on $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$ (cf. equations 5.11 and 5.31). However, simply using $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$ in place of the independent variables $\boldsymbol{X}$ in a (generalized) linear model usually results in *overfitting*. This means that the regression function strongly adheres to observed values $\boldsymbol{Y}^{\mathsf{s}}$ but often generalizes poorly to unobserved values, e.g. when used for out-of-sample prediction. Therefore, this strategy typically "leads to a wiggly fit" (Ruppert, Wand and Carroll, 2003, p. 65) because even small random fluctuations in the data are represented in the model (cf. Hastie, Tibshirani and Friedman, 2008, pp. 151 ff; Ruppert, Wand and Carroll, 2003, pp. 58 ff).

To overcome this issue, obtain smoother regression functions and predictions of higher stability, it is therefore quite common to use a penalty term when fitting regression splines. This is an example of the shrinkage methods discussed more generally in section 5.1.11. In case of B-splines, the typical penalty is

$$\boldsymbol{p}\left(\boldsymbol{\beta}\right) \coloneqq \lambda \cdot \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{V} \boldsymbol{\beta} \qquad , \tag{5.64}$$

where $\boldsymbol{V} \in \mathbb{R}^{||\boldsymbol{b}||_1 \times ||\boldsymbol{b}||_1}$ is a *penalty matrix*. The *smoothing (or penalty) parameter* $\lambda \in \mathbb{R}_{\geq 0}$ may be considered a fixed predetermined constant but is more commonly estimated by using optimality criteria (cf. Hastie, Tibshirani and Friedman, 2008, pp. 156 ff; Wood, 2017, pp. 126 ff). Penalties for non-smoothness of the regression function in form of equation 5.64 are added to the distance function that is used for model fitting (cf. sections 5.1.2 to 5.1.5) to achieve predictions that are less volatile. A more general and detailed discussion of penalization and the choice of λ is provided in section 5.1.11.

The type of smoothness that is achieved by using the penalties defined in equation 5.64 depends on the exact specification of $\boldsymbol{V}$. Going back to the work of Reinsch (1967), this penalty matrix for regression splines is often based on the second derivatives of the

transformations with respect to $\boldsymbol{\beta}$ (cf. e.g. Boor, 2001, p. 207; Hastie, Tibshirani and Friedman, 2008, p. 154; Wood, 2017, pp. 126 ff), such that

$$\boldsymbol{V} \coloneqq \begin{bmatrix} \boldsymbol{V}^{(1)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{V}^{(p)} \end{bmatrix} \in \mathbb{R}^{||\boldsymbol{b}||_1 \times ||\boldsymbol{b}||_1} \quad . \tag{5.65}$$

Elements of the component matrices $\boldsymbol{V}^{(m)} \in \mathbb{R}^{b_m \times b_m}$ for all $m = 1, \dots p$ are then defined by

$$v_{jk}^{(m)} = \int \frac{\partial^2 \left(\mathbf{B}_j^l \left(c, \boldsymbol{K}^{\boldsymbol{x}_{\cdot m}} \right) \right)}{\partial \left(c \right) \partial \left(c \right)} \frac{\partial^2 \left(\mathbf{B}_k^l \left(c, \boldsymbol{K}^{\boldsymbol{x}_{\cdot m}} \right) \right)}{\partial \left(c \right) \partial \left(c \right)} \, d \, c \tag{5.66}$$

for all $j, k = 1, \dots, b_m$ and $m = 1, \dots, p$ (cf. Hastie and Tibshirani, 1990, p. 28). In that regard, it is convenient that the derivative of a B-spline of order l with respect to the input variable $\boldsymbol{x}_{\cdot j}$ is determined by a B-spline transformation of order $l - 1$ through

$$\frac{\partial \left(\mathbf{B}_k^l \left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}} \right) \right)}{\partial \left(\boldsymbol{x}_{\cdot j} \right)} = l \cdot \left(\frac{\mathbf{B}_k^{l-1} \left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}} \right)}{K_{k+l}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+1}^{l-1} \left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}} \right)}{K_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \right) \quad . \tag{5.67}$$

Derivatives of higher order follow by recursion (cf. appendix B.4.4.2; de Boor, 1972; 1978, p. 138; Piegl and Tiller, 1997, pp. 59 ff). An important benefit when using equation 5.66 is that the result for base functions of degree $l = 3$ is a *natural cubic spline*, implying that the regression function is linear outside the interval of observed values (cf. Hastie and Tibshirani, 1990, pp. 27 ff). Since this strategy aims at a smoother regression line than in the non-penalized model, it is commonly referred to as *'smoothing splines'* (cf. e.g. Eilers and Marx, 1996, p. 89; Hastie and Tibshirani, 1990, p. 27; Wood, 2017, p. 144). However, the use of second derivatives in equality 5.66 imposes considerable computational burden in optimization (cf. e.g. Wood, 2017, p. 144). An important computationally cheaper alternative are *penalized splines (P-splines)* introduced by Eilers and Marx (1996). In this case, multiplication with $\boldsymbol{V}$ is used to construct differences of the values in $\boldsymbol{\beta}$. For example, using

$$\boldsymbol{V}^{(m)} \coloneqq \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 & 0 \\ & \vdots & & \ddots & & & \vdots & \\ 0 & 0 & 0 & \dots & 0 & 0 & -1 & 1 \end{bmatrix} \tag{5.68}$$

as components for equation 5.65 constitutes squared first differences, but the penalty's order is easily adjustable (cf. Wood, 2017, pp. 149 f).

In many cases, the knots that are used to define the B-spline base functions in equations 5.58 are predefined or evenly spaced over the range or quantiles of the independent variables (cf. Wood, 2017, pp. 133, 149). Nevertheless, knot positioning can be subject to optimization as well, e.g. by using the gradient-based methods described in section 4.2 (cf. section 5.1.9; de Boor, 1978, pp. 181, 271 f; de Boor and Rice, 1968a,b; Laube, Franz and Umlauf, 2018). The derivative of a base function $\mathbf{B}_k^l \left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}} \right)$ with respect to the knots that can be used for this purpose is defined by

$$\frac{\partial \left(\mathbf{B}_k^l \left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}} \right) \right)}{\partial \left(\boldsymbol{K}^{\boldsymbol{x}_{\cdot j}} \right)} = \left[\frac{\partial \left(\mathbf{B}_k^l \left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}} \right) \right)}{\partial \left(K_1^{\boldsymbol{x}_{\cdot j}} \right)} \dots \frac{\partial \left(\mathbf{B}_k^l \left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}} \right) \right)}{\partial \left(K_s^{\boldsymbol{x}_{\cdot j}} \right)} \right] \quad . \tag{5.69}$$

Its components are

$$\frac{\partial\left(\mathbf{B}_k^l\left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right)}{\partial\left(K_m^{\boldsymbol{x}_{\cdot j}}\right)} = \begin{cases} 0 & , \text{ if } \quad 0 < k < m-l-1 \\ \dfrac{\mathbf{B}_{k+1}^l\left(\boldsymbol{x}_{\cdot j}, \overline{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right)}{K_m^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} & , \text{ if } \quad k = m-l-1 \\ \dfrac{\mathbf{B}_{k+1}^l\left(\boldsymbol{x}_{\cdot j}, \overline{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \dfrac{\mathbf{B}_k^l\left(\boldsymbol{x}_{\cdot j}, \overline{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+l}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} & , \text{ if } \quad m-l \leq k \leq m-1 \\ -\dfrac{\mathbf{B}_k^l\left(\boldsymbol{x}_{\cdot j}, \overline{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+l}^{\boldsymbol{x}_{\cdot j}} - K_m^{\boldsymbol{x}_{\cdot j}}} & , \text{ if } \quad k = m \\ 0 & , \text{ if } \quad m < k < u \quad . \end{cases} \tag{5.70}$$

These are based on the modified knot vector

$$\overline{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}} \coloneqq \begin{bmatrix} K_1^{\boldsymbol{x}_{\cdot j}} & \cdots & K_m^{\boldsymbol{x}_{\cdot j}} & K_m^{\boldsymbol{x}_{\cdot j}} & \cdots & K_u^{\boldsymbol{x}_{\cdot j}} \end{bmatrix} \quad , \tag{5.71}$$

i.e. the original knot vector $\boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}$ where the respective knot $K_m^{\boldsymbol{x}_{\cdot j}}$ is duplicated (cf. appendix B.4.4.1; Piegl and Tiller, 1998, p. 931).

These results are particularly important also for section 5.1.9, where gradient-based knot selection in artificial neural networks is proposed. Beforehand, multivariate adaptive regression splines are introduced in the following section 5.1.7. This established approach aims at finding optimal knot locations and interaction terms for regression splines using a trial-and-error strategy (cf. Hastie and Tibshirani, 1990, p. 249).

5.1.7 Multivariate Adaptive Regression Splines and Regression Trees

Multivariate adaptive regression splines (MARS) constitute an extension to the regression splines introduced in the previous section 5.1.6. To obtain a possibly better prediction model for $\boldsymbol{Y}$, the main idea of MARS is to make an optimal choice for the number and location of knots $\boldsymbol{K}^{\boldsymbol{X}}$, as well as for interaction terms between spline transformations of $\boldsymbol{X}$. The optimization parameters in this context are

$$\boldsymbol{\Theta} \coloneqq \begin{bmatrix} \boldsymbol{K}^{\boldsymbol{X}} & \boldsymbol{\beta} \end{bmatrix} \quad \in \quad \mathbb{R}^u \quad , \tag{5.72}$$

with dimension u not being predetermined. These parameters are again chosen such that a distance function $\delta : \mathbb{R}^u \to \mathbb{R}_{\geq 0}$ is minimized. The regression parameters $\boldsymbol{\beta}$ are usually optimized using the fitting methods discussed in the previous sections 5.1.2 to 5.1.6. In contrast, the knot locations and B-spline interaction terms in this context are selected by a greedy trial-and-error strategy. MARS models and the outlined fitting strategy are proposed by Friedman (1991b, p. 17). Considering such a model for $\boldsymbol{y}_{\cdot l}$, the steps for determining the model's parameters from data set s are described by algorithm 12.

In this algorithm, c is the predefined number of utilized transformations. The output is the combination of these transformations $\widetilde{\boldsymbol{X}}^{(j)} \in \mathbb{R}^{n^{\mathsf{s}} \times v_j}$ for $j = 1, \ldots, c$, of which all but $\widetilde{\boldsymbol{X}}^{(1)}$ and $\widetilde{\boldsymbol{X}}^{(2)}$ are potentially constructed as interaction terms from multiple columns of $\boldsymbol{X}^{\mathsf{s}}$. To that end, $\mathscr{V}^{(a)}$ denotes the set of indices for all variables that are used in the a-th

Algorithm 12: MARS: forward stepwise selection

Input: $\boldsymbol{X}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}} \times p}$; $\boldsymbol{y}^{\mathsf{s}}_{\cdot l} \in \mathbb{R}^{n^{\mathsf{s}}}$; $\boldsymbol{w}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}}}$; $\delta : \mathbb{R}^{u} \to \mathbb{R}_{\geq 0}$; $c \in \mathbb{N}$

Initialize the intercept $\widetilde{\boldsymbol{X}}^{(1)} \coloneqq \mathbf{1}_{n^{\mathsf{s}} \times 1}$, $\boldsymbol{K} \coloneqq \mathbf{0}_{0 \times 0}$ $\boldsymbol{d} \coloneqq \boldsymbol{d}^{*} \coloneqq \begin{bmatrix} \infty & \dots & \infty \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{c}_{\geq 0}$, and set $\mathscr{V}^{(1)} \coloneqq \emptyset$

for $b = 2, \dots, c$ **do**

  for $a = 1$ to $(b-1)$ **do**

    for $j \notin \mathscr{V}^{(a)}$ **do**

      for $\boldsymbol{K}^{\boldsymbol{x}^{\mathsf{s}}_{\cdot j}} \in \mathscr{K}^{\boldsymbol{x}^{\mathsf{s}}_{\cdot j}}$ **do**

        Construct the candidate matrix $\widetilde{\boldsymbol{X}}^{*} \in \mathbb{R}^{n^{\mathsf{s}} \times (h \cdot u)}$, containing all interactions of columns in $\widetilde{\boldsymbol{X}}^{(a)} \in \mathbb{R}^{n^{\mathsf{s}} \times h}$ and candidate transformation $\boldsymbol{t}\left(\boldsymbol{x}^{\mathsf{s}}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}^{\mathsf{s}}_{\cdot j}}\right) \in \mathbb{R}^{n^{\mathsf{s}} \times u}$:

$$\widetilde{\boldsymbol{X}}^{*} \coloneqq \begin{bmatrix} \widetilde{\boldsymbol{x}}^{(a)}_{\cdot 1} \circ \left[\boldsymbol{t}\left(\boldsymbol{x}^{\mathsf{s}}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}^{\mathsf{s}}_{\cdot j}}\right)\right]_{\cdot 1} & \dots & \widetilde{\boldsymbol{x}}^{(a)}_{\cdot h} \circ \left[\boldsymbol{t}\left(\boldsymbol{x}^{\mathsf{s}}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}^{\mathsf{s}}_{\cdot j}}\right)\right]_{\cdot u} \end{bmatrix} \quad .$$

        Calculate the prediction error when using $\widetilde{\boldsymbol{X}}^{*}$ in addition to the already chosen transformations, such that the matrix of independent variables is given by $\begin{bmatrix} \widetilde{\boldsymbol{X}}^{(1)} & \dots & \widetilde{\boldsymbol{X}}^{(b-1)} & \widetilde{\boldsymbol{X}}^{*} \end{bmatrix}$:

$$d^{*}_{b} \coloneqq \min_{\boldsymbol{\beta}} \left(\delta \left(\begin{bmatrix} \boldsymbol{K}^{\mathsf{T}} & \left(\boldsymbol{K}^{\boldsymbol{x}^{\mathsf{s}}_{\cdot j}}\right)^{\mathsf{T}} & \boldsymbol{\beta}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \right) \right)$$

        if $d^{*}_{b} < d_{b}$ **then**

          Choose $\widetilde{\boldsymbol{X}}^{*}$ as (intermediate) b-th transformation:

          Set $\widetilde{\boldsymbol{X}}^{(b)} \coloneqq \widetilde{\boldsymbol{X}}^{*}$, $\boldsymbol{K}^{*} \coloneqq \boldsymbol{K}^{\boldsymbol{x}^{\mathsf{s}}_{\cdot j}}$, $\mathscr{V}^{(b)} \coloneqq \mathscr{V}^{(a)} \cup j$ and $d_{b} \coloneqq d^{*}_{b}$

        end if

      end for

    end for

  end for

  Update $\boldsymbol{K} \leftarrow \begin{bmatrix} \boldsymbol{K}^{\mathsf{T}} & (\boldsymbol{K}^{*})^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$

end for

Calculate

$$\boldsymbol{\beta}^{*} \coloneqq \operatorname*{argmin}_{\boldsymbol{\beta}} \left(\delta \left(\begin{bmatrix} \boldsymbol{K}^{\mathsf{T}} & \boldsymbol{\beta}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \right) \right)$$

Return: $\widetilde{\boldsymbol{X}}^{\mathsf{s}} \coloneqq \begin{bmatrix} \widetilde{\boldsymbol{X}}^{(1)} & \dots & \widetilde{\boldsymbol{X}}^{(c)} \end{bmatrix}$ and $\boldsymbol{\Theta} \coloneqq \begin{bmatrix} \boldsymbol{K}^{\mathsf{T}} & (\boldsymbol{\beta}^{*})^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$

base function. Further, $\boldsymbol{t}\left(\boldsymbol{x}^{\mathsf{s}}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}^{\mathsf{s}}_{\cdot j}}\right)$ is a candidate transformation of $\boldsymbol{x}^{\mathsf{s}}_{\cdot j}$ (cf. definition 5.61), and $\mathscr{K}^{\boldsymbol{x}^{\mathsf{s}}_{\cdot j}}$ denotes the set of all possible knot combinations for $\boldsymbol{x}^{\mathsf{s}}_{\cdot j}$. In addition, $\boldsymbol{K}^{*}$ is an intermediate storage for potential knots, and $\boldsymbol{K}$ contains the final knots. The latter is constructed by successively concatenating the chosen knots, starting from the empty vector $\mathbf{0}_{0 \times 0}$. The output transformations are found by evaluating the interaction terms of candidate and previously chosen transformations with regard to their reduction of the distance function δ. The lowest possible value for δ is sought by comparison of $\boldsymbol{d}$ and $\boldsymbol{d}^{*}$. A distance function that is commonly used in this context is a modified version of the

generalized cross-validation (gcv) criterion proposed by Craven and Wahba (1979; cf. also Friedman, 1991b, p. 20). It is defined by

$$\delta\left(\boldsymbol{\Theta}\right) \;=\; \left(1 - \frac{\dim\left(\boldsymbol{\beta}\right) + \lambda \cdot \dim\left(\boldsymbol{K}\right)}{n^{\mathsf{s}}}\right)^{-2} \cdot \left(\boldsymbol{w}^{\mathsf{s}}\right)^{\mathsf{T}} \left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}} - \widehat{\boldsymbol{y}}_{\cdot l}^{\mathsf{s}}\right)^{\circ 2} \qquad , \tag{5.73}$$

which is the weighted residual sum of squares multiplied by a penalty for model-complexity. A general justification of this criterion is given in section 5.1.11, but in the present context, the penalty is defined by the number of parameters used to generate $\widehat{\boldsymbol{Y}}$. This number is determined by $\dim\left(\boldsymbol{\beta}\right)$ and $\dim\left(\boldsymbol{K}\right)$, which respectively denote the number of regression parameters and knots. In this regard, $\lambda \in \mathbb{R}_{\geq 0}$ is a penalty constant, which is usually set to three for models with and two for models without interaction terms (cf. Hastie and Tibshirani, 1990, p. 275; Hastie, Tibshirani and Friedman, 2008, p. 325). Other loss functions can be applied as well, e.g. the negative log-likelihood for binomial variables in conjunction with a link function (cf. section 5.1.4; Friedman, 1991b, p. 47). Starting from an intercept-only model, every iteration of the outer loop (starting in step 3) in algorithm 12 chooses new base functions as the interactions of an already chosen and an additional transformation, respectively denoted by $\widetilde{\boldsymbol{X}}^{(a)}$ and $\boldsymbol{t}\left(\boldsymbol{x}_{\cdot j}^{\mathsf{s}}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}^{\mathsf{s}}}\right)$. The three inner loops (beginning in steps 4 to 6) determine the best choice for these new base functions with respect to δ, but only interactions using distinct variables are considered.

However, letting each of these loops iterate over all possible values tremendously increases computation time, especially for larger data sets. To overcome this issue, Friedman (1991a,b; 1993) proposes different strategies to evaluate only a subset of values in each loop. First of all, two-sided truncated power base functions rather than B-splines are used in most MARS implementations. For splines of order l, the former are defined by

$$\boldsymbol{t}^{l}\left(\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \;\coloneqq\; \left[\mathbb{I}\left(\boldsymbol{x}_{\cdot j} \leq \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \cdot \left(\boldsymbol{x}_{\cdot j} - \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)^{l} \quad \mathbb{I}\left(\boldsymbol{x}_{\cdot j} \geq \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \cdot \left(\boldsymbol{x}_{\cdot j} - \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)^{l}\right] \tag{5.74}$$

and constitute an alternative but in numerical terms less appealing spline representation when compared to B-splines (cf. de Boor, 1972, p. 50; Friedman, 1991b, p. 29). Nevertheless, their usage is justified by reducing the complexity of algorithm 12: each of these base function is defined by a single knot, such that the set of possible knot-combinations $\mathscr{K}^{\boldsymbol{x}_{\cdot j}}$ reduces to the set of unique values of $\boldsymbol{x}_{\cdot j}$. Furthermore, the number of resulting candidate base functions is two (cf. Bakin, Hegland and Osborne, 1997; Friedman, 1991b; Hastie and Tibshirani, 1990, p. 275). The second approach to improve the computation speed of the MARS model is to reduce the number of possible parent functions $\widetilde{\boldsymbol{x}}_{\cdot a}$ in the loop beginning in step 4. This can be done by using a "parent priority queue" (Friedman, 1993, p. 6). Only parent functions (identified by the index a) that substantially decreased the distance in the preceding iteration are considered in this strategy. To avoid diminishing a parent function for all later steps by giving them a very low priority in one step and never re-evaluating them, the natural "'aging' property" (Friedman, 1993, p. 8) of this approach is assisted through adjusting the priority by the number of iterations they were not evaluated. A similar strategy, conditional on the values in the superordinate loops, is applied to the subsequent loops beginning in steps 5 and 6 (cf. Friedman, 1993, pp. 9 f).

A model generated by algorithm 12 is usually overfit and rather complex. Therefore, it typically needs some reduction to be used for actual predictions, which is achieved by a procedure similar to algorithm 12. Through successively deleting columns of $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$, only

those that provide the lowest respective value of δ are kept. Since the base functions can overlap due to the interaction terms, this can be done by deleting a single column in every step, as formalized in algorithm 13 (cf. Friedman, 1991b, p. 17).

Algorithm 13: MARS: backward stepwise selection

1: **Input:** $\widetilde{\boldsymbol{X}}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}} \times h}$; $\boldsymbol{y}^{\mathsf{s}}_{\cdot l} \in \mathbb{R}^{n^{\mathsf{s}}}$; $\boldsymbol{w}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}}}$; $\delta : \mathbb{R}^u \to \mathbb{R}_{\geq 0}$
2: Initialize $\mathcal{I} \coloneqq \mathcal{I}^{(h+1)} \coloneqq \{1, \ldots, h\}$, $d \coloneqq \min_{\boldsymbol{\Theta}} (\delta(\boldsymbol{\Theta}))$ and $\boldsymbol{d} \coloneqq \boldsymbol{d}^* \coloneqq \begin{bmatrix} \infty & \ldots & \infty \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^h_{\geq 0}$
3: **for** $b = h$ to 2 **do**
4: **for** $a = 2$ to k **do**
5: Set $\mathcal{M}^{(a)} \coloneqq \mathcal{I}^{(b+1)} \setminus \{a\}$ and calculate the loss function value when using only the subset $\widetilde{\boldsymbol{X}}^{\mathsf{s}}_{\cdot \mathcal{M}^{(a)}}$ defined by $\mathcal{M}^{(a)}$ as predictors:

$$d_a^* \coloneqq \min_{\boldsymbol{\beta}} \left(\delta \left(\begin{bmatrix} \boldsymbol{K}^{\mathsf{T}} & \left(\boldsymbol{K}^{\boldsymbol{x}^{\mathsf{s}}_{\cdot j}} \right)^{\mathsf{T}} & \boldsymbol{\beta}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \right) \right)$$

6: **if** $d_a^* < d$ **then**
7: Set $d \coloneqq d_a^*$ and $\mathcal{I} \coloneqq \mathcal{M}^{(a)}$
8: **end if**
9: **if** $d_a^* < d_b$ **then**
10: Set $d_b \coloneqq d_a^*$ and $\mathcal{I}^{(b)} \coloneqq \mathcal{M}^{(a)}$
11: **end if**
12: **end for**
13: **end for**
14: **Return:** $\mathcal{I}$

As before, $\widetilde{\boldsymbol{X}}^{\mathsf{s}}_{\cdot \mathcal{M}^{(a)}}$ denotes a subset of columns in $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$ that is defined by the set $\mathcal{M}^{(a)}$. Consequently, each iteration of the outer for-loop beginning in step 3 of algorithm 13 deletes one single column of $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$. The inner loop that begins in step 4 determines the best choice for this deletion with respect to δ, but the intercept column can never be removed. The model with the lowest distance value is sought, using comparisons of d, $\boldsymbol{d}$ and $\boldsymbol{d}^*$. It is determined by the set of column-indices $\mathcal{I}$, which defines the output (cf. Friedman, 1991a,b; Hastie, Tibshirani and Friedman, 2008, pp. 321 ff).

MARS can be interpreted as a continuous generalization of *regression trees*. The latter emerge from using truncated power functions of order zero, such that transformations are given by

$$\boldsymbol{t}^0 (\boldsymbol{x}_{\cdot j}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}) \coloneqq \begin{bmatrix} \mathbb{I}(\boldsymbol{x}_{\cdot j} \leq \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}) & \mathbb{I}(\boldsymbol{x}_{\cdot j} \geq \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}) \end{bmatrix} \quad . \tag{5.75}$$

In this case, each base function is a binary partitioning of the covariate space (cf. Friedman, 1991b, pp. 10 ff; Hastie and Tibshirani, 1990, pp. 275 f).

The idea of MARS is to find optimal knots and interaction terms for regression splines. This is achieved by combining the fitting methods outlined in the previous sections with a greedy trial-and-error strategy for knots and interactions. Although the strategies for acceleration and simplification discussed above considerably reduce computation times for the MARS algorithm, evaluating a large number of potential knots and interactions is

still computationally expensive. This is especially true when the output is non-linear in $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$, e.g. when using a link function to predict categorical variables (cf. Friedman, 1991b, p. 47). Application for more complex transformations, such as B-splines, is therefore typically hardly feasible (cf. Bakin, Hegland and Osborne, 2000, p. 468). An alternative approach that is frequently compared to MARS (cf. e.g. Ahmadi et al., 2019; Friedman, 1993, p. 1; Mirabbasi et al., 2019; Zhang and Goh, 2016) and can even be used to perform knot selection for B-splines (cf. section 5.1.9; Laube, Franz and Umlauf, 2018) is constituted by artificial neural networks. Similar to MARS, these can perform flexible optimization regarding the underlying transformations of input variables. They even allow for combinations of multiple transformations, but typically do not select interaction terms. Artificial neural networks are introduced in the following section 5.1.8.

5.1.8 Artificial Neural Networks

Generalized additive models are quite flexible in describing various relations between independent and dependent variables. However, if the underlying transformation functions are not considered fixed, their specification, e.g. with respect to the knots in case of B-splines, is typically rather difficult and costly (cf. section 5.1.7; Friedman, 1991b; Hastie and Tibshirani, 1990, pp. 235 ff; Wood, 2017, pp. 119, 150 ff).

Artificial neural networks (ANNs) are an alternative way to transform predictor variables $\boldsymbol{X}$ for building a potentially non-linear prediction model for $\boldsymbol{Y}$ (cf. Bishop, 1995, p. 6). Although these models do not incorporate an optimal selection of interaction terms as in the MARS algorithm 12, predetermined interaction terms that are represented in $\boldsymbol{X}$ are feasible, just as for (generalized) linear and additive models. Furthermore, it can be shown that ANNs are able to approximate any functional relationship to any desired degree of precision (cf. Hornik, Stinchcombe, White et al., 1989). Artificial neural networks can therefore be used to model arbitrary relationships between the independent and dependent variables $\boldsymbol{X}$ and $\boldsymbol{Y}$. Continuing the notation of the preceding sections, their structure resembles a system of chained non-linear regression equations that are based on transformation functions. In this manner, ANNs keep up with the flexibility of GAMs by using a potentially larger number of transformations that are chained but often simpler than those discussed before (cf. Bishop, 1995, pp. 136 f; Ripley, 1996, pp. 143 ff).

An ANN with $A-1$ hidden layers predicting $\boldsymbol{Y}^{\mathsf{t}}$ from $\boldsymbol{X}^{\mathsf{t}}$ is defined by

$$\begin{aligned} \widetilde{\boldsymbol{X}}^{(0)} &:= \boldsymbol{X}^{\mathsf{t}} \\ \widetilde{\boldsymbol{X}}^{(i)} &:= \boldsymbol{t}_{\mathrm{i}}\left(\sum_{j=0}^{A} \widetilde{\boldsymbol{X}}^{(j)}\boldsymbol{\beta}^{(ij)}\right) \qquad \text{for all} \quad i = 1, \ldots, A \\ \widehat{\boldsymbol{Y}}^{\mathsf{t}} &:= \widetilde{\boldsymbol{X}}^{(A)} \quad . \end{aligned} \tag{5.76}$$

The output of each layer $i = 1, \ldots, A$ is a matrix $\widetilde{\boldsymbol{X}}^{(i)} \in \mathbb{R}^{n^{\mathsf{t}} \times c_i}$ with columns $\widetilde{\boldsymbol{x}}_{\cdot m}^{(i)}$ and possibly an intercept column $\widetilde{\boldsymbol{x}}_{\cdot 1}^{(i)} := \mathbf{1}_{n\times 1}$. In the literature relating to neural networks, intercept coefficients are commonly referred to as the *bias* (cf. e.g. Bishop, 1995, p. 81; Hagan et al., 1996, p. 2-12; Hastie, Tibshirani and Friedman, 2008, pp. 392 f). In this context, the vector $\boldsymbol{c} = \begin{bmatrix} c_0 & \ldots & c_A \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{A+1}$ contains the number of columns for each layer's output, using $c_0 := p$ for notational coherence to denote the input dimension, which corresponds to the output of the zeroth (or input) layer.

For $A \geq i > 0$, each column $\widetilde{\boldsymbol{x}}_{\cdot m}^{(i)}$ that does not represent an intercept is called a *derived feature*. Such a single column is commonly interpreted as the output of a hidden unit m in layer i of the ANN, which is termed an *artificial neuron*. This column $\widetilde{\boldsymbol{x}}_{\cdot m}^{(i)}$ represents a transformation of independent variables $\boldsymbol{X}^{\mathsf{t}}$, computed as a linear combination of the outputs $\widetilde{\boldsymbol{X}}^{(j)}$ from any layer j for $0 \leq j \leq A$. These outputs are combined using neuron-specific parameter vectors $\boldsymbol{\beta}_{\cdot m}^{(ij)} \in \mathbb{R}^{c_j}$, and the linear combination is transformed by a potentially non-linear prespecified layer-specific transformation $\boldsymbol{t}_{\mathrm{i}} : \mathbb{R}^{n^{\mathsf{t}} \times h} \to \mathbb{R}^{n^{\mathsf{t}} \times s}$ for arbitrary given dimensions $h, s \in \mathbb{N}$. Correspondingly, calculation of the complete layer can be written in matrix notation using $\boldsymbol{\beta}^{(ij)} = \left[\boldsymbol{\beta}_{\cdot 1}^{(ij)} \quad \ldots \quad \boldsymbol{\beta}_{\cdot c_i}^{(ij)}\right]^{\mathsf{T}} \in \mathbb{R}^{c_j \times c_i}$ as outlined in definition 5.76. In the context of artificial neural networks, $\boldsymbol{t}_{\mathrm{i}}$ is called an *activation function*, and the predictions are simply given by the output of the last layer (cf. Bishop, 1995, p. 82; Hagan et al., 1996, p. 2-2 ff; Hastie, Tibshirani and Friedman, 2008, pp. 392 ff).

In a *feed-forward* neural network, only information from preceding layers is used in each layer i, implying that many of the coefficients are restricted to zero and thereby reducing model complexity:

$$\boldsymbol{\beta}^{(ij)} \stackrel{!}{=} \mathbf{0} \qquad \text{for all} \quad j \geq i \qquad . \tag{5.77}$$

In contrast, *recurrent* ANNs may also use information $\widetilde{\boldsymbol{X}}^{(j)}$ from layers $j \geq i$ in layer i and, thus, do not impose this restriction. The number of free parameters is often reduced even further, by using only the output $\widetilde{\boldsymbol{X}}^{(i-1)}$ from the immediately preceding layer, such that

$$\boldsymbol{\beta}^{(ij)} \stackrel{!}{=} \mathbf{0} \qquad \text{for all} \quad j \neq i - 1 \qquad . \tag{5.78}$$

However, this is not necessarily the case, and using information from more than the immediately preceding layer is called a *skip layer connection*. Additional equality constraints can be imposed for certain parameters, resulting in *convolutional* ANNs (cf. Hastie, Tibshirani and Friedman, 2008, p. 407; Goller and Kuchler, 1996; Hagan et al., 1996, p. 2-13 ff; Ripley, 1996, pp. 143 ff).

An artificial neural network is fit by minimizing a distance function with respect to the model parameters, which as in the previous sections is denoted for a data set s. To represent this optimization while avoiding tedious tensor or array notation, the following paragraphs use vectorized (flattened) coefficient-matrices. These are defined by a function $\mathbf{vec} : \mathbb{R}^{c_j \times c_i} \to \mathbb{R}^{c_j \cdot c_i}$, such that

$$\mathbf{vec}\left(\boldsymbol{\beta}^{(ij)}\right) \coloneqq \begin{bmatrix} \boldsymbol{\beta}_{\cdot 1}^{(ij)} \\ \vdots \\ \boldsymbol{\beta}_{\cdot c_i}^{(ij)} \end{bmatrix} \quad \in \quad \mathbb{R}^{c_i \cdot c_j} \tag{5.79}$$

for all $i = 1, \ldots, A$ and $j = 0, \ldots, A$ as well as

$$\boldsymbol{\Theta} \coloneqq \begin{bmatrix} \mathbf{vec}\left(\boldsymbol{\beta}^{(10)}\right) \\ \vdots \\ \mathbf{vec}\left(\boldsymbol{\beta}^{(AA)}\right) \end{bmatrix} \quad \in \quad \mathbb{R}^{\dim(\boldsymbol{\Theta})} \tag{5.80}$$

simply concatenate all columns of (each) $\boldsymbol{\beta}^{(jk)}$ to a vector of length $\dim(\boldsymbol{\Theta}) \coloneqq \boldsymbol{c}^{\mathsf{T}} \mathbf{1}_{A \times A} \boldsymbol{c}$.

Apart from this flattening, the coefficient matrices (and their entries in particular) do not change. Since the dimension of $\boldsymbol{\Theta}$ grows quickly, it becomes evident why many of the coefficients need to be restricted to zero for fitting the model (cf. equations 5.77 and 5.78; Hagan et al., 1996, p. 12-22 ff; Ripley, 1996, pp. 148 ff).

Determining the parameters $\boldsymbol{\Theta}$ from data set s is again based on minimizing a distance function $\delta_{\mathrm{m}} : \mathbb{R}^{\dim(\boldsymbol{\Theta})} \to \mathbb{R}_{\geq 0}$, such that

$$\boldsymbol{\Theta}^* \;=\; \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left(\delta_{\mathrm{m}} \left(\boldsymbol{\Theta} \right) \right) \tag{5.81}$$

for different possible types of distance functions that are indicated by using subscripts (m in this general case). These loss functions δ_{m} are usually a weighted sum over the loss of each individual observation and variable, i.e.

$$\delta_{\mathrm{m}} \left(\boldsymbol{\Theta} \right) \;=\; \sum_{k=1}^{n^{\mathsf{s}}} w_k^{\mathsf{s}} \cdot \widetilde{\delta}_{\mathrm{m}} \left(\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}} \right) \qquad . \tag{5.82}$$

Here, $\widetilde{\delta}_{\mathrm{m}} : \mathbb{R}^{1 \times o} \to \mathbb{R}_{\geq 0}$ is a non-negative function expressing the prediction error for the o variables $\boldsymbol{y}_{k\cdot}^{\mathsf{s}} \in \mathbb{R}^{1 \times o}$ for observation $k \in \mathcal{S}^{\mathsf{s}}$. Neural networks are usually fit using gradient-based methods described in chapter 4 for optimization. To that end, it is assumed that all required derivatives exist, which implies that activation functions $\boldsymbol{t}_{\mathrm{i}}$ for all layers $i = 1, \ldots, A$ need to be differentiable. However, due to the large number of free parameters, it is common to avoid using Hessian matrices, and even Fisher and Quasi-Newton approximations are often considered too costly. The common practice is to apply gradient descent methods for fitting ANNs, in which case the Hessian is substituted by an identity matrix. This strategy is labeled *backpropagation* of errors in the context of neural networks. Computing the transformations to construct $\widetilde{\boldsymbol{X}}^{(i)}$ for all $i = 1, \ldots, A$ is similarly labeled the *forward pass* (cf. section 4.2.3; Bishop, 1995, pp. 140 ff, 287; Hagan et al., 1996, p. 9-1 ff; Hastie, Tibshirani and Friedman, 2008, pp. 396 f; Ripley and Venables, 2016).

For this setting, only the Jacobian matrix $\mathbf{J}_{\delta_{\mathrm{m}}} \left(\boldsymbol{\Theta} \right)$ of distance function 5.82 is required for optimization. For the feed-forward neural networks exclusively used throughout the following chapters, it is defined by

$$\mathbf{J}_{\delta_{\mathrm{m}}} \left(\boldsymbol{\Theta} \right) \;=\; \sum_{k=1}^{n^{\mathsf{s}}} w_k^{\mathsf{s}} \cdot \mathbf{J}_{\widetilde{\delta}_{\mathrm{m}}} \left(\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}} \right) \mathbf{J}_{\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}}} \left(\boldsymbol{\Theta} \right) \qquad . \tag{5.83}$$

Expressing each predicted row $\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}}$ as a *response function* $\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}} : \mathbb{R}^{\dim(\boldsymbol{\Theta})} \to \mathbb{R}^{o}$ of parameters $\boldsymbol{\Theta}$, equality 5.83 is determined by multiplying the Jacobian matrices of distance and response function of the ANN, which are respectively denoted by $\mathbf{J}_{\widetilde{\delta}_{\mathrm{m}}}$ and $\mathbf{J}_{\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}}}$. While the first Jacobian depends on the chosen distance function (examples of which are discussed below), the second is only dependent on the transformation functions since

$$\mathbf{J}_{\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}}} \left(\boldsymbol{\Theta} \right) \;=\; \left[\frac{\partial (\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}})}{\partial \left(\mathbf{vec} \left(\boldsymbol{\beta}^{(10)} \right) \right)} \quad \cdots \quad \frac{\partial (\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}})}{\partial \left(\mathbf{vec} \left(\boldsymbol{\beta}^{(AA)} \right) \right)} \right] \tag{5.84}$$

holds due to definition 5.80. The convenient thing about neural networks is that the entries of this Jacobian can be found by recursively applying the chain rule. Because

equality 5.83 is a sum over all observations, each row of $\widehat{\boldsymbol{Y}}^{\mathsf{s}}$ can be considered separately in equation 5.84. For an efficient representation of derivatives, the factors

$$\boldsymbol{d}_k^{(ij)} \coloneqq \frac{\partial(\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}})}{\partial\left(\widetilde{\boldsymbol{x}}_{k\cdot}^{(j)}\right)} = \begin{cases} \mathbf{J}_{\boldsymbol{t}_i}\left(\sum_{a=0}^{A}\widetilde{\boldsymbol{x}}_{k\cdot}^{(a)}\boldsymbol{\beta}^{(ia)}\right)\left(\boldsymbol{\beta}^{(ij)}\right)^{\mathsf{T}} & \text{if} \quad i = A \\ \left(\sum_{a=1}^{A}\boldsymbol{d}_k^{(ai)}\right)\mathbf{J}_{\boldsymbol{t}_i}\left(\sum_{a=0}^{A}\widetilde{\boldsymbol{x}}_{k\cdot}^{(a)}\boldsymbol{\beta}^{(ia)}\right)\left(\boldsymbol{\beta}^{(ij)}\right)^{\mathsf{T}} & \text{if} \quad 0 < i < A \end{cases} \tag{5.85}$$

are defined for all elements $k \in \mathcal{S}^{\mathsf{s}}$. As a direct consequence of definition 5.76, these factors can be passed backwards from layer i to j to determine equation 5.84, which is why this approach is called backpropagation. The partial derivatives of the predictions $\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}}$ with respect to the parameters $\boldsymbol{\beta}^{(ij)}$ required for this purpose are then given by

$$\frac{\partial(\widehat{\boldsymbol{y}}_{k\cdot}^{\mathsf{s}})}{\partial\left(\mathbf{vec}\left(\boldsymbol{\beta}^{(ij)}\right)\right)} = \begin{cases} \mathbf{J}_{\boldsymbol{t}_i}\left(\sum_{a=0}^{A}\widetilde{\boldsymbol{x}}_{k\cdot}^{(a)}\boldsymbol{\beta}^{(ia)}\right)\otimes\widetilde{\boldsymbol{x}}_{k\cdot}^{(j)} & \text{if} \quad i = A \\ \left(\sum_{a=1}^{A}\boldsymbol{d}_k^{(ai)}\right)\left(\mathbf{J}_{\boldsymbol{t}_i}\left(\sum_{a=0}^{A}\widetilde{\boldsymbol{x}}_{k\cdot}^{(a)}\boldsymbol{\beta}^{(ia)}\right)\otimes\widetilde{\boldsymbol{x}}_{k\cdot}^{(j)}\right) & \text{if} \quad 0 < i < A\ . \end{cases} \tag{5.86}$$

Here, $\otimes$ denotes the Kronecker product, and $\mathbf{J}_{\boldsymbol{t}_i}$ is the Jacobian matrix of the i-th layer's activation function, examples of which are discussed below (cf. Bishop, 1995, pp. 263 ff; Goller and Kuchler, 1996; Hagan et al., 1996, p. 11-7 ff; Ripley, 1996, pp. 150 ff; Hastie, Tibshirani and Friedman, 2008, p. 396).

For neural networks that are not strictly feeding forward, solutions can be found by very similar strategies, e.g. by backpropagation through time or structure (cf. Mozer, 1995, pp. 354 f; Goller and Kuchler, 1996) as well as real-time recurrent learning. These approaches strongly resemble the method presented above, but unfold (iterate through) the recursive connection a finite number of times while calculating the parameter updates (cf. Hagan et al., 1996, p. 14-11 ff).

Equations 5.76 to 5.86 describe structure and estimation techniques for general ANNs. A brief overview of selected distance and activation functions specifically used throughout this thesis is given in the following paragraphs. Common choices for loss functions to be applied to each observation (cf. equation 5.82) are the squared (or quadratic) loss

$$\widetilde{\delta}_{\mathrm{Q}}\left(\widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}}\right) \coloneqq ||\boldsymbol{y}_{i\cdot}^{\mathsf{s}} - \widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}}||_2^2 \tag{5.87}$$

in case of continuous dependent variables. For categorical outcomes, the cross-entropy (or deviance)

$$\widetilde{\delta}_{\mathrm{D}}\left(\widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}}\right) \coloneqq -\boldsymbol{y}_{i\cdot}^{\mathsf{s}} \circ \log\left(\widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}}\right) \tag{5.88}$$

can be used, which is e.g. motivated by the estimated Kullback-Leibler distance (negative log-likelihood) used in sections 5.1.3 and 5.1.4. The derivatives of these loss functions are given by

$$\mathbf{J}_{\widetilde{\delta}_{\mathrm{Q}}}\left(\widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}}\right) = -2\cdot\left(\boldsymbol{y}_{i\cdot}^{\mathsf{s}} - \widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}}\right) \tag{5.89}$$

and

$$\mathbf{J}_{\widetilde{\delta}_{\mathrm{D}}}\left(\widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}}\right) = -\boldsymbol{y}_{i\cdot}^{\mathsf{s}} \oslash \widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}} \tag{5.90}$$

(cf. Bishop, 1995, pp. 89 ff, 237 ff; Hastie, Tibshirani and Friedman, 2008, pp. 395 ff).

Typical choices for activation functions $\boldsymbol{t}_{\mathrm{i}}$ that are relevant in the present context are the linear and the softmax function. To introduce these functions, preliminarily note that equations 5.85 and 5.86 are calculated separately for each element $k = 1, \ldots, n^{\mathsf{s}}$. Therefore, tedious tensor or array notation can once again be avoided by defining the transformations in a row-wise fashion, i.e. applied to a vector $\boldsymbol{v} \in \mathbb{R}^h$, such that their derivatives are again given in form of Jacobian matrices. Computationally, an array containing the n^{s} resulting Jacobians is used in backpropagation. By multiplication with the Jacobian of the distance function and summation over all observations as introduced in equation 5.83, this array is reduced to an aggregate Jacobian matrix (cf. Bishop, 1995, pp. 140 ff). Following these considerations, the linear activation function is defined by

$$\boldsymbol{t}^{(\mathrm{l})}(\boldsymbol{v}) \coloneqq \boldsymbol{v} \tag{5.91}$$

with corresponding Jacobian matrix

$$\mathbf{J}_{\boldsymbol{t}^{(\mathrm{l})}}(\boldsymbol{v}) = \mathbf{I}_h \quad . \tag{5.92}$$

The softmax function is defined by

$$\boldsymbol{t}^{(\mathrm{s})}(\boldsymbol{v}) \coloneqq \mathbf{softmax}(\boldsymbol{v}) = \exp(\boldsymbol{v}) \cdot \left|\left|\left(\exp(\boldsymbol{v})\right)^{\circ(-1)}\right|\right|_1 \quad . \tag{5.93}$$

Its Jacobian is

$$\mathbf{J}_{\boldsymbol{t}^{(\mathrm{s})}}(\boldsymbol{v}) = -\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a} + \mathbf{diag}\left(\boldsymbol{a}^{\circ 2} + \boldsymbol{a} \circ (\mathbf{1}_{h\times 1} - \boldsymbol{a})\right) \quad , \tag{5.94}$$

where $\boldsymbol{a} \coloneqq \mathbf{softmax}(\boldsymbol{v})$ (cf. Bishop, 1995, p. 64; Hagan et al., 1996, p. 24-6 ff). Note that if a bias (intercept) column is added to the transformation, a leading row of zeros is inserted in the Jacobian matrices because the constant is not determined by $\boldsymbol{v}$. These functions are just a small subset of the many activation functions that can be used for ANNs. Further examples include the hyperbolic tangent, radial basis or rectified linear unit activation functions (cf. e.g. Bishop, 1995, pp. 121 ff; Hastie, Tibshirani and Friedman, 2008, p. 392; Schmidt-Hieber et al., 2020).

An important advantage of ANNs is that many of the approaches presented before can be seen as special cases thereof, which are determined by the choice of distance and activation functions. In particular, this holds for the (generalized) linear and additive regression models, as long as the transformation functions are fixed and differentiable (cf. Hagan et al., 1996, p. 22-17 ff; Hastie, Tibshirani and Friedman, 2008, pp. 392 ff; Venables and Ripley, 2002, pp. 243 ff). Examples thereof are the linear regression (cf. section 5.1.2), which uses the squared loss and linear activation function, as well as GLMs and GAMs for binary data, which typically use special cases of the cross-entropy loss and softmax activation function. Furthermore, the multinomial regression is an ANN without any hidden layers, using the softmax activation and cross-entropy distance function. In case of categorical variables, $\boldsymbol{Y}$ represents one indicator (dummy) variable for each of the o possible values (cf. Hastie, Tibshirani and Friedman, 2008, pp. 389 ff; Venables and Ripley, 2002, pp. 243 ff). Further special cases of ANNs include certain types of the projection pursuit regression, but the latter usually uses fewer but more complex transformation functions (cf. Hastie, Tibshirani and Friedman, 2008, pp. 394 f).

As described above, the most common fitting method for neural networks is backpropagation (gradient descent), mainly because ANNs usually have a large number of free parameters. Backpropagation typically results in algorithms requiring more but cheaper iterations than methods using the actual Hessian matrix or more refined substitutes. The latter are applied for most of the special cases discussed above, which may lead to deviating results in parameter estimates. However, ANNs themselves can likewise be estimated e.g. by Newton or Quasi-Newton approaches, provided that all required derivatives exist (cf. Bishop, 1995, pp. 287 ff; Hagan et al., 1996, p. 9-10 ff; Ripley, 1996, pp. 158 ff).

Established activation functions already offer a tremendous flexibility. Nevertheless, a further convenient feature of artificial neural networks that are fit by backpropagation is that it is rather straightforward to introduce new transformation functions. Incorporating a new activation function in an ANN merely requires implementation of this function and its Jacobian, i.e. as in equations 5.91 to 5.94 (cf. e.g. Bishop, 1995, pp. 121 ff). This property is used in the following section 5.1.9 to further integrate the ideas of artificial neural networks and generalized additive models. In particular, a non-parametric component based on B-spline layers with optimized knot positioning is introduced.

5.1.9 Semi-parametric Artificial Neural Networks

The connection between semi-parametric GAMs on the one and ANNs on the other side is regularly emphasized and deepened (cf. e.g. Brath, Montanari and Toth, 2002; Breidt and Opsomer, 2009, p. 106; Insua and Müller, 1998; Ripley, 1996, pp. 182 ff; Schmidt-Hieber et al., 2020; Specht et al., 1991). As outlined in section 5.1.6, B-splines are the perhaps most common transformation used in the context of GAMs and have various further applications (cf. e.g. Böhm, Farin and Kahmann, 1984; Hastie and Tibshirani, 1986; Wood, 2017, pp. 142 ff). At the same time, artificial neural networks introduced in section 5.1.8 are very flexible and can be used to represent many important types of models. As a consequence, there are various approaches to use splines as transformations for artificial neural networks (cf. e.g. Folgheraiter, 2016; Guarnieri, Piazza and Uncini, 1999; Hong and Chen, 2011; Lin et al., 2006; Raya-Armenta, Lozano-Garcia and Avina-Cervantes, 2018; Wang and Lei, 2001; Zhang et al., 2017).

One problem in this regard is that spline base functions generally depend on the input variables as well as the vectors of knots. For the regression context, these knots are often considered to be prespecified and fix (cf. section 5.1.6; Hastie and Tibshirani, 1990, p. 247; Wood, 2017, pp. 122 ff), a limitation that is e.g. tackled by the MARS algorithm (cf. section 5.1.7; Friedman, 1991a,b).

When considering the application of spline transformations as general activation functions in artificial neural networks, both the fixed knots approach as well as the trial-and-error strategy of the MARS algorithm appear sensible only for transformations of the input variables $\widetilde{\boldsymbol{X}}^{(0)}$ but not of derived features $\widetilde{\boldsymbol{X}}^{(i)}$ $(i > 0)$. On the one hand, splines require knots that span the entire range of possible input values to yield adequate results. This is because the respective base functions become zero for values beyond the first or the last knot and, thus, perform poorly for extrapolation outside the interval between these knots (cf. equations 5.58 and 5.74; Hastie, Tibshirani and Friedman, 2008, p. 144; Piegl and Tiller, 1997, p. 58). Since the realized values of the derived features in an ANN may change in every iteration of the fitting procedure (cf. definition 5.76), an adequate spline transformation of these derived features is, hence, not possible when using predetermined

knots that are not adjusted to the changes in the range of $\widetilde{\boldsymbol{X}}^{(i)}$. On the other hand, the approach proposed for the MARS algorithm is capable of adjusting the knots to the changes in the derived features between iterations of the fitting procedure. However, it cannot be incorporated in backpropagation, which heavily relies on recursive applications of the chain rule to represent the interdependencies between parameters in an ANN. This would no longer be possible if the knots were, as in the MARS algorithm, updated independently from the remaining parameters since the gradient information used to update all other coefficients would not account for any changes in the knots (cf. sections 5.1.7 and 5.1.8). Backpropagation as the most prevalent fitting method for ANNs is hence not compatible with the MARS approach because the dependencies between knots and regression coefficients in the ANN cannot be represented correctly for updating the parameters in this case. Basically the same argument holds when considering the use of knots that are evenly spaced over the range or quantiles of the input variables because these knots would require adjustments for every update of the derived features as well.

As a consequence, current extensions and applications of splines in context of artificial neural networks are limited to transformations of the input variables alone. In some cases, the knots used for this purpose are optimized by heuristic approaches, such as simulated annealing and evolutionary algorithms, which again leads to trial-and-error components in conjunction with random permutations of the input (cf. e.g. Santos Coelho and Guerra, 2008; Van To and Kositviwat, 2005; Yiu et al., 2001). Such strategies rely on comparably costly iterative procedures for each update of the knots, which is why a set of prespecified knots is used in most cases (cf. e.g. Folgheraiter, 2016, p. 8; Hong and Chen, 2011, p. 820; Lin et al., 2006, p. 1447; Raya-Armenta, Lozano-Garcia and Avina-Cervantes, 2018, p. 2805; Wang and Lei, 2001, p. 6; Zhang et al., 2017, p. 12), or a fixed transformation is applied to approximate a spline's behavior (cf. Guarnieri, Piazza and Uncini, 1999).

However, neural networks can be used to perform knot selection for B-splines by means of gradient-based optimization. For example, Laube, Franz and Umlauf (2018) employ two separate ANNs to choose knots as well as regression coefficients for graphical curve and surface approximations via B-splines.

As a generalization of this approach, the following paragraphs introduce *semi-parametric artificial neural networks*. The basic idea is to incorporate general B-spline layers in an ANN, a strategy that allows for adaptive optimized B-spline transformations of input variables and derived features while maintaining the feasibility of backpropagation. By helping to overcome the limitations outlined above, this new methodological proposal extends and generalizes the use of non-parametric components in artificial neural networks within and beyond the regression context.

When including a B-spline layer as the i-th layer in an ANN, the output of this layer is defined by equation 5.63, i.e. as B-spline base functions of degree l applied to the input $\widetilde{\boldsymbol{X}}^{(j)}$ obtained from an arbitrary layer $j \neq i$:

$$\begin{aligned} \widetilde{\boldsymbol{X}}^{(i)} &:= \boldsymbol{t}_{\mathrm{i}}\left(\widetilde{\boldsymbol{X}}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{X}}^{(j)}}\right) \\ &= \left[\mathbf{B}_1^l\left(\widetilde{\boldsymbol{x}}_{\cdot 1}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot 1}^{(j)}}\right) \cdots \mathbf{B}_s^l\left(\widetilde{\boldsymbol{x}}_{\cdot c_j}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot c_j}^{(j)}}\right)\right] \quad . \end{aligned} \tag{5.95}$$

As before is $c_j = \text{ncol}\left(\widetilde{\boldsymbol{X}}^{(j)}\right)$ the number of columns in $\widetilde{\boldsymbol{X}}^{(j)}$, while $s = \left|\boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot c_j}^{(j)}}\right| + l - 1$ in this context denotes the number of B-spline base functions resulting for the last input column $\widetilde{\boldsymbol{x}}_{\cdot c_j}^{(j)}$ in $\widetilde{\boldsymbol{X}}^{(j)}$. The vectors of knots that are used to transform all c_j input columns $\widetilde{\boldsymbol{x}}_{\cdot 1}^{(j)} \dots \widetilde{\boldsymbol{x}}_{\cdot c_j}^{(j)}$ are specific for output layer i and, as in equation 5.62, concatenated in

$$\boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{X}}^{(j)}} := \left[\left(\boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot 1}^{(j)}}\right)^{\mathsf{T}} \quad \dots \quad \left(\boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot c_j}^{(j)}}\right)^{\mathsf{T}}\right]^{\mathsf{T}} \quad . \tag{5.96}$$

Therefore, $\boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{X}}^{(j)}}$ is considered as the vector of neural network parameters for B-spline layer i. Note that in order to simplify notation, $\widetilde{\boldsymbol{X}}^{(j)}$ is the sole input in this representation. This does not limit generality because transformations of different input layers are independent, such that multiple B-spline layers can be used to transform the output of multiple other layers. By using concatenation layers, the output from an arbitrary number of B-spline (or any other) layers can then be combined if necessary.[4]

For optimizing the B-spline knots when fitting an ANN, the same logic that is used for equations 5.85 and 5.86 can be applied. For this purpose, the backpropagation factors for spline layer i with respect to input layer j are defined as

$$\boldsymbol{d}_k^{(ij)} := \begin{cases} \dfrac{\partial\left(\boldsymbol{t}_{\text{i}}\left(\widetilde{\boldsymbol{x}}_{k\cdot}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{X}}^{(j)}}\right)\right)}{\partial\left(\widetilde{\boldsymbol{x}}_{k\cdot}^{(j)}\right)} & \text{if} \quad i = A \\ \left(\sum_{a=1}^{A} \boldsymbol{d}_k^{(ai)}\right) \dfrac{\partial\left(\boldsymbol{t}_{\text{i}}\left(\widetilde{\boldsymbol{x}}_{k\cdot}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{X}}^{(j)}}\right)\right)}{\partial\left(\widetilde{\boldsymbol{x}}_{k\cdot}^{(j)}\right)} & \text{if} \quad 0 < i < A \end{cases} \tag{5.97}$$

for all $k \in \mathcal{S}^{\mathsf{s}}$. Calculating these factors requires the block-diagonal matrix

$$\frac{\partial\left(\boldsymbol{t}_{\text{i}}\left(\widetilde{\boldsymbol{x}}_{k\cdot}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{X}}^{(j)}}\right)\right)}{\partial\left(\widetilde{\boldsymbol{x}}_{k\cdot}^{(j)}\right)} = \begin{bmatrix} \boldsymbol{D}^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D}^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{D}^{(c_j)} \end{bmatrix} \quad . \tag{5.98}$$

Its elements

$$\begin{aligned} \boldsymbol{D}^{(m)} &:= \frac{\partial\left(\boldsymbol{t}_{\text{i}}\left(\widetilde{x}_{km}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot m}^{(j)}}\right)\right)}{\partial\left(\widetilde{x}_{km}^{(j)}\right)} \\ &= \left[\frac{\partial\left(\mathbf{B}_1^l\left(\widetilde{x}_{km}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot m}^{(j)}}\right)\right)}{\partial\left(\widetilde{x}_{km}^{(j)}\right)} \cdots \frac{\partial\left(\mathbf{B}_h^l\left(\widetilde{x}_{km}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot m}^{(j)}}\right)\right)}{\partial\left(\widetilde{x}_{km}^{(j)}\right)}\right]^{\mathsf{T}} \end{aligned} \tag{5.99}$$

for all $m = 1, \dots, c_j$ can be determined using equation 5.67. In this context denoted by $h = \left|\boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot m}^{(j)}}\right| + l - 1$ is the number of resulting base functions of degree l for input column $\widetilde{\boldsymbol{x}}_{\cdot m}^{(j)}$.

[4] Concatenation layers simply have an identity link and fixed identity matrices as coefficients.

The corresponding derivatives of the predictions with respect to the vector of knots are given by

$$\frac{\partial(\widehat{\boldsymbol{y}}_{i\cdot})}{\partial\left(\boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot m}^{(j)}}\right)} = \begin{cases} \dfrac{\partial\left(\boldsymbol{t}_{\mathrm{i}}\left(\widetilde{x}_{km}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot m}^{(j)}}\right)\right)}{\partial\left(\boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot m}^{(j)}}\right)} & \text{if} \quad i = A \\ \left(\sum_{a=1}^{A} \boldsymbol{d}_i^{(ai)}\right) \dfrac{\partial\left(\boldsymbol{t}_{\mathrm{i}}\left(\widetilde{x}_{km}^{(j)}, \boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot m}^{(j)}}\right)\right)}{\partial\left(\boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{x}}_{\cdot m}^{(j)}}\right)} & \text{if} \quad 0 < i < A \end{cases}, \tag{5.100}$$

which can be computed using equation 5.69.

The above considerations allow including general B-spline layers with optimized knot positioning in the structure and fitting procedure of artificial neural networks. The backward pass is defined by using coefficients $\boldsymbol{K}_{(i)}^{\widetilde{\boldsymbol{X}}^{(j)}}$ instead of $\boldsymbol{\beta}^{(ij)}$ in equations 5.80 and 5.86, which requires application of equalities 5.97 and 5.100 in backpropagation. The forward pass is determined by equations 5.76 and 5.95. Since current pre-existing software packages for ANNs do not consider such spline layers, a custom-made `C++` implementation of semi-parametric ANNs is developed in the context of this thesis. An outline is given in section 6.1.2, more details can be found in appendix C.2. Among others, one purpose of this implementation is to evaluate such ANNs in the Monte Carlo simulations and the application study in chapters 6 and 7.

An alternative machine learning approach that is proposed for prediction from non-probability samples and sometimes compared to ANNs is constituted by support vector machines (cf. e.g. Buelens, Burger and van den Brakel, 2018, p. 323; Xu et al., 2013, p. 33). These models are likewise based on non-linear transformations of the independent variables, but rely on a single transformation. In that way, they can efficiently apply high-dimensional transformations by using a specialized optimization approach. Support vector machines are introduced in the following section 5.1.10.

5.1.10 Support Vector Machines

To fit a non-linear prediction model for $\boldsymbol{Y}$, the basic idea of support vector machines (SVMs) is to apply linear methods in a transformed space of the input variables. Therefore, they are non-linear in the original space of $\boldsymbol{X}^{\mathtt{s}}$ and can be seen as ANNs with a single hidden layer (cf. Boser, Guyon and Vapnik, 1992, p. 144). The main difference between SVMs and the models discussed so far lies in the fitting method: by focusing on influential observations, optimization for SVMs is considered advantageous in case of certain high-dimensional transformations (cf. Hastie, Tibshirani and Friedman, 2008, p. 431). Since support vector machines can be used for classification (cf. Boser, Guyon and Vapnik, 1992) as well as continuous regression tasks (cf. Drucker et al., 1997), both cases are introduced in the following overview.

To start with classification tasks, the *support vector classifier* considers a single binary variable of interest, $\boldsymbol{y}_{\cdot l}^{\mathtt{t}} \in \{-1; 1\}^{n^{\mathtt{t}}}$. Similar as in the previous sections, a matrix of transformed predictors $\widetilde{\boldsymbol{X}}^{\mathtt{t}} \coloneqq \boldsymbol{t}\left(\boldsymbol{X}^{\mathtt{t}}\right) \in \mathbb{R}^{n^{\mathtt{t}} \times h}$ is constructed by means of a transformation $\boldsymbol{t} : \mathbb{R}^{n^{\mathtt{t}} \times p} \to \mathbb{R}^{n^{\mathtt{t}} \times h}$, where $\widetilde{\boldsymbol{x}}_{\cdot 1}^{\mathtt{t}} \coloneqq \mathbf{1}_{n^{\mathtt{t}} \times 1}$ constitutes an intercept column. To obtain

predictions based on a vector of parameters $\boldsymbol{\beta} \in \mathbb{R}^h$, a decision boundary

$$\mathrm{F}\left(\boldsymbol{\beta}\right) \coloneqq \beta_1 + \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{t}} \cdot \beta_j = \widetilde{\boldsymbol{X}}^{\mathsf{t}} \boldsymbol{\beta} \tag{5.101}$$

is constructed. For $\mathrm{F}\left(\boldsymbol{\beta}\right) = 0$, this boundary constitutes a hyperplane which is intended to separate the two classes in an optimal way. Correspondingly, the predictions are defined by

$$\widehat{\boldsymbol{y}}_{\cdot l}^{\mathsf{t}} \coloneqq \operatorname{sign}\left(\widetilde{\boldsymbol{X}}^{\mathsf{t}} \boldsymbol{\beta}\right) \quad , \tag{5.102}$$

where the sign-function $\operatorname{sign} : \mathbb{R}^{n^{\mathsf{t}}} \to \{-1; 1\}^{n^{\mathsf{t}}}$ is applied element-wise (cf. Boser, Guyon and Vapnik, 1992, p. 145).

The main difference between SVMs and the methods discussed before lies in the formulation of the optimization problem. The underlying idea for classification is to determine the coefficients $\boldsymbol{\beta} \in \mathbb{R}^h$ from data set s, such that the minimal distance $e \in \mathbb{R}_{\geq 0}$ between the hyperplane and observations of each group is maximized. For cases when perfect separation is not feasible, a slack variable $\boldsymbol{\xi} \in \mathbb{R}_{\geq 0}^{n^{\mathsf{s}}}$ similar to those used in section 4.2.2.2 is required to quantify the violation of this ideal classification. A penalty for such violations is added to the distance function by multiplying $\boldsymbol{\xi}$ with a given vector of individual penalty (or cost) parameters $\boldsymbol{c} \in \mathbb{R}_{\geq 0}^{n^{\mathsf{s}}}$. Note that $\boldsymbol{c}$, for example, can be a vector of survey weights. The preliminary optimization problem is then given by

$$\begin{aligned} \underset{(\boldsymbol{\beta}, \boldsymbol{\xi})}{\operatorname{argmin}} & \left(\frac{1}{2} \cdot \sum_{j=2}^{h} \beta_j^2 + \boldsymbol{c}^{\mathsf{T}} \boldsymbol{\xi} \right) \\ \text{s. t.} \quad & \boldsymbol{y}_{\cdot l}^{\mathsf{s}} \circ \widehat{\boldsymbol{y}}_{\cdot l}^{\mathsf{s}} + \boldsymbol{\xi} \geq \mathbf{1}_{n^{\mathsf{s}} \times 1} \\ & \boldsymbol{\xi} \geq \mathbf{0}_{n^{\mathsf{s}} \times 1} \quad , \end{aligned} \tag{5.103}$$

where inequalities are applied component-wise. Note that the minimal margin e does not occur in this problem since it is fixed by

$$e \coloneqq \left(\sqrt{\sum_{j=2}^{h} \beta_j^2} \right)^{-1} \quad . \tag{5.104}$$

This can be arbitrarily defined, as the hyperplane is orthogonal to $\left[\beta_2, \ldots, \beta_h\right]^{\mathsf{T}}$. Therefore, any scalar multiple of this vector constitutes the same hyperplane for an adjusted value of the intercept β_1. Definition 5.104 merely selects one of an infinite number of possible solutions for the same hyperplane. The *support vectors* in problem 5.103 are given by the values $\left[\boldsymbol{x}_{i\cdot}^{\mathsf{s}} \quad y_{il}^{\mathsf{s}}\right]$ for all $i \in \{j : \xi_j > 0\}$. These contain the observed values that are relevant for obtaining the final parameters, all other observations (for which $\xi_j = 0$) are disregarded in that respect (cf. Boser, Guyon and Vapnik, 1992, p. 146; Hastie, Tibshirani and Friedman, 2008, pp. 129 ff, 418 ff).

Problem 5.103 can be solved by means of the methods discussed in chapter 4. However, the reason that SVMs gain advantages when the number of transformations $h = \operatorname{ncol}\left(\widetilde{\boldsymbol{X}}^{\mathsf{s}}\right)$ is large lies in the fact that the problem can be reformulated in dual form (cf. appendix B.4.5.1; Boser, Guyon and Vapnik, 1992; Cortes and Vapnik, 1995; Geiger and Kanzow, 2002, pp. 314 ff). In that case, optimization can be done solely for the Lagrange multipliers

$\boldsymbol{\alpha} \in \mathbb{R}^{n^{\mathsf{s}}}$ that are used to enforce the margin (first inequality) constraint in problem 5.103. Defining the optimization parameters as $\boldsymbol{\Theta} \coloneqq \boldsymbol{\alpha}$, the dual optimization problem is then defined by

$$\begin{aligned} \boldsymbol{\Theta}^* \; &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left(\frac{1}{2} \boldsymbol{\Theta}^{\mathsf{T}} \boldsymbol{Q} \boldsymbol{\Theta} - \mathbf{1}_{1 \times n^{\mathsf{s}}} \boldsymbol{\Theta} \right) \\ &\text{s. t.} \qquad \left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}\right)^{\mathsf{T}} \boldsymbol{\Theta} \overset{!}{=} 0 \\ &\qquad\qquad \boldsymbol{\Theta} \geq \mathbf{0}_{n^{\mathsf{s}} \times 1} \\ &\qquad\qquad \boldsymbol{\Theta} \leq \boldsymbol{c} \qquad , \end{aligned} \tag{5.105}$$

where inequalities are again applied element-wise. The quadratic multiplier is defined by

$$\boldsymbol{Q} \;\coloneqq\; \left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l} \left(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}\right)^{\mathsf{T}}\right) \circ \left(\widetilde{\boldsymbol{X}}^{\mathsf{s}}_{\cdot \mathcal{J}} \left(\widetilde{\boldsymbol{X}}^{\mathsf{s}}_{\cdot \mathcal{J}}\right)^{\mathsf{T}}\right) \quad \in \quad \mathbb{R}^{n^{\mathsf{s}} \times n^{\mathsf{s}}} \quad , \tag{5.106}$$

for which $\widetilde{\boldsymbol{X}}^{\mathsf{s}}_{\cdot \mathcal{J}}$ using $\mathcal{J} \coloneqq \{2, \ldots, h\}$ denotes all variables in $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$ except the intercept column. This approach can reduce the number of optimization parameters to the number of observations in case of a large number of transformations $h > n^{\mathsf{s}}$.

The convex quadratic problem 5.105 has exactly the same form as problem 4.13 because $\boldsymbol{\Theta} \geq \mathbf{0}_{n^{\mathsf{s}} \times 1}$ can be equivalently written as $-\boldsymbol{\Theta} \leq \mathbf{0}_{n^{\mathsf{s}} \times 1}$. Problem 5.105 may therefore be solved using the methods discussed in section 4.2.2.1 (cf. Hastie, Tibshirani and Friedman, 2008, pp. 420 f). However, $\boldsymbol{Q}$ grows quadratically with increasing sample size, and the constraints each concern only a single observation. Therefore, it is more common to use decomposition methods in this context. These iteratively select a subset of observations as working set, and solve problem 5.105 for this subset. Different strategies are available for this purpose, an overview is given by Chang and Lin (2011) as well as Fan, Chen and Lin (2005). The support vectors in the dual problem are defined by $\left[\boldsymbol{x}^{\mathsf{s}}_{i\cdot} \quad y^{\mathsf{s}}_{il}\right]$ for all $i \in \{j : \alpha_j > 0\}$.

Once the Lagrange multipliers are found, regression parameters $\boldsymbol{\beta}$ are determined to define the separating hyperplane and predictions. Using the KKT-conditions (cf. definition 4.12; Karush, 1939, quoted in Kjeldsen, 2000; Kuhn and Tucker, 1951), $\boldsymbol{\beta}$ is found as

$$\beta_j \;=\; \left(\boldsymbol{\alpha} \circ \boldsymbol{y}^{\mathsf{s}}_{\cdot l}\right)^{\mathsf{T}} \widetilde{\boldsymbol{x}}^{\mathsf{s}}_{\cdot j} \quad \text{for all} \quad j = 2, \ldots, h \tag{5.107a}$$

and

$$\beta_1 \;=\; \mathrm{E}\left(y^{\mathsf{s}}_{il} - \sum_{j=2}^{h} x^{\mathsf{s}}_{ij} \cdot \beta_j \middle| 0 < \alpha_i < c_i \right) \qquad . \tag{5.107b}$$

If the condition in equation 5.107b is not met by any α_i, the intercept is determined as

$$\begin{aligned} \beta_1 \;\approx\; \frac{1}{2} \cdot \Bigg(&\operatorname{Max}\left(y^{\mathsf{s}}_{il} - \sum_{j=2}^{h} \widetilde{x}^{\mathsf{s}}_{ij} \cdot \beta_j \middle| \alpha_i = c_i, y^{\mathsf{s}}_{il} = -1 \right) \\ &+ \operatorname{Min}\left(y^{\mathsf{s}}_{il} - \sum_{j=2}^{h} \widetilde{x}^{\mathsf{s}}_{ij} \cdot \beta_j \middle| \alpha_i = c_i, y^{\mathsf{s}}_{il} = 1 \right) \Bigg) \end{aligned} \tag{5.107c}$$

(cf. appendix B.4.5.1; Chang and Lin, 2011, p. 10; Smola and Schölkopf, 2004, p. 201). Extensions of the support vector classifier to more than two classes are typically based on solving multiple binary classification problems (cf. e.g. Friedman, 1996; Hastie and Tibshirani, 1998).

To apply similar ideas for regression of continuous dependent variables $\boldsymbol{y}^{\mathsf{t}}_{\cdot k} \in \mathbb{R}^{n^{\mathsf{t}}}$, *support vector regression* can be used. Just like in the least-squares model for transformations $\widetilde{\boldsymbol{X}}^{\mathsf{t}}$ (cf. sections 5.1.2 and 5.1.6), the predictions are defined by

$$\widehat{\boldsymbol{y}}^{\mathsf{t}}_{\cdot i} \coloneqq \widetilde{\boldsymbol{X}}^{\mathsf{t}}\boldsymbol{\beta} \quad . \tag{5.108}$$

The parameters $\boldsymbol{\beta}$ are determined from data set s similarly as for the support vector classifier. Following a closely related reasoning, a regression line (hyperplane) that is linear in the transformed but non-linear in the original space of $\boldsymbol{X}^{\mathsf{s}}$ is sought. As the prediction error here is constituted by the absolute distance of observations to this regression surface, the *maximal* permissible distance of points to the regression line is constrained by $e \in \mathbb{R}_{\geq 0}$. Thus, support vector regression reduces the number of observations needed to compute the parameters, by taking into account only residuals larger than e. Again, slack-variables $\boldsymbol{\xi}, \boldsymbol{\xi}^* \in \mathbb{R}^{n^{\mathsf{s}}}_{\geq 0}$ are used to quantify the violation of the given boundary e. As for the support vector classifier, $\boldsymbol{c} \in \mathbb{R}^{n^{\mathsf{s}}}_{\geq 0}$ is a vector of observation-specific cost parameters attributed to these slack variables, which can be survey weights. The primal and preliminary optimization problem in this setting is given by

$$\begin{aligned} &\underset{(\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\xi}^*)}{\operatorname{argmin}} \left(\frac{1}{2} \cdot \sum_{j=2}^{h} \beta_j^2 + \boldsymbol{c}^{\mathsf{T}} \left(\boldsymbol{\xi} + \boldsymbol{\xi}^* \right) \right) \\ &\text{s. t.} \qquad \boldsymbol{y}^{\mathsf{s}}_{\cdot l} - \widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot l} - \boldsymbol{\xi} \leq e \\ &\qquad\qquad \widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot l} - \boldsymbol{y}^{\mathsf{s}}_{\cdot l} - \boldsymbol{\xi}^* \leq e \\ &\qquad\qquad\qquad \boldsymbol{\xi},\ \boldsymbol{\xi}^* \geq \boldsymbol{0}_{n^{\mathsf{s}} \times 1} \quad . \end{aligned} \tag{5.109}$$

Similar as before, observed values for elements with absolute prediction errors above the boundary e constitute the support vectors $\begin{bmatrix} \boldsymbol{x}^{\mathsf{s}}_{i\cdot} & y^{\mathsf{s}}_{il} \end{bmatrix}$ for all $i \in \left\{ (j : \xi_j > 0) \vee \left(\xi_j^* > 0 \right) \right\}$ that are relevant for determining the final parameters, and inequalities are applied component-wise (cf. Hastie, Tibshirani and Friedman, 2008, p. 436).

Just like for the support vector classifier, problem 5.109 can be solved by using the methods discussed in chapter 4 but is again more conveniently reformulated in dual form to reduce the problem's dimension if the number of columns in $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$ is large. For Lagrange parameters $\boldsymbol{\alpha}, \boldsymbol{\alpha}^* \in \mathbb{R}^{n^{\mathsf{s}}}$ corresponding to the margin (first two inequality) constraints in problem 5.109, the optimization parameters are given by $\boldsymbol{\Theta} \coloneqq \begin{bmatrix} \boldsymbol{\alpha}^{\mathsf{T}} & (\boldsymbol{\alpha}^*)^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$. The resulting optimization problem is then

$$\begin{aligned} \boldsymbol{\Theta}^* \ = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} &\left(\frac{1}{2} \cdot \boldsymbol{\Theta}^{\mathsf{T}} \begin{bmatrix} \boldsymbol{Q} & -\boldsymbol{Q} \\ -\boldsymbol{Q} & \boldsymbol{Q} \end{bmatrix} \boldsymbol{\Theta} - \boldsymbol{\Theta}^{\mathsf{T}} \begin{bmatrix} e \cdot \boldsymbol{1}_{n^{\mathsf{s}} \times 1} + \boldsymbol{y}^{\mathsf{s}}_{\cdot l} \\ e \cdot \boldsymbol{1}_{n^{\mathsf{s}} \times 1} - \boldsymbol{y}^{\mathsf{s}}_{\cdot l} \end{bmatrix} \right) \\ \text{s. t.} \qquad &\begin{bmatrix} \boldsymbol{1}_{1 \times n^{\mathsf{s}}} & -\boldsymbol{1}_{1 \times n^{\mathsf{s}}} \end{bmatrix} \boldsymbol{\Theta} \overset{!}{=} 0 \\ &\boldsymbol{\Theta} \geq \ \boldsymbol{0}_{(2 \cdot n^{\mathsf{s}}) \times 1} \\ &\boldsymbol{\Theta} \leq \ \begin{bmatrix} \boldsymbol{c}^{\mathsf{T}} & \boldsymbol{c}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \quad . \end{aligned} \tag{5.110}$$

The quadratic multiplier in this case is constituted by elements

$$\boldsymbol{Q} \coloneqq \widetilde{\boldsymbol{X}}^{\mathsf{s}}_{\cdot \mathcal{J}} \left(\widetilde{\boldsymbol{X}}^{\mathsf{s}}_{\cdot \mathcal{J}} \right)^{\mathsf{T}} \quad \in \quad \mathbb{R}^{n^{\mathsf{s}} \times n^{\mathsf{s}}} \quad , \tag{5.111}$$

again denoting by $\widetilde{\boldsymbol{X}}^{\mathsf{s}}_{\cdot \mathcal{J}}$ for $\mathcal{J} \coloneqq \{2, \ldots, h\}$ all transformed variables in $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$ except the intercept column. The equality constraint in problem 5.110 simply requires that $||\alpha_i - \alpha_i^*||_1 = 0$,

and the inequalities constitute box-constraints for the Lagrange multipliers, such that $0 \leq \alpha_i, \alpha_i^* \leq c_i$ holds for all $i \in \mathcal{S}^{\mathsf{s}}$ (cf. appendix B.4.5.2; Chang and Lin, 2011, p. 8). Again, the convex quadratic problem 5.110 is exactly in the form of problem 4.13 and may readily be solved using the methods discussed in section 4.2.2.1 (cf. Hastie, Tibshirani and Friedman, 2008, pp. 420 f). As for the support vector classifier, however, it is commonly decomposed into smaller problems of the same form to improve computability for large samples (cf. Chang and Lin, 2011; Fan, Chen and Lin, 2005).

Once the Lagrange multipliers are found, regression parameters can be calculated from the support vectors by using the KKT-conditions (cf. definition 4.12; Karush, 1939, quoted in Kjeldsen, 2000; Kuhn and Tucker, 1951). The parameters are determined by

$$\beta_j \;=\; (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^{\mathsf{T}}\, \widetilde{\boldsymbol{x}}^{\mathsf{s}}_{\cdot j} \quad \text{for all} \quad j = 2, \ldots, h \tag{5.112a}$$

and

$$\begin{aligned} \beta_1 \;= \frac{1}{2} \cdot \Bigg(& \mathrm{E}\left(y^{\mathsf{s}}_{il} + e - \sum_{j=2}^{h} \widetilde{x}^{\mathsf{s}}_{ij} \cdot \beta_j \middle| 0 < \alpha_i^* < c_i \right) \\ & + \mathrm{E}\left(y^{\mathsf{s}}_{il} - e - \sum_{j=2}^{h} \widetilde{x}^{\mathsf{s}}_{ij} \cdot \beta_j \middle| 0 < \alpha_i \; < c_i \right) \Bigg) \quad . \end{aligned} \tag{5.112b}$$

If the conditions in equation 5.112b are not met by any α_i or α_i^*, the intercept is determined as

$$\begin{aligned} \beta_1 \;\approx \frac{1}{2} \cdot \Bigg(& \mathrm{Max}\left(y^{\mathsf{s}}_{il} + e - \sum_{j=2}^{h} \widetilde{x}^{\mathsf{s}}_{ij} \cdot \beta_j \middle| \alpha_i^* = c_i \right) \\ & + \mathrm{Min}\left(y^{\mathsf{s}}_{il} - e - \sum_{j=2}^{h} \widetilde{x}^{\mathsf{s}}_{ij} \cdot \beta_j \middle| \alpha_i = c_i \right) \Bigg) \end{aligned} \tag{5.112c}$$

(cf. appendix B.4.5.2; Chang and Lin, 2011, p. 10; Smola and Schölkopf, 2004, p. 201).

As outlined in the previous discussion, the quadratic multipliers $\boldsymbol{Q}$ in the dual problems 5.105 and 5.110 are determined by using the matrix product $\widetilde{\boldsymbol{X}}^{\mathsf{s}}_{\cdot\mathcal{J}} \left(\widetilde{\boldsymbol{X}}^{\mathsf{s}}_{\cdot\mathcal{J}} \right)^{\mathsf{T}}$ that includes all columns in $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$ except the intercept column (cf. equations 5.106 and 5.111). At the same time, the number of optimization parameters is not influenced by the number of columns h in $\widetilde{\boldsymbol{X}}^{\mathsf{s}}$. Therefore, solving the dual optimization problems for SVMs solely requires the above matrix product, but not the actual transformation $\boldsymbol{t}$. In fact, only the *kernel function*

$$\mathrm{K}\left(\boldsymbol{x}^{\mathsf{s}}_{i\cdot}, \boldsymbol{x}^{\mathsf{s}}_{j\cdot} \right) \;\coloneqq\; \left(\boldsymbol{t}\left(\boldsymbol{x}^{\mathsf{s}}_{j\cdot} \right) \right)^{\mathsf{T}} \boldsymbol{t}\left(\boldsymbol{x}^{\mathsf{s}}_{i\cdot} \right) = \left(\widetilde{\boldsymbol{x}}^{\mathsf{s}}_{i\cdot} \right)^{\mathsf{T}} \widetilde{\boldsymbol{x}}^{\mathsf{s}}_{j\cdot} \tag{5.113}$$

that returns the inner product of two transformed observations needs to be specified and evaluated for SVMs (cf. Boser, Guyon and Vapnik, 1992, pp. 147 f; Chang and Lin, 2011, p. 3; Cortes and Vapnik, 1995, p. 283). As a result, SVMs provide an approach for using transformations to very high and even infinite dimensional spaces, as long as these are represented by such a kernel function K that can be evaluated (cf. Cortes and Vapnik, 1995, p. 276; Hastie, Tibshirani and Friedman, 2008, p. 423). There are various choices for K, such as radial (Gaussian), polynomial and hyperbolic tangent functions, for which an overview with discussion is given by Smola and Schölkopf (2004, pp. 201 ff).

Support vector machines strongly depend on the penalization of parameter volatility in their loss function, which is expressed by the sum of squared slope parameters in equations 5.103 and 5.109 (cf. Hastie, Tibshirani and Friedman, 2008, p. 424). Similar as for smoothing and P-splines (cf. section 5.1.6), this penalty constitutes an essential component of SVMs. However, such penalties can also be applied for any of the other models discussed throughout the current section 5.1. When used for this purpose, penalization is also referred to as 'shrinkage' (cf. e.g. Berk, 2008, pp. 70 ff; Hastie, Tibshirani and Friedman, 2008, pp. 61 ff), for which an overview is given in the following section 5.1.11.

5.1.11 Shrinkage Methods

Various prediction models are described in the previous sections 5.1.1 to 5.1.10. They all rely on different assumptions about the relationships (and partially the distributions) of dependent and independent variables. Yet, model fitting is mostly done in quite similar ways, i.e. through minimization of a loss function by means of the gradient based optimization methods described in section 4.2.

Shrinkage or regularization methods constitute an extension to this approach. As such, they can be viewed as an adaptation of the fitting techniques presented so far rather than defining new types of models. These methods are particularly useful in cases of ill-posed problems, for which a (unique) solution to the optimization may not even exist without shrinkage. Examples can be found in cases where the number of (potential) predictors b is high, where 'high' is usually expressed in relation to the number of observations n^{s} in data set s to which the model is fit. The rationale behind this idea is that if the ratio b/n^{s} approaches one, the variability of parameters $\boldsymbol{\Theta}$ and predictions $\widehat{\boldsymbol{Y}}$ usually increases heavily. In cases where this ratio exceeds one, a unique solution to the optimization problem does often not even exist (cf. e.g. Breidt and Opsomer, 2017, p. 202; Hastie, Tibshirani and Friedman, 2008, pp. 61 ff; Friedman, Hastie and Tibshirani, 2010, p. 3).

Often, it is sensible to make a qualified choice for including certain potential explanatory variables and excluding others, e.g. on grounds of theoretical justification and/or cross-validation. An overview is provided by Hastie, Tibshirani and Friedman (2008, pp. 219 ff) and James et al. (2013, pp. 202 ff). For cases where such a choice is not feasible, shrinkage methods address this difficulty by incorporating a measure of variability for the optimization parameters $\boldsymbol{\Theta} \in \mathbb{R}^h$ directly in the model fitting procedure. The variability is quantified by a non-negative penalty function $\boldsymbol{p} : \mathbb{R}^h \to \mathbb{R}_{\geq 0}$ and limited by including an (additional) constraint of the form

$$\boldsymbol{p}\left(\boldsymbol{\Theta}\right) \leq b \tag{5.114}$$

for some prespecified constant $b \geq 0$. The KKT-conditions 4.12 state that an optimal solution $\begin{bmatrix}(\boldsymbol{\Theta}^*)^{\mathsf{T}} & \alpha\end{bmatrix}^{\mathsf{T}}$ fulfills

$$\begin{aligned} \mathbf{J}_{\delta}\left(\boldsymbol{\Theta}^*\right) + \alpha \cdot \mathbf{J}_{\boldsymbol{p}}\left(\boldsymbol{\Theta}^*\right) &= \mathbf{0} \\ \boldsymbol{p}\left(\boldsymbol{\Theta}^*\right) - b &\leq \mathbf{0} \\ \alpha &\geq 0 \\ \alpha \cdot \left(\boldsymbol{p}\left(\boldsymbol{\Theta}^*\right) - b\right) &= \mathbf{0} \qquad , \end{aligned} \tag{5.115}$$

where $\delta : \mathbb{R}^h \to \mathbb{R}$ is the model's original loss function and $\alpha \in \mathbb{R}$ is a Lagrange multiplier for constraint 5.114.

In principle, SQP-methods as presented in section 4.2.2 can be used to find a solution for this problem. However, it is often considered to be computationally advantageous to perform unconstrained optimization and use a penalized version of δ: by choosing $b \coloneqq \boldsymbol{p}(\boldsymbol{\Theta}^*)$, it is easy to see that

$$\boldsymbol{\Theta}^* \;=\; \boldsymbol{\Theta}(\lambda) \coloneqq \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left(\delta(\boldsymbol{\Theta}) + \lambda \cdot \boldsymbol{p}(\boldsymbol{\Theta})\right) \tag{5.116}$$

fulfills conditions 5.115 for $\lambda = \alpha$ (cf. Kloft et al., 2011, pp. 990 f). To denote their dependency on the *penalty or shrinkage parameter* $\lambda \in \mathbb{R}_{\geq 0}$, the optimal parameters are expressed as a function $\boldsymbol{\Theta} : \mathbb{R}_{\geq 0} \to \mathbb{R}^h$ of λ in this specification. Since b is a monotonically decreasing function of λ, it is usually computationally simpler to minimize the penalized version of the original loss function δ defined in equation 5.116 rather than solving an optimization problem under constraint 5.114 (cf. Fan and Li, 2006; Hastie, Tibshirani and Friedman, 2008, pp. 61 ff; Hoerl and Kennard, 1970; Tibshirani, 1996; 1997). Strategies for choosing λ in this penalization context are described below. As long as $\boldsymbol{p}$ is differentiable once or twice (as required by the applied optimization algorithm), penalization simply results in adding λ times the respective Jacobian $\mathbf{J}_{\boldsymbol{p}}(\boldsymbol{\Theta})$ and, where required, Hessian $\mathbf{H}_{\boldsymbol{p}}(\boldsymbol{\Theta})$ to the ones derived for the unpenalized model. By means of this extension, optimization is guided towards lower variation in parameters (and hence predictions) as expressed by $\boldsymbol{p}$. The downside is that penalty terms can induce bias to the parameter estimates and predictions, a problem which is commonly referred to as the *bias-variance trade-off* (cf. e.g. Berk, 2008, pp. 70 ff; Hoerl and Kennard, 1970, pp. 60 f; James et al., 2013, pp. 217 ff).

Regularization is, for example, used in support vector machines to achieve a unique solution and for smoothing and P-splines to reduce volatility of the regression function (cf. sections 5.1.6 and 5.1.10). In the general regression context, $\boldsymbol{p}$ is commonly chosen to be a vector norm, for which Fan and Li (2006) show that it includes many model selection criteria used throughout different scientific disciplines as special cases. The two most prominent shrinkage methods in this field are ridge regression proposed by Hoerl and Kennard (1970) and LASSO ('least absolute shrinkage and selection operator') introduced by Tibshirani (1996; cf. e.g. Berk, 2008, pp. 70 ff; James et al., 2013, pp. 214 ff; Ruppert, Wand and Carroll, 2003, pp. 65 ff). Note that intercept parameters are typically not penalized to prevent bias in the overall mean of the predictions. Therefore, a vector $\boldsymbol{c} = \begin{bmatrix} c_1 & \dots & c_h \end{bmatrix}^\mathsf{T} \in \{0;1\}^h$ is used to exclude these elements of $\boldsymbol{\Theta}$ in the following discussion and simplify notation, i.e. it is one only for non-intercept parameters and zero else: $c_i \coloneqq \mathbb{I}(\Theta_i \text{ is not an intercept parameter})$.

In ridge regression, $\boldsymbol{p}$ is chosen to be the squared 2-norm, i.e.

$$\boldsymbol{p}_{\mathrm{r}}(\boldsymbol{\Theta}) \;\coloneqq \boldsymbol{c}^\mathsf{T}\boldsymbol{\Theta}^{\circ 2} = ||\boldsymbol{\Theta}_{\mathcal{I}}||_2^2 \qquad , \tag{5.117}$$

where $\mathcal{I} \coloneqq \{i : c_i = 1\}$ selects the non-intercept elements of $\boldsymbol{\Theta}$. In this case, the Jacobian and Hessian matrix are given by

$$\mathbf{J}_{\boldsymbol{p}_{\mathrm{r}}}(\boldsymbol{\Theta}) \;= 2 \cdot (\boldsymbol{c} \circ \boldsymbol{\Theta})^\mathsf{T} \tag{5.118}$$

and

$$\mathbf{H}_{\boldsymbol{p}_{\mathrm{r}}}(\boldsymbol{\Theta}) \;= 2 \cdot \mathbf{diag}(\boldsymbol{c}) \qquad . \tag{5.119}$$

In contrast, the LASSO uses the 1-norm, such that

$$\boldsymbol{p}_{\text{l}}\left(\boldsymbol{\Theta}\right) \coloneqq \boldsymbol{c}^{\mathsf{T}} \operatorname{Abs}\left(\boldsymbol{\Theta}\right) = \left|\left|\boldsymbol{\Theta}_{\mathcal{I}}\right|\right|_1 \quad , \tag{5.120}$$

with $\mathcal{I}$ as before. The first and second derivatives respectively are

$$\begin{aligned} \frac{\partial\left(\boldsymbol{p}_{\text{l}}\left(\boldsymbol{\Theta}\right)\right)}{\partial\left(\Theta_j\right)} &= c_j \circ \operatorname{sign}\left(\Theta_j\right) \quad \text{for all} \quad \Theta_j \neq 0 \\ \frac{\partial^2\left(\boldsymbol{p}_{\text{l}}\left(\boldsymbol{\Theta}\right)\right)}{\partial\left(\Theta_j\right)\partial\left(\Theta_j\right)} &= 0 \quad\quad\quad\quad \text{for all} \quad \Theta_j \neq 0 \quad , \end{aligned} \tag{5.121}$$

but the Jacobian and Hessian matrix of $\boldsymbol{p}_{\text{l}}$ do not exist because the absolute value function is not differentiable at zero. Therefore, it is difficult to include the LASSO shrinkage into a general unconstrained optimization scenario in form of equation 5.116. For the context of (generalized) linear regression models, however, a number of specialized and efficient solutions are available, such as the least angle regression (cf. Efron et al., 2004) and the coordinate descent algorithm (cf. Friedman et al., 2007; Friedman, Hastie and Tibshirani, 2010).

Even though the outlined non-differentiability prevents using a LASSO penalty in general unconstrained optimization problems, this is still possible in the *constrained* optimization procedures discussed in section 4.2.2. Despite the fact that Jacobian and Hessian matrix of penalty 5.120 do not exist when optimizing w.r.t. $\boldsymbol{\Theta}$, the penalty can be expressed by decomposing $\boldsymbol{\Theta}$ into its positive and negative parts, which are denoted by $\boldsymbol{\xi}^+, \boldsymbol{\xi}^- \in \mathbb{R}^h_{\geq 0}$ and included in the optimization problem. Using element-wise inequalities, the constraints

$$\begin{aligned} \boldsymbol{\xi}^+ - \boldsymbol{\xi}^- &\overset{!}{=} \boldsymbol{\Theta} \\ \boldsymbol{\xi}^+, \boldsymbol{\xi}^- &\geq \mathbf{0}_{(2 \cdot h) \times 1} \end{aligned} \tag{5.122}$$

allow reformulating equation 5.120 as

$$\boldsymbol{p}_{\text{l}}\left(\boldsymbol{\Theta}\right) = \boldsymbol{p}_{\text{l}}\left(\left[\left(\boldsymbol{\xi}^+\right)^{\mathsf{T}} \quad \left(\boldsymbol{\xi}^-\right)^{\mathsf{T}}\right]^{\mathsf{T}}\right) = \boldsymbol{c}^{\mathsf{T}}\left(\boldsymbol{\xi}^+ + \boldsymbol{\xi}^-\right) \quad , \tag{5.123}$$

which is differentiable with respect to $\boldsymbol{\xi}^+$ and $\boldsymbol{\xi}^-$. Using an *extended* vector of optimization parameters $\widetilde{\boldsymbol{\Theta}} = \left[\boldsymbol{\Theta}^{\mathsf{T}} \quad \left(\boldsymbol{\xi}^+\right)^{\mathsf{T}} \quad \left(\boldsymbol{\xi}^-\right)^{\mathsf{T}}\right]^{\mathsf{T}}$, the Jacobian and Hessian matrix of the original optimization problem can be complemented by the partial derivatives

$$\begin{aligned} \frac{\partial\left(\boldsymbol{p}_{\text{l}}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)}{\partial\left(\boldsymbol{\xi}^+\right)} &= \frac{\partial\left(\boldsymbol{p}_{\text{l}}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)}{\partial\left(\boldsymbol{\xi}^-\right)} = \mathbf{1}_{1 \times h} \\ \frac{\partial\left(\boldsymbol{p}_{\text{l}}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)}{\partial\left(\boldsymbol{\xi}^+\right)\partial\left(\boldsymbol{\xi}^+\right)} &= \frac{\partial\left(\boldsymbol{p}_{\text{l}}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)}{\partial\left(\boldsymbol{\xi}^-\right)\partial\left(\boldsymbol{\xi}^-\right)} = \mathbf{0}_{h \times h} \quad . \end{aligned} \tag{5.124}$$

The cost of doing so is having $2 \cdot h$ additional parameters and $3 \cdot h$ supplementary constraints in the optimization problem (cf. He, 2011, pp. 9 ff).

Therefore, the LASSO is computationally more demanding than the ridge penalty. However, its advantage is that for sufficiently large values of λ, it will cause some of the regression coefficients to become exactly zero, thereby providing "a kind of continuous subset selection" (Hastie, Tibshirani and Friedman, 2008, p. 69) for the auxiliary variables.

To combine this appealing property with the better computational tractability of ridge penalties, Zou and Hastie (2005) introduce the *elastic-net* penalty, which is a convex combination of both. For regression models, an efficient solution can again be based on least angle regression (cf. Efron et al., 2004; Zou and Hastie, 2005).

The shrinkage parameter λ may be considered as a fixed constant, but it is more commonly determined by minimizing a distance measure $\tilde{\delta} : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$, i.e.

$$\lambda^* \; = \; \underset{\lambda}{\operatorname{argmin}} \left(\tilde{\delta} \left(\lambda \right) \right) \qquad . \tag{5.125}$$

Typically, this distance is an estimate of the expected prediction or *generalization error* when the model is used to predict observations that are independent from those considered for fitting the model. When the distance $\delta \left(\boldsymbol{\Theta} \right)$ that is used to quantify the prediction error for model fitting depends on the predictions obtained from parameters $\boldsymbol{\Theta}$ (cf. sections 5.1.2 to 5.1.6), an estimate for the expected prediction error of λ for the whole population is

$$\tilde{\delta} \left(\lambda \right) = \widehat{\mathrm{E}} \left(\delta \left(\boldsymbol{\Theta} \left(\lambda \right) \right) \right) \tag{5.126}$$

for $\boldsymbol{\Theta} \left(\lambda \right)$ as the solution of problem 5.116. Unfortunately, the actual distance value $\delta \left(\boldsymbol{\Theta} \left(\lambda \right) \right)$ in data set s to which the model is fit is not a good estimate for the generalization error in equality 5.126. Since it is used to estimate parameters $\boldsymbol{\Theta} \left(\lambda \right)$, $\delta \left(\boldsymbol{\Theta} \left(\lambda \right) \right)$ can be driven to zero if the model is of sufficient complexity, which leads to overfitting. Different more adequate ways to estimate the generalization error exist, of which (generalized) cross-validation techniques are the most common ones (cf. e.g. Craven and Wahba, 1979, p. 379; Hastie, Tibshirani and Friedman, 2008, pp. 219 ff; McLachlan, 2004, pp. 337 ff; Wood, 2017, p. 169). Cross-validation for fitting a model in data set s is described in algorithm 14.

Algorithm 14: Cross-validation algorithm

Input: $\mathcal{S}^{\mathsf{s}} \in \mathbb{S}$; $\boldsymbol{X}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}} \times p}$; $\boldsymbol{y}^{\mathsf{s}}_{\cdot a} \in \mathbb{R}^{n^{\mathsf{s}} \times o}$; $\boldsymbol{w}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}}}$; $\tilde{\delta} : \mathbb{R}^{\mathbb{R}_{\geq 0}} \rightarrow \mathbb{R}_{\geq 0}$; $\boldsymbol{\lambda} \in \mathbb{R}^{c}_{\geq 0}$; $b \in \mathbb{N}$

Initialize $\boldsymbol{d} = \boldsymbol{0}_{c \times 1}$ and $\widehat{\boldsymbol{y}}_{\cdot a} = \boldsymbol{0}_{n^{\mathsf{s}} \times 1}$

Randomly partition the set of observations $\mathcal{S}^{\mathsf{s}}$ into b mutually exclusive subsets $\mathcal{S}^{(l)}$ of (up to integer rounding precision) size $\left| \mathcal{S}^{(l)} \right| \approx n^{\mathsf{s}} / b$ for all $l = 1, \ldots, b$, such that $\mathcal{S}^{\mathsf{s}} = \bigcup_{l=1}^{J} \mathcal{S}^{(l)}$ and $\emptyset = \mathcal{S}^{(k)} \cap \mathcal{S}^{(l)}$ for all $k \neq l$

for $k = 1, \ldots, c$ **do**

  for $l = 1, \ldots, b$ **do**

    Determine $\boldsymbol{\Theta}^{(\mathrm{l})} \left(\lambda_k \right)$ as the solution to problem 5.116 by fitting a model $\mathbf{m}_{\mathrm{l}} \left(\boldsymbol{X}_{\mathcal{J} \cdot}, \boldsymbol{\Theta}^{(\mathrm{l})} \left(\lambda_k \right) \right)$ for $\boldsymbol{y}^{\mathsf{s}}_{\cdot a}$ to all observations $i \in \mathcal{J}$ that are not in the l-th subset, i.e. $\mathcal{J} := \mathcal{S}^{\mathsf{s}} \setminus \mathcal{S}^{(l)}$

    Use parameters $\boldsymbol{\Theta}^{(\mathrm{l})} \left(\lambda_k \right)$ to predict $\boldsymbol{y}_{\mathcal{S}^{(l)} a}$ by $\widehat{\boldsymbol{y}}_{\mathcal{S}^{(l)} a} \leftarrow \mathbf{m}_{\mathrm{l}} \left(\boldsymbol{X}_{\mathcal{S}^{(l)} \cdot}, \boldsymbol{\Theta}^{(\mathrm{l})} \left(\lambda_k \right) \right)$ for all observations $i \in \mathcal{S}^{(l)}$ that are not considered for fitting the model

  end for

  Calculate $d_k \leftarrow \tilde{\delta} \left(\lambda_k \right)$ as the expected loss over all predictions in $\widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot a}$, which are uniquely defined since $\mathcal{S}^{\mathsf{s}}$ is the union of all mutually exclusive subsets $\mathcal{S}^{(l)}$

end for

Return: λ_k for $k = \mathrm{Min} \left(\{ l : d_l = \mathrm{Min} \left(\boldsymbol{d} \right) \} \right)$

To determine an optimal value λ^* of the shrinkage parameter in the sense of problem 5.125 by means of this algorithm, various prespecified candidate values that are combined in a vector $\boldsymbol{\lambda} \in \mathbb{R}^u_{\geq 0}$ are evaluated in a grid search. For each of these values, estimation of the generalization error requires a sample that is used for fitting the model. To achieve this requirement, cross-validation removes certain observations $i \in \mathcal{S}^{(l)}$ from data set s before model fitting. The remaining observations $j \in \left\{\mathcal{S}^{\mathsf{s}} \setminus \mathcal{S}^{(l)}\right\}$ are called the *training set* because they are used for fitting the model. The generalization error is evaluated only for observations $i \in \mathcal{S}^{(l)}$ that are not in the training set and hence referred to as the *validation set*. This is repeated for all $l = 1, \ldots, b$ subsets $\mathcal{S}^{(l)}$, such that the estimated generalization error can be obtained for the whole data set s. The element of $\boldsymbol{\lambda}$ which results in the lowest estimated generalization error is returned as the optimal solution λ^* (cf. Allen, 1974; Hastie, Tibshirani and Friedman, 2008, pp. 219 ff; Stone, 1974; 1977).

The formulation of this cross-validation algorithm 14 is quite general but can be computationally demanding. This is especially the case when multiple shrinkage parameters have to be estimated for different shrinkage terms, which is e.g. common for smoothing and P-splines that use multiple independent variables (cf. Hastie and Tibshirani, 1990, p. 159; Wood, 2017, p. 170). Important simplifications of leave-one-out cross-validation are possible if predictions are linear in $\widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot a}$, i.e.

$$\widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot a} \;=\; \boldsymbol{A}\left(\boldsymbol{\Theta}\left(\lambda\right)\right)\boldsymbol{y}^{\mathsf{s}}_{\cdot a} \qquad , \tag{5.127}$$

where the *smoother matrix* $\boldsymbol{A}\left(\boldsymbol{\Theta}\left(\lambda\right)\right)$ for $\boldsymbol{A} : \mathbb{R}^h \to \mathbb{R}^{n^{\mathsf{s}} \times n^{\mathsf{s}}}$ depends on $\boldsymbol{\Theta}\left(\lambda\right)$ and hence on λ but not on $\boldsymbol{y}^{\mathsf{s}}_{\cdot a}$. Leave-one-out cross-validation results from setting $b = 1$ in algorithm 14. Consequently, a model $\mathbf{m}_{\mathrm{i}}\left(\boldsymbol{x}^{\mathsf{s}}_{i\cdot}, \boldsymbol{\Theta}^{(\mathrm{i})}\left(\lambda\right)\right)$ that excludes only a single observation i is fit in step 6 of algorithm 14 for all $i = 1, \ldots, n^{\mathsf{s}}$ and each candidate value for λ. If the predictions are linear in $\boldsymbol{y}^{\mathsf{s}}_{\cdot a}$ as in equality 5.127, all these models can be obtained directly from the full model's smoother matrix $\boldsymbol{A}\left(\boldsymbol{\Theta}\left(\lambda\right)\right)$ through

$$\boldsymbol{B}\boldsymbol{y}^{\mathsf{s}}_{\cdot a} \;=\; \begin{bmatrix} \mathbf{m}_1\left(\boldsymbol{x}^{\mathsf{s}}_{1\cdot}, \boldsymbol{\Theta}^{(1)}\left(\lambda\right)\right) \\ \vdots \\ \mathbf{m}_{\mathrm{n}^{\mathsf{s}}}\left(\boldsymbol{x}^{\mathsf{s}}_{n^{\mathsf{s}}\cdot}, \boldsymbol{\Theta}^{(\mathrm{n}^{\mathsf{s}})}\left(\lambda\right)\right) \end{bmatrix} \qquad . \tag{5.128a}$$

In this context,

$$\boldsymbol{B} \;\coloneqq\; \left(\boldsymbol{A}\left(\boldsymbol{\Theta}\left(\lambda\right)\right) - \boldsymbol{C}\right) \oslash \left(\left(\mathbf{I}_{n^{\mathsf{s}}} - \boldsymbol{C}\right)\mathbf{1}_{n^{\mathsf{s}} \times n^{\mathsf{s}}}\right) \quad \in \mathbb{R}^{n^{\mathsf{s}} \times n^{\mathsf{s}}} \tag{5.128b}$$

is used to construct the predictions $\widehat{y}^{\mathsf{s}}_{ia}$ obtained from a model that is fit excluding element i for all $i \in \mathcal{S}^{\mathsf{s}}$, similarly as $\boldsymbol{A}\left(\boldsymbol{\Theta}\left(\lambda\right)\right)$ constitutes the predictions of the full model in equality 5.127. This matrix $\boldsymbol{B}$ is defined by means of

$$\boldsymbol{C} \coloneqq \mathbf{diag}\left(\mathbf{diag}\left(\boldsymbol{A}\left(\boldsymbol{\Theta}\left(\lambda\right)\right)\right)\right) \in \mathbb{R}^{n^{\mathsf{s}} \times n^{\mathsf{s}}} \qquad , \tag{5.128c}$$

which denotes a diagonal matrix that contains only the diagonal elements of $\boldsymbol{A}\left(\boldsymbol{\Theta}\left(\lambda\right)\right)$ (cf. appendix B.4.6.2; Craven and Wahba, 1979; Hastie and Tibshirani, 1990, pp. 46 ff; Golub, Heath and Wahba, 1979; Wood, 2017, pp. 169 ff).

For loss functions that are based on the residual sum of squares (cf. e.g. equalities 5.12 and 5.87), Craven and Wahba (1979) as well as Golub, Heath and Wahba (1979) propose a further simplification, which is referred to as *generalized cross-validation*. From equalities 5.128, it directly follows that the generalization error for element i when excluding it from

model fitting is given by

$$\mathbf{m}_{\mathrm{i}}\left(\boldsymbol{x}^{\mathsf{s}}_{i\cdot}, \boldsymbol{\Theta}^{(\mathrm{i})}(\lambda)\right) - y^{\mathsf{s}}_{ia} = (1 - a_{ii})^{-1} \cdot (\widehat{y}^{\mathsf{s}}_{ia} - y^{\mathsf{s}}_{ia}) \tag{5.129a}$$

$$\approx (1 - \operatorname{tr}(\boldsymbol{A})/n^{\mathsf{s}})^{-1} \cdot (\widehat{y}^{\mathsf{s}}_{ia} - y^{\mathsf{s}}_{ia}) \quad . \tag{5.129b}$$

The generalized cross-validation (gcv) criterion is therefore

$$\operatorname{gcv}(\lambda) \coloneqq n^{\mathsf{s}} \cdot ||\widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot a} - \boldsymbol{y}^{\mathsf{s}}_{\cdot a}||_2^2 \Big/ (n^{\mathsf{s}} - \operatorname{tr}(\boldsymbol{A}))^2 \quad . \tag{5.130}$$

The rationale of approximation 5.129b is to replace diagonal element a_{ii} in equality 5.129a by its expectation $\mathrm{E}(\mathbf{diag}(\boldsymbol{A})) = \operatorname{tr}(\boldsymbol{A})/n^{\mathsf{s}}$ to obtain an asymptotically optimal estimate for λ. In linear models, $\operatorname{tr}(\boldsymbol{A})$ is simply the number of model parameters, such that $\operatorname{tr}(\boldsymbol{A})$ is often referred to as the *effective number of parameters or degrees of freedom* used for a general smoother matrix $\boldsymbol{A}$. In addition, the use of generalized cross-validation further simplifies calculation of squared generalization errors in algorithm 14, which can be obtained from the full model's residuals (cf. appendix B.4.6.2; Craven and Wahba, 1979; Hastie and Tibshirani, 1990, pp. 46 ff; Golub, Heath and Wahba, 1979; Wood, 2017, pp. 16, 166 ff).

However, the simplifications obtained from equalities 5.128 to 5.130 are based on equality 5.127 and, hence, do not hold for models that are non-linear in $\widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot a}$, e.g. for generalized linear or additive models. Following O'Sullivan, Yandell and Raynor (1986), Hastie and Tibshirani (1990, p. 159) nevertheless propose using approximation 5.129b and replace the residual sum of squares $\left(n^{\mathsf{s}} \cdot ||\widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot a} - \boldsymbol{y}^{\mathsf{s}}_{\cdot a}||_2^2\right)$ by the deviance $(-n^{\mathsf{s}} \cdot ||\boldsymbol{y}^{\mathsf{s}}_{\cdot a} \circ \log(\widehat{\boldsymbol{y}}^{\mathsf{s}}_{\cdot a})||_1)$ obtained from the converged GLM or GAM (cf. equation 5.88). The rationale behind this approach is that a quadratic approximation of the deviance is used in each step of the iteratively reweighted least squared (or backfitting) procedure, such that the actual deviance can be used in place of the approximation in the final model. In contrast, Gu (1990; 1992) proposes optimizing penalty parameters for each single working model that is fit in the IRWLS procedure. This is motivated by the fact that approximation 5.129b can be used directly for each of the linear models that are fit to the adjusted dependent variables to implement Fisher scoring (cf. equalities 5.27 and 5.36). This method can be computationally faster than using the deviance of the final model because it does not require performing the entire IRWLS sequence for each potential value of λ, but it may also lead to convergence issues. A detailed overview and discussion of both approaches is provided by Wood (2017, pp. 173 ff).

Similarly as for choosing shrinkage parameters, (generalized) cross-validation can be used to compare the generalization error of different models. Therefore, it can also be used for variable and model selection as well as for choosing other hyper-parameters based on data set s. The gcv criterion defined in equality 5.130 is approximately identical to Akaike's (1973) information criterion, which is also a common tool for this purpose (cf. Hastie and Tibshirani, 1990, p. 158; Wood, 2017, p. 174).

All the model-based methods discussed throughout the current section 5.1 rely on the fact that under conditional independence assumption 5.1, unbiased estimation of $\mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y}_{i\cdot} \mid \boldsymbol{x}_{i\cdot})$ from the non-probability sample is possible. Generalization to a population or probability sample in this context is then achieved by using the distribution of $\boldsymbol{X}$ (cf. equation 5.8). A different approach is considered in the following section 5.2, where pseudo-design-based methods for weighting non-probability samples are discussed.

5.2 Pseudo-design-based Methods: Weighting

For non-probability samples, the unknown and/or uncontrolled nature of the sample selection process can result in systematically differing distributions of target variables $\boldsymbol{Y}$ in population and sample. From equation 2.24, it is evident that both are related by

$$\frac{\mathrm{f}_{\boldsymbol{Y}^{\mathsf{nps}}}(\boldsymbol{y}_{i\cdot})}{\mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y}_{i\cdot})} = \frac{\mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot})}{\mathrm{P}(r_i^{\mathsf{nps}} = 1)} \quad , \tag{5.131}$$

where $\mathrm{f}_{\boldsymbol{Y}^{\mathsf{nps}}}(\boldsymbol{y}_{i\cdot})$ and $\mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y}_{i\cdot})$ respectively denote the density of $\boldsymbol{Y}$ in the non-probability sample and the population. If sampling mechanism and variables of interest are related, this ratio is typically different from one, which may occur in a probability sample as well. To account for potential bias in this case (cf. section 2.3), one major way is constituted by model-based approaches discussed in the previous section 5.1. Classical design-based estimation, however, resolves this matter using design weights that compensate for the sample selection process (cf. section 2.2 and the references cited therein). This is achieved by relying on the fact that $\mathrm{P}(r_i^{\mathsf{ps}} = 1 \,|\, \boldsymbol{y}_{i\cdot})/\mathrm{P}(r_i^{\mathsf{ps}} = 1) = \mathrm{E}(w_i^{\mathsf{ps}} \,|\, r_i^{\mathsf{ps}} = 1)/\mathrm{E}(w_i^{\mathsf{ps}} \,|\, \boldsymbol{y}_{i\cdot}, r_i^{\mathsf{ps}} = 1)$ constitutes the right-hand side of equality 5.131 for a probability sample ps (cf. appendix B; Pfeffermann and Sverchkov, 1999, p. 185).

Therefore, one major issue for estimation from non-probability samples can be seen in the lack of analogous design weights. As outlined in figure 5.1, pseudo-design-based methods for non-probability samples deal with this problem by generating some surrogate weights, which are referred to as *pseudo-design weights* and denoted by $\widetilde{\boldsymbol{w}} \in \mathbb{R}^{n^{\mathsf{nps}}}$. These pseudo-design weights are then substituted for design weights in classical design-based estimation, i.e. the non-probability sample is treated as if it was obtained by probability sampling with the corresponding design weights $\widetilde{\boldsymbol{w}}$ (cf. Buelens, Burger and van den Brakel, 2018, pp. 331 f; Elliott and Valliant, 2017, p. 259). These ideas resemble those of weighting adjustment procedures that are frequently applied for probability samples, e.g. to compensate for non-response or construct more efficient estimators (cf. e.g. Deville and Särndal, 1992; Little, 1986; Särndal and Lundström, 2005; Kott, 2006).

In contrast to model-based approaches that rely on the conditional distribution of $\boldsymbol{Y}$ to compensate for the sample's selectivity, the basic idea in the pseudo-design-based framework is to focus on equality 5.131 to achieve the same objective. Since the true data generating process is typically unknown for non-probability samples, computation of pseudo-design weights has to rely on some sort of assumptions, of which two main realizations can be found. First, the right-hand side of equation 5.131 highlights the relevance of representing the non-probability sample selection process to obtain proper weights. This suggests the use of *response propensities* introduced in section 3.6 to express the relation of conditional and unconditional probabilities of elements being observed in the non-probability sample. To estimate these propensities, different prediction models discussed in the previous section 5.1 are adopted. Second, the left-hand side of equality 5.131 requires that weighted sample and population distribution coincide if selectivity is ignorable. *Calibration* is therefore based on explicitly adopting constraints that enforce conformity in certain aspects of distributions (mostly means or totals) for weighting a non-probability sample. As in the preceding sections, both of these ideas use auxiliary variables, for which information in- and outside the non-probability sample is required (cf. e.g. Buelens, Burger and van den Brakel, 2018, p. 332; Dever, Rafferty and Valliant, 2008, p. 60; Pfeffermann, 2015, pp. 443 f; Valliant and Dever, 2011, p. 108).

In comparison to the model-based approaches discussed in section 5.1, the pseudo-design-based framework has an appealing property: if a single set of weights can be found to perfectly describe the sample selection mechanism, it can be applied to estimate any quantity of interest, just like design weights. Even though this ideal case is rarely realistic, it is still worthwhile to strive for (cf. Baker et al., 2013b, p. 102; Buelens et al., 2012, p. 10). Nevertheless, it may often be easier to obtain good prediction models for a single or few target variables than a good weighting scheme to account for the entire selectivity of a non-probability sample (cf. Steinmetz, Tijdens and Pedraza, 2009, p. 16).

To summarize, compare and extend methods for non-probability samples, the subsequent discussion is structured as follows. Weighting based on estimated response propensities is discussed in section 5.2.1, followed by established calibration methods in section 5.2.2. In section 5.2.3, both ideas are combined and extended by introducing response models in the general form of semi-parametric neural networks that can incorporate calibration constrains for totals, covariances and/or correlations. Sub-sampling constitutes a further strategy that can be framed as pseudo-design-based (cf. Posner and Ash, 2012, p. 4) and is described in section 5.2.4.

5.2.1 Response Propensity Weighting

The concept of obtaining weights for non-probability samples is inherently related to classical design-based estimation. As outlined above (cf. equality 5.131), the first idea is to mimic the implied but usually unknown inclusion probabilities $\pi_i^{\mathsf{nps}} = \mathrm{P}\left(i \in \mathcal{S}^{\mathsf{nps}}\right)$ for a non-probability sample nps. Unfortunately, one sample constitutes only a single realization of $\boldsymbol{r}^{\mathsf{nps}}$. Consequently, the individual probabilities π_i^{nps} cannot be obtained as the expected value over all possible samples, which would be the case in probability sampling (cf. definition 2.3; Schouten, Cobben and Bethlehem, 2009, p. 105). Therefore, auxiliary variables are used to describe the response process, which are denoted by $\boldsymbol{Z}$ as before (cf. figure 5.1). If these variables perfectly describe the sampling process, the true inclusion probabilities $\boldsymbol{\pi}^{\mathsf{nps}}$ could be obtained from $\boldsymbol{Z}$, just as for probability samples. Consequently, the response propensity $p_i^{\mathrm{nps}} = \mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right)$ for all $i \in \mathcal{S}^{\mathrm{P}}$ (cf. definition 3.27) can be seen as an approximation for π_i^{nps} when assuming probability sampling by a design that is solely determined through $\boldsymbol{Z}$ (cf. Biffignandi and Pratesi, 2003, p. 8; Enderle, Münnich and Bruch, 2013, p. 92; Schouten, Shlomo and Skinner, 2009, p. 11). This corresponds to assuming conditional independence $\left(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{r}^{\mathsf{nps}}\right) | \, \boldsymbol{Z}$ in the form of assumption 5.1, from which it follows that

$$p_i^{\mathrm{nps}} = \mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right) = \mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}\right) \quad . \tag{5.132}$$

Under this condition, weighting by the inverse of p_i^{nps} achieves unbiasedness for design linear estimators if $p_i^{\mathrm{nps}} > 0$ for all $i \in \mathcal{S}^{\mathrm{P}}$, similarly to the Horvitz-Thompson estimator (cf. appendix B; Dawid, 1979, p. 3; Horvitz and Thompson, 1952, pp. 667 ff; Imbens, 2000, p. 708; Lunceford and Davidian, 2004, p. 2941):

$$\begin{aligned} \mathrm{E}\left(\sum_{i \in \mathcal{S}^{\mathsf{nps}}} \frac{\boldsymbol{y}_{i\cdot}}{p_i^{\mathrm{nps}}}\right) &= \sum_{i \in \mathcal{S}^{\mathrm{P}}} \mathrm{E}\left(\mathrm{E}\left(\frac{r_i^{\mathsf{nps}} \cdot \boldsymbol{y}_{i\cdot}}{p_i^{\mathrm{nps}}} \,\middle|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}\right)\right) \\ &= \sum_{i \in \mathcal{S}^{\mathrm{P}}} \mathrm{E}\left(\boldsymbol{y}_{i\cdot} \cdot \frac{\mathrm{E}\left(r_i^{\mathsf{nps}} \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}\right)}{p_i^{\mathrm{nps}}}\right) = \sum_{i \in \mathcal{S}^{\mathrm{P}}} \boldsymbol{y}_{i\cdot} \quad . \end{aligned} \tag{5.133}$$

However, the true response propensities are usually unknown because $\boldsymbol{r}^{\mathsf{nps}}$ is only a single realization obtained from a non-probability sampling mechanism that is itself not known. Therefore, it is typically necessary to rely on a model to obtain estimated response propensities $\widehat{\boldsymbol{p}}^{\mathrm{nps}}$ (cf. section 3.6; Rosenbaum and Rubin, 1983, p. 47). In this case, $\widehat{\boldsymbol{p}}^{\mathrm{nps}}$ is a prediction from a *response model* $\mathbf{m}$ that is assumed to describe the non-probability sample selection process, such that

$$\widehat{p}_i^{\mathrm{nps}} = \widehat{\mathrm{P}}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right) \coloneqq \mathbf{m}\left(\boldsymbol{z}_{i\cdot}, \boldsymbol{\Theta}\right) \quad \text{for all} \quad i \in \mathcal{S}^{\mathrm{P}} \qquad . \tag{5.134}$$

In most applications, one of the models described in section 5.1 is fit to the binary dependent variable $\boldsymbol{r}^{\mathsf{nps}}$ (cf. section 5.1) for this purpose. Generalized linear (logit) models are of particular relevance in this context, but additive and machine learning models are increasingly applied as well (cf. section 3.6; Lee, Lessler and Stuart, 2010, pp. 337 ff; Rosenbaum and Rubin, 1983, p. 47; Schonlau et al., 2009, p. 294; Schouten, Cobben and Bethlehem, 2009, p. 105). In any case, it is required that observed values of $\boldsymbol{Z}$ in- and outside the non-probability sample are available because fitting such a model requires measured variability in $\boldsymbol{r}^{\mathsf{nps}}$. As before, observations outside the non-probability sample may, for example, come from a full population register or a reference sample (cf. also sections 3.2 and 3.6).

A closely related but slightly more specialized adaptation is proposed by Elliott (2009, pp. 2 f) as well as Elliott and Valliant (2017, pp. 256 f), to which the authors refer as estimation of *'pseudo-weights'*.[5] For this approach, it is assumed that the observed values outside the non-probability sample required to model $\boldsymbol{p}^{\mathrm{nps}}$ come from a reference probability sample res with corresponding inclusion indicator variable $\boldsymbol{r}^{\mathsf{res}}$. Furthermore, the sampling design generating $\boldsymbol{r}^{\mathsf{res}}$ is considered known and either perfectly or at least very well described by the variables $\boldsymbol{Z}$. The indicator for whether an element is part of the non-probability and/or the reference sample is denoted by $r_i^{\mathsf{u}} \coloneqq \mathbb{I}\left(i \in (\mathcal{S}^{\mathsf{nps}} \cup \mathcal{S}^{\mathsf{res}})\right) \in \{0; 1\}$ for all $i \in \mathcal{S}^{\mathrm{P}}$. The true propensity can then be written as

$$p_i^{\mathrm{nps}} \propto \frac{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot} \,|\, r_i^{\mathsf{nps}} = 1\right)}{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot} \,|\, r_i^{\mathsf{res}} = 1\right)} \cdot \mathrm{P}\left(r_i^{\mathsf{res}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right) \tag{5.135}$$

for all $i \in \mathcal{S}^{\mathrm{P}}$. The fraction applied in relation 5.135 is then determined by

$$\frac{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot} \,|\, r_i^{\mathsf{nps}} = 1\right)}{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot} \,|\, r_i^{\mathsf{res}} = 1\right)} \approx \frac{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot} \,|\, r_i^{\mathsf{nps}} = 1, r_i^{\mathsf{u}} = 1\right)}{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot} \,|\, r_i^{\mathsf{nps}} = 0, r_i^{\mathsf{u}} = 1\right)} \propto \frac{\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{u}} = 1\right)}{\mathrm{P}\left(r_i^{\mathsf{nps}} = 0 \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{u}} = 1\right)} \quad , \tag{5.136}$$

assuming that the intersection $(\mathcal{S}^{\mathsf{nps}} \cap \mathcal{S}^{\mathsf{res}})$ is negligible, such that $\boldsymbol{r}^{\mathsf{u}} \approx \boldsymbol{r}^{\mathsf{nps}} + \boldsymbol{r}^{\mathsf{res}}$ (cf. appendix B). The reason for using such an approximation is that it eases estimation of response propensities since numerator and denominator in the right-hand side of relation 5.136 are then based on complementary events. The authors propose using $\widehat{\mathrm{P}}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{u}} = 1\right) = 1 - \mathrm{P}\left(r_i^{\mathsf{nps}} = 0 \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{u}} = 1\right) = \mathbf{m}\left(\boldsymbol{z}_{i\cdot}, \boldsymbol{\Theta}\right)$ for estimating the probability of these events. As in equality 5.134, a model $\mathbf{m}$ for $\boldsymbol{r}^{\mathsf{nps}}$ is required for this purpose, but the modeled probability is now conditional on being in the combined sample. Therefore, the model can be fit without using design weights $\boldsymbol{w}^{\mathsf{res}}$ for the reference sample,

[5] The phrase 'pseudo-weights' is used specifically for the approach proposed by Elliott (2009) as well as Elliott and Valliant (2017). It is not to be confused with the more general term 'pseudo-design weights', which can refer to any of the weights discussed in the current section 5.2.

which differentiates it from typical propensity score models (cf. Elliott and Valliant, 2017, p. 257). From relations 5.135 and 5.136, the estimated propensity scores for all $i \in \mathcal{S}^{\mathrm{P}}$ follow as

$$\widehat{p}_i^{\,\mathrm{nps}} = \frac{\mathbf{m}\left(\boldsymbol{z}_{i\cdot}, \boldsymbol{\Theta}\right)}{1 - \mathbf{m}\left(\boldsymbol{z}_{i\cdot}, \boldsymbol{\Theta}\right)} \cdot \mathrm{P}\left(r_i^{\mathsf{res}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right) \quad . \tag{5.137}$$

The model $\mathbf{m}$ is thus used for predicting non-probability sample membership for element i, conditional on i being in the combined sample.

For identifying $\widehat{\boldsymbol{p}}^{\mathrm{nps}}$ based on this model $\mathbf{m}$, it is still necessary to determine the conditional probability $\mathrm{P}\left(r_i^{\mathsf{res}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right)$ of element i being in the reference sample given its value of $\boldsymbol{z}_{i\cdot}$. If $\boldsymbol{Z}$ includes all design variables used for selecting res, this conditional probability may be directly obtained from $\boldsymbol{z}_{i\cdot}$ and the reference sample's design. However, it is more likely that this is not the case, such that $\boldsymbol{Z}$ does not perfectly describe the sampling design generating $\boldsymbol{r}^{\mathsf{res}}$. In this scenario, an estimate $\widehat{\mathrm{P}}\left(r_i^{\mathsf{res}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right)$ can be obtained from a model as well. If this model coincides with $\mathbf{m}$, i.e. by defining $\widehat{\mathrm{P}}\left(r_i^{\mathsf{res}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right) \coloneqq 1 - \mathbf{m}\left(\boldsymbol{z}_{i\cdot}, \boldsymbol{\Theta}\right)$, equations 5.134 and 5.137 are equivalent (cf. equations 5.135 and 5.136). However, a central property of probability sample res is that design weights $\boldsymbol{w}^{\mathsf{res}}$ are known. Using this information, Elliott and Valliant (2017, p. 257) propose modeling $\boldsymbol{\pi}^{\mathsf{res}} = \left(\boldsymbol{w}^{\mathsf{res}}\right)^{\circ(-1)}$ as a function of $\boldsymbol{Z}$, e.g. by means of beta regression (cf. section 2.2; Ferrari and Cribari-Neto, 2004). Predictions obtained from this model for $\boldsymbol{\pi}^{\mathsf{res}}$ correspond to the estimated conditional probabilities $\widehat{\mathrm{P}}\left(r_i^{\mathsf{res}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right)$ required to determine equation 5.137. In any case, the quality of this approach highly depends on the sample selection of both samples being well described by variables $\boldsymbol{Z}$ (cf. Elliott and Valliant, 2017, p. 256).

Once response propensities are estimated from equalities 5.134 or 5.137, they are commonly used as an approximation for inclusion probabilities in classical design-based estimation (cf. equation 5.133 and section 2.2). In analogy to definition 2.13, the pseudo-design weights

$$\widetilde{w}_i \coloneqq \frac{w_i^{\mathsf{nps}}}{\widehat{p}_i^{\,\mathrm{nps}}} \quad \text{for all} \quad i \in \mathcal{S}^{\mathsf{nps}} \tag{5.138}$$

in this case result as the inverse estimated response propensities, multiplied by some initial weights w_i^{nps}. The latter are typically set to ones ($\boldsymbol{w}^{\mathsf{nps}} \coloneqq \mathbf{1}_{n^{\mathsf{nps}} \times 1}$) if there is no prior information available to specify them (cf. e.g. Biffignandi and Pratesi, 2000, p. 1528; Chang and Kott, 2008, p. 556; Lee, Lessler and Stuart, 2010, p. 340; Little, 1988b, p. 293; Pfeffermann, 2015, p. 444; Särndal and Lundström, 2005, pp. 51, 106; Schonlau et al., 2009, p. 294; Valliant and Dever, 2011, p. 109).

Despite the validity of equalities 5.133, an issue with propensity weights is that they can lead to instabilities and high variances due to their reliance on the underlying model. To reduce the variability of weights and resulting estimates as well as the potential impacts of a misspecified model, the estimated propensities are often stabilized by creating groups and using the group-averaged estimated propensity score for weighting. To that end, the population $\mathcal{S}^{\mathrm{P}}$ is partitioned into J mutually exclusive subsets, such that $\mathcal{S}^{\mathrm{P}} = \bigcup_{j=1}^{J} \mathcal{S}^{(j)}$ and $\emptyset = \mathcal{S}^{(j)} \cap \mathcal{S}^{(k)}$ for all $j \neq k$, where the assignment of element i to one of these subsets is determined by the magnitude of $\widehat{p}_i^{\,\mathrm{nps}}$ (cf. e.g. Brick, 2013, p. 336; Little, 1986, p. 147; Rosenbaum and Rubin, 1983, pp. 51 ff). In this case, weights are calculated using the

mean propensity score within each subset observed in the non-probability sample, i.e.

$$\widetilde{w}_i \coloneqq w_i^{\mathsf{nps}} \cdot \left(\sum_{j=1}^{J} \frac{\mathbb{I}\left(i \in \left(\mathcal{S}^{\mathsf{nps}} \cap \mathcal{S}^{(j)}\right)\right)}{|(\mathcal{S}^{\mathsf{nps}} \cap \mathcal{S}^{(j)})|} \cdot \sum_{\substack{k \in \mathcal{S}^{\mathsf{nps}}, \\ k \in \mathcal{S}^{(j)}}} \widehat{p}_k^{\mathrm{nps}} \right)^{-1} \quad \text{for all} \quad i \in \mathcal{S}^{\mathsf{nps}} \quad , \tag{5.139}$$

where $\left(\mathcal{S}^{\mathsf{nps}} \cap \mathcal{S}^{(j)}\right)$ identifies elements in the non-probability sample that belong to the j-th group (cf. Little, 1986, p. 147; Valliant and Dever, 2011, p. 115).

Besides their application as approximate inclusion probabilities for non-probability samples, an additional use of propensity scores occurs in *propensity post-stratification*. Similar as before, the observations are split into subsets depending on the propensity scores for this purpose, but the actual weights are determined by aligning the subsets' weighted proportions with some external benchmarks. In that case, a matrix of auxiliary variables defined by J indicators for membership in subsets $\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(J)}$ as above, i.e.

$$\widetilde{\boldsymbol{X}} \coloneqq \begin{bmatrix} \mathbb{I}\left(1 \in \mathcal{S}^{(1)}\right) & \cdots & \mathbb{I}\left(1 \in \mathcal{S}^{(J)}\right) \\ \vdots & \ddots & \vdots \\ \mathbb{I}\left(\mathrm{N} \in \mathcal{S}^{(1)}\right) & \cdots & \mathbb{I}\left(\mathrm{N} \in \mathcal{S}^{(J)}\right) \end{bmatrix} \tag{5.140}$$

is defined but usually not fully observed (cf. e.g. Valliant and Dever, 2011, pp. 116 f). The weighted estimates of each subset's size in the non-probability sample are then adapted to meet the respective size in the reference sample. This adaptation can be achieved by using $\widetilde{\boldsymbol{X}}$ as defined in equality 5.140 as calibration variable for the calibration methods that are discussed in the following section 5.2.2.

5.2.2 Calibration Weighting

Calibration weighting is considered the second essential pseudo-design-based approach for dealing with non-probability samples (cf. e.g. Bianchi and Biffignandi, 2013, p. 39; Buelens, Burger and van den Brakel, 2018, p. 332; Enderle, Münnich and Bruch, 2013, p. 94). It includes various methods to calculate weights for non-probability samples, of which propensity post-stratification (cf. section 5.2.1) is only a small subset. A summary of important concepts and strategies for calibration is given in the current section 5.2.2. The aim is to provide an overview and unifying notation of different but related ideas, serving as a motivation and foundation for the subsequent discussion. To that end, important aspects regarding the (numerical) solution of the calibration approaches are deferred to the following section 5.2.3.

From the previous discussion (cf. sections 2.2 and 5.2.1), it is evident that ideal design or pseudo-design weights compensate for any systematic differences between sample and population distribution of $\boldsymbol{Y}$ and therefore allow for unbiased estimates (cf. equation 5.131). Since an unbiased estimator and the true value (or another unbiased estimator) coincide in expectation, the basic idea of calibration is to reverse these arguments and find weights that explicitly enforce such conformity.

> "A calibration estimator uses calibrated weights, which are as close as possible, according to a given distance measure, to the original sampling design weights π_k^{-1} while also respecting a set of constraints, the calibration equations."
>
> (Deville and Särndal, 1992, p. 376)

Assuming that some external benchmarks are available for the auxiliary variables, the *calibration constraints* ensure adaptation of the weighted sample estimates towards these benchmarks. A set of weights can, thus, be determined by means of constrained optimization, where a distance measure is minimized with respect to these constraints (cf. also Deming and Stephan, 1940, p. 428; Zhang, 2000, p. 178).

For application to non-probability samples, the basic justification of this approach is similar to that outlined in equations 5.131 to 5.133 for propensity weights. Under conditional independence assumption 5.1, it holds that $\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{x}_{i\cdot}\right) = \mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{x}_{i\cdot}\right)$. To determine this conditional probability and construct pseudo-design weights, calibration utilizes the fact that $\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{x}_{i\cdot}\right) \propto \mathrm{f}_{\boldsymbol{X}^{\mathsf{nps}}}\left(\boldsymbol{x}_{i\cdot}\right) / \mathrm{f}_{\boldsymbol{X}}\left(\boldsymbol{x}_{i\cdot}\right)$, which is evident from equation 5.131. Using the inverse of $\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{x}_{i\cdot}\right)$ for weighting, unbiased design linear estimators can then be obtained as in equalities 5.133 if $\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{x}_{i\cdot}\right) > 0$ for all $i \in \mathcal{S}^{\mathrm{P}}$. Following these considerations, ideal calibration constraints would enforce congruence of the weighted non-probability sample and the known population density of $\boldsymbol{X}$. Unless $\boldsymbol{X}$ contains only a small number of categorical variables, however, this is usually not feasible. On the one hand, full population densities are rarely ever available as auxiliary information if the number of possible values in $\boldsymbol{X}$ is not very limited. On the other hand, a solution would anyhow be impossible in most cases where this number of possible values is large because at least some realizations of $\boldsymbol{X}$ would typically be not at all observed in the sample (cf. e.g. Chen, Valliant and Elliott, 2019, p. 665; Deville and Särndal, 1992, pp. 379 f; Deville, Särndal and Sautory, 1993, pp. 1014 f; Zhang, 2000, pp. 179 f).

As a simplification that is more practicable, calibration constraints are commonly specified with regard to only totals (or means) of the auxiliary variables $\boldsymbol{X}$, which are also referred to as the *calibration variables* in this context. The calibration estimators formalized on this basis by Deville and Särndal (1992) are the presumably most commonly cited and used ones. The aim of these authors is to find a vector of *calibration weights*

$$\widetilde{\boldsymbol{w}} \;\coloneqq\; \boldsymbol{w}^{\mathsf{nps}} \circ \boldsymbol{g} \quad \in \quad \mathbb{R}^{n^{\mathsf{nps}}} \qquad , \tag{5.141}$$

where $\boldsymbol{g} \in \mathbb{R}^{n^{\mathsf{nps}}}$ is a vector of rescaling factors, which are termed *correction weights*. These are used to calibrate the original weights $\boldsymbol{w}^{\mathsf{nps}}$, such that weighted total estimates for the non-probability sample nps coincide with known or estimated totals for a *calibration benchmark data source* cal:

$$\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right) \;\overset{!}{=}\; \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right) \qquad . \tag{5.142}$$

Equation 5.142 constitutes the calibration constraints in this case. As before, a non-informative value for $\boldsymbol{w}^{\mathsf{nps}}$ is a vector of ones, and $\boldsymbol{w}^{\mathsf{cal}}$ denotes weights for data set cal. Correspondingly, $\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)$ and $\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)$ respectively are weighted total estimates from the non-probability and benchmark sample, the latter being referred to as *calibration targets* or, as above, benchmarks (cf. section 2.2). To induce as little change as possible to $\boldsymbol{w}^{\mathsf{nps}}$ while achieving compliance with equality 5.142, a distance function $\delta : \mathbb{R}^{n^{\mathsf{nps}}} \to \mathbb{R}_{\geq 0}^{n^{\mathsf{nps}}}$ is used. Calibration of total estimates from a non-probability sample to those from

the calibration benchmark data set for auxiliary variables $\boldsymbol{X}$ is thus determined by the optimization problem

$$\begin{aligned} &\underset{\boldsymbol{g}}{\operatorname{argmin}} \left(\left(\boldsymbol{w}^{\mathsf{nps}}\right)^{\mathsf{T}} \delta\left(\boldsymbol{g}\right) \right) \\ &\text{s. t.} \qquad \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right) \;=\; \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right) \qquad , \end{aligned} \tag{5.143}$$

which leads to minimization of a weighted sum of distances w.r.t. calibration constraints (cf. Deville and Särndal, 1992, pp. 376 ff).

The most prominent example of this approach is the *generalized regression estimator (GREG)*, which minimizes the quadratic relative distance between $\widetilde{\boldsymbol{w}}$ and $\boldsymbol{w}^{\mathsf{nps}}$:

$$\delta\left(\boldsymbol{g}\right) \;=\; \frac{1}{2} \cdot \left(\mathbf{1}_{n^{\mathsf{nps}} \times 1} - \boldsymbol{g}\right)^{\circ 2} \qquad . \tag{5.144}$$

A closed-form solution is available for this special case, such that the calibration weights can be determined by

$$\boldsymbol{g} \;= 1 + \boldsymbol{X}^{\mathsf{nps}} \left(\left(\boldsymbol{X}^{\mathsf{nps}}\right)^{\mathsf{T}} \mathbf{diag}\left(\boldsymbol{w}^{\mathsf{nps}}\right) \boldsymbol{X}^{\mathsf{nps}} \right)^{-1} \left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right) - \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{nps}}\right) \right)^{\mathsf{T}} \quad . \tag{5.145}$$

The estimator's name is due to the fact that the resulting totals $\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)$ can be written as a function of predictions and residuals from a regression model (cf. equation 5.14; Breidt and Opsomer, 2017, pp. 195 f; Cassel, Särndal and Wretman, 1976; Rupp, 2018, p. 20; Särndal, 2007, p. 103). Although frequently treated as a separate calibration technique (cf. e.g. Brick, 2013, p. 334; Loosveldt and Sonck, 2008, p. 94; Pedraza, Tijdens and Bustillo, 2007, p. 21; Steinmetz and Tijdens, 2009, p. 29), post-stratification is a special case of the GREG: by using an $\boldsymbol{X}$-matrix that only contains indicator variables (interaction terms) for the cross-classification of post-stratification variables, the weighted frequencies in cross tables of $\boldsymbol{X}$ are adjusted to those of the benchmark data set cal (cf. Zhang, 2000, p. 181).

Even though the GREG is the presumably most popular calibration approach, it suffers from certain limitations. In particular, the resulting calibration weights can be smaller than zero, making their justification in a design-based context difficult (cf. Deville and Särndal, 1992, p. 376; Fuller, 2002, p. 16; Park and Fuller, 2005, p. 85). Consequently, a number of alternatives are discussed in the relevant literature: Deville and Särndal (1992, p. 378) as well as Deville, Särndal and Sautory (1993, p. 1014) discuss different distance functions which lead to results that are asymptotically equivalent to the GREG, for example to limit the possible range of resulting correction weights (cf. also Münnich, Sachs and Wagner, 2012, p. 472). An additional benefit in this case is that the ratio of the largest to the smallest calibration weight, which is termed "Gelman factor" by Münnich et al. (2012, p. 27), can be restricted. Generally speaking, the MSE of resulting weighted estimators typically increases rapidly when this ratio becomes larger (cf. Gelman, 2007; Meng et al., 2009). To incorporate lower and upper boundaries for the weights, problem 5.143 can be extended by the additional inequality constraints

$$L_g \;\leq g_i \leq U_g \tag{5.146}$$

for all $i = 1, \ldots, n^{\mathsf{nps}}$, where $L_g, U_g \in \mathbb{R}$ respectively represent a single scalar that serves as lower and upper boundary for all values in $\boldsymbol{g}$. Deville and Särndal (1992, p. 378) include constraints 5.146 by means of a penalized distance function rather than adding inequality constraints to problem 5.143.

Based on these ideas, Estevao and Särndal (2000) demonstrate that different distance functions often lead to very similar results. Questioning the advantage of distance functions in general, they propose the *functional form approach*, which is also called "instrument vector approach" (Estevao and Särndal, 2006, p. 129; cf. Folsom and Singh, 2000; Guggemos and Tillé, 2010, pp. 3201 ff; Kott, 2003; 2006). The idea is to determine correction weights as a function $\boldsymbol{g} : \mathbb{R}^s \rightarrow \mathbb{R}^{n^{\mathsf{nps}}}$ of a vector of weighting parameters $\boldsymbol{\omega} \in \mathbb{R}^s$, which is of arbitrary given length $s \in \mathbb{N}$. In this formulation, definition 5.141 is re-written as a function $\widetilde{\boldsymbol{w}} : \mathbb{R}^s \rightarrow \mathbb{R}^{n^{\mathsf{nps}}}$:

$$\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}(\boldsymbol{\omega}) \coloneqq \boldsymbol{w}^{\mathsf{nps}} \circ \boldsymbol{g}(\boldsymbol{\omega}) \quad . \tag{5.147}$$

In this approach, the parameters $\boldsymbol{\omega}$ are determined exclusively from calibration constraints for $\boldsymbol{X}$, but the function $\boldsymbol{g}(\boldsymbol{\omega})$ can depend on variables $\boldsymbol{Z}$. With $\boldsymbol{g}(\boldsymbol{\omega})$ being defined by parameters the $\boldsymbol{\omega}$ and a matrix $\boldsymbol{Z}$, definition 5.147 provides a link between propensity and calibration weighting. In particular, $\boldsymbol{g}(\boldsymbol{\omega})$ can be defined as the reciprocals of some estimated propensity scores, such that the functional form of calibration weights is the same as for the propensity weights discussed in the previous section 5.2.1. For example, this is the case when choosing $\boldsymbol{g}(\boldsymbol{\omega}) = (\mathbf{m}(\boldsymbol{Z}^{\mathsf{nps}}, \boldsymbol{\omega}))^{\circ -1}$ to achieve coincidence with equations 5.134 and 5.138, such that the parameters $\boldsymbol{\omega}$ play the same role for calibration as for propensity weighting. Nevertheless, the optimization problems to determine these parameters with respect to calibration constraints are clearly different from fitting typical response propensity models (cf. e.g. section 5.1.3; Chang and Kott, 2008, p. 555; Estevao and Särndal, 2000, pp. 382 ff; Kott, 2006, pp. 134 ff).

Following these considerations, Folsom and Singh (2000) as well as Kott (2006) introduce the generalized raking

$$\boldsymbol{g}(\boldsymbol{\omega}) \coloneqq \exp(\boldsymbol{Z}^{\mathsf{nps}}\boldsymbol{\omega}) \tag{5.148}$$

and the logit model

$$\boldsymbol{g}(\boldsymbol{\omega}) \coloneqq 1 + \exp(-\boldsymbol{Z}^{\mathsf{nps}}\boldsymbol{\omega}) \tag{5.149}$$

in the context of calibration (cf. also Berkson, 1944; Deming and Stephan, 1940; Deville and Särndal, 1992, p. 378). The close link between functional form approach and propensity weighting is emphasized by equation 5.149. The reciprocal of the logistic function used for computing calibration weights is the same by which propensity weights are computed from a generalized linear logit model (cf. equations 5.21 and 5.138; Folsom and Singh, 2000, p. 591). Alternative concepts for bringing together propensity and calibration weighting are mainly based on a two-step procedure. In that case, response propensities are estimated in a first step, and the resulting propensity weights are then calibrated in a second step, i.e. by choosing $\boldsymbol{w}^{\mathsf{nps}} = 1/\widehat{\boldsymbol{p}}^{\mathrm{nps}}$ in definition 5.141 (cf. e.g. Enderle, Münnich and Bruch, 2013, p. 94; Lee and Valliant, 2009, p. 335; Särndal and Lundström, 2005, pp. 51 f; Valliant and Dever, 2011, p. 109).

For the calibration approaches discussed so far, the benchmark data set cal ideally corresponds to the whole population. The calibration targets indeed may come from known population information, e.g. from registers, but, especially in the context of non-probability samples, it is more common to use calibration benchmarks that are subject to uncertainty themselves (cf. e.g. Bethlehem, 2008b, p. 34; Schouten, 2007, p. 55; Steinmetz et al., 2014, pp. 282 f). For example, this is the case when cal is a probability sample rather than the full population. In such settings, as well as when using a large number of calibration variables, enforcing exact compliance of weighted estimates and benchmarks may be impossible, unreasonable from a theoretical point of view or lead to highly instable

weights (cf. Burgard, Münnich and Rupp, 2020, p. 12; Deville and Särndal, 1992, p. 380; Deville, Särndal and Sautory, 1993, p. 1015; Guggemos and Tillé, 2010, p. 3199). As a consequence, further extensions to the ideas discussed above concern the quality of the calibration targets determined from cal.

To relax exact calibration, there is a growing discussion on methods that apply *soft calibration constraints*, where exact compliance (as stated in equation 5.143) is substituted by a sufficient closeness to the calibration targets. Chang and Kott (2008, p. 557) suggest solving

$$\begin{aligned} &\underset{\boldsymbol{\omega},\boldsymbol{\epsilon}}{\operatorname{argmin}} \left(\sum_{k=1}^{p} \mathrm{v}_k \cdot \frac{(1-\epsilon_k)^2}{2} \right) \\ &\text{s. t.} \qquad \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}} \left(\widetilde{\boldsymbol{w}} \right) \;=\; \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}} \left(\boldsymbol{w}^{\mathsf{cal}} \right) \circ \boldsymbol{\epsilon}^{\mathsf{T}} \end{aligned} \tag{5.150}$$

for the previously defined raking or logit model, such that calibration weights $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}(\boldsymbol{\omega})$ depend on $\boldsymbol{\omega}$ and are, thus, based on the functional form approach (cf. equations 5.147 to 5.149). In problem 5.150, $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 & \dots & \epsilon_p \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^p$ represents a vector of multiplicative error terms to be minimized. These terms quantify the deviations of the estimated from the benchmark totals for all p calibration variables. Denoted by $\mathbf{v} \in \mathbb{R}^p$ is a corresponding vector of predefined importance weights to combine these errors into a weighted sum (cf. also Jahn, 2011, pp. 292 ff). In contrast, the 'penalized calibration' method proposed by Guggemos and Tillé (2010) does not apply the functional form approach and is stronger related to the GREG (cf. definitions 5.141 and 5.147). The optimization problem in this case is

$$\begin{aligned} &\underset{\boldsymbol{g},\boldsymbol{\epsilon}}{\operatorname{argmin}} \left(\sum_{j=1}^{n^{\mathsf{nps}}} w_j^{\mathsf{nps}} \cdot \frac{(1-g_j)^2}{2} \;+\; \sum_{k=1}^{p} \mathrm{v}_k \cdot \frac{(1-\epsilon_k)^2}{2} \right) \\ &\text{s. t.} \qquad \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}} \left(\widetilde{\boldsymbol{w}} \right) \;=\; \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}} \left(\boldsymbol{w}^{\mathsf{cal}} \right) \circ \boldsymbol{\epsilon}^{\mathsf{T}} \\ &\qquad\qquad L_{\epsilon_k} \;\leq\; \epsilon_k \;\leq\; U_{\epsilon_k} \quad \text{for all} \quad k \;=\; 1, \dots, p \qquad . \end{aligned} \tag{5.151}$$

In problem 5.151, a vector of error multipliers $\boldsymbol{\epsilon}$ is used as before, but its values are constrained by lower and upper boundaries $L_{\epsilon_k}, U_{\epsilon_k} \in \mathbb{R}$ for all $k = 1, \dots, p$. These boundaries are used by Guggemos and Tillé (2010, p. 3204) to enforce exact calibration as in equality 5.142 by setting $L_{\epsilon_k} = U_{\epsilon_k} = 1$ for some auxiliary variables. For other variables, the calibration errors are left completely unconstrained, which results in penalizing deviations from the corresponding calibration benchmarks as in problem 5.150 and adding these penalties to the GREG's distance function (cf. equality 5.144). Burgard, Münnich and Rupp (2019, p. 5) as well as Rupp (2018, p. 126) extend these ideas and consider box-constraints defined by arbitrary boundaries L_{ϵ_k} and U_{ϵ_k} for the calibration errors. This idea is based on the work of Münnich, Sachs and Wagner (2012, p. 473) as well as Wagner (2013, pp. 102 ff), where the approach described in equation 5.151 is supplemented by arbitrary boundaries $L_{\widetilde{w}_i}, U_{\widetilde{w}_i} \in \mathbb{R}$ for the weights, such that $L_{\widetilde{w}_i} \leq \widetilde{w}_i \leq U_{\widetilde{w}_i}$ for all $i \in \mathcal{S}^{\mathsf{nps}}$. In case of these two latter extensions, efficient computational implementations for solving the calibration problems are proposed.

Note that the present notation is based on the work of Burgard, Münnich and Rupp (2019), Münnich, Sachs and Wagner (2012), Rupp (2018) as well as Wagner (2013). Aim of the current formulation is to provide a coherent description of the various calibration approaches as well as to make the problem independent from different scales of $\boldsymbol{X}$-variables. For most references cited above, the adaptation merely concerns the unification

of mathematical symbols. The notation of problems 5.150 and 5.151 does, however, differ slightly from the original ones used by Guggemos and Tillé (2010) as well as Chang and Kott (2008). These authors use weighted absolute instead of relative deviances from the calibration targets and have a clearer depiction of interactions between calibration errors. However, their representation can be equivalently reformulated as outlined above by appropriate selection of importance weights and interaction effects in the $\boldsymbol{X}$-matrix.

Except for the GREG, there is usually no (general) analytical solution for the presented calibration methods. As a consequence, numerical solutions based on the methods discussed in chapter 4 are typically used (cf. e.g. Chang and Kott, 2008, pp. 558 f; Deville and Särndal, 1992; Deville, Särndal and Sautory, 1993; Guggemos and Tillé, 2010; Kott, 2006, p. 141; Münnich, Sachs and Wagner, 2012; Särndal, 2007, p. 106). Introducing a framework for unification and extension of the pseudo-design-based methods discussed so far, a detailed formulation of a possible numerical solution strategy is provided as part of the following section 5.2.3.

5.2.3 Calibrated Semi-parametric Artificial Neural Networks

As summarized in the previous sections 5.2.1 and 5.2.2, the pseudo-design-based framework for estimation from non-probability samples encompasses methods for propensity and calibration weighting. A deeper integration of these ideas appears therefore desirable to establish a comprehensive weighting framework for non-probability samples (cf. e.g. Baker et al., 2010, p. 47; Kott, 2006, p. 564; Lee and Valliant, 2009, p. 341; Valliant and Dever, 2011, p. 109). Such an integration is proposed in the current section 5.2.3 and supplemented by particular extensions, which are motivated below.

Based on the functional form approach (cf. equation 5.147 and the related discussion), different proposals for a synthesis of propensity and calibration weighting exist (cf. e.g. Chang and Kott, 2008; Estevao and Särndal, 2006; Kim, Kwon and Paik, 2016; Kim and Park, 2010; Kott, 2006). However, all current methods that can be used to obtain pseudo-design weights for non-probability samples suffer from one or more of the following drawbacks that are already discussed in the scientific debate (cf. e.g. Chang and Kott, 2008, pp. 556 ff; Deville, Särndal and Sautory, 1993, pp. 1014 ff; Guggemos and Tillé, 2010, p. 3199; Kott and Liao, 2017, p. 161; Kott, 2006, pp. 136 ff):

a) the variables that determine the response propensities need to be observed in the non-probability and a reference sample,

b) a solution can only be found for certain relations between the numbers of calibration constraints and response model variables,

c) the range of the resulting weights cannot be restricted,

d) compliance with calibration benchmarks is necessarily exact, which imposes strong assumptions with regard to their accuracies, and/or

e) response and calibration model are not integrated but used sequentially.

A further shortcoming is hardly ever explicitly discussed in the academic literature: pseudo-design weights should ideally provide unbiasedness with respect to the whole distribution, not only its first moments (cf. equation 5.131). Hence, a weighting procedure should be able to consider aspects of the relevant distributions beyond means and totals (cf. Lenau and Münnich, 2017, pp. 62 f) since there can also be biases in higher moments and other aspects of the distribution (cf. Elliott and Valliant, 2017, p. 262; Groves and

Couper, 1998, pp. 10 f; Schouten, 2007, p. 67; Weisberg, 2005, p. 190). However, calibration constraints are usually limited to means or totals, even though the statistics to be estimated from most samples are not (cf. sections 2.2 and 5.2.2). When considering typical uses of non-probability samples, measures of association are of substantial relevance. In particular, covariances and correlations are common statistics of interest and at the same time serve as foundation for various more complex statistics and models (cf. chapter 2 and section 5.1; Baker et al., 2013a, p. 22; Groves and Couper, 1998, pp. 10 f; Japec et al., 2015, p. 850). In many cases, it is simply assumed that there is no bias in covariances or correlations (cf. e.g. Andridge et al., 2019, p. 1481; Pasek, 2016, p. 283; Steinmetz, Tijdens and Pedraza, 2009, p. 27) or that total calibration adjusts for these and other aspects of the distribution as well (cf. e.g. Rubin, 1979, p. 319; Schouten, 2007, p. 67).

Since these assumptions are not necessarily valid, it seems worthwhile to consider additional types of calibration constraints to better account for possible selection biases. Such an approach is proposed as part of the following discussion. To integrate the different ideas of pseudo-design weights for non-probability samples and overcome the limitations discussed above, a new *calibrated response model* is introduced. The structural form implied for the pseudo-design weights is outlined first, followed by the specification of calibration constraints. These constraints allow for soft and exact calibration not only of totals but also of covariances or correlations, which is motivated by the above-mentioned relevance of these quantities for many applications of non-probability sampling. Considering a rather general distance function, a fitting procedure based on the methods summarized in chapter 4 is then introduced. The discussion concludes with illustrating the proposed method's potential to integrate many existing weighting methods as special cases. Yet, the suggested approach is not limited to these special cases and provides a number of extensions.

Structure of the Model

As a foundation for integrating pseudo-design-based ideas for non-probability samples, the functional form approach establishes a close relation between propensity and calibration weighting. For both of these approaches, the respective pseudo-design weights $\widetilde{\boldsymbol{w}}$ are determined through multiplying some (often non-informative) prior weights $\boldsymbol{w}^{\mathsf{nps}}$ by a factor $\boldsymbol{g} \coloneqq \boldsymbol{g}(\boldsymbol{\omega})$. It is defined as a function of weighting parameters $\boldsymbol{\omega} \in \mathbb{R}^s$, often with regard to some response model variables $\boldsymbol{Z}$ (cf. equation 5.147). Therefore, $\boldsymbol{g}(\boldsymbol{\omega})$ can be interpreted as a vector of model predictions or their reciprocals. Various strategies for choosing the form of this function $\boldsymbol{g}$ to determine the pseudo-design weights are available. In principle, these options include all types of model specifications presented in section 5.1, as long as they are suitable for binary dependent variables and allow predicting the corresponding probabilities (cf. sections 5.2.1 and 5.2.2).

To make a choice for $\boldsymbol{g}$ that is useful for integrating propensity and calibration weighting, it is necessary to consider the differences between both strategies, which primarily concern the computation of $\boldsymbol{\omega}$. As discussed in section 5.2.1, response propensity models rely on describing the data generating process of the non-probability sample, while calibration weighting determines weights such that they align the non-probability sample with externally obtained benchmarks (cf. section 5.2.2). To combine both ideas, these considerations suggest an integration and trade-off between modeling the sample inclusion and relying on calibration constraints to determine the weights. In any case, it appears sensible to allow for restrictions in the possible range of the resulting weights (cf. Burgard, Münnich and Rupp, 2019, p. 4; Kott, 2006, p. 142; Little, 1986, p. 147).

Most of the models discussed in section 5.1 that are suitable for propensity modeling are fit by means of gradient-based optimization. At the same time, the general calibration and box-constraints discussed in section 5.2.2 can be non-linear in the parameters. For fitting a calibrated response model, optimization by means of sequential quadratic programming (cf. section 4.2) is therefore an apparent choice, and similar numerical strategies are applied for the majority of existing calibration methods summarized in section 5.2.2 (cf. e.g. Chang and Kott, 2008, p. 558; Deville and Särndal, 1992, p. 380; Guggemos and Tillé, 2010, p. 3210). However, some of the models presented in section 5.1 are not compatible with `SQP`. These models on the one hand encompass MARS, which only partially rely on gradient information for optimization. On the other hand, the parameters for SVMs are determined from the support vectors alone, while the nature of calibration constraints requires jointly considering all elements of the non-probability sample.

To propose a class of calibrated response models that is as flexible as possible and feasible for the gradient-based optimization methods presented in section 4.2.2, the subsequent discussion formally introduces *calibrated semi-parametric artificial neural networks*. Such ANNs are versatile and therefore allow integrating various established weighting methods, an advantage which is illustrated after elaborating the details of the proposed method, at the end of the current section 5.2.3. For establishing the weighting model, let $\boldsymbol{\omega}$ represent the vectorized neural network coefficients as in equation 5.80. Combining the functional form approach for calibration with propensity weighting, the pseudo-design weights are defined in correspondence to equations 5.138 and 5.147:

$$\widetilde{\boldsymbol{w}} \; = \widetilde{\boldsymbol{w}}\left(\boldsymbol{\omega}\right) \coloneqq \boldsymbol{w}^{\mathsf{nps}} \circ \boldsymbol{g}\left(\boldsymbol{\omega}\right) \; = \boldsymbol{w}^{\mathsf{nps}} \circ \left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}\left(\boldsymbol{\omega}\right)\right)^{\circ(-1)} \qquad , \tag{5.152}$$

where $\boldsymbol{g}\left(\boldsymbol{\omega}\right) \coloneqq \left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}\left(\boldsymbol{\omega}\right)\right)^{\circ(-1)}$ is a vector of inverse estimated propensities obtained from an artificial neural network (cf. sections 5.1.8 and 5.1.9). These predictions are expressed as a function $\widehat{\boldsymbol{p}}^{\mathrm{nps}} : \mathbb{R}^{s} \to \mathbb{R}^{n^{\mathsf{nps}}}$ of the weighting parameters $\boldsymbol{\omega}$. Similar as in linear probability models, the codomain of $\widehat{\boldsymbol{p}}^{\mathrm{nps}}$ is $\mathbb{R}^{n^{\mathsf{nps}}}$ rather than $[0,1]^{n^{\mathsf{nps}}}$ (cf. Greene, 2008, pp. 772 ff; Wooldridge, 2012, pp. 248 ff). This definition allows for general real-valued pseudo-design weights and achieves coherence with the definitions of $\boldsymbol{g}$ and $\widetilde{\boldsymbol{w}}$ in section 5.2.2. For example, this is required for expressing the GREG as a special case of formulation 5.152. In the end, pseudo-design weights are treated as a function $\widetilde{\boldsymbol{w}} : \mathbb{R}^{s} \to \mathbb{R}^{n^{\mathsf{nps}}}$ of the weighting parameters. These parameters are then determined such that calibration constraints depending on $\widetilde{\boldsymbol{w}}$ are fulfilled.

Such constraints, which are expressed for calibration variables $\boldsymbol{X}$ as before, define the adjustment of weighted estimates in the non-probability sample nps to those in the calibration benchmark data set cal. To e.g. allow for targets of different quality (cf. section 5.2.2), exact as well as soft calibration are considered. In correspondence to problems 5.150 and 5.151, the total calibration constraints are constituted by

$$\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right) \; \overset{!}{=} \; \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right) \circ \boldsymbol{\epsilon}^{\mathsf{T}} \tag{5.153}$$

(cf. Chang and Kott, 2008, p. 557; Guggemos and Tillé, 2010, p. 3204; Burgard, Münnich and Rupp, 2019, p. 5). As before, $\boldsymbol{\epsilon} \in \mathbb{R}^{p}$ is a vector of multiplicative calibration errors that quantifies the deviations of the estimates from the benchmark values.

As motivated above, calibration constraints for covariance and correlation matrices are additionally introduced. Note that calibration of first as well as second moments may in principle be achieved by using only total constraints in form of equation 5.153 if main and interaction terms are included in the matrix of calibration variables $\boldsymbol{X}$. Maximum likelihood covariance estimates, which are a function of estimated first and second moments, can be calibrated in that way. The same argument holds for correlations when all relevant (co-)variances are calibrated in this manner. Post-stratification or raking to cross-tables are examples where this is the case (cf. appendix B.5.2; Lenau and Münnich, 2017, pp. 62 f). However, this strategy suffers from major limitations in the present context. On the one hand, errors of main as well as interaction terms (different columns in $\boldsymbol{X}$) have an influence on the error of a calibrated covariance or correlation. Consequently, there is no straightforward extension for soft calibration of these quantities in analogy to equation 5.153. On the other hand, calibration of covariances by means of total constraints is possible only for ML estimation of (co-)variances but not for the bias corrected estimates. Although the magnitude of bias when using ML estimates is negligible for large simple random samples (cf. equation 2.18f), it generally depends on the pseudo-design weights themselves in the weighted case. Therefore, calibration of the corrected covariance estimates appears preferable. A third issue when including interaction or squared terms in total calibration can be numerical instability, which is illustrated in section 6.2.

Calibration constraints for covariance and correlation matrices are therefore imposed autonomously and in analogy to equation 5.153. As these matrices are symmetric, their number of unique elements is given by $r = \sum_{l=1}^{p} l = (p^2 + p)/2$, with diagonal entries being all ones for correlation matrices. Using a function $\mathbf{vec} : \mathbb{R}^{p \times p} \to \mathbb{R}^r$ that returns the vector of unique elements in these matrices, the corresponding calibration equations are

$$\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right) \overset{!}{=} \mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right) \circ \boldsymbol{\varepsilon}^{\mathsf{T}} \tag{5.154a}$$

$$\mathbf{vec}\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right) \overset{!}{=} \mathbf{vec}\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right) \circ \boldsymbol{\varepsilon}^{\mathsf{T}} \tag{5.154b}$$

$$\mathbf{vec}\left(\widehat{\boldsymbol{\rho}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right) \overset{!}{=} \mathbf{vec}\left(\widehat{\boldsymbol{\rho}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right) \circ \boldsymbol{\varepsilon}^{\mathsf{T}} \tag{5.154c}$$

for ML and bias corrected covariance as well as correlation estimates, respectively. In analogy to $\boldsymbol{\epsilon}$ in equation 5.153, a vector $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ is used to quantify the deviations of estimates from benchmarks, i.e. it represents the multiplicative calibration errors for covariances or correlations. For the sake of notational brevity, the three cases of equations 5.154 are only distinguished in the following discussion where required for mathematical correctness. In all other cases, the corresponding equations and references to these are interchangeable.

Note that $\boldsymbol{\epsilon}$ and $\boldsymbol{\varepsilon}$ can be interpreted as vectors containing ratios of estimated to benchmark quantities that are used for calibration. When $\epsilon_k = 1$ or $\varepsilon_l = 1$, the corresponding calibration target is met exactly. In case of $\epsilon_k < 1$ or $\varepsilon_l < 1$, there is an underestimation of the target quantity, while $\epsilon_k > 1$ or $\varepsilon_l > 1$ indicate overestimation. To limit the possible degree of deviation, these values can be restricted by box-constraints in the form of $L_{\epsilon_k} \leq \epsilon_k \leq U_{\epsilon_k}$ and $L_{\varepsilon_l} \leq \varepsilon_l \leq U_{\varepsilon_l}$ for some or all constraints. Setting both the lower and upper boundary to one enforces exact calibration for the respective quantity (cf. Burgard, Münnich and Rupp, 2019, p. 5; Münnich, Sachs and Wagner, 2012, p. 473; Rupp, 2018, p. 126; Wagner, 2013, p. 102). In the following discussion, L_h and U_h generally denote prespecified lower and upper boundaries for arbitrary parameters h.

Whether a distance function should be used to determine calibration weights is subject to scientific debates. To achieve a general weighting model that is able to integrate a variety of existing pseudo-design-based approaches, such a function nevertheless has to be applied because many of these existing methods do rely on distance measures (cf. sections 5.2.1 and 5.2.2). With regard to these established methods, different reasons to use a distance function can be identified, which determine three main components that appear worth considering when fitting a calibrated response model:

a) All models presented in section 5.1 that are suitable and commonly used in the context of propensity weighting, such as GLMs and GAMs, make use of a loss function to represent the estimated propensities' quality. Since the dependent variable in these models is typically the binary sample inclusion indicator, a predominant choice in response models is to use the binomial log-likelihood as a special case of the cross-entropy (or deviance; cf. sections 5.1 and 5.2.1; Skinner et al., 2009, p. 6; Valliant and Dever, 2011, p. 133).

b) Following the discussion in section 5.2.2, the application of soft calibration constraints is motivated above. This concept urges the need for measuring closeness to the calibration targets when determining the weights, such that deviations of the estimates from their corresponding benchmark values are typically included in the distance metric. A weighted sum of the squared relative deviations is a popular choice for this purpose. Including such a component in the distance function is of particular importance for fitting a weighting model by using only calibration constraint, e.g. in absence of a reference sample to be used for the above component ***a)*** (cf. problems 5.150 and 5.151; equations 5.153 and 5.154; Chang and Kott, 2008; Guggemos and Tillé, 2010; Wagner, 2013, p. 102; Rupp, 2018, p. 117).

c) Some weighting methods include a squared loss component that penalizes the deviation of parameters $\boldsymbol{\omega}$ from a vector of constants (cf. equations 5.144 and 5.151). This strategy is closely related to the shrinkage methods for model fitting presented in section 5.1.11 and can be used to reduce the variability of the resulting weights. It is particularly relevant to achieve a unique solution for weighting methods that employ a large number of parameters, e.g. in case of the GREG where this number equals the sample size (cf. Deville and Särndal, 1992, p. 377).

The most common approach to consider such different types of criteria for optimization is by combining them using a weighted sum (cf. Ehrgott, 2005, pp. 55 ff; Jahn, 2011, pp. 292 ff; Marler and Arora, 2004, p. 375). This idea is frequently employed for calibration (cf. e.g. problems 5.150 and 5.151; Chang and Kott, 2008, p. 557; Guggemos and Tillé, 2010, p. 3204; Münnich, Sachs and Wagner, 2012, p. 473; Burgard, Münnich and Rupp, 2019, p. 5) and adopted in the present context as well. To that end, a vector of predetermined importance weights $\mathbf{v} \in \mathbb{R}_{\geq 0}^{s+p+r+1}$ is introduced to achieve a scalar-valued distance function for optimization that combines the above components ***a)*** to ***c)*** and allows controlling each of these components' relevance. Based on the outlined considerations, the distance function $\delta : \mathbb{R}^s \times \mathbb{R}^p \times \mathbb{R}^r \to \mathbb{R}_{\geq 0}$ for fitting the proposed calibrated semi-parametric artificial neural networks is hence

$$\begin{aligned} \delta\left(\boldsymbol{\omega}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}\right) \;=\; & \mathrm{v}_1 \cdot \delta_{\mathrm{m}}\left(\boldsymbol{\omega}\right) \;+\; \sum_{j=1}^{s} \mathrm{v}_{(j+1)} \cdot \left(\omega_j - \mathrm{C}_{\omega_j}\right)^2 \Big/ 2 \\ & + \sum_{k=1}^{p} \mathrm{v}_{(s+k+1)} \cdot \left(\epsilon_k - 1\right)^2 \Big/ 2 \;+\; \sum_{l=1}^{r} \mathrm{v}_{(s+p+l+1)} \cdot \left(\varepsilon_l - 1\right)^2 \Big/ 2 \quad . \end{aligned} \tag{5.155}$$

The first component ***a)*** is constituted by the ANN's distance function $\delta_{\mathrm{m}}(\boldsymbol{\omega})$, which is defined in equation 5.82. It is used to incorporate a propensity model's loss function and only meaningful when reference data for modeling the selection process is available that provides observations outside the non-probability sample, just as for common propensity models (cf. section 5.2.1). In cases where no such observations are available, this component can be disregarded in equation 5.155 by setting $\mathrm{v}_1 = 0$.

The second component ***b)*** is incorporated by means of the second and third sum (second row) in equation 5.155. These sums are used to penalize deviations from total- and covariance or correlation benchmarks, which are respectively expressed by $\boldsymbol{\epsilon}$ and $\boldsymbol{\varepsilon}$. Since calibration targets are met exactly when $\boldsymbol{\epsilon}$ and $\boldsymbol{\varepsilon}$ are vectors of ones, their respective values are penalized for deviating from ones to be as close as possible to the benchmarks while considering the other parts of the distance function (cf. equations 5.153 and 5.154).

In a similar manner, the third component ***c)*** is represented by the first sum in equation 5.155. It is used for penalizing the distance between weighting parameters $\boldsymbol{\omega}$ and a vector of centering constants $\mathbf{C}_{\boldsymbol{\omega}} \in \mathbb{R}^s$, with element $\mathrm{C}_{\omega_j} \coloneqq [\mathbf{C}_{\boldsymbol{\omega}}]_j$ corresponding to the constant for parameter ω_j. Adequate choices for these constants depend on the selected functional form for $\boldsymbol{g}$ (i.e. the ANN) and can incorporate potential assumptions about the response process. Adapting the discussion in section 5.1.11, a non-informative option for $\mathrm{C}_{\boldsymbol{\omega}}$ can be a vector of weighting parameters that leads to constant weights. For example, the centering constants can be chosen to penalize deviation from the inverse sampling fraction, i.e. such that $\widetilde{\boldsymbol{w}}(\mathbf{C}_{\boldsymbol{\omega}}) = \mathrm{N}/n^{\mathsf{nps}} \cdot \mathbf{1}_{n^{\mathsf{nps}} \times 1}$ assigns each element the inverse average response propensity (cf. equation 5.152; Folsom and Singh, 2000, p. 599). Penalization of intercept parameter(s) can be avoided by setting the respective importance weights to zero (cf. Hastie, Tibshirani and Friedman, 2008, p. 64).

Depending on the choice of the importance weights $\mathbf{v}$ and the centering values $\mathbf{C}_{\boldsymbol{\omega}}$, different types of distance functions evolve from equation 5.155. Important special cases are discussed at the end of the current section 5.2.3. While importance weights $\mathbf{v}$ are assumed to be predetermined in the present context, more detailed considerations and illustrations regarding the choice of these weights are discussed in section 6.2. In any case, this distance function is minimized by means of constrained optimization to fit the weighting model. A detailed description of the fitting procedure is provided below.

Fitting Calibrated Semi-parametric Artificial Neural Networks

To summarize the preceding considerations, the objective is to determine pseudo-design weights $\widetilde{\boldsymbol{w}}$ for the non-probability sample by fitting a calibrated response model. To that end, an optimal vector of model parameters

$$\boldsymbol{\Theta} \;=\; \left[\boldsymbol{\omega}^{\mathsf{T}}\; \boldsymbol{\epsilon}^{\mathsf{T}}\; \boldsymbol{\varepsilon}^{\mathsf{T}}\right]^{\mathsf{T}} \quad \in \quad \mathbb{R}^u \tag{5.156}$$

needs to be determined. It includes the vector of weighting parameters $\boldsymbol{\omega} \in \mathbb{R}^s$ in conjunction with the vectors of calibration ratios for totals and covariances (or correlations), respectively denoted by $\boldsymbol{\epsilon} \in \mathbb{R}^p$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^r$. Consequently, $u = s + p + r$ denotes the cumulated number of weighting parameters, total and covariance constraints. The following overview describes the numerical strategy that is used to determine these parameters $\boldsymbol{\Theta}$.

Following the definition of Deville and Särndal (1992, p. 376), $\boldsymbol{\Theta}$ is found by minimizing a distance function under constraints, as outlined in the previous paragraphs (cf. equalities 5.152 to 5.155). The optimization problem is hence defined by

$$\begin{aligned}
\boldsymbol{\Theta}^* \; &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left(\delta\left(\boldsymbol{\omega}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}\right)\right) \\
&\text{s. t.} \quad \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right) \;=\; \widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right) \circ \boldsymbol{\epsilon}^{\mathsf{T}} \\
&\qquad\; \mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right) \;=\; \mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right) \circ \boldsymbol{\varepsilon}^{\mathsf{T}} \\
&\qquad\; L_{\omega_j} \;\leq \omega_j \;\leq U_{\omega_j} \quad \text{for all} \quad j \;= 1, \ldots, s \\
&\qquad\; L_{\epsilon_k} \;\leq \epsilon_k \;\leq U_{\epsilon_k} \quad \text{for all} \quad k \;= 1, \ldots, p \\
&\qquad\; L_{\varepsilon_l} \;\leq \varepsilon_l \;\leq U_{\varepsilon_l} \quad \text{for all} \quad l \;= 1, \ldots, r \qquad ,
\end{aligned} \tag{5.157}$$

taking into account soft calibration constraints for totals and covariances as well as box-constraints for all optimization parameters. Modifications for considering correlation and bias-corrected covariance estimates are discussed below. Solving problem 5.157 represents fitting a calibrated semi-parametric artificial neural network. In this context, the distance function δ and the pseudo-design weights $\widetilde{\boldsymbol{w}}$ can take different forms, but the weights are generally determined by inverse participations propensities $\widehat{\boldsymbol{p}}^{\,\mathrm{nps}}\left(\boldsymbol{\omega}\right)$ predicted from an ANN (cf. definitions 5.152 and 5.155).

As indicated above, a numerical solution via sequential quadratic programming seems advantageous in this case. An important reason is that ANNs in general allow for a quite modifiable model structure and, thus, rely heavily on utilizing the chain rule to optimize parameters. As a consequence, there is no general formulation for Hessian matrices when considering ANNs of arbitrary structures. The prevalent fitting technique for ANNs (backpropagation) is therefore solely based on gradient information and cannot account for potentially non-linear constraints. Nevertheless, neural networks can alternatively be fit by Quasi-Newton approaches. In particular, it is straightforward to apply the `BFGS` method (cf. algorithms 5 and 10) for fitting any ANN that is feasible for backpropagation because the required first differences in parameters and gradient must be available for gradient descent methods as well (cf. section 5.1.8; Bishop, 1995, pp. 287 ff; Hagan et al., 1996, p. 9-10 ff). An approximated rather than exact Hessian matrix is used for optimization in this case, for which `SQP` typically offers numerical advantages in comparison to other constrained optimization methods (cf. section 4.2; Boggs and Tolle, 1995; Geiger and Kanzow, 2002, pp. 256 ff; Jarre and Stoer, 2004, p. 337).

For applying sequential quadratic programming as defined in algorithm 8, problem 5.157 is rewritten more compactly in the form of problem 4.10:

$$\begin{aligned}
\boldsymbol{\Theta}^* \;&=\; \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left(\delta\left(\boldsymbol{\Theta}\right)\right) \\
&\qquad \text{s. t.} \quad \bar{\mathbf{g}}\left(\boldsymbol{\Theta}\right) = \mathbf{0}_{(p+r)\times 1} \\
&\qquad\qquad\; \widetilde{\mathbf{g}}\left(\boldsymbol{\Theta}\right) \leq \mathbf{0}_{u\times 1} \qquad ,
\end{aligned} \tag{5.158}$$

where the relational operators are again applied element-wise. The components of problem 5.158 and their respective Jacobian matrices required for optimization are discussed below.

Corresponding to the compact notation of problem 5.158, lower- and upper boundaries as well as centering constants $L_{\Theta_j}, U_{\Theta_j}, \mathrm{C}_{\Theta_j}$ for all $j = 1, \ldots, u$ elements of $\boldsymbol{\Theta}$ are denoted

by vectors

$$\begin{aligned} \boldsymbol{L}_{\boldsymbol{\Theta}} &:= \begin{bmatrix} \boldsymbol{L}_{\boldsymbol{\omega}}^{\mathsf{T}} & \boldsymbol{L}_{\boldsymbol{\epsilon}}^{\mathsf{T}} & \boldsymbol{L}_{\boldsymbol{\varepsilon}}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} & \in & \mathbb{R}^{u} \\ \boldsymbol{U}_{\boldsymbol{\Theta}} &:= \begin{bmatrix} \boldsymbol{U}_{\boldsymbol{\omega}}^{\mathsf{T}} & \boldsymbol{U}_{\boldsymbol{\epsilon}}^{\mathsf{T}} & \boldsymbol{U}_{\boldsymbol{\varepsilon}}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} & \in & \mathbb{R}^{u} \\ \mathbf{C}_{\boldsymbol{\Theta}} &:= \begin{bmatrix} \mathbf{C}_{\boldsymbol{\omega}}^{\mathsf{T}} & \mathbf{1}_{\mathbf{1}\times \boldsymbol{p}} & \mathbf{1}_{\mathbf{1}\times \boldsymbol{r}} \end{bmatrix}^{\mathsf{T}} & \in & \mathbb{R}^{u} \quad . \end{aligned} \tag{5.159}$$

Using importance weights $\mathbf{v} \in \mathbb{R}_{\geq 0}^{u+1}$ as before, the distance function to be minimized is thus defined by

$$\delta\left(\boldsymbol{\Theta}\right) := \mathrm{v}_1 \cdot \delta_{\mathrm{m}}\left(\boldsymbol{\omega}\right) + \frac{1}{2} \cdot \mathbf{v}_{\mathcal{J}}^{\mathsf{T}} \left(\boldsymbol{\Theta} - \mathbf{C}_{\boldsymbol{\Theta}}\right)^{\circ 2} \tag{5.160}$$

as in equation 5.155, using $\mathcal{J} = \{2, \ldots, u+1\}$ to subset all elements of $\mathbf{v}$ but the first. The corresponding Jacobian matrix is

$$\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}\right) = \mathrm{v}_1 \cdot \begin{bmatrix} \mathbf{J}_{\delta_{\mathrm{m}}}\left(\boldsymbol{\omega}\right) & \mathbf{0}_{1\times p} & \mathbf{0}_{1\times r} \end{bmatrix} + \left(\mathbf{v}_{\mathcal{J}} \circ \left(\boldsymbol{\Theta} - \mathbf{C}_{\boldsymbol{\Theta}}\right)\right)^{\mathsf{T}} \quad . \tag{5.161}$$

As above, $\delta_{\mathrm{m}}\left(\boldsymbol{\omega}\right)$ is the distance function of the artificial neural network defined in equation 5.82, and its Jacobi matrix $\mathbf{J}_{\delta_{\mathrm{m}}}\left(\boldsymbol{\omega}\right)$ is determined by equation 5.83.

Furthermore, the pseudo-design weights are defined by $\widetilde{\boldsymbol{w}}\left(\boldsymbol{\omega}\right) = \boldsymbol{w}^{\mathsf{nps}} \circ \left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}\left(\boldsymbol{\omega}\right)\right)^{\circ(-1)}$ (cf. definition 5.152). Since $\widehat{\boldsymbol{p}}^{\mathrm{nps}}\left(\boldsymbol{\omega}\right)$ is the vector of response propensities predicted from the artificial neural network, it holds that $\mathbf{J}_{\widehat{\boldsymbol{p}}^{\mathrm{nps}}}\left(\boldsymbol{\omega}\right)$ is the Jacobian matrix of ANN predictions defined in equation 5.84. Therefore, the Jacobian matrix of $\widetilde{\boldsymbol{w}}$ can be obtained as

$$\mathbf{J}_{\widetilde{\boldsymbol{w}}}\left(\boldsymbol{\omega}\right) = -\left(\mathbf{1}_{1\times s} \otimes \left(\boldsymbol{w}^{\mathsf{nps}} \circ \left(\widehat{\boldsymbol{p}}^{\mathrm{nps}}\left(\boldsymbol{\omega}\right)\right)^{\circ(-2)}\right)\right) \circ \mathbf{J}_{\widehat{\boldsymbol{p}}^{\mathrm{nps}}}\left(\boldsymbol{\omega}\right) \quad , \tag{5.162}$$

which can be used to incorporate inequality and equality constraints. These are respectively constituted by the functions $\widetilde{\mathbf{g}} : \mathbb{R}^{u} \to \mathbb{R}^{2 \cdot u}$ and $\overline{\mathbf{g}} : \mathbb{R}^{u} \to \mathbb{R}^{p+r}$, which are defined as

$$\widetilde{\mathbf{g}}\left(\boldsymbol{\Theta}\right) := \begin{bmatrix} \boldsymbol{L}_{\boldsymbol{\Theta}} \\ \boldsymbol{\Theta} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Theta} \\ \boldsymbol{U}_{\boldsymbol{\Theta}} \end{bmatrix} \tag{5.163}$$

and

$$\overline{\mathbf{g}}\left(\boldsymbol{\Theta}\right) := \begin{bmatrix} \left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right)^{\mathsf{T}} \\ \left(\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right)\right)^{\mathsf{T}} \end{bmatrix} - \begin{bmatrix} \left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right)^{\mathsf{T}} \\ \left(\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right)\right)^{\mathsf{T}} \end{bmatrix} \circ \begin{bmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\varepsilon} \end{bmatrix} \quad . \tag{5.164}$$

The corresponding Jacobi matrix of the box-constraints $\widetilde{\mathbf{g}}\left(\boldsymbol{\Theta}\right)$ is given by

$$\mathbf{J}_{\widetilde{\mathbf{g}}}\left(\boldsymbol{\Theta}\right) = \begin{bmatrix} -\mathbf{I}_u \\ \mathbf{I}_u \end{bmatrix} \quad . \tag{5.165}$$

For equality constraints $\overline{\mathbf{g}}\left(\boldsymbol{\Theta}\right)$, the Jacobian matrix is determined by

$$\mathbf{J}_{\overline{\mathbf{g}}}\left(\boldsymbol{\Theta}\right) = \begin{bmatrix} \left(\boldsymbol{X}^{\mathsf{nps}}\right)^{\mathsf{T}} \mathbf{J}_{\widetilde{\boldsymbol{w}}}\left(\boldsymbol{\omega}\right) & -\mathbf{diag}\left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right) & \mathbf{0}_{p\times r} \\ \left(\dfrac{\partial\left(\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right)\right)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\right)^{\mathsf{T}} \mathbf{J}_{\widetilde{\boldsymbol{w}}}\left(\boldsymbol{\omega}\right) & \mathbf{0}_{r\times p} & -\mathbf{diag}\left(\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right)\right) \end{bmatrix}, \tag{5.166}$$

where $\mathbf{J}_{\widetilde{\boldsymbol{w}}}\left(\boldsymbol{\omega}\right)$ is the Jacobian of pseudo-design weights $\widetilde{\boldsymbol{w}}$, as defined in equation 5.162.

Furthermore is

$$\frac{\partial\Big(\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right)\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} = \left[\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{11}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} \quad \frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{12}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} \quad \cdots \quad \frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{pp}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\right] \tag{5.167}$$

the matrix containing the gradients for each unique entry of the estimated covariance matrix with respect to $\widetilde{\boldsymbol{w}}$.

As discussed with regard to equations 5.154, equivalent expressions for calibrating bias-corrected covariance as well as correlation estimates follow by respectively substituting $\widehat{\boldsymbol{\Sigma}}$ or $\widehat{\boldsymbol{\rho}}$ for $\widetilde{\boldsymbol{\Sigma}}$ in equations 5.157 to 5.167. The derivatives required to solve problem 5.158 for each of these options are provided below. In equation 5.167, the weighted ML estimate for the covariance of any two columns $\boldsymbol{x}_{\cdot k}$ and $\boldsymbol{x}_{\cdot l}$ is denoted by $\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl}$. The corresponding derivatives are

$$\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} = \left(\mathbf{e}\left(\boldsymbol{x}_{\cdot k} \circ \boldsymbol{x}_{\cdot l}\right) - \boldsymbol{x}_{\cdot k}^{\mathsf{nps}} \cdot \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}_{\cdot 1}}\left(\widetilde{\boldsymbol{w}}\right) - \boldsymbol{x}_{\cdot l}^{\mathsf{nps}} \cdot \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}_{\cdot \mathrm{k}}}\left(\widetilde{\boldsymbol{w}}\right)\right) \cdot \left(\widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}\right)\right)^{-1}, \tag{5.168}$$

where $\mathbf{e} : \mathbb{R}^u \to \mathbb{R}^u$ represents centering around the weighted mean as in equation 3.6a. Based on this expression, and using a corresponding notation for elements of bias-corrected covariance as well as correlation matrices, the derivatives of the former are defined by

$$\begin{aligned}\frac{\partial\Big(\left[\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} &= \\ \frac{1}{1-\nu\left(\widetilde{\boldsymbol{w}}\right)} \cdot &\left(\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} + 2 \cdot \left(\widetilde{\boldsymbol{w}} \cdot \widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}\right) - \widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}^{\circ 2}\right)\right) \cdot \frac{\left[\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl}}{\left(\widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}\right)\right)^3}\right).\end{aligned} \tag{5.169}$$

As before, $\nu\left(\widetilde{\boldsymbol{w}}\right)$ denotes the bias correction factor for covariances defined in equation 2.18g. Differentiation of the correlation coefficient yields

$$\begin{aligned}\frac{\partial([\widehat{\boldsymbol{\rho}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)]_{kl})}{\partial\left(\widetilde{\boldsymbol{w}}\right)} = \quad & \frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} \cdot \left(\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk} \cdot \left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}}\right)^{-1} \\ & - \frac{[\widehat{\boldsymbol{\rho}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)]_{kl}}{2} \cdot \frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} \cdot \left(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}\right)^{-1} \\ & - \frac{[\widehat{\boldsymbol{\rho}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)]_{kl}}{2} \cdot \frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} \cdot \left(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}\right)^{-1} \quad .\end{aligned} \tag{5.170}$$

For obtaining pseudo-design weights for non-probability samples, calibrated semi-parametric neural networks can be fit through sequential quadratic programming by using the above components (cf. problem formulations 5.157 and 5.158 as well as section 4.2). Detailed derivations of the presented Jacobian matrices are provided in appendix B.5.1. As for the semi-parametric artificial neural networks in general (cf. section 5.1.9), a calibrated version thereof is apparently not implemented in any pre-existing software. To make this approach employable for the simulation study and practical use (cf. chapters 6 and 7), a custom-made implementation in `C++` is developed in the context of this thesis. An outline is given in section 6.1.2, and more details can be found in appendix C.3.

Integration of Existing Approaches

The purpose of the current section 5.2.3 is to suggest an integrative weighting framework that combines relevant ideas of pseudo-design-based approaches for non-probability samples. Illustrating that common weighting methods (cf. e.g. Deville and Särndal, 1992; Deville, Särndal and Sautory, 1993; Folsom and Singh, 2000; Kott, 2006 and the further references given in the previous sections) are readily incorporated as special cases, the integrative potential of the proposed calibrated semi-parametric artificial neural networks is summarized in the following discussion.

As outlined in sections 5.1.8 and 5.1.9, ANNs include highly relevant special cases, such as generalized linear and additive regression models, which are typical choices for response propensity models (cf. Biffignandi and Pratesi, 2000, p. 1530; Brookhart et al., 2006, p. 1151; Enderle, Münnich and Bruch, 2013, p. 94). Hence, propensity weighting based on a generalized linear or additive *logit model* is equivalent to fitting an ANN without constraints when applying the softmax activation and cross-entropy distance function (cf. Berkson, 1944; Deville and Särndal, 1992, p. 378). The *generalized raking model* (cf. Deming and Stephan, 1940; Deville and Särndal, 1992, p. 378) defined in equation 5.148 is closely related to the logit model but uses a different activation function. Note that to incorporate this function in a neural network, one has to additionally define the inverse of this transformation as activation function. In accordance with equations 5.92 to 5.94, the raking transformation applied to a vector $\boldsymbol{v} \in \mathbb{R}^h$ of arbitrary given size h is defined by

$$\boldsymbol{t}^{(\mathrm{r})}(\boldsymbol{v}) \;=\; \exp(-\boldsymbol{v}) \tag{5.171}$$

with corresponding Jacobian matrix

$$\mathbf{J}_{\boldsymbol{t}^{(\mathrm{r})}}(\boldsymbol{v}) \;=\; -\,\boldsymbol{t}^{(\mathrm{r})}(\boldsymbol{v}) \qquad . \tag{5.172}$$

Furthermore, the *GREG* can be written as a single layer neural network with linear activation function and response model variables $\boldsymbol{Z}^{\mathsf{nps}} = \mathbf{I}_{n^{\mathsf{nps}}}$ set to an identity matrix. In this case, the loss function (cf. equation 5.144) reduces to the penalty component for the weighting parameters, with centering constants all being one. Importance weights are thus equal to $\boldsymbol{w}^{\mathsf{nps}}$ for this component and zero for all others. The total constraints have to be met exactly, and there are no covariance constraints in this case (cf. problem 5.143 and equation 5.144).

The further extensions discussed in section 5.1.8 can likewise be seen as special cases of the proposed framework. The approach introduced by Chang and Kott (2008) corresponds to a single layer ANN with raking or softmax activation function. In this case, the loss function reduces to the weighted quadratic distance of the total errors $\boldsymbol{\epsilon}$ by setting all other importance weights to zero, and there are no covariance or box-constraints. The result is calibration problem 5.150. 'Penalized calibration' introduced by Guggemos and Tillé (2010) can be represented as a calibrated ANN in analogy to the GREG described above. The only required adaptation is that some of the box-constraints for total estimates are discarded while the weighted quadratic calibration errors for these quantities are added to the loss function, which results in problem 5.151. As specified before, interactions of calibration errors $(\epsilon_k \cdot \epsilon_l)$ can be accounted for by extending the matrix of calibration variables by the corresponding interactions of variables. The extended box-constraints introduced by Burgard, Münnich and Rupp (2019), Rupp (2018), Münnich, Sachs and Wagner (2012) as well as Wagner (2013) are directly adopted in problem 5.157.

The outlined special cases emphasize the integrative potential of the proposed calibrated semi-parametric artificial neural networks. However, highly specialized algorithms to perform optimization are available for some of those special cases (cf. e.g. Münnich, Sachs and Wagner, 2012, p. 474; Rupp, 2018, pp. 121 ff). These specialized algorithms are hardly adaptable to the present context because all calibration constraints can be non-linear in the weighting parameters $\boldsymbol{\omega}$ (cf. definition 5.152; Burgard, Münnich and Rupp, 2019, p. 7). Consequently, a rather general `SQP` routine is applied for the proposed weighting model to allow for versatile specifications of weights and loss function. In that regard, numerical optimization is performed differently than for the methods discussed above in order to not limit the tremendous flexibility of calibrated neural networks to these special cases. In principle, the definition of ANNs (cf. equations 5.76 to 5.95) allows for far more sophisticated and complex model specifications. For example, multi-layer non-parametric weighting models are also feasible. This allows for a flexible representation of the non-probability sample selection or participation process that is approximated by the neural network. Moreover, soft calibration of covariances and correlations provides an extension to established total calibration methods (not only) for non-probability samples that are potentially biased in the estimated dependencies between variables.

A further proposal for compensating selectivity of non-probability samples is sub-sampling (cf. e.g. Kim and Wang, 2019, p. 181; Meng, 2018, p. 710). Since this approach can be interpreted in the pseudo-design-based paradigm (cf. Posner and Ash, 2012, p. 4), it is described in the following section 5.2.4.

5.2.4 Sub-sampling from Non-probability Samples

Sub-sampling is a further approach for estimation from non-probability samples. The basic idea is to select a subset of observations from a non-probability sample and rely on this sub-sample for estimation. This strategy is particularly common to maintain computability in the context of Big Data, e.g. when fitting prediction models. At the same time, it can be used to reduce biases under certain conditions (cf. section 2.1; Ai et al., 2018; Kim and Wang, 2019; Ma and Sun, 2015; Pfeffermann, 2015, p. 430; Schouten et al., 2016, p. 745; Wang et al., 2016). Sub-sampling is often treated as standalone approach that is considered to be unrelated to model- or pseudo-design-based paradigms for non-probability samples. Nevertheless, it is closely linked to the latter. A sub-sample from the set of observations $\mathcal{S}^{\mathsf{nps}}$ constituting a non-probability sample simply excludes certain elements from this set, potentially adjusting the weights of the remaining observations. As for resampling (cf. algorithm 1), sub-sampling can be done either with or without replacement (cf. Genuer et al., 2017; Wang, Yang and Stufken, 2019, p. 395). By multiplying an initial weight by the number of times an element occurs in the sub-sample, ideas of sub-sampling can be represented in the pseudo-design-based framework (cf. Ai et al., 2018, pp. 3 f; Kim and Wang, 2019, p. 180; Posner and Ash, 2012, p. 4).

Closely related to the ideas of the previous sections 5.2.1 to 5.2.3, there are different ways to select sub-samples and determine weights, even though all of them define certain weights to be zero. In general, sub-sampling is constituted by the steps which are described in the following algorithm 15 (cf. Ai et al., 2018, p. 4; Kim and Wang, 2019, p. 184; Ma and Sun, 2015).

Algorithm 15: General sub-sampling algorithm

1: **Input:** $\mathcal{S}^{\mathsf{nps}} \in \mathbb{S}$; $\boldsymbol{b} : [0,1]^{n^{\mathsf{nps}}} \to \mathbb{R}^{n^{\mathsf{nps}}}$; $n^{\mathsf{sub}} \in \mathbb{N}$
2: Assign each element $i \in \mathcal{S}^{\mathsf{nps}}$ a probability π_i^{sub} of being selected into the sub-sample.
3: Draw a sub-sample of size n^{sub} from $\mathcal{S}^{\mathsf{nps}}$, where elements are selected according to their assigned probability π_i^{sub}.
4: Determine the vector $\boldsymbol{c} \in \mathbb{N}^{n^{\mathsf{nps}}}$, with elements c_i denoting the number of times that unit i is selected for the sub-sample for all $i \in \mathcal{S}^{\mathsf{nps}}$.
5: Calculate the pseudo-design weights $\widetilde{\boldsymbol{w}}$

$$\widetilde{\boldsymbol{w}} \coloneqq \boldsymbol{w}^{\mathsf{nps}} \circ \boldsymbol{b}\left(\boldsymbol{\pi}^{\mathsf{sub}}\right) \circ \boldsymbol{c} \quad \in \quad \mathbb{R}^{n^{\mathsf{nps}}} \tag{5.173}$$

that depends on $\boldsymbol{\pi}^{\mathsf{sub}}$ by some prespecified function $\boldsymbol{b}$.
6: **Return:** $\widetilde{\boldsymbol{w}}$

In analogy to equations 2.16 and 5.133, the expected value of $\widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{w}}\right)$ can then be written as

$$\begin{aligned} \mathrm{E}\left(\sum_{i \in \mathcal{S}^{\mathsf{sub}}} \widetilde{w}_i \cdot \boldsymbol{y}_{i\cdot}\right) &= \sum_{i \in \mathcal{S}^{\mathsf{nps}}} \mathrm{E}\left(r_i^{\mathsf{sub}} \cdot \widetilde{w}_i\right) \cdot \boldsymbol{y}_{i\cdot} \\ &= \sum_{i \in \mathcal{S}^{\mathrm{P}}} \mathrm{E}\left(r_i^{\mathsf{nps}} \cdot r_i^{\mathsf{sub}} \cdot \widetilde{w}_i\right) \cdot \boldsymbol{y}_{i\cdot} = \sum_{i \in \mathcal{S}^{\mathrm{P}}} \mathrm{E}\left(r_i^{\mathsf{nps}} \cdot \widetilde{w}_i\right) \cdot \boldsymbol{y}_{i\cdot} \quad , \end{aligned} \tag{5.174}$$

using the vector of inclusion indicators $\boldsymbol{r}^{\mathsf{sub}}$ for the sub-sample as before (cf. definition 2.2). Note that equalities 5.174 follow from definition 5.173 simply because $\widetilde{w}_i = 0$ if $r_i^{\mathsf{sub}} = 0$. The general representation of sub-sampling in algorithm 15 encompasses different realizations which depend on the choice of the probabilities $\boldsymbol{\pi}^{\mathsf{sub}}$ as well as the selection of the sub-samples and the corresponding weights. In the simplest form, sub-sampling is done by simple random sampling with or without replacement. In this case, selection probabilities $\pi_i^{\mathsf{sub}} = \mathrm{E}\left(r_i^{\mathsf{sub}} \cdot c_i\right) \coloneqq n^{\mathsf{sub}} \big/ n^{\mathsf{nps}}$ are constant for all $i \in \mathcal{S}^{\mathsf{nps}}$, and weights are defined by $\boldsymbol{b}\left(\boldsymbol{\pi}^{\mathsf{sub}}\right) \coloneqq \left(\boldsymbol{\pi}^{\mathsf{sub}}\right)^{\circ(-1)}$. As long as π_i^{sub} is positive, this approach solely reduces the number of observations but does not alter the bias since, conditional on the sample sizes, it holds that $\mathrm{E}\left(r_i^{\mathsf{sub}} \cdot \widetilde{w}_i\right) = \mathrm{E}\left(r_i^{\mathsf{nps}} \cdot w_i^{\mathsf{nps}}\right) \cdot \pi_i^{\mathsf{sub}} \big/ \pi_i^{\mathsf{sub}} = \mathrm{E}\left(r_i^{\mathsf{nps}} \cdot w_i^{\mathsf{nps}}\right)$ (cf. National Research Council of the United States, 2013, p. 108; Varian, 2014, p. 4).

For the purpose of bias reduction, more refined sub-sampling approaches therefore attempt to find $\boldsymbol{\pi}^{\mathsf{sub}}$ such that it ideally holds that

$$\pi_i^{\mathsf{sub}} \;\propto\; 1/\pi_i^{\mathsf{nps}} \qquad . \tag{5.175}$$

If this is the case, and if $\pi_i^{\mathsf{nps}} > 0$ for all $i \in \mathcal{S}^{\mathrm{P}}$, pseudo-design weights $\widetilde{w}_i = \mathrm{N} \big/ n^{\mathsf{sub}} \cdot c_i$ for all $i \in \mathcal{S}^{\mathsf{nps}}$ are sufficient to achieve approximately unbiased linear statistics when assuming that $\mathrm{E}\left(r_i^{\mathsf{nps}} \cdot r_i^{\mathsf{sub}}\right) \approx \mathrm{E}\left(r_i^{\mathsf{nps}}\right) \cdot \mathrm{E}\left(r_i^{\mathsf{sub}}\right)$ (cf. Kim and Wang, 2019, pp. 179 ff). This idea goes back to approaches for obtaining a simple random sample from a complex probability survey design, which use sub-sampling from a realized probability sample to revert its complex but known selection mechanism. Therefore, it is referred to as *inverse sampling* (cf. e.g. Hinkins, Oh and Scheuren, 1997; Rao, Scott and Benhin, 2003). In contrast to probability sampling, however, the true inclusion probability π_i^{nps} is typically unknown for non-probability samples (cf. section 2.3). Thus, the challenge is to determine

π_i^{sub} such that it fulfills relation 5.175. In some cases, this is integrated as part of the non-probability sampling procedure itself, e.g. by using prior response rates to adjust the gross sample sizes of certain groups (cf. Loosveldt and Sonck, 2008, p. 96), or by applying quota sampling (cf. Mercer et al., 2017, p. 260) or sample matching (cf. section 3.5; Rivers, 2007, p. 1) when drawing from a non-probability panel. In this setting, it is assumed that the non-probability sampling mechanism is adequately described by the auxiliary variables which are used for grouping, quotation or matching. As for other pseudo-design-based approaches, this is valid only when conditional independence assumption 5.1 holds.

This assumption is equally relevant when sub-sampling from an already obtained data set rather than a panel. If conditional independence holds, propensity or calibration weights are valid approximations for the inverse probability $1/\pi_i^{\mathsf{nps}}$ (cf. sections 5.2.1 to 5.2.3). Consequently, Kim and Wang (2019, p. 181) propose choosing π_i^{sub} proportional to such previously determined pseudo-design weights. By using this strategy, sub-sampling approximates propensity or calibration weighting. Hence, an alternative but highly similar approach is to draw the sub-sample by simple random sampling as above and determine the weights by choosing the function $\boldsymbol{b}$ in correspondence to the respective weighting method. In particular, $\boldsymbol{b}\left(\boldsymbol{\pi}^{\mathsf{sub}}\right) \circ \boldsymbol{c}$ may constitute the inverse propensity scores or correction weights, as in equations 5.138 and 5.141 (cf. Kim and Wang, 2019, pp. 184 f). In this case, sub-sampling itself is again a tool solely for reducing the size and complexity of computations while other pseudo-design-based methods are used to account for selectivity.

Further strategies are used to construct multiple sub-samples instead of a single one. In this case, algorithm 15 is applied $a \in \mathbb{N}$ times, and estimation is done using the a different vectors of pseudo-design weights. The estimates resulting from these weights are then combined, i.e. by (weighted) averaging or majority rules. This idea closely resembles that of resampling (cf. algorithm 1), a detailed overview of such algorithms for Big Data is provided by Wang et al. (2016). Different strategies for sub-sampling, weighting and aggregation of the a estimates are subsumed under this general approach. They encompass many ideas for fitting a model through ensemble learning, such as bootstrap aggregation (bagging), boosting or stacking. As a further extension, random selection of the independent variables in a model can be added as well (cf. Genuer et al., 2017; Hastie, Tibshirani and Friedman, 2008, pp. 588, 616 ff; Varian, 2014, p. 14).

In summary, sub-sampling methods choose and potentially re-weight a subset of observations. Under certain conditions, applying sub-sample selection and weighting techniques can be used to address the challenges of non-probability sampling discussed in section 2.3. In addition, elements that are excluded from the sub-sample do not need to be observed. When applied during the actual non-probability sampling stage, sub-sampling may therefore allow for better allocation of resources if adequate auxiliary information is available (cf. Rivers, 2007, p. 2). For similar reasons, it is applied for estimation from Big Data sources, where computational power is typically the limited resource that needs to be used efficiently (cf. Pfeffermann, 2015, p. 433). Because it can be represented by assignment of (partially zero) pseudo-design weights, sub-sampling is considered a pseudo-design-based approach in this thesis.

However, the distinction between the model- and the pseudo-design-based paradigm for estimation from non-probability samples anyhow suggests a strict separation between both frameworks that is not necessarily meaningful in real applications. Approaches to combine both lines of thought are therefore discussed in the following section 5.3.

5.3 Synthesis of Model- and Pseudo-design-based Methods

In the previous sections, an overview of prediction and weighting methods (not only) for non-probability samples is provided, subdivided into the model- and the pseudo-design-based paradigm. Both lines of thought attempt to solve the challenges of non-probability samples outlined in section 2.3 and can be seen as realizations of the framework for informative sampling discussed by Pfeffermann (2011; cf. also Pfeffermann and Sverchkov, 1999). In linking population and sample distribution as described in equation 2.24, the ideal case for both paradigms occurs if $\boldsymbol{Y}$ and $\boldsymbol{r}$ are conditionally independent given a set of auxiliary variables, for now uniformly denoted by $\boldsymbol{X}$ solely to facilitate a coherent discussion. Under this conditional independence assumption 5.1, it holds that

$$\mathrm{P}\left(r_i = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{x}_{i\cdot}\right) = \mathrm{P}\left(r_i = 1 \,|\, \boldsymbol{x}_{i\cdot}\right) \quad . \tag{5.176}$$

The model- and the pseudo-design-based paradigm both seek to account for potential selectivity of non-probability samples by using equality 5.176, although in different ways (cf. equation 5.132; appendix B.3). Model-based methods rely on the fact that equality 5.176 implies that $\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\boldsymbol{x}_{i\cdot}\right) = \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\boldsymbol{x}_{i\cdot}, r_i = 1\right)$. Using a model for this conditional distribution, estimation is then based on external information about the distribution of $\boldsymbol{X}$, most commonly by predicting $\boldsymbol{Y}$ for the population or a reference data set (cf. equations 5.6 to 5.8). In contrast, pseudo-design-based approaches derive estimates $\widehat{\mathrm{P}}\left(r_i = 1 \,|\, \boldsymbol{x}_{i\cdot}\right)$ for the right-hand side of equation 5.176, mostly by means of propensity models and/or calibration. Weighting by the inverse of this estimated probability is then used to compensate for selection bias (cf. equalities 5.131 and 5.133).

In many publications on methods for non-probability samples, ideas from the model- and the pseudo-design-based paradigm are considered separately, and even joint discussions of both paradigms are rather rare (for exceptions, cf. e.g. Buelens, Burger and van den Brakel, 2018; Elliott and Valliant, 2017). This separation is presumably attributable to the historical rivalry between design- and model-based inference in survey statistics, which goes back to the influential papers of Neyman (1934) and Royall (1970; cf. Buelens, Burger and van den Brakel, 2018, p. 325; Magnussen, 2015, p. 317; Särndal, 1978). Yet, such a twofold classification implies a strict separation between both paradigms that is not necessarily of major importance in actual applications of non-probability sampling. Indeed, the approaches discussed in the preceding sections 5.1 and 5.2 already exhibit some overlap and synthesis of both worlds, and "boundaries between design-based and predictive inference have been fading for some time" (Buelens et al., 2012, p. 18).

Important cases of such an overlap between the two paradigms are 'model-assisted' estimators, which make use of prediction models but remain in the (pseudo-)design-based framework. The presumably best known example for such an estimator is the GREG introduced in section 5.2.2, for which resulting design linear estimates can alternatively be written as a function of predictions and residuals from a regression model (cf. e.g. equations 5.14 and 5.145; Breidt and Opsomer, 2017, pp. 191 ff; Cassel, Särndal and Wretman, 1976; Wu and Sitter, 2001, p. 185). Moreover, all methods to obtain pseudo-design weights for non-probability samples heavily rely on models for $\mathrm{P}\left(r_i = 1 \,|\boldsymbol{x}_{i\cdot}\right)$. From alternating points of view, such weighting techniques may themselves be viewed as prediction models and are, thus, clearly different from classical design weights (cf. sections

2.2 and 5.2; Baker et al., 2013a, p. 67; Buelens et al., 2012, p. 8; Little, 1993, p. 1002; Valliant and Dever, 2011, p. 109). Furthermore, the prediction models described in section 5.1 can be fit by using survey weights. This strategy is commonly applied in probability samples, where inverse inclusion probabilities can be used to represent $\mathrm{P}\left(r_i = 1 \,|\boldsymbol{y}_{i\cdot}\right)$. Optimizing a HT-estimate of the population's loss function accounts for cases where the inclusion probabilities may be related to the target variables (cf. e.g. equations 5.12 and 5.22; section 2.3; Binder, 1983, p. 282; Pfeffermann, 2011, p. 122). Therefore, weighted prediction models make use of concepts which are inherently related to the (pseudo-)design-based framework. Considering pseudo-design weights as a surrogate for design weights in case of non-probability sampling, a straightforward extension is to use weighted loss functions for model fitting as well (cf. e.g. Beaumont, 2000; Breidt and Opsomer, 2017; Fuller, 2009, p. 378; Pfeffermann and Sverchkov, 1999; Ripley and Venables, 2016; Särndal, Swensson and Wretman, 1992, pp. 192 ff; Wood, 2017, p. 181).

Although most publications on non-probability sampling tend to strictly separate pseudo-design and model-based approaches, it appears sensible to consider such a combined use of both paradigms since conditional independence is rarely perfect in real applications. In extension to the connections between both schools of thoughts already established in the previous sections and summarized above, there are some alternative approaches. These tackle the challenges of non-probability samples and closely related fields in a slightly more specialized way. A brief overview is provided in the following discussion. To that end, section 5.3.1 introduces a joint model for response indicator and target variable, while section 5.3.2 focuses on weighted aggregation of predictions.

5.3.1 Integration of Response and Outcome Model

A first alternative to the estimation methods discussed so far is based on explicitly modeling the joint distribution of $\boldsymbol{r}^{\mathsf{nps}}$ and $\boldsymbol{Y}$. As in section 5.1, the goal is to find some predictions $\widehat{\boldsymbol{y}}_{\cdot k}$ for $\boldsymbol{y}_{\cdot k}$ as a function of $\boldsymbol{X}$ and parameters $\boldsymbol{\Theta}$. The conditional distribution expressed by a statistical model is that of the residuals $\boldsymbol{e}_{\cdot k} = \boldsymbol{y}_{\cdot k} - \widehat{\boldsymbol{y}}_{\cdot k}$,

$$\mathrm{f}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(y_{ik} \,|\boldsymbol{x}_{i\cdot}, \boldsymbol{\Theta}\right) \;=\; \mathrm{f}_{\boldsymbol{e}_{\cdot\mathrm{k}}}\left(e_{ik}\right) \qquad . \tag{5.177}$$

In case of any dependency between $\boldsymbol{y}_{\cdot k}$ and $\boldsymbol{r}^{\mathsf{nps}}$ that is not explained by $\boldsymbol{X}$, the conditional distribution in the non-probability sample

$$\mathrm{f}_{\boldsymbol{e}_{\cdot\mathrm{k}}}\left(e_{ik} \,|r_i^{\mathsf{nps}} = 1\right) \;=\; \frac{\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|e_{ik}\right)}{\mathrm{P}\left(r_i^{\mathsf{nps}} = 1\right)} \cdot \mathrm{f}_{\boldsymbol{e}_{\cdot\mathrm{k}}}\left(e_{ik}\right) \tag{5.178}$$

differs from that in the population if $\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|e_{ik}\right) \neq \mathrm{P}\left(r_i^{\mathsf{nps}} = 1\right)$ (cf. equality 2.24; Pfeffermann, 2011, p. 116). To find a solution for estimating $\boldsymbol{\Theta}$ from equation 5.178, a bivariate model for $\boldsymbol{y}_{\cdot k}$ and r_i^{nps} can be specified.

The probably most popular example thereof (cf. Brick, 2013, p. 335; van Buuren, 2018, p. 97) is the model proposed by Heckman (1976; 1979). The assumed conditional distribution is $\boldsymbol{e}_{\cdot k} \sim \mathrm{N}\left(0, \boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot\mathrm{k}}}\right)$ under the linear regression model $\widehat{\boldsymbol{y}}_{\cdot k} = \mathrm{E}\left(\boldsymbol{y}_{\cdot k} \,|\, \boldsymbol{X}\right) = \boldsymbol{X}\boldsymbol{\beta}_{\cdot k}$ for $\boldsymbol{\Theta} = \boldsymbol{\beta}_{\cdot k}$, as introduced in section 5.1.2. For the full population, the model can be written as

$$y_{ik} \;= \boldsymbol{x}_{i\cdot}\boldsymbol{\beta}_{\cdot k} + e_{ik} \quad \text{for all} \quad i \in \mathcal{S}^{\mathrm{P}} \qquad . \tag{5.179}$$

Therefore, the model equation for the non-probability sample nps from which the parameters are estimated is

$$y_{ik}^{\mathsf{nps}} = \boldsymbol{x}_{i\cdot}^{\mathsf{nps}}\boldsymbol{\beta}_{\cdot k} + \mathrm{E}\left(e_{ik}\,|r_i^{\mathsf{nps}} = 1\right) + \widetilde{e}_{ik} \quad \text{for all} \quad i \in \mathcal{S}^{\mathsf{nps}} \quad . \tag{5.180}$$

Here, $\widetilde{\boldsymbol{e}}_{\cdot k} \in \mathbb{R}^{n^{\mathsf{nps}}}$ is a new residual variable with mean zero, which is used to differentiate sample and population density (cf. equation 5.178). When estimating $\boldsymbol{\beta}_{\cdot k}$ from equation 5.180, dependencies between $\boldsymbol{y}_{\cdot k}$ and $\boldsymbol{r}^{\mathsf{nps}}$ which are not explained by $\boldsymbol{X}$ result in $\mathrm{E}\left(e_{ik}\,|r_i^{\mathsf{nps}} = 1\right) \neq \mathrm{E}\left(e_{ik}\,|r_i^{\mathsf{nps}} = 0\right)$ because conditional independence assumption 5.1 does not hold. In this case, omitting $\mathrm{E}\left(e_{ik}\,|r_i^{\mathsf{nps}} = 1\right)$ leads to biased parameter estimates due to violation of the Gauss-Markov theorem (cf. Greene, 2008, p. 44; Heckman, 1976, p. 478; Wooldridge, 2012, pp. 83 ff).

For fitting a prediction model that accounts for this issue, Heckman (1976; 1979) introduces a latent (unobserved) variable to express an underlying response or participation tendency, which is defined by a second regression equation

$$y_{il}^{*} = \boldsymbol{z}_{i\cdot}\boldsymbol{\beta}^{*} + e_{il}^{*} \quad \text{for all} \quad i \in \mathcal{S}^{\mathrm{P}} \quad . \tag{5.181}$$

It is assumed that the non-probability sample's membership indicator can be expressed by this tendency through

$$r_i^{\mathsf{nps}} = \mathbb{I}\left(y_{il}^{*} \geq 0\right) = \mathbb{I}\left(e_{il}^{*} \geq -\boldsymbol{z}_{i\cdot}\boldsymbol{\beta}^{*}\right) \quad \text{for all} \quad i \in \mathcal{S}^{\mathrm{P}} \quad . \tag{5.182}$$

Furthermore assuming that the residuals in equations 5.179 and 5.181 are realizations from a bivariate normal distribution, i.e.

$$\begin{bmatrix}\boldsymbol{e}_{\cdot k} & \boldsymbol{e}_{\cdot l}^{*}\end{bmatrix} \sim \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \boldsymbol{\rho}_{\boldsymbol{e}_{\cdot\mathrm{k}}\boldsymbol{e}_{\cdot\mathrm{l}}^{*}} \cdot \sqrt{\boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot\mathrm{k}}}}\\ \boldsymbol{\rho}_{\boldsymbol{e}_{\cdot\mathrm{k}}\boldsymbol{e}_{\cdot\mathrm{l}}^{*}} \cdot \sqrt{\boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot\mathrm{k}}}} & \boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot\mathrm{k}}}\end{pmatrix}\right) \quad , \tag{5.183}$$

allows replacing the conditional expectation in equation 5.180 for all $i \in \mathcal{S}^{\mathsf{nps}}$ by

$$\begin{aligned}\mathrm{E}\left(e_{ik}\,|r_i^{\mathsf{nps}} = 1\right) &= \boldsymbol{\rho}_{\boldsymbol{e}_{\cdot\mathrm{k}}\boldsymbol{e}_{\cdot\mathrm{l}}^{*}} \cdot \sqrt{\boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot\mathrm{k}}}} \cdot \mathrm{E}\left(e_{il}^{*}\,|e_{il}^{*} \geq -\boldsymbol{z}_{i\cdot}\boldsymbol{\beta}^{*}\right)\\ &= \boldsymbol{\rho}_{\boldsymbol{e}_{\cdot\mathrm{k}}\boldsymbol{e}_{\cdot\mathrm{l}}^{*}} \cdot \sqrt{\boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot\mathrm{k}}}} \cdot \frac{\Phi'\left(\boldsymbol{z}_{i\cdot}\boldsymbol{\beta}^{*}\right)}{\Phi\left(\boldsymbol{z}_{i\cdot}\boldsymbol{\beta}^{*}\right)} \quad .\end{aligned} \tag{5.184}$$

Here $\Phi' : \mathbb{R} \to (0,1)$ and $\Phi : \mathbb{R} \to (0,1)$ respectively denote density and distribution function of the standard normal distribution, and their ratio is the inverse Mill's ratio (cf. Greene, 2008, p. 866; Johnson and Kotz, 1972, pp. 112 f; Mills, 1926; Wooldridge, 2012, p. 618). Consequently, this ratio can be used as additional auxiliary variable to rewrite equation 5.180 as

$$y_{ik}^{\mathsf{nps}} = \boldsymbol{x}_{i\cdot}^{\mathsf{nps}}\boldsymbol{\beta}_{\cdot k} + \widetilde{\beta}_k \cdot \frac{\Phi'\left(\boldsymbol{z}_{i\cdot}\boldsymbol{\beta}^{*}\right)}{\Phi\left(\boldsymbol{z}_{i\cdot}\boldsymbol{\beta}^{*}\right)} + \widetilde{e}_{ik} \quad \text{for all} \quad i \in \mathcal{S}^{\mathsf{nps}} \quad , \tag{5.185}$$

where $\widetilde{\beta}_k = \boldsymbol{\rho}_{\boldsymbol{e}_{\cdot\mathrm{k}}\boldsymbol{e}_{\cdot\mathrm{l}}^{*}} \cdot \sqrt{\boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot\mathrm{k}}}}$. This is basically an adaptation of the Tobit model developed by Tobin (1958).

The difficulty in identifying equation 5.185 is that $\boldsymbol{y}^*_{\cdot l}$ is not observed to estimate $\boldsymbol{\beta}^*$ from equality 5.181. However, the outcome $\boldsymbol{r}^{\mathsf{nps}}$ is observed. By assumptions 5.181 to 5.183, the conditional probability to participate is defined by the *probit model*

$$\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 | \boldsymbol{z}_{i\cdot}\right) \;=\; \Phi\left(\boldsymbol{z}_{i\cdot}\boldsymbol{\beta}^*\right) \quad \text{for all} \quad i \in \mathcal{S}^{\mathrm{P}} \tag{5.186}$$

(cf. Amemiya, 1985, pp. 268 ff; Maddala, 1983, p. 269; Simonoff, 2003, pp. 393 f). This formulation constitutes a generalized linear model, which can be used to estimate coefficients $\boldsymbol{\beta}^*$ and thus determine the inverse Mill's ratio (cf. section 5.1.3; Heckman, 1979, p. 157; Toomet and Henningsen, 2008). In comparison to the generalized linear logit model, the only modification lies in the use of an alternative (non-canonical) link function Φ. The remaining discussion of section 5.1.3 is directly applicable, in particular the use of Fisher scoring for finding the coefficients $\boldsymbol{\beta}^*$ and the necessity to have observations in- and outside the non-probability sample (cf. also section 5.2.1). Once $\boldsymbol{\beta}^*$ is obtained from a GLM, equation 5.185 can be used to estimate $\boldsymbol{\beta}$ and $\widetilde{\beta}_k$ by means of a linear model. However, this strategy suffers from some severe limitations. On the one hand, it requires strong distributional assumptions that are easily violated. On the other hand, the model is designed for cases where some variables $\boldsymbol{z}_{\cdot l}$ exclusively predict sample inclusion but are not included in $\boldsymbol{X}$ for predicting the target variable. If this is not the case, the inverse Mill's ratio is typically almost collinear to $\boldsymbol{X}$ (cf. Little, 1988b, p. 290; Rubin, 2006, p. 11; Wooldridge, 2012, p. 619; Weisberg, 2005, pp. 151 ff).

To incorporate dependencies between sample inclusion indicator and variable(s) of interest, information about the (approximated) selection process is included as additional auxiliary variable in the Heckman model introduced in the current section 5.3.1. This is an alternative way of conditioning on the propensities other than matching or weighting (cf. sections 3.5 and 5.2.1; Lee, Lessler and Stuart, 2010, p. 337). Note that differences between predictions from logit and probit models are usually negligible, and coefficients are even approximately convertible between the two link functions (cf. e.g. Amemiya, 1981, p. 1488; Hahn and Soyer, 2005, p. 1; Simonoff, 2003, pp. 393 f). Therefore, it may be reasonable to use response propensities from logit rather than probit models for this purpose as well.[6] In any case, modeling the joint distribution of $\boldsymbol{r}^{\mathsf{nps}}$ and $\boldsymbol{Y}$ constitutes a link between model- and pseudo-design-based methods. As an alternative strategy to establish such a link, weighted aggregation of predictions is introduced in the following section 5.3.2.

[6] In some applications, the predicted probabilities from logit models are even used directly as auxiliary variables for predicting $\boldsymbol{Y}$, rather than calculating and using the inverse Mill's ratio for this purpose (cf. e.g. Nassimbeni, 2001, p. 258; Xu, Wong and Choi, 2014, p. 8).

5.3.2 Weighted Aggregation of Predictions

An alternative way to integrate model- and pseudo-design-based approaches is to use a model for the conditional distribution $\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\boldsymbol{z}_{i\cdot}\right)$ of the target variable as in section 5.1, and a separate model for $\mathrm{P}\left(r_i = 1\right)/\mathrm{P}\left(r_i = 1 \,|\, \boldsymbol{z}_{i\cdot}\right)$ to represent the sampling process as in section 5.2. When assuming conditional independence of $\boldsymbol{r}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$, both models can be fit by jointly using a non-probability and a reference sample (cf. equations 5.7 and 5.132 as well as the related discussion). Through application of Bayes' theorem for estimating the distribution of $\boldsymbol{Z}$ as outlined in equation 2.24, the unconditional distribution of $\boldsymbol{Y}$ is then estimated by means of equality 5.8. This strategy can be implemented by using the model predictions $\widehat{\boldsymbol{Y}}^{\mathsf{nps}}$ in place of the observed values $\boldsymbol{Y}^{\mathsf{nps}}$ when performing pseudo-design weighted estimation in the non-probability sample, where weights are proportional to the inverse estimated propensities $1\big/\widehat{\mathrm{P}}\left(r_i = 1 \,|\, \boldsymbol{z}_{i\cdot}\right)$. It is therefore referred to as *weighted aggregation of predictions* in the context of this thesis. The advantage of this approach in comparison to the Heckman model discussed in the previous section 5.3.1 is that no further assumptions for the underlying bivariate distribution of $\boldsymbol{Y}$ and $\boldsymbol{r}^{\mathsf{nps}}$ are required (cf. assumption 5.183).

A form of weighted aggregation of predictions that is of notable interest (cf. e.g. Elliott and Valliant, 2017, p. 260; Lax and Phillips, 2009; Mercer et al., 2017, pp. 264 f) is *'multilevel regression and post-stratification' (MRP)*, proposed for non-probability samples by Wang et al. (2015; cf. also Gelman and Little, 1997; Gelman et al., 2016a,b; Park, Gelman and Bafumi, 2004). As its name suggests, this approach represents $\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\boldsymbol{z}_{i\cdot}\right)$ by means of a mixed model, most commonly a (generalized) linear one (cf. section 5.1.5). Post-stratification weights that calibrate estimated and known population totals of $\boldsymbol{Z}$ (cf. section 5.2.2) are then applied to estimate the relevant statistics of the unconditional distribution of $\boldsymbol{Y}$ from the mixed model's predictions $\widehat{\boldsymbol{Y}}^{\mathsf{nps}}$ in the non-probability sample. In this setting, $\boldsymbol{Z}$ is generally defined to be a matrix of J indicator variables. These variables represent the membership in mutually exclusive subsets $\mathcal{S}^{(j)}$ for $j = 1, \ldots, J$ of the population, such that $\mathcal{S}^{\mathrm{P}} = \bigcup_{j=1}^{J} \mathcal{S}^{(j)}$, $\emptyset = \mathcal{S}^{(j)} \cap \mathcal{S}^{(k)}$ for all $j \neq k$ and

$$\boldsymbol{Z} = \begin{bmatrix} \mathbb{I}\left(1 \in \mathcal{S}^{(1)}\right) & \cdots & \mathbb{I}\left(1 \in \mathcal{S}^{(J)}\right) \\ \vdots & \ddots & \vdots \\ \mathbb{I}\left(\mathrm{N} \in \mathcal{S}^{(1)}\right) & \cdots & \mathbb{I}\left(\mathrm{N} \in \mathcal{S}^{(J)}\right) \end{bmatrix} \quad . \tag{5.187}$$

For example, $\boldsymbol{Z}$ may express the cross-combinations of socio-demographic and geographical information (cf. Park, Gelman and Bafumi, 2004, p. 376). The design matrix for the random components of the mixed model is likewise determined by this matrix $\boldsymbol{Z}$, optionally in conjunction with additional random slope terms containing selected interactions of model variables $\boldsymbol{X}$ and response variables $\boldsymbol{Z}$ (cf. section 5.1.5; Ghitza and Gelman, 2013, p. 766).

Weighted aggregation of predictions in general constitutes a synthesis of model- and pseudo-design-based ideas. On the one hand, it is different but still closely related to mass-imputation, where predictions are made for observations in the reference sample (cf. section 5.1). On the other hand, it resembles pseudo-design-based estimation but relies on model predictions $\widehat{\boldsymbol{Y}}^{\mathsf{nps}}$ in place of actually observed values for target variables $\boldsymbol{Y}^{\mathsf{nps}}$ in the non-probability sample (cf. section 5.2). As summarized in the introduction to

the current section 5.3, an alternative approach to jointly use weighting and prediction methods for non-probability samples is to apply weights when fitting the model. This can be done by minimizing a weighted loss function for fitting a prediction model for $\boldsymbol{Y}$ (cf. e.g. Beaumont, 2000; Breidt and Opsomer, 2017; Fuller, 2009, p. 378; Pfeffermann and Sverchkov, 1999; Ripley and Venables, 2016; Särndal, Swensson and Wretman, 1992, pp. 192 ff; Wood, 2017, p. 181). In principle, both of these approaches allow combining all of the discussed weighting and prediction models and are, therefore, quite comprehensive and adaptive. An important advantage of MRP is that calibration benchmarks suffice as auxiliary information, which is rarely the case for weighted prediction models (cf. section 5.1; Pfeffermann, 2011). In contrast, a major benefit of fitting weighted prediction models is that this approach allows encompassing purely model- and design-based methods as border cases: when weights are constant over all observations, the fully model-based approach emerges, while the purely pseudo-design-based framework uses the observed distribution of $\boldsymbol{Y}$ in place of the model (cf. sections 2.2, 5.1 and 5.2).

In sections 5.1 to 5.3, various approaches for estimation from non-probability samples are discussed. These are mainly proposed for reducing biases in point estimation methods, which is the major focus of the scientific debate around non-probability samples. Inference is then predominantly based on classical design- or model-based strategies, yet partially with extensions as those used for multiple imputation (cf. e.g. Buelens, Burger and van den Brakel, 2018; Elliott and Valliant, 2017; Kim et al., 2018; Rafei, Flannagan and Elliott, 2020). An overview is given in the following section 5.4.

5.4 Inference

A major advantage of probability sampling is that it provides a known randomization process through sample selection. For measurable probability sampling designs, an estimator's quality can therefore be assessed from a realized sample, and valid inference can be carried out, e.g. by means of variance estimates, confidence intervals or statistical tests. Such inferential methods that are valid for arbitrary variables of interest without making any assumptions about their distribution require design-based concepts of repeated sampling variation of (asymptotically) unbiased estimates around the true population statistic of interest (cf. section 2.2; Breidt and Opsomer, 2017, p. 191; Kalton, 1983, p. 90).

Due to unknown and/or uncontrolled selection processes, it is difficult to perfectly transfer these concepts and hence classical design-based inference to non-probability samples. This is especially the case when unbiasedness of point estimates is not even asymptotically guaranteed (cf. section 2.3; Buelens, Burger and van den Brakel, 2018, p. 327; Japec et al., 2015, p. 862) and one reason why the discussion on estimation methods for non-probability samples mainly focuses on bias reduction in the referred literature as well as the previous sections 5.1 to 5.3. Yet, the ways in which inference from non-probability samples is attempted strongly adhere to concepts discussed in these preceding sections, and likewise utilize prediction and weighting models. Indeed, there is no perfectly valid way for inference from non-probability samples in general, unless the respective modeling assumptions discussed throughout the previous sections hold and, thus, allow compensating for selection bias. If those assumptions are violated, estimating the bias component of the MSE requires even stronger knowledge or assumptions about the selection process (cf. e.g. chapter 3). If such additional premises appear available and reasonable for inference, they typically could be used to actually compensate for the bias, rather than estimating

it solely for inferential purposes (cf. e.g. Baker et al., 2010, p. 47; 2013a, p. 107; Mercer et al., 2017, p. 258; Pfeffermann, 2015, pp. 441 ff; Schouten, 2007). As a consequence, methods for inference are typically limited to the same assumptions and information that are already used to facilitate point estimation by means of prediction or weighting models (cf. sections 5.1 to 5.3). Inference for non-probability samples is then typically based exclusively on variance estimation because MSE and variance coincide if an adequate weighting or prediction model is found (cf. Buelens, Burger and van den Brakel, 2018, p. 330; Chen, Valliant and Elliott, 2019, p. 673; Elliott and Valliant, 2017, p. 257; Kim et al., 2018, p. 10; Rafei, Flannagan and Elliott, 2020, p. 159). A summary of inferential approaches proposed for non-probability samples is provided in the current section. Such methods typically refer to either classical design- or model-based inference. Therefore, the following presentation is merely a summary of central ideas in each of these paradigms that can be applied for non-probability samples. Comprehensives overviews for the general context of survey sampling are e.g. given by Lohr (2010), Pfeffermann and Rao (2009), Särndal (1978), Valliant (2009) as well as Wolter (2007).

As outlined in the previous sections, model- as well as pseudo-design-based approaches in general obtain estimators $\widehat{\boldsymbol{\vartheta}}$ based on parameters $\boldsymbol{\Theta}$, e.g. through models for predicting target variables or response propensities. By the law of total variance (cf. e.g. Blitzstein and Hwang, 2013, p. 401), the variance of an estimator $\widehat{\boldsymbol{\vartheta}}$ can, thus, be decomposed as

$$\mathbf{V}\left(\widehat{\boldsymbol{\vartheta}}\right) = \mathbf{E}\left(\mathbf{V}\left(\widehat{\boldsymbol{\vartheta}}\,\middle|\,\boldsymbol{\Theta}\right)\right) + \mathbf{V}\left(\mathbf{E}\left(\widehat{\boldsymbol{\vartheta}}\,\middle|\,\boldsymbol{\Theta}\right)\right) \tag{5.188}$$

into a within and between component. These can be interpreted as the variability due to sample selection given the model and the uncertainty about the actual model (cf. Binder and Roberts, 2009, pp. 43 ff; Opsomer, 2009, p. 7; Rafei, Flannagan and Elliott, 2020, pp. 159 f). This decomposition is commonly used when inferential methods for non-probability samples are proposed (cf. e.g. Elliott and Valliant, 2017, p. 259; Isaksson and Lee, 2005, p. 3145; Kim et al., 2018, pp. 10 f).

If $\boldsymbol{\Theta}$ is considered predetermined and hence fix rather than estimated, $\mathbf{E}\left(\widehat{\boldsymbol{\vartheta}}\,\middle|\,\boldsymbol{\Theta}\right)$ is constant and equality 5.188 thus reduces to the within variance $\mathbf{V}\left(\widehat{\boldsymbol{\vartheta}}\,\middle|\,\boldsymbol{\Theta}\right) = \mathbf{E}\left(\mathbf{V}\left(\widehat{\boldsymbol{\vartheta}}\,\middle|\,\boldsymbol{\Theta}\right)\right)$. In this case, estimating the variance from a single prediction or weighting model with a fixed set of parameters is feasible. This occurs e.g. in the pseudo-design-based framework when pseudo-design weights are assumed to perfectly describe the non-probability sampling process, such that $\widetilde{\boldsymbol{w}}$ constitutes an ideal substitution for some hypothetical design weights. Under this condition and full coverage of the target population, classical design-based methods can be applied for inference in non-probability samples, just as for probability samples. Typically, it is additionally assumed in this context that non-probability sample inclusion of two elements is independent to achieve further simplification. Treating the weights as fixed, the resulting approaches encompass closed-form variance estimates for linear and linearized non-linear statistics as well as general non-parametric resampling techniques (cf. e.g. Barratt, Ferris and Lenton, 2015; Faas and Schoen, 2006; Guarte and Barrios, 2006; Spijkerman et al., 2009). An introductory overview to general design-based inference is given in section 2.2, together with further references which provide an exhaustive presentation (cf. e.g. Wolter, 2007). However, treating a pseudo-design weighted non-probability sample as if it was a probability sample with known design weights can lead to considerable underestimation of actual variances (cf. Breidt and Opsomer, 2009, pp. 117 f; Lee and Valliant, 2009, p. 341; Rafei, Flannagan and Elliott, 2020, p. 160).

Nevertheless, similar ideas are likewise applied when using mass imputation for reference samples. In such cases, the imputed variable of interest is treated as if it was an actually observed variable (cf. e.g. Yang and Kim, 2018, p. 5). For the model-based framework, alternatives to this somewhat naive form of inference emerge when assuming that a known statistical model with fixed parameters $\boldsymbol{\Theta}$ describes $\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \mid \boldsymbol{x}_{i\cdot}\right)$ for the population. When this is the case, inference can be based on knowledge about the distribution of $\boldsymbol{X}$, which can e.g. be obtained from a reference sample (cf. section 5.1; Pfeffermann, 2011). A core element for inference in this setting typically is the residual variance $\boldsymbol{\Sigma}_{\boldsymbol{E}}$, which corresponds to $\mathbf{V}\left(\boldsymbol{Y} \mid \boldsymbol{X}\right)$ under the model. As an example for this model-based inference approach, consider the estimator for $\boldsymbol{\tau}_{\boldsymbol{Y}}$. Since $\boldsymbol{\tau}_{\boldsymbol{Y}} = \mathrm{N} \cdot \mathbf{E}\left(\boldsymbol{Y}\right)$, it can be estimated as $\widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}} = \mathrm{N} \cdot \mathbf{E}\left(\widehat{\mathbf{E}}\left(\boldsymbol{Y} \mid \boldsymbol{X}\right)\right)$ by using a known population distribution of $\boldsymbol{X}$ and a model for the conditional expectation. If this model holds, the estimator's variance is given by $\mathbf{V}\left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}}\right) = \mathrm{N}^2 \cdot \mathbf{V}\left(\mathbf{E}\left(\widehat{\mathbf{E}}\left(\boldsymbol{Y} \mid \boldsymbol{X}\right)\right)\right)$ and can be estimated from the model as well. In certain combinations of estimators and models, this expression may be solved analytically after the model is fit. For example, under the linear model's assumptions 5.18, the result is $\widehat{\mathbf{V}}\left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{Y}}\right) = \mathrm{N}^2 / n^{\mathsf{nps}} \cdot \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{E}}$, where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{E}}$ is the residual variance estimated from the model (cf. e.g. Breidenbach, McRoberts and Astrup, 2016, p. 276; Valliant, Dorfman and Royall, 2000, p. 145). Similar arguments apply to other estimators and models as well (cf. Buelens, Burger and van den Brakel, 2018, p. 330). In many situations, however, no such analytic expression for the model-based variance is available. In such cases, parametric resampling methods can be applied, which draw resamples from the model distribution rather than real observations (cf. algorithm 1; Efron and Tibshirani, 1998, pp. 53 ff, 296 ff). Exhaustive overviews and details for model-based inference are e.g. given by Royall (1970; 1992), Särndal (1978), Valliant, Dorfman and Royall (2000) or Valliant (2009). Besides assuming that the model is true, a further limitation of these approaches is that they require the conditional distribution to be obtainable from the model. Therefore, they are not applicable to most machine learning models, which typically do not explicitly consider distributional assumptions (cf. Hastie, Tibshirani and Friedman, 2008, pp. 261 ff).

The strategies discussed so far exclusively rely on $\mathbf{V}\left(\widehat{\boldsymbol{\vartheta}} \,\middle|\, \boldsymbol{\Theta}\right)$ to estimate $\mathbf{V}\left(\widehat{\boldsymbol{\vartheta}}\right)$ and consider $\boldsymbol{\Theta}$ as fix for this purpose. However, this is rarely a realistic assumption. Even if the presumed structural form of the prediction or weighting model is true, fitting parameters $\boldsymbol{\Theta}$ to a sample still induces some degree of model uncertainty (cf. Binder and Roberts, 2009, p. 45; Opsomer, 2009, p. 7). More refined methods for inference therefore incorporate the variation of $\boldsymbol{\Theta}$ and in particular the second component $\mathbf{V}\left(\mathbf{E}\left(\widehat{\boldsymbol{\vartheta}} \,\middle|\, \boldsymbol{\Theta}\right)\right)$ of equation 5.188. This is the variation between different sets of parameters $\boldsymbol{\Theta}$, which are now considered random due to the sampling variance (cf. Binder and Roberts, 2009, p. 54; Kim et al., 2018, pp. 8 f). However, closed-form expressions to estimate equation 5.188 are usually not available for general combinations of models and estimators (cf. Buelens, Burger and van den Brakel, 2018, p. 330). Consequently, it is common practice to approximate the distribution of $\boldsymbol{\Theta}$ through resampling methods, which typically apply strategies as represented by algorithm 1 (cf. e.g. Buelens, Burger and van den Brakel, 2018; Elliott and Valliant, 2017; Guarte and Barrios, 2006, p. 279; Kim et al., 2018; Lee and Valliant, 2009). From each non-probability sample, a resamples are drawn as described below, and a prediction or weighting model is fit to each of these. For each resample $j = 1, \ldots, a$, estimates $\widehat{\boldsymbol{\vartheta}}^{(j)}$ are then obtained exactly as for the original sample, i.e. by weighting the resample or using it to impute for the reference sample.

The estimated between variance is then determined from the variation over all resamples, i.e.

$$\widehat{\mathbf{V}}_{\mathbf{b}} = \widehat{\mathbf{V}}\left(\widehat{\mathbf{E}}\left(\hat{\boldsymbol{\vartheta}} \,\middle|\, \boldsymbol{\Theta}\right)\right) \coloneqq \frac{1}{a-1} \cdot \sum_{j=1}^{a} \left(\hat{\boldsymbol{\vartheta}}^{(j)} - \left(\frac{1}{a} \cdot \sum_{k=1}^{a} \hat{\boldsymbol{\vartheta}}^{(k)}\right)\right)^2 \quad . \tag{5.189a}$$

Furthermore, estimates $\widehat{\mathbf{V}}\left(\hat{\boldsymbol{\vartheta}}^{(j)} \,\middle|\, \boldsymbol{\Theta}^{(j)}\right)$ can be obtained for each of the iterations by means of the strategies described above. This may be done by using model-based strategies that rely on the residual variance or parametric bootstrapping. However, the more typical case is to naively apply classical design-based variance estimation in each iteration, treating the pseudo-design weighted non-probability or the imputed reference sample as if they were probability samples with known design weights and target variables (cf. Rafei, Flannagan and Elliott, 2020, p. 160; Yang and Kim, 2018, p. 5). In whatever way these values are calculated, Rafei, Flannagan and Elliott (2020, p. 160) propose using them for estimating the expectation for the within variance by

$$\widehat{\mathbf{V}}_{\mathbf{w}} = \widehat{\mathbf{E}}\left(\widehat{\mathbf{V}}\left(\hat{\boldsymbol{\vartheta}} \,\middle|\, \boldsymbol{\Theta}\right)\right) \coloneqq \frac{1}{a} \cdot \sum_{j=1}^{a} \widehat{\mathbf{V}}\left(\hat{\boldsymbol{\vartheta}}^{(j)} \,\middle|\, \boldsymbol{\Theta}^{(j)}\right) \tag{5.189b}$$

to jointly consider within and between variance components. Rubin's (1987, pp. 89 f) rules are then used to combine estimates for both by

$$\begin{aligned} \widehat{\mathbf{V}}_{\mathbf{t}} = \widehat{\mathbf{V}}\left(\hat{\boldsymbol{\vartheta}}\right) &\coloneqq \widehat{\mathbf{E}}\left(\widehat{\mathbf{V}}\left(\hat{\boldsymbol{\vartheta}} \,\middle|\, \boldsymbol{\Theta}\right)\right) + \left(1 + \frac{1}{a}\right) \cdot \widehat{\mathbf{V}}\left(\widehat{\mathbf{E}}\left(\hat{\boldsymbol{\vartheta}} \,\middle|\, \boldsymbol{\Theta}\right)\right) \\ &= \widehat{\mathbf{V}}_{\mathbf{w}} + \left(1 + \frac{1}{a}\right) \cdot \widehat{\mathbf{V}}_{\mathbf{b}} \end{aligned} \tag{5.189c}$$

as an estimator for equality 5.188. In equation 5.189c, the factor $(1 + 1/a)$ is used to achieve unbiasedness regarding the between variance estimator due to a finite number of resampling iterations. These ideas correspond to methods for inference that are applied in multiple imputation (cf. e.g. van Buuren, 2018, p. 43; Little and Rubin, 2019; Rubin, 1987). Since the distribution of $\boldsymbol{\Theta}$ depends on the unknown non-probability sampling process in the present context, such approaches still have to rely on assumptions to implement a resampling procedure in the context of non-probability samples (cf. Elliott and Valliant, 2017, p. 257). Most commonly, it is assumed that simple random sampling is an adequate approximation to the non-probability sampling mechanism for the purpose of resampling (cf. e.g. Buelens, Burger and van den Brakel, 2018, p. 330; Kim et al., 2018, p. 10; Lee and Valliant, 2009, p. 330). As for the naive application of classical design-based variance estimation to obtain estimator 5.189b, this assumption is typically a simplification rather than a realistic representation of the actual selection mechanism. Nevertheless, additionally or exclusively considering the between component of decomposition 5.188 can yield variance estimates that are less sensitive to miss-specified models and provide better inference than the even more simplifying approaches discussed above, which assume the estimated model to be true (cf. Kim et al., 2018, pp. 20 ff; Lee and Valliant, 2009, p. 341).

Another foundation for estimating the potential deviation from the true population parameter, which even includes the bias component, is based on the work of Meng (2018) as well as Schouten (2007) and discussed in section 3.8. As described in that context, $\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}$ for design linear estimators can be expressed as a function of correlation, variance and sampling fraction, additionally incorporating the variation of weights where applicable. If a third variable is available that is strongly correlated to both the variables of interest and the inclusion indicator $\boldsymbol{r}^{\mathsf{nps}}$, an accuracy interval for $\widehat{\boldsymbol{\vartheta}}$ can be obtained from using non-probability and reference sample. Yet, the high magnitude of both correlations that is required to make this interval narrow enough to facilitate meaningful inference is rarely available in practical applications (cf. Schouten, 2007, p. 66). Therefore, this approach rather constitutes an indication for selectivity of non-probability samples in most cases, similar as the other methods discussed in chapter 3. Nevertheless, all of those techniques for assessing selectivity may provide at least some guidance for the accuracy of estimates obtained from a non-probability sample.

In summary, inference for non-probability samples is typically based on prediction or weighting models, corresponding to the model- or the pseudo-design-based approaches discussed in the preceding sections 5.1 to 5.3. Using the decomposition denoted in equations 5.188 and 5.189, methods that incorporate the between component $\mathbf{V}\left(\mathbf{E}\left(\widehat{\boldsymbol{\vartheta}} \middle| \boldsymbol{\Theta}\right)\right)$ are often considered as providing better results. These approaches commonly rely on resampling to approximate the variation of $\boldsymbol{\Theta}$ between possible non-probability samples. Nevertheless, a general inferential framework for non-probability sampling seems to be hardly feasible because all techniques proposed for this purpose have to rely on modeling assumptions (cf. Baker et al., 2013a, p. 105). Therefore, the possibility to assess the quality of a point estimator from a single non-probability sample is clearly more situational than in probability sampling and depends on whether these assumptions hold or not.

The degree to which the outlined inferential methods are suitable for different settings of estimation from non-probability samples is examined as part of the simulation studies discussed in the following chapter 6. The main purpose of these simulations is to evaluate and compare the performance of methods discussed throughout chapters 3 to 5 for assessing and correcting potential issues of non-probability samples.

6 Monte Carlo Simulation Studies

Various approaches proposed for assessing and compensating the potential issues of non-probability samples (cf. chapter 2) are discussed in the preceding chapters 3 and 5. In the current chapter 6, the aim is to investigate the performance and limitations of these methods. Typically, an estimator's quality is defined with respect to its distribution over all possible samples and the population statistic to be estimated (cf. e.g. equations 2.10 to 2.12). For an appropriate evaluation and comparison of the proposed methods, it is therefore necessary that the truth is known about both the underlying non-probability sample selection mechanism and the target population (cf. e.g. Buelens, Burger and van den Brakel, 2018, p. 327).

This kind of knowledge is typically not available in case of real data obtained from non-probability sampling. When nevertheless used for methodological evaluations, such realized data sets often pose further challenges (besides the selection mechanism) that limit the reliability of results. For example, differences between a non-probability and a reference (benchmark) sample in terms of survey mode and questionnaire design can introduce additional sources of potential errors. Such differences are not inherently linked to non-probability sampling, but typically not distinguishable from the bias caused by sample selection if only one non-probability and one reference sample are available (cf. chapters 2 and 7; Bethlehem and Biffignandi, 2012, pp. 97 ff, 242; Biffignandi and Artaz, 2012, p. 368; Elliott and Valliant, 2017, pp. 252 ff; Groves, 1989, pp. 295 ff; Weisberg, 2005).

For a fair and reliable evaluation and comparison of the methods proposed for non-probability samples, simulation studies are, therefore, typically more appropriate than applying methods to a single real non-probability sample (cf. Buelens, Burger and van den Brakel, 2018, p. 327; Enderle, Münnich and Bruch, 2013, p. 95). In the context of survey sampling and estimation, Monte Carlo simulations are a common and versatile way to analyze the properties of various estimators (cf. e.g. Bethlehem and Biffignandi, 2012, pp. 407 ff; Kim et al., 2018, pp. 12 ff; Münnich, Burgard and Vogt, 2013, pp. 178 ff; Rafei, Flannagan and Elliott, 2020, pp. 160 ff; Särndal, Swensson and Wretman, 1992, pp. 276 ff; Wolter, 2007, pp. 315 ff). A comprehensive summary of Monte Carlo simulations in survey statistics is provided by Burgard, Dörr and Münnich (2020), of which only some important aspects are outlined below. The general idea is to approximate the repeated sampling distribution of an estimator $\widehat{\boldsymbol{\vartheta}}$ by a large number of draws from a finite population and/or a super-population model. Different types of Monte Carlo simulation studies can be distinguished in dependence of how and what is drawn. In any case, such simulations allow strictly controlling all relevant influences on estimation and, hence, prevent contamination by other sources of errors when evaluating methods for non-probability samples (cf. Rafei, Flannagan and Elliott, 2020, p. 154). Furthermore, the true population and quantities to be estimated as well as the non-probability sampling process are known. To mimic the information that is available in real non-probability samples, this truth is not used for estimation purposes but nevertheless required for measuring an estimator's performance (cf. also Burgard, 2013, pp. 92 ff; Münnich et al., 2003; Rafei, Flannagan and Elliott, 2020, p. 161).

With regard to the differentiation of Monte Carlo simulations summarized by Burgard, Dörr and Münnich (2020, p. 17), all simulation studies in the current chapter 6 can be classified as *(quasi) design-based*. In each case, a fixed population is considered, and

randomness is only due to sample selection. This setting corresponds to the basic design-based framework introduced in sections 2.2 and 2.3. However, it is hardly ever possible to use real populations in simulations because full register or census data for these typically either does not exist or is not available due to confidentiality reasons (cf. Burgard et al., 2017b, p. 235; Merkle, Burgard and Münnich, 2016, p. 6). As a consequence, synthetic populations are used for the subsequent simulation studies. Yet, there are manifold types and characteristics of non-probability samples, e.g. in terms of highly different target populations, sample selection mechanisms and variables of interest. Depending on these characteristics, each of the various methods for non-probability samples may be more or less adequate because all these methods incorporate different assumptions (cf. chapters 3 and 5). Therefore, using only a single synthetic population and sampling mechanism limits generalizability to any other cases of non-probability sampling unless they strongly resemble the simulated one (cf. chapter 2; Burgard, Dörr and Münnich, 2020, pp. 16 ff; Kim et al., 2018, p. 12). To overcome this limitation and cover a variety of possible application settings for the methods under consideration, a set of fixed and finite scenario populations and sampling mechanisms is used in the following sections. For each of these populations, the methods are evaluated in a coinciding manner to achieve comparable results. To still not exclusively rely on synthetic data and simulated samples, an application study using a real non-probability sample is presented in the following chapter 7.

A summary of the software that is used in the simulation and application studies is given in section 6.1. For the methods that are newly proposed or extended in the context of this thesis, novel computational implementations are provided and then evaluated in a preliminary simulation study, which is discussed in section 6.2. To assess all methods for non-probability samples discussed in chapters 3 and 5, a larger Monte Carlo simulation is presented in section 6.3.

6.1 Software for the Considered Methods

Simulation and application studies throughout the subsequent sections are implemented in the statistical computing software `R` (cf. R Core Team, 2018). Many of the methods discussed in chapters 3 to 5 are readily accessible in `R` or one of the various existing packages that can be used to extend it. Available packages used for the simulation studies are summarized in section 6.1.1. For the methods that are newly proposed in the context of this thesis, custom-made implementations are developed. A corresponding overview of these implementations and the utilized software libraries is provided in section 6.1.2.

6.1.1 Pre-existing Software for Established Methods

Tests for selectivity discussed in section 3.4 constitute a first group of methods for which implementations are required. While ANOVA as well as (univariate) Kolmogorov-Smirnov and Kruskal-Wallis test are implemented in the elementary `R`-software as part of the `stats` package (cf. R Core Team, 2018), equivalence tests for means (in the univariate case) can be performed using the `R`-package `TOSTER` (cf. Lakens, 2017). The tests proposed by Hawkins (1981) and Anderson and Darling (1952) are implemented in the `R`-package `MissMech` (cf. Jamshidian, Jalal and Jansen, 2014). An alternative and faster implementation for the Anderson-Darling test is available in the `C`-implementation provided by Marsaglia and Marsaglia (2004). As a prerequisite for applying Rosenblatt's

(1952) theorem, the package `uniftest` (cf. Melnik and Pusev, 2015) provides tests for uniformly distributed variables, e.g. the one proposed by Neyman (1937; cf. Ledwina, 1994). An implementation for matching (cf. section 3.5) is provided in the package `Matching` (cf. Diamond and Sekhon, 2013).

In addition, implementations for prediction models discussed in section 5.1 are required. Generalized additive (mixed) models are available from the package `mgcv` (cf. Wood, 2011; 2017). The MARS algorithms are implemented in the `R`-Packages `earth` for the MARS model (Milborrow, 2019) and `rpart` (cf. Therneau and Atkinson, 2018) for the special case of regression trees. Support vector machines are available from the package `e1071` (cf. Meyer et al., 2019). Since this package does not allow for individual weights but is only a user interface for the `C++` library `LIBSVM` (cf. Chang and Lin, 2011), an available extension for `LIBSVM` (cf. Chang et al., n.d.) is used to allow for such individual weights. The `R`-package `glmnet` (cf. Simon et al., 2011) provides penalized versions of generalized linear regression models (LASSO, ridge and elastic net). Multivariate sample selection and dependent variable models, such as the Heckman model (cf. section 5.3.1), are available from package `sampleSelection` (cf. Toomet and Henningsen, 2008). Further packages which are used in the simulations include `data.table` (cf. Dowle and Srinivasan, 2019), `mvtnorm` (cf. Genz and Bretz, 2009) and `survey` (cf. Lumley, 2004). For the methods that are newly proposed in the context of this thesis, custom-made implementations are introduced in the following section 6.1.2 because pre-existing software does not exist.

6.1.2 Software Implementation of Newly Proposed Methods

The two main methodological novelties proposed in the context of this thesis are semi-parametric neural networks (cf. section 5.1.9) and a calibrated version thereof (cf. section 5.2.3). As these methods are newly suggested, there is no pre-existing software implementation. Inter alia to make these ideas assessable in the following simulation and application studies, custom-made computational implementations in `C++` are developed as part of the following `R`-packages: package `sqp` (cf. appendix C.1; Lenau, 2020) provides the basic routines for sequential quadratic programming as well as unconstrained optimization using the `BFGS` algorithm (cf. section 4.2). It serves as a foundation to be used in the other packages. Semi-parametric artificial neural networks are implemented in package `ann` (cf. appendix C.2). Among other options, this package provides the possibility for B-spline layers and corresponding fitting routines (backpropagation as well as `BFGS`), which are partially based on the `sqp` package. Both of these packages constitute the foundation for package `calmod` (cf. appendix C.3), which facilitates calibrated semi-parametric artificial neural networks as response models with calibration constraints. Its optimization routines are entirely based on `sqp`.

Implementation of these three packages is largely based on the linear algebra library `Armadillo` (cf. Sanderson and Curtin, 2016; 2018), which provides a "high-level application programming interface" (Sanderson and Curtin, 2016, p. 1). It serves as a wrapper for the linear algebra package `LAPACK` (cf. Anderson et al., 1999; Demmel, 1997), which can use any adequate implementation of basic linear algebra subprograms (`BLAS`), such as `RBLAS` (cf. R Core Team, 2018) or `Open BLAS` (cf. Wang et al., 2013). The latter is used for the simulations. For integration of `C++`, `Armadillo` and `R`, the `R`-packages `Rcpp` (cf. Eddelbuettel et al., 2011) and `RcppArmadillo` (cf. Eddelbuettel and Sanderson, 2014) are used.

As described in chapter 4, an elementary component of the applied optimization techniques is to solve systems of linear equations. With a large variety of freely available software libraries implementing linear solvers, a selection of them is available and can be used for this purpose in `sqp` (and thus `ann` and `calmod`). The default solver, which is implemented in the software library `SuperLU` (cf. Demmel, Gilbert and Li, 1999; Li, 2005), uses LU-factorization (Gaussian elimination) as presented in section 4.1. By choice of the user, it can be interchanged with other approaches, for instance Cholesky- (cf. Martin, Peters and Wilkinson, 1965; 1966) or QR-factorization (cf. Businger and Golub, 1965). Further available options are the conjugate gradient (cf. Fletcher and Reeves, 1964; Hestenes and Stiefel, 1952) and generalized minimal residuals (cf. Saad and Schultz, 1986) methods. To incorporate these options, the `C++` libraries `ViennaCL` (cf. Rupp et al., 2016) and `Eigen` (cf. Guennebaud, Jacob et al., 2010) are used, the latter in conjunction with its `R`-integration provided in package `RcppEigen` (cf. Bates and Eddelbuettel, 2013). In addition, `calmod` contains one complete out-of-the-box implementation of `SQP`, which is available from the `NLOPT`-library for non-linear-optimization (cf. Johnson, n.d.). This is the sequential least squares quadratic programming (`SLSQP`) approach proposed by Kraft (1988; 1994). As further utility functions closely related to the pseudo-design weights, `calmod` provides implementations of weighted estimates and R-indicator (cf. sections 2.2 and 3.7). The remaining methods discussed in chapters 3 to 5 but not explicitly mentioned throughout the current section 6.1 are either readily available in `R` itself or can be straightforwardly implemented based on its utility.

6.2 Prior Applicability Studies for the Developed Methods and Software

As described in the previous section 6.1.2, three `R`-packages are developed in the context of this thesis to provide computational implementations for (calibrated) semi-parametric neural networks. In the current section 6.2, two introductory simulation studies are presented. Their purpose is to test the proper functionality of the custom-made implementations and illustrate basic ideas and potential advantages of the newly proposed methods and developed `R`-packages.

6.2.1 Prior Evaluation of Semi-parametric Artificial Neural Networks

The proposed semi-parametric neural networks (cf. section 5.1.9) apply B-spline transformations and optimize the respective knot locations. To evaluate this basic idea, a first Monte Carlo simulation is presented. In summary, the purpose of this study is to compare the following options to select B-spline knots for regression contexts:

a) knots are evenly spaced over the range $\boldsymbol{X}$,

b) knots are located at the quantiles of $\boldsymbol{X}$, and

c) knots are optimized as parameters of the model.

Strategies ***a)*** and ***b)*** are commonly used for regression splines, such that they constitute classical additive models (cf. section 5.1.6). Option ***c)*** can be realized in form of a semi-parametric artificial neural network and corresponds to a regression spline with joint optimization of knots and regression parameters (cf. section 5.1.9). These three options are compared with regard to their predictive performance.

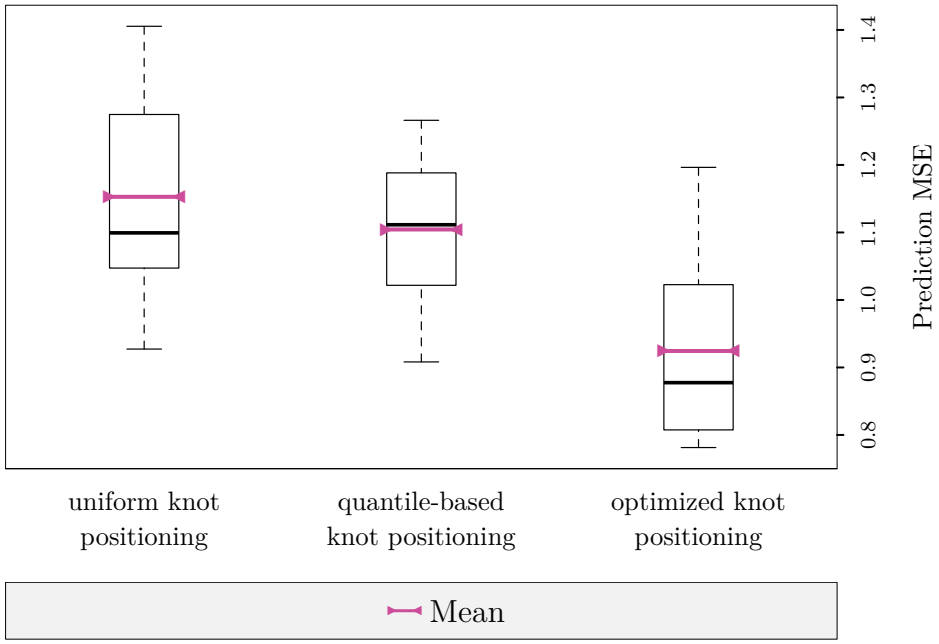


Figure 6.1: Comparison of knot positioning methods for out-of-sample predictions

The aim in this preliminary study is solely to evaluate the idea of knot optimization in ANNs and the provided computational implementation. This strategy is compared to the outlined alternatives that choose predetermined knots. Therefore, the following simulation setup is rather simplistic and does not represent a realistic scenario of survey statistics. Non-probability sampling is not considered here, but incorporated in the following simulations. In the current setting, two variables are generated in a finite population of size $N = 10\,000$. The auxiliary variable $\boldsymbol{X}$ is log-normally distributed, such that $\log(\boldsymbol{X}) \sim N(0, 1)$. The target variable $\boldsymbol{Y}$ follows a conditional normal distribution given $\boldsymbol{X}$, such that $\boldsymbol{Y} \sim N(0.5 \cdot \boldsymbol{X}, 1)$, and the correlation between both variables is $\boldsymbol{\rho_{XY}} = 0.5$. From this finite population, 10 000 simple random samples of size $n = 100$ are drawn.

Each of these samples is used to fit three prediction models. All of these use $\boldsymbol{X}$ as input variable, apply a B-spline transformation to it and determine predictions for the population as a linear combination of the B-spline base functions, as discussed in section 5.1.9. The only difference between the three models that are fit to each sample lies in the selection of four knots $\boldsymbol{K^X} \in \mathbb{R}^4$, which is done in accordance to strategies ***a)*** to ***c)*** outlined above. Considering the population values of $\boldsymbol{X}$ as known auxiliary information, these three options are compared with respect to their predictive power. The main relevance of prediction models in the context of non-probability sampling and this thesis is to provide out-of-sample predictions (cf. figure 5.1; Buelens, Burger and van den Brakel, 2018, p. 325). Therefore, the prediction error made by each approach is measured by the prediction MSE over all elements of the population, i.e. the mean-squared-error when the model is fit on the sample and predictions are made for the whole population.

The results are depicted in figure 6.1, with boxplots representing the distribution of prediction MSEs over all 10 000 samples. Means over all iterations are indicated in purple. The comparison reveals that knots which are evenly spaced over the range of $\boldsymbol{X}$ yield the highest minimal, average and maximal MSE of the three knot selection strategies. Placing the knots at quantiles of $\boldsymbol{X}$ performs slightly better. In this case, minimum,

average and maximum of the resulting prediction MSEs are lower than for evenly spaced knots, and the same holds for the variance. Even better results are achieved when using knot optimization in an ANN. When applied for prediction in the current setting, this strategy is able to further reduce the minimal, mean and maximal MSE.

Under the rather simple conditions in this simulation and in comparison to both approaches for choosing predetermined knots, the proposed optimal knot selection strategy, thus, allows reducing the average prediction error for semi-parametric models. Therefore, the method and its implementation perform as expected. Altering the conditional distribution for $\boldsymbol{Y}|\boldsymbol{X}$, e.g. by using a conditional log-normal instead of a normal distribution, yields highly similar results when comparing the three knot selection strategies. Any further tuning of hyper-parameters, e.g. in form of cross-validation to determine optimal shrinkage parameters, is not considered in this preliminary study but may be used to improve the results for all of the three options. Furthermore, no conclusions regarding the relative performance of other prediction models (cf. section 5.1) can be drawn. This limitation is tackled in the more realistic and comprehensive simulation study presented in section 6.3. As a preparation for that simulation, incorporation of calibration constraints when fitting semi-parametric ANNs is subject to further preliminary evaluations, which are presented in the following section 6.2.2.

6.2.2 Prior Evaluation of Calibrated Semi-parametric Artificial Neural Networks

A calibrated extension to semi-parametric neural networks is proposed in section 5.2.3, in particular to facilitate response (propensity) models that can incorporate calibration constraints. The basic idea and custom-made implementation of such models for pseudo-design weighting is tested and evaluated in the current section 6.2.2 by means of a second preliminary Monte Carlo simulation. In a first step, three options for pseudo-design weights are considered:

a) Propensity weights based on a generalized linear logit model.

b) Propensity weights from option ***a)*** are calibrated in a second step using the GREG.

c) A calibrated neural network with soft calibration and logit (softmax) activation function is used to obtain propensity weights.

These methods represent a bandwidth of those discussed in section 5.2. The first two approaches are well-established ones (cf. e.g. Enderle, Münnich and Bruch, 2013, p. 94; Särndal and Lundström, 2005, pp. 51 f) and included for illustrative reasons only. The main purpose of the following study is to assess the applicability of the newly proposed option ***c)***. This calibrated ANN basically constitutes a logit model that is fit w.r.t. soft calibration constraints. Therefore, it integrates the structural form of the response propensity model outlined in option ***a)*** with calibration constraints as added in option ***b)***, using a single rather than two separate steps to determine the weights.

Because the simulation study is designed solely for testing the proposed method and software, it is again based on a simple model rather than a realistic scenario. Two variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ for a finite population of size $\mathrm{N} = 100\,000$ are generated from a bivariate normal distribution, such that

$$\begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu_X} \\ 10 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma_X} & 0.3 \cdot \sqrt{\boldsymbol{\Sigma_X}} \\ 0.3 \cdot \sqrt{\boldsymbol{\Sigma_X}} & 1 \end{bmatrix} \right) \quad . \tag{6.1}$$

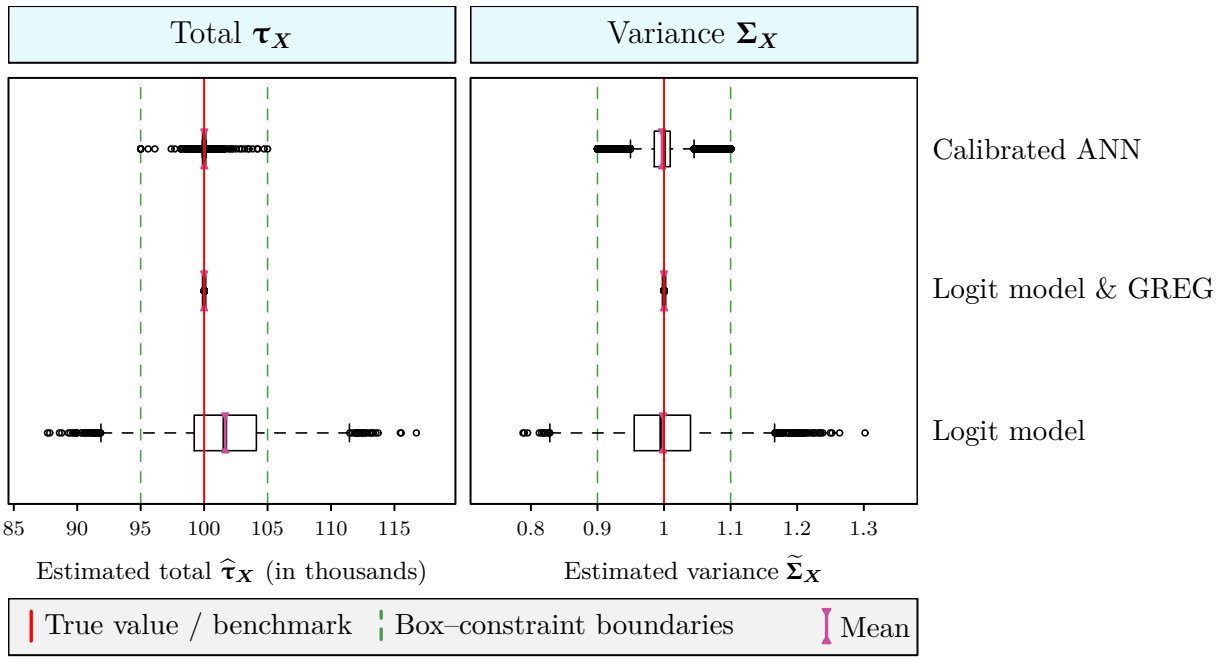


Figure 6.2: Compliance with total and (co-)variance benchmarks when combining response and calibration weighting for $\boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{\Sigma}_{\boldsymbol{X}} = 1$

From this population, 10 000 non-probability samples of size $n^{\mathsf{nps}} = 500$ are drawn. Sample selection is implemented by using unequal probability sampling without replacement, where selection probabilities $\boldsymbol{\pi}^{\mathsf{nps}}$ are correlated with $\boldsymbol{X}$ and $\boldsymbol{Z}$ by $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.3$ and $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$. These probabilities are not used for estimation but only for selecting the samples, which is a common way to simulate a non-probability sampling process (cf. section 6.3; Valliant and Dever, 2011, p. 122). For fitting response propensity models as outlined above, a complementary reference sample res of size $n^{\mathsf{res}} = 250$ is drawn by simple random sampling without replacement for each of the 10 000 non-probability samples.

To evaluate the three weighting techniques summarized above in options ***a)*** to ***c)***, $\boldsymbol{X}$ is used as calibration variable, for which total $\boldsymbol{\tau}_{\boldsymbol{X}}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ are available as calibration benchmarks. The GREG is calibrated to $\boldsymbol{\tau}_{\boldsymbol{X}}$ and $\boldsymbol{\tau}_{\boldsymbol{X}^{\circ 2}}$ (the total of the squared $\boldsymbol{X}$-variable) to meet both constraints (cf. section 5.2.3). For the response (propensity) model, $\boldsymbol{X}$ and $\boldsymbol{Z}$ observed in the non-probability and the reference samples are used as auxiliary variables, additionally including an intercept in the model. A parametric logit GLM is used as propensity model. The calibrated ANN is specified accordingly, with the only modification that it incorporates soft calibration constraints. Since the aim is to evaluate the compliance of calibrated ANNs with calibration constraints when using the `R`-package `calmod`, estimators for the total and variance of $\boldsymbol{X}$ are considered. For the three weighting methods, the distribution of resulting estimates from all 10 000 samples is shown as boxplots in figure 6.2 for the case of $\boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{\Sigma}_{\boldsymbol{X}} = 1$. Estimates and (soft) calibration constraints for the total $\boldsymbol{\tau}_{\boldsymbol{X}}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ are respectively shown in the left and right panel of the figure.

As the logit model in option ***a)*** does not incorporate calibration, the resulting propensity weighted estimates are not limited by any constraint. As a consequence, the variation of the resulting estimates for total and variance of $\boldsymbol{X}$ is the highest of all three weighting approaches. Furthermore, there is some bias when estimating the total $\boldsymbol{\tau}_{\boldsymbol{X}}$, as the Monte Carlo mean of the estimates (purple line) is different from the true population value (red

line). Because the theoretical foundation of propensity weighting for eliminating bias holds for true rather than estimated response propensities, this remaining bias is caused by the small sampling fractions and, thus, vanishes with increasing sample sizes n^{nps} and n^{res} (cf. also King and Nielsen, 2019, p. 443). In the present context, however, it illustrates the use of calibration to further reduce bias. In contrast to pure propensity weighting, applying the GREG in a second step to calibrate propensity weights towards true total and variance of $\boldsymbol{X}$ (option $\boldsymbol{b}$)) results in estimates that are unbiased and constant over all replications. All estimates coincide with the calibration targets because the GREG enforces exact compliance with these benchmarks. The calibrated neural network outlined in option $\boldsymbol{c}$) applies soft calibration, permitting 5% deviation from the population total and 10% from the variance. Variation of estimates is permissible and occurs within the interval defined by these boundary constraints, which are shown as green dashed lines. In that regard, calibrated neural networks constitute a middle course, with limiting cases corresponding to either one of the other two represented methods. By choosing infinitely small boxes, exact calibration can be enforced, as in case of the GREG. Choosing infinitely large boxes can lead to plain propensity weights obtained from the logit model.

To exactly achieve these limiting cases, the applied loss function has to be determined correspondingly to incorporate deviance as well as penalization for soft calibration (cf. section 5.2.3). To that end, a vector of importance weights $\mathbf{v}$ is introduced in equation 5.155. These weights facilitate a flexible combination of the ANN's distance function (i.e. the deviance / negative log-likelihood of the response model) with penalties for soft calibration and parameter shrinkage. Therefore, the distance measure used for fitting the weighting model is smoothly adaptable between these components. One aim of this approach is to represent existing weighting methods as special cases of the proposed calibrated ANNs, which is e.g. the case for weights obtained from the logit propensity model or generalized regression estimation depicted in figure 6.2.

Beyond representing such special cases, calibrated semi-parametric neural networks facilitate a smooth transition between the ideas of propensity weighting on the one and calibration weighting on the other side, facilitating combinations and trade-offs between both approaches. For this purpose, importance weights $\mathbf{v}$ have to be chosen to represent the desired combination and degree of penalization in the distance measure. An open question in this regard is how to select the respective importance weights for the different elements of the loss function (cf. section 5.2.3; Chang and Kott, 2008, p. 559; Guggemos and Tillé, 2010, p. 3205; Marler and Arora, 2004, p. 375; Rupp, 2018, pp. 150 f). Figure 6.3 illustrates the influence of these importance weights and their default values in R-package `calmod`.

To that end, six out of the 10 000 samples represented in figure 6.2 are selected for illustrative purposes, which are labeled sample #1 to #6 for simplicity. Estimation of the total $\boldsymbol{\tau}_{\boldsymbol{X}}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ in each of the six samples is depicted in figure 6.3. The resulting estimates are represented in dependency of the importance weights that control the degree of penalization for soft calibration towards the corresponding benchmarks. In correspondence to equation 5.155, these importance weights for soft calibration of total and variance are respectively denoted by v_5 and v_6. All other importance weights, i.e. for the deviance or parameter shrinkage, are kept constant at their default values in `calmod`, which are discussed below. Also, the penalty for variance soft calibration is kept fixed for total estimates, such that only the penalization for soft calibration of the total is adjusted in the first row of figure 6.3. The reverse holds for variance estimates in the second row.

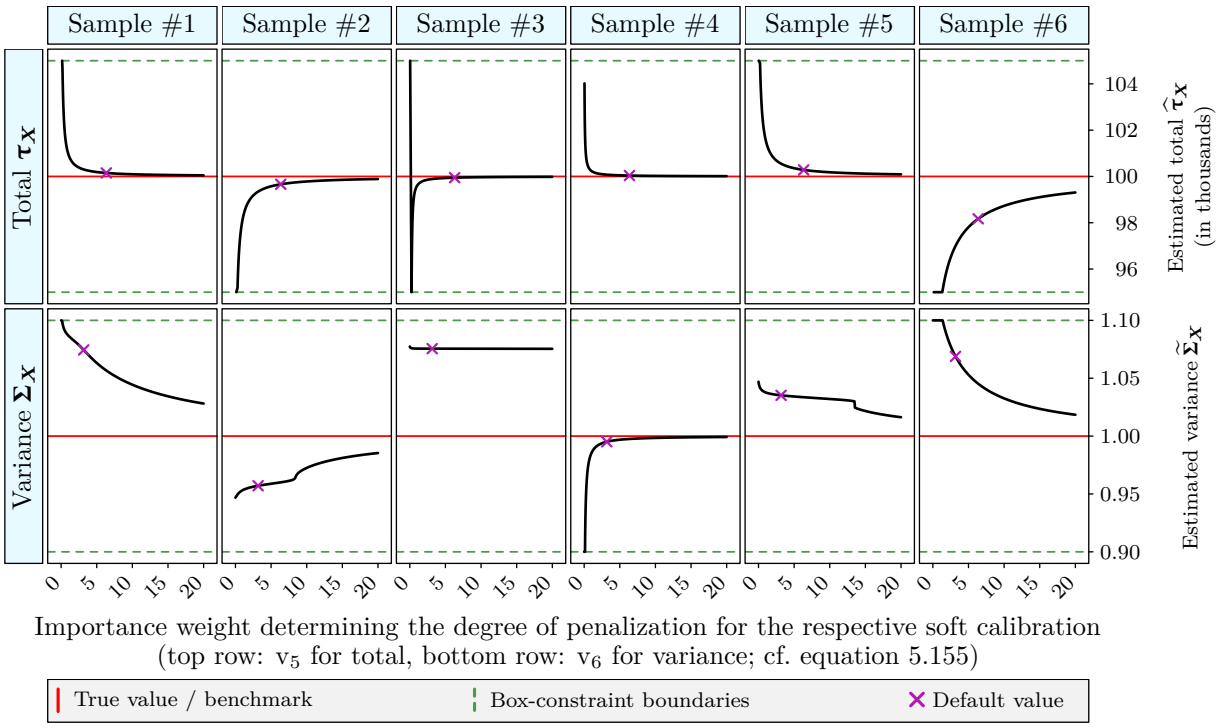


Figure 6.3: Influence of importance weights when fitting calibrated ANNs: importance weights for the soft calibration distance components

In most of the six samples, it is evident that the distance between estimated and true calibration benchmark is a monotonically decreasing function of the respective importance weight. This is the desired behavior and main rationale for including soft calibration penalties for optimization (cf. also Rupp, 2018, pp. 150 f). Sample #3 constitutes a somewhat inconvenient exception because values of v_6 show almost no visible influence on estimates of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. Since calibration is non-linear in the response model's parameters and only three of such parameters are used in the present model, this can occur because the set of feasible solutions does not allow any other result. Even more important, the above-mentioned monotonicity does not hold in sample #3 if soft calibration of the total is nearly unpenalized, i.e. when v_5 is close to zero. The reason is that in this case, the size of the response model's deviance overshadows all other components in the distance metric. To avoid soft calibration penalties which are nearly negligible for the overall distance measure, the respective importance weights should therefore not be too small. A general rule to determine these weights is, however, not directly evident. This finding is underlined by the results obtained from the remaining samples. For sample #4, comparably small values of v_5 and v_6 suffice to bring estimates close to the total and variance benchmark. In contrast, penalties for both soft calibration constraints need to be relatively large for sample #6 to achieve estimates that are at least fairly close to the calibration targets. Results for the other samples mostly lie somewhere in between. In general, the calibrated estimates resulting from coinciding degrees of penalization expressed by the values of v_5 or v_6 can be highly different between samples. Except for the inconvenient exception in sample #3 mentioned above, the influence of v_5 on meeting the benchmark total $\boldsymbol{\tau}_{\boldsymbol{X}}$ at least seems to be of similar shape across the samples, but the strength of this influence differs. For estimates of the variance, the differences between samples are even more severe and benchmarks are not necessarily met even for strong penalization. As discussed above, this is caused by the low number of parameters in relation to the number of non-linear constraints.

These findings suggest that predetermining importance weights v_5 or v_6 for soft calibration without considering the actual sample may not be an adequate strategy. A sample-dependent choice for the degree of penalization appears more reasonable for two reasons. On the one hand, samples can strongly differ regarding their similarity to the population even if they are realizations from the same sample selection mechanism. This holds already for perfect probability sampling and may be even more severe in case of non-probability samples (cf. chapter 3; Kish, 1965, p. 403; Särndal, Swensson and Wretman, 1992, p. 41). Therefore, closeness of estimates and calibration benchmarks can be very different already before fitting a weighting model. The fact that this closeness influences the initial conditions for optimization, e.g. in terms of starting values and the need for slack variables, should be accounted for when choosing the degree of penalization (cf. algorithm 8). On the other hand, importance weights for the response model's deviance (negative log-likelihood) are kept constant in figure 6.3. The deviance as distance component does not only depend on the non-probability sample's composition but also on that of the reference sample. Therefore, the choice of importance weights should generally consider that distance components can take quite different values for distinct samples (cf. equation 5.155; sections 5.1.3 and 5.1.8)

The default implementation of importance weights in `calmod` when combining calibration and propensity weighting is based on the above considerations. When no user-defined importance weights are provided, the strategy is to determine elements of $\mathbf{v}$ such that each component of the distance function (deviance, penalties for total and covariance soft calibration as well as parameter shrinkage) defined in equation 5.155 has the same maximum. Since only the relation of importance weights is relevant for minimizing the loss function (up to a scaling constant; cf. Chang and Kott, 2008, p. 569), it is arbitrary to define this maximum for each component to be one. All default choices of importance weights are defined to meet this maximum.

Considering the general case and the notation introduced in section 5.2.3, the vector of optimization parameters $\boldsymbol{\Theta}$ contains s weighting parameters as well as p total and r covariance multipliers for calibration. The response model's (ANN's) maximum deviance is assumed to be that of the null model, which is the response model where all non-intercept parameters are zero. The first default entry v_1 is consequently given by

$$v_1 \;=\; 1 \Big/ \delta_{\mathrm{m}}\left(\boldsymbol{\omega}^{(0)}\right) \qquad , \tag{6.2a}$$

where $\delta_{\mathrm{m}}\left(\boldsymbol{\omega}^{(0)}\right)$ is the ANN's distance function for the initial weighting coefficients $\boldsymbol{\omega}^{(0)}$, which are all zero except for the intercept(s), i.e. the deviance of the null model.[7] Nevertheless, note that it is possible to construct cases where the final deviance of the response model after optimization is actually higher than that of the null model. The simplest example occurs when using box-constraints for the parameters that are highly contradictory to the actual data. Such cases, however, are of a theoretical rather than practical relevance when a plausible response model is specified. The deviance typically does not exceed $\delta_{\mathrm{m}}\left(\boldsymbol{\omega}^{(0)}\right)$, such that it seems reasonable to assume this as an upper

[7] The intercept parameters for the null model can by determined from the initial population size estimates from non-probability and reference sample, e.g. as $\log\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{nps}}\right) \Big/ \widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{res}}\right)\right)$ for the logit model used in figures 6.2 and 6.3. The vector of initial weights $\boldsymbol{w}^{\mathsf{nps}}$ for the non-probability sample can be scaled in advance to meet potential population size calibration constraints if necessary.

bound. For all other distance components, the maximum value depends on the vectors of centering constants $\mathbf{C}_{\boldsymbol{\Theta}}$ as well as lower and upper boundaries of the box-constraints $\boldsymbol{L}_{\boldsymbol{\Theta}}$ and $\boldsymbol{U}_{\boldsymbol{\Theta}}$. The maximum contribution of each element of the optimization parameters $\boldsymbol{\Theta}$ can therefore be expressed by a vector

$$\boldsymbol{d} \coloneqq \frac{1}{2} \cdot \mathbf{Rowmax}\left(\left[(\boldsymbol{U}_{\boldsymbol{\Theta}} - \mathbf{C}_{\boldsymbol{\Theta}})^{\circ 2} \quad (\mathbf{C}_{\boldsymbol{\Theta}} - \boldsymbol{L}_{\boldsymbol{\Theta}})^{\circ 2}\right]\right) \quad \in \mathbb{R}_{\geq 0}^{s+p+r} \quad . \tag{6.2b}$$

Based on definition 6.2b, default importance weights for all other distance components are hence determined by

$$\mathbf{v}_{\mathcal{J}} \;=\; 1/||\boldsymbol{d}_{\mathcal{J}}||_1 \tag{6.2c}$$

for all $\mathcal{J} \in \{\{2, \ldots, s+1\}, \{s+2, \ldots, s+p+1\}, \{s+p+2, \ldots, s+p+r+1\}\}$, i.e. for parameter shrinkage, total and covariance calibration penalties. This ensures that the highest possible contribution of each of these components to the overall distance function is equal to one.

Note that the shrinkage penalty for weighting parameters in equation 5.155 is

$$\sum_{j=1}^{s} \frac{\mathrm{v}_{(j+1)}}{2} \cdot \left(\omega_j - \mathrm{C}_{\omega_j}\right)^2 \;= \frac{1}{2} \cdot \mathbf{v}_{\mathcal{J}}^{\mathsf{T}} \left(\boldsymbol{\Theta}_{\mathcal{J}} - \mathbf{C}_{\boldsymbol{\Theta}_{\mathcal{J}}}\right)^{\circ 2} \quad \text{for} \quad \mathcal{J} = \{2, \ldots, s+1\} \quad , \tag{6.3}$$

for which the maximum value is set to be the same as for the other components of the distance function when using the default importance weights defined in equations 6.2. It may seem that this leads to over-shrinkage of weighting parameters ω_j because penalization of parameters deviating from centering values C_{ω_j} is too strong. However, calibration constraints are non-linear in $\boldsymbol{\omega}$ and initial values $\boldsymbol{\omega}^{(0)}$ are usually relatively far from an optimal solution. Therefore, the feasible range $\boldsymbol{U}_{\boldsymbol{\omega}} - \boldsymbol{L}_{\boldsymbol{\omega}}$ between lower and upper bounds for $\boldsymbol{\omega}$ typically needs to be much larger than that for the soft calibration multipliers $\boldsymbol{\epsilon}$ and $\boldsymbol{\varepsilon}$ to be adequate for numerical optimization. By considering these boundaries in equality 6.2b, penalization of weighting parameters when using default values for $\mathbf{v}$ is therefore much lower than for soft calibration and typically does not lead to over-shrinkage (cf. also Hastie, Tibshirani and Friedman, 2008, pp. 156 ff; Wood, 2017, p. 128).

A further benefit of equalities 6.2 is that they help to prevent certain distance components from becoming overly dominant in the optimization problem, which would make all other components negligible in relation. As discussed with regard to sample #3 in figure 6.3, this problem can occur if e.g. soft calibration penalties or the response model's deviance are much higher than the remaining components. The issue of one component being dominant for optimization is at least partially counterbalanced when using importance weights which enforce all distance components to have the same maximum (cf. Craven and Wahba, 1979, p. 379; Hastie, Tibshirani and Friedman, 2008, pp. 228 ff; Wood, 2011, p. 8).

Note that for the related but simpler special cases in problems 5.150 and 5.151, Chang and Kott (2008, p. 559) as well as Guggemos and Tillé (2010, p. 3205) propose alternative strategies for finding importance weights that are based on some quality measure (e.g. the estimated variance of a point estimator). However, optimizing importance weights with respect to a single estimator has the general drawback that it somewhat limits the use of weights for multi-purpose estimation (cf. section 5.2). An even stronger argument against such an approach in the present context is that suitable methods for estimating variances or MSEs in non-probability samples are currently at best situationally available

(cf. section 5.4) and, therefore, hardly usable to determine $\mathbf{v}$. Consequently, default importance weights in `calmod` are based on equalities 6.2 rather than being optimized with regard to a quality measure.

In figure 6.3, the default importance weights and corresponding estimated values are represented as purple crosses. It is evident that in samples where it is relatively cheap to come close to the calibration targets, a high degree of alignment with these benchmarks is obtained when using the default values. In cases where it is relatively costly, only some adjustment towards the targets is made. The results for the calibrated ANNs in figure 6.2 are entirely based on the default values. For most samples, the resulting estimates are quite close to the calibration benchmarks. However, this becomes too costly some cases, such that estimates deviate farther from the benchmarks but box-constraints still limit the estimates' feasible range and prevent results that are too far off from the calibration targets. Nevertheless, the above proposal solely concerns the computation of default values in `R`-package `calmod`. These are used only when no choice for $\mathbf{v}$ is made by the user. Arbitrary user-defined importance weights can nevertheless be specified (cf. appendix C.3).

As outlined with regard to the estimated variance for sample #3 in figure 6.3, constraints in calibrated neural networks can be quite restrictive. Especially when using a parametric calibrated response model as in the above examples, the number of constraints is often not much smaller than that of the weighting model's parameters ($\boldsymbol{\omega}$) and calibration is typically non-linear in these parameters. In some cases, this can lead to the feasible set of solutions being empty, such that the problem is unsolvable. In figure 6.2, roughly 2.8% of the samples are excluded, for which infeasible constraints prevent finding valid solutions. This problem is not specific to the implementation in `calmod`, and it is very difficult to detect such cases a priori. Possible remedies can be the relaxation of constraints or to rely on a response model that is more flexible, e.g. due to a larger number of free parameters (cf. Johnson, n.d.; Rupp, 2018, pp. 150 f).

An example for using a high number of model parameters is the GREG. In contrast to the three weighting parameters employed by logit model and calibrated ANN in the above example, the GREG optimizes the weights themselves, and thus employs one optimization parameter for each observation in the non-probability sample (cf. section 5.2.2). As the number of free weighting model parameters is equal to the sample size, the corresponding calibration problem is solvable for all samples considered in figure 6.2. For a fair comparison, it is useful that the GREG can be interpreted as special case of a calibrated artificial neural network for which a single linear activation layer, an identity matrix as independent variables and exact calibration are used (cf. section 5.2.3). As a second part of the above simulation, computational performance of the GREG (cf. equation 5.145) is compared with such an equivalent calibrated ANN. Both weighting methods constitute alternate formulations for solving the same optimization problem and hence offer the same degree of flexibility. The GREG is calculated using the `R`-package `survey`, the calibrated ANN using `calmod`. Exact calibration is enforced in both cases, i.e. exact rather than soft constraints are considered for the ANN. Because one weighting parameter is used for each observation in the non-probability sample, the variable $\boldsymbol{Z}$ and the reference sample are irrelevant for the optimization problem and therefore not used in this context. The remaining setup of the simulation is the same as before, considering a range of values for the mean $\boldsymbol{\mu}_{\boldsymbol{X}}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ of $\boldsymbol{X}$ for data generation model 6.1.

Table 6.1: Numerical stability and coincidence of the GREG and an equivalent calibrated ANN (each using one parameter per observation) for different values of $\boldsymbol{\mu_X}$ and $\boldsymbol{\Sigma_X}$

		Total calibration			**Total and variance calibration**		
		Valid solutions		Max. diff.	Valid solutions		Max. diff.
$\boldsymbol{\mu_X}$	$\sqrt{\boldsymbol{\Sigma_X}}$	GREG	cal. ANN		GREG	cal. ANN	
10^0	10^0	100	100	0	100	100	0
	10^1	100	100	0	100	100	0
	10^2	100	100	0	100	100	0
	10^3	100	100	0	100	100	0
	10^4	100	100	0	0	100	
10^2	10^0	100	100	0	18	100	0
	10^1	100	100	0	100	100	0
	10^2	100	100	0	100	100	0
	10^3	100	100	0	100	100	0
	10^4	100	100	0	0	100	
10^4	10^0	0	100		0	25	
	10^1	100	100	0	0	0	
	10^2	100	100	0	0	100	
	10^3	100	100	0	0	100	
	10^4	100	100	0	0	100	

All numbers are in (rounded) percentage points.
cal. ANN: Calibrated artificial neural network
Valid solutions: Share of samples for which calibration constraints are met
Max. diff.: Maximal unsigned relative difference between valid weights from both methods in all 10 000 samples. Empty cells denote cases where at least one of the methods does not result in any valid solutions.

The shares of valid solutions found with the GREG and the calibrated ANN are represented in table 6.1. These shares represent the fractions of samples for which the resulting weights actually fulfill the imposed calibration constraints. Although some scenarios are rather artificial, it turns out that the classical formulation of the GREG often provides less numerical stability than fitting an equivalent calibrated neural network. For $\boldsymbol{\mu_X} = 1$, both options perform similar unless the variance of $\boldsymbol{X}$ is high and used for calibration. In this case, the GREG needs to calibrate the total $\boldsymbol{\tau_X}$ of $\boldsymbol{X}$ as well as that of the squared variable, $\boldsymbol{\tau_{X^{\circ 2}}}$. It fails to meet the benchmarks because the numerical solution becomes unstable when one of the targets is much larger than the other. Therefore, valid weights are obtained for none of the samples by the classical GREG-formulation when $\boldsymbol{\mu_X} = 1$ and $\sqrt{\boldsymbol{\Sigma_X}} = 10\,000$. In contrast, the calibrated ANN still complies with all constraints in this case because the variance is calibrated as an actual central moment, rather than using the ordinary moment of the squared variable. The results are similar for $\boldsymbol{\mu_X} = 100$, except for the GREG with variance constraint being unreliable also for the case of $\boldsymbol{\Sigma_X} = 1$. Due to the fact that the ratio $\boldsymbol{\tau_X}/\boldsymbol{\tau_{X^{\circ 2}}}$ is again highly different from one in this case, calibration constraints are met in only 18% of the samples. For these first two choices of $\boldsymbol{\mu_X}$ discussed so far, total constraints alone do not cause any issues with both methods, and additional calibration of the variance is always feasible for the calibrated ANN formulation. Both

changes when the mean of $\boldsymbol{X}$ is increased even further, for $\boldsymbol{\mu_X} = 10\,000$. In case of low variance $\left(\sqrt{\boldsymbol{\Sigma_X}} = 1\right)$, even total calibration alone is not feasible with the classical GREG formula. The same holds when adding a variance constraint, regardless of the size of this variance. In contrast, the ANN with only a total constraint yields valid weights for all samples and values of $\boldsymbol{\Sigma_X}$. When incorporating calibration of the variance, however, the share of valid solutions drops to 25% for $\sqrt{\boldsymbol{\Sigma_X}} = 1$ and 0% for $\sqrt{\boldsymbol{\Sigma_X}} = 10$. Calibration of higher variances performs well, such that this is presumably for the same reasons that hinder the GREG from finding solutions in many scenarios, i.e. a coefficient of variation that is very different from one. In all situations where both methods find valid solutions to the calibration problem, the respective weights coincide. This is denoted in the column representing the maximum unsigned relative difference between valid weights from the two compared procedures ('max. diff.'). Clearly, this quantity can only be calculated for cases where both methods yield any valid solutions.

Although the current simulation hardly represents a realistic example for non-probability sampling, the above results nevertheless illustrate the applicability of the proposed calibrated neural networks. Such ANNs facilitate a middle course between pure propensity on the one and calibration weighting on the other side, allowing for flexible adjustments between these two border cases (cf. section 5.2.3). Although the combination of distance function components is based on a rather simple strategy (cf. equalities 6.2), the realized implementation for numerical optimization in presence of soft calibration under box-constraints appears suitable and yields adequate results in most cases. As any other computational implementation of statistical methods, the `SQP` optimization algorithm has its limits in terms of numerical precision, and the feasible set must be non-empty to find a valid solution. When compared to the GREG as a well-established and widely used weighting method, the results in table 6.1 nevertheless stress numerical advantages of the proposed calibrated neural networks and corresponding optimization routines. Especially in challenging cases, the examples show that calibrated ANNs can be more reliable than the classical GREG formulation and achieve coinciding results for all considered situations where both can be applied. These differences occur when using the common routines of the `survey` package but can be validated for an independent implementation of equation 5.145 as well. The above results are therefore not specific to a particular implementation of the GREG weights. Overall, the advantages of calibrated neural networks are particularly evident when not only the totals are used as calibration benchmarks.

The simulation studies discussed in the current section 6.2 serve as a prior assessment and illustration of the proposed (calibrated) semi-parametric artificial neural networks as well as of the corresponding custom-made implementations in scope of the three `R`-packages `sqp`, `ann` and `calmod`. For prediction, semi-parametric neural networks yield on average lower population MSEs in comparison to fixed predetermined knot selection strategies. Calibrated ANNs constitute an integration and extension to existing weighting methods and provide higher numerical stability than the GREG as a frequently applied calibration approach. Therefore, the newly proposed methods exhibit advantages over some well-established and widely used ones. Nevertheless, the above simulations are intended for preliminary evaluation and testing only and therefore follow an admittedly rather simplifying and artificial setup. To examine and compare the full bandwidth of methods discussed in chapters 3 and 5, especially for cases of non-probability sampling that are closer to reality, a more comprehensive Monte Carlo simulation is presented in the following section 6.3.

6.3 Evaluation of Methods for Non-probability Samples

Various methods for assessing and compensating the issues and challenges of non-probability samples are discussed in chapters 3 and 5. An evaluation and comparison of the proposed methods with regard to their benefits and pitfalls in the context of non-probability samples is essential, especially when considering the rising amount and relevance of data obtained from such samples (cf. also chapter 2; Daas et al., 2015, p. 249; Groves, 2011, p. 869; Japec et al., 2015, p. 860). Although Buelens, Burger and van den Brakel (2015; 2018) already cover as least a considerable range of the presented estimation methods, most other publications exclusively focus on a single or at best a few methods for non-probability samples (cf. e.g. Chen, Valliant and Elliott, 2019; Elliott, 2009; Kim et al., 2018; Rafei, Flannagan and Elliott, 2020; Yang and Kim, 2018). Despite its importance for methodological developments, an overarching comparison of the large variety of methods proposed for non-probability samples is hardly to be found in studies published so far (cf. Elliott and Valliant, 2017, p. 262). One purpose of this thesis is to tackle this gap. To that end, a simulation study evaluating the presented methods under different scenarios of non-probability sampling is conducted.

In contrast to classical probability sampling designs, non-probability selection mechanisms are manifold and often hard to identify (cf. chapter 2). As a consequence, setting up simulations in the context of non-probability sampling is typically less standardized than for studies applied in more traditional fields of survey statistics (cf. e.g. Buelens, Burger and van den Brakel, 2018, p. 328; Chen, Valliant and Elliott, 2019, p. 672; Kim et al., 2018, pp. 12 ff; Rafei, Flannagan and Elliott, 2020, p. 160; Yang and Kim, 2018, p. 10). A thorough description of setup and work-flow of the simulation study is therefore given in the following section 6.3.1, before results are discussed in section 6.3.2.

6.3.1 Setup of the Simulation Study

As discussed in chapter 2, non-probability samples can arise from various populations and selection mechanisms and are, therefore, subject to manifold and often unknown degrees of selectivity. At the same time, the available auxiliary information and its potential to assess and compensate this selectivity is likewise highly diverse. The same holds for the respective variables of interest when considering the different types of non-probability samples. Therefore, a simulation study focusing on methods for non-probability samples has to consider a variety of potential options.

Central challenges of non-probability sampling resemble the impact of complex probability sampling designs (cf. chapter 2). This immediately suggests the use of (quasi) design-based simulations, which are the typical studies conducted in the context of non-probability sampling (cf. e.g. Buelens, Burger and van den Brakel, 2018, p. 328; Chen, Valliant and Elliott, 2019, p. 672; Rafei, Flannagan and Elliott, 2020, p. 160). However, the real data generating mechanism in non-probability sampling is usually unknown. Even in cases where assessment of the selection process is possible, it is often limited by the availability of external information and depends on (modeling) assumptions (cf. chapter 3; Biffignandi and Pratesi, 2002; Biffignandi et al., 2002; Steinmetz et al., 2014, pp. 278 ff). Implementing a realistic non-probability sampling procedure for a simulation study is therefore hardly possible without making assumptions about the selection process (cf. Enderle, Münnich and Bruch, 2013, p. 95; Lee and Valliant, 2009, p. 331).

When assessing and comparing methods in the general context of non-probability sampling, varying degrees of selectivity and auxiliary information in real data suggest incorporating simulation scenarios that represent relevant settings of these aspects. As a consequence, simulations that evaluate methods for non-probability samples in many publications are either

a) focused on a specific real data set, using a realistic population that is closely related to this data and a selection model that is assumed to hold for the particular setting (cf. e.g. Buelens, Burger and van den Brakel, 2018, p. 327; Chen, Valliant and Elliott, 2019, pp. 672 f; Valliant and Dever, 2011, p. 120), or

b) relying on an explicit statistical model to create synthetic target variables and/or non-probability selection mechanisms under different scenarios (cf. e.g. Kim et al., 2018, p. 12; Rafei, Flannagan and Elliott, 2020, p. 160; Yang and Kim, 2018, p. 10).

Both of these approaches have particular (dis-)advantages. In the context of option ***a)***, only a single population, selectivity scenario and set of target variables are considered. These characteristics are intended to mimic those of the real non-probability sample for which the simulation is designed. If set up correctly, conclusions from such simulations are well suited for this specific setting. However, focusing on a single scenario, even if it is based on a real-life example, is too narrow to draw conclusions that generalize well to different types and applications of non-probability sampling. In contrast, models used for option ***b)*** are more suitable for constructing and comparing different scenarios. However, the use of explicit structural assumptions that is required in statistical models can lead to simulation results that are again hardly generalizable to any other (real) settings for which these assumptions may be invalid (cf. chapter 2; Buelens, Burger and van den Brakel, 2018; Burgard, Dörr and Münnich, 2020, pp. 16 ff). For both options, a prevalent problem occurs when variables and/or participation mechanisms are constructed from a model that is later evaluated with respect to its quality for estimation (cf. e.g. Chen, Valliant and Elliott, 2019, pp. 672 f; Lee and Valliant, 2009, p. 331). Such a strategy can cause issues and limitations when interpreting simulation results, and may be prone to circular reasoning (cf. Kim et al., 2018, pp. 12 ff; Setoguchi et al., 2008, p. 548; Stürmer et al., 2007, p. 1111). A realistic population and a participation process which both do not follow such a specific model considered for estimation are therefore desirable. Consequently, the design of the Monte Carlo simulation in the present section 6.3 is a combination and trade-off between the two approaches. This goal is pursued by using an authentic population and implementing different scenarios of sample selection and available auxiliary information.

When a realistic population is required for simulations, a common problem is that real population (e.g. register or census) data is usually not available, because it either does not exist or is not accessible due to confidentiality reasons (cf. Burgard et al., 2017b, p. 235; Merkle, Burgard and Münnich, 2016, p. 6). Therefore, the following simulation is based on the fully synthetic AMELIA data set, "which provides a realistic framework for open and reproducible research" (Burgard et al., 2017b, p. 235). AMELIA is designed to mimic the EU-SILC (European Union Statistics on Income and Living Conditions) population, explicitly incorporating marginal distributions and fundamental interactions of the real EU-SILC variables. Its main purpose is to serve as a sound artificial population of households and persons that can be used for evaluation of statistical methods in Monte Carlo simulations. The synthetic population is a research outcome of the AMELI

(Advanced Methodology for European Laeken Indicators) project, with extensions being developed under the InGRID (Inclusive Growth Research Infrastructure Diffusion) and InGRID2 project (cf. Graf et al., 2011; Merkle, Burgard and Münnich, 2016). Since it is synthetically constructed, the population contains realistic but not real persons or households. This helps to overcome disclosure risks, such that the data set is publicly available. Detailed descriptions of the data and its generation are provided by Alfons et al. (2011), Kolb (2012) as well as Burgard et al. (2017a,b).

As motivated above, it seems sensible to apply a bandwidth of scenarios to not only represent a single setting of non-probability sampling. This is done by altering strictly controlled factors to represent different degrees of selectivity and utility of available auxiliary information. The objective is to assess and compare the methods proposed for non-probability samples under variation of those factors. Since the performance of these methods is assumed to be influenced by these controlled conditions, insights into the relative advantages and pitfalls of the different methods can be obtained (cf. Bethlehem, 2008a, pp. 36 ff; Buelens, Burger and van den Brakel, 2015, p. 14; 2018, p. 327).

The degree of selectivity and utility of auxiliary information both are mainly determined by the dependencies between non-probability sampling mechanism, auxiliary and target variables (cf. chapters 3 and 5). Consequently, these dependencies are varied across the simulated scenarios, which is a common strategy when simulating non-probability sampling (cf. e.g. Bethlehem, 2010, p. 181; Buelens, Burger and van den Brakel, 2018, p. 328; Chen, Valliant and Elliott, 2019, p. 672; Kim et al., 2018, p. 12; Rafei, Flannagan and Elliott, 2020, pp. 160 f; Yang and Kim, 2018, p. 10). To initialize the scenario populations, a random subset of size $N = 20\,000$ is drawn from AMELIA.[8] It is then restructured as outlined below to establish a number of patterns representing cross-combinations of the factors to be varied. The following AMELIA variables are used in the simulation study:

- The main target variable $\boldsymbol{y}_{\cdot 1}$ is personal income. To be able to consider not only univariate statistics of target variables, a second variable of interest is used. This second variable is denoted by $\boldsymbol{y}_{\cdot 2}$ and generated using the same subsequently described strategy and univariate (i.e. personal income) distribution as for $\boldsymbol{y}_{\cdot 1}$. The two target variables are correlated by $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{y}_{\cdot 2}} = 0.5$.
- Household income serves as auxiliary variable $\boldsymbol{X}$. It is used as predictor in model-based and as calibration variable in pseudo-design-based methods.
- Age is considered as additional auxiliary variable $\boldsymbol{Z}$ that is used as independent variables for the response model in addition to $\boldsymbol{X}$.
- Selection probabilities $\boldsymbol{\pi}^{\mathsf{nps}}$ are used to select non-probability samples for the simulation. The *initial* probabilities are constructed proportional to draws from a conditional standard normal distribution given $\boldsymbol{X}$, $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{Z}$. By re-ordering as described below, the conditional distribution is no longer Gaussian.

The correlations between these variables are then adjusted to obtain different scenarios of selectivity and available auxiliary information. For a fair comparison of methods, not only linear but also non-linear dependencies have to be considered, which is achieved by including quadratic terms (cf. Buelens, Burger and van den Brakel, 2015, pp. 15 f; Kim

[8] Restricting the population size is necessary to account for the rather slow computational performance in some pre-existing `R`-packages that implement the examined methods.

Table 6.2: Settings for the simulation study

Setting 1	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\in$	$\{0.0;\ 0.3;\ 0.6\}$
	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	$\in$	$\left\{0.0;\ 0.3;\ 0.6;\ \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}\right\}$
	$\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}}$	$\in$	$\left\{0.6;\ \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathsf{nps}}}\right\}$
Setting 2	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\in$	$\{0.6\}$
	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	$\in$	$\{0.3\}$
	$\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}}$	$\in$	$\{0.0;\ 0.3;\ 0.6\}$
	$\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}^{\circ 2}\boldsymbol{\pi}^{\mathsf{nps}}}$	$\in$	$\left\{0.0;\ 0.3;\ 0.6;\ \boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}} \cdot \boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{y}_{\cdot 1}^{\circ 2}}\right\}$
	$\boldsymbol{\rho}_{\boldsymbol{Z}^{\circ 2}\boldsymbol{\pi}^{\mathsf{nps}}}$	$\in$	$\{0.6;\ \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{Z}^{\circ 2}} \cdot \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathsf{nps}}}\}$

Further fixed correlations in both settings are $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$. The remaining dependencies result as products of the specified ones, i.e. are determined by conditional independence. Conditional independence also holds in case of products of two correlations specified as elements of the respective sets. Both settings are combined with 100% and 80% coverage of the target population.

et al., 2018, p. 12; Rafei, Flannagan and Elliott, 2020, p. 160; Yang and Kim, 2018, p. 10). The correlations that define the simulation scenarios consequently are those

- of $\boldsymbol{Z}$ with $\boldsymbol{X}$ and $\boldsymbol{X}^{\circ 2}$,
- of $\boldsymbol{y}_{\cdot 1}$ with $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{X}^{\circ 2}$ and $\boldsymbol{Z}^{\circ 2}$, as well as
- of $\boldsymbol{\pi}^{\mathsf{nps}}$ with $\boldsymbol{y}_{\cdot 1}$, $\boldsymbol{Z}$, $\boldsymbol{y}_{\cdot 1}^{\circ 2}$ and $\boldsymbol{Z}^{\circ 2}$.

The original dependencies between AMELIA variables are modified by reordering the existing values of $\boldsymbol{Z}$, $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{\pi}^{\mathsf{nps}}$ to adjust the correlations between the variables. In this way, different scenario populations and participation patterns are generated without using any explicit model, and the marginal distributions of variables in AMELIA are not altered, thereby retaining their similarity to real ones. Furthermore, these marginal distributions coincide over all scenarios, which fosters comparability between the different settings. In general, values from $\{0.0; 0.3; 0.6\}$ are considered for the correlations outlined above. In addition, it seems important to contrast purely linear with non-linear dependencies between the variables and to compare the methods' performance under fulfillment and violation of conditional independence (cf. chapter 5). This is achieved by additionally including scenarios that constitute conditional independence between $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{\pi}^{\mathsf{nps}}$ given $\boldsymbol{Z}$ and between any squared variable and all other quantities given the respective linear variable. An example for the latter case is $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$, which corresponds to $\left(\boldsymbol{y}_{\cdot 1} \perp\!\!\!\perp \boldsymbol{X}^{\circ 2}\right) \big| \boldsymbol{X}$ (cf. Dawid, 1979, p. 3). The possible combinations of these values constitute a mixture and trade-off between values that are commonly chosen in other simulation studies (cf. e.g. Andridge et al., 2019, p. 1471; Elliott, 2009, p. 4; Kim et al., 2018, p. 12; Rafei, Flannagan and Elliott, 2020, pp. 160 f; Yang and Kim, 2018, p. 10).

With a total of $3^4 \cdot 4^6 = 331\,776$ possible populations being computationally infeasible for a simulation study, selected cross-combinations are chosen from the outlined values.[9] The simulation scenarios therefore result from two settings, which are outlined in table 6.2. In

[9] For the four correlations to be varied between $\boldsymbol{X}$ and $\boldsymbol{Z}$ on the one and all other variables on the other hand, 3^4 combinations are possible since each of these correlations is chosen from $\{0.0; 0.3; 0.6\}$. Considering the six remaining correlations to be varied, the number of possible combinations is 4^6 because conditional independence constitutes a fourth option to choose.

each of these settings, different correlations are varied, while all other relations are kept fix or correspond to conditional independence. These variations are chosen to represent a bandwidth of possible selectivity with respect to $\boldsymbol{y}_{\cdot 1}$ as well as different potentials of the independent variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ to predict the target variable $\boldsymbol{y}_{\cdot 1}$ and the non-probability sampling process ($\boldsymbol{\pi}^{\mathsf{nps}}$). In the first setting, the capability of $\boldsymbol{X}$ and its squared values to explain $\boldsymbol{y}_{\cdot 1}$ is varied. This is combined with two different degrees of selectivity, expressed by the correlation between $\boldsymbol{\pi}^{\mathsf{nps}}$ and $\boldsymbol{y}_{\cdot 1}$. In the second setting, the amount of selectivity is altered further, by choosing different correlations of $\boldsymbol{\pi}^{\mathsf{nps}}$ with $\boldsymbol{y}_{\cdot 1}$ and its quadratic term. In addition, the non-linear relationship between response variables $\boldsymbol{Z}$ and selection probability $\boldsymbol{\pi}^{\mathsf{nps}}$, as expressed by $\boldsymbol{\rho}_{\boldsymbol{Z}^{\circ 2}\boldsymbol{\pi}^{\mathsf{nps}}}$, is varied.

Besides this dependency structure, incorporating possible coverage issues in a simulation study is closely related to the generation of participation probabilities $\boldsymbol{\pi}^{\mathsf{nps}}$. In simulating non-probability samples, there are two perspectives concerning potential coverage errors: some authors (e.g. Bethlehem, 2010, p. 181; Lee and Valliant, 2009, pp. 331 f) use fixed indicator variables for whether an element is in the sampling frame. Hence, π_i^{nps} is zero for some elements i of the population. In contrast, others (e.g. Valliant and Dever, 2011, p. 123) tend to treat coverage as random. A new indicator is generated in each sampling step. This scenario does not systematically exclude parts of the population over all simulation runs and is therefore representable by full coverage with adjusted values of $\boldsymbol{\pi}^{\mathsf{nps}}$ (cf. Bethlehem, 1988, p. 253; Särndal and Lundström, 2005, pp. 49 f). Both points of view are represented in the simulation, by additionally considering the cases where 100% and 80% of the population are covered. Under-coverage is constructed by using a threshold value for $\boldsymbol{\pi}^{\mathsf{nps}}$ such that the lowest 20% of the inclusion probabilities are set to zero.

Combining each of the possible combinations from table 6.2 with 100% and 80% coverage, the result is a total number of 94 populations. To avoid ambiguity due to this still large number of cross-combinations, the corresponding parameters defining the scenario populations are provided in the context of the simulation results as well. Note that these scenarios are not designed to necessarily mirror realistic dependencies for the variables chosen as $\boldsymbol{X}$, $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{Z}$ (e.g. personal and household income). As motivated above, restructuring the original AMELIA constitutes a trade-off between realistic data, comparability and the ability to represent different scenarios for the degree of sample selectivity and usefulness of auxiliary information.

Sampling for the simulation consists of drawing non-probability as well as reference samples. From each scenario population, 1 000 samples of each kind are independently selected. Non-probability samples are drawn by means of unequal probability sampling, using inclusion probabilities $\boldsymbol{\pi}^{\mathsf{nps}}$ to mimic selection processes in real samples. Poisson sampling without replacement is applied for this purpose (cf. Tillé, 2006, pp. 76 ff), using an expected sample size $\mathrm{E}\left(n^{\mathsf{nps}}\right) = 500$. Once the sample is drawn, selection probabilities are assumed to be unknown (i.e. for estimation), which is the common case in non-probability sampling (cf. chapter 3). Even though reference samples often originate from complex survey designs in reality, these designs are of negligible interest for evaluating methods in the context of non-probability samples. Reference samples in this simulation are therefore selected by simple random sampling without replacement for the sake of simplicity. Since reference samples are typically smaller than non-probability samples in real applications, their size in the simulation is $n^{\mathsf{res}} = 200$. As depicted in figures 2.1 and 5.1, only auxiliary variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ are observed in the reference sample, but $\boldsymbol{y}_{\cdot 1}$

is not. The outlined sampling procedures are widespread in published simulation studies focusing on non-probability samples (cf. e.g. Andridge et al., 2019, p. 1472; Bethlehem and Biffignandi, 2012, pp. 406 ff; Chen, Valliant and Elliott, 2019, p. 673; Rafei, Flannagan and Elliott, 2020, p. 161; Valliant and Dever, 2011, p. 122).

The selected samples constitute the foundation to evaluate the methods proposed for examining and compensating the issues of non-probability selection processes. As described in chapters 3 and 5, there are different archetypal settings of auxiliary information which are typically considered for these purposes. In summary, these settings correspond to availability of

***a*)** micro-data for the reference sample,

***b*)** calibration benchmarks for the population, or

***c*)** a combination of both.

To limit the computational burden in the already large-scale study, the simulation is restricted to these three settings because they are rather typical. Nevertheless, further scenarios could be constructed, e.g. by considering availability of certain types of micro-data for the whole population or by determining calibration benchmarks from the reference sample rather than the population (cf. e.g. chapter 7; Buelens, Burger and van den Brakel, 2018, p. 329; Chen, Valliant and Elliott, 2019, p. 672; Rafei, Flannagan and Elliott, 2020, pp. 160 f).

To calculate weights for the simulated non-probability samples, the pseudo-design-based methods discussed in section 5.2 are applied. These can be based on scenarios ***a*)** to ***c*)** of available auxiliary information and assume particular functional forms and loss functions for determining the weights. For each non-probability sample, 41 distinct weighting vectors are computed, one of which is defined to be non-informative, i.e. equal to a vector of ones, such that the non-probability sample is unweighted in this case. As outlined above, auxiliary variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ are used in case of response (propensity) models. Where calibration benchmarks are used, these correspond to population statistics of $\boldsymbol{X}$. For weighting methods that use soft calibration (i.e. calibrated ANNs), the permissible maximum absolute deviation from targets is set to 2.5% for totals and 10% for variances.

The model-based (prediction) approaches under consideration are those discussed in section 5.1 and apply different types of loss functions, fitting methods and assumptions about the relationship between independent and dependent variable. In total, 14 different types of prediction models are examined in the simulation. Because $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{Z}$ are not inherently related, $\boldsymbol{X}$ is used as the only predictor for almost all of these models. For mixed models, however, the common use for non-probability samples is to incorporate variables that affect sample inclusion as random effects (cf. sections 5.1.5 and 5.3.2; Gelman et al., 2016a; Wang et al., 2015). Therefore, a classified version of $\boldsymbol{Z}$ is used to specify the random effects for mixed models, which has the further benefit that the use of additionally incorporating $\boldsymbol{Z}$ for prediction can be evaluated. Classes of $\boldsymbol{Z}$ for this purpose are generated by splitting the observations into 10 groups, according to their values in $\boldsymbol{Z}$. In case of matching, an average of the five nearest neighbors in terms of the Mahalanobis distance is used as prediction (cf. section 5.1.1). Once the models are fit, they can be used to impute the target variable $\boldsymbol{y}_{\cdot 1}$ for the reference sample, where classical design-based methods are employed for estimation (cf. section 2.2).

To jointly apply weights and prediction models, the most common strategy is to rely on a weighted loss-function (cf. e.g. Beaumont, 2000; Breidt and Opsomer, 2017; Fuller, 2009, p. 378). Purely model- or pseudo-design-based strategies are special cases thereof, respectively using non-informative weights or the weighted distribution of $\boldsymbol{y}_{.1}$ in place of the model. In between these border cases, numerous cross-combinations of weighting and prediction models are possible. An alternative, which is sometimes considered in the context of non-probability sampling, is to obtain weighted estimates from the modeled distribution (or predictions) for the non-probability sample itself, rather than imputing for a reference data set. The main example of this strategy is multilevel regression and post-stratification (Wang et al., 2015). Joint models for selection process and target variables, like the one proposed by Heckman (1976; 1979), constitute further alternatives to combine the model- and the pseudo-design-based paradigm. All these options are discussed in section 5.3 and applied in the simulation.

The determined (and potentially non-informative) pseudo-design weights are furthermore used in methods that assess the selectivity of non-probability samples. The approaches discussed in chapter 3, which include statistical tests, matching, R-indicators and MSE-intervals, are considered in the simulation. In total, these result in 59 different measures to examine the selectivity in each non-probability sample.

Estimation is performed using the weighted non-probability or the imputed reference samples. With regard to the considered types of estimators, existing simulation studies in the context of non-probability sampling mainly focus on means and totals (cf. e.g. Bethlehem and Biffignandi, 2012, p. 295; Buelens, Burger and van den Brakel, 2018, p. 328; Chen, Valliant and Elliott, 2019, p. 672; Valliant and Dever, 2011, p. 124; Rafei, Flannagan and Elliott, 2020, p. 161). However, the literature review in section 2.1 reveals that multivariate statistics are of considerable importance for many real-life applications of non-probability samples as well. To cover a bandwidth of uni- and multivariate statistics that appear relevant in that review, the point estimators considered for the simulation study include totals and means as well as covariances, correlations and regression coefficients.

Even though approaches for assessing selectivity can provide some indication about the bias of estimates, there is no general theory for quantifying it from a single non-probability sample. For inference, actually estimating the MSE's bias component would require even stronger assumptions and/or better auxiliaries than those used for point estimation. Such additional variables or assumptions could typically be used to compensate the bias, rather than to quantify it only for inference. Therefore, inference in the simulation study is based on variance estimation methods, which is the typical case considered in the context of non-probability samples (cf. chapter 5; Buelens, Burger and van den Brakel, 2018, p. 330; Chen, Valliant and Elliott, 2019, p. 673; Elliott and Valliant, 2017, p. 257; Kim et al., 2018, p. 10; Rafei, Flannagan and Elliott, 2020, p. 159). As described in section 5.4, a first naive approach is to apply classical design-based variance estimates for this purpose (cf. e.g. equations 2.20). To that end, a pseudo-design weighted non-probability or an imputed reference sample is treated as if it was a probability sample with known design weights and target variables. These methods are considered in the simulation because they are frequently used for real non-probability samples (cf. e.g. Barratt, Ferris and Lenton, 2015; Faas and Schoen, 2006; Guarte and Barrios, 2006; Spijkerman et al., 2009). Nevertheless, such approaches ignore the fact that estimation from non-probability samples typically requires weighting and/or prediction models. This can lead to seriously biased inference, which is why resampling techniques are typically considered to be more

adequate in this context (cf. Baker et al., 2010, p. 47; Bethlehem, 2008b, p. 21; Lee and Valliant, 2009, p. 341; Rafei, Flannagan and Elliott, 2020, p. 160). To consider resampling methods, the Monte Carlo and rescaling bootstrap are applied in the simulation, from which variance estimates are obtained as described in equations 5.189. Because the within and between component of the variance are considered either separately or jointly in different publications, each of these options is evaluated for both resampling methods under consideration in the simulation. Note that for all of these variance estimates, it is typically assumed that the non-probability sample inclusion of two distinct elements is independent. This assumption holds for the Poisson sampling strategy applied for the simulation, but may be violated in real cases of non-probability sampling (cf. Buelens, Burger and van den Brakel, 2018; Chen, Valliant and Elliott, 2019; Elliott and Valliant, 2017; Guarte and Barrios, 2006; Kim et al., 2018; Tillé, 2006, p. 77).

In a final step, the simulation's output is used to evaluate and compare the performance of all the discussed assessment and estimation methods for non-probability samples. This is done by means of graphical illustration as well as with regard to the relative bias

$$\mathbf{RBias}\left(\widehat{\boldsymbol{\vartheta}}\right) = \mathbf{Bias}\left(\widehat{\boldsymbol{\vartheta}}\right) \oslash \boldsymbol{\vartheta} \tag{6.4}$$

and relative root mean squared error

$$\mathbf{RRMSE}\left(\widehat{\boldsymbol{\vartheta}}\right) = \sqrt{\mathbf{MSE}\left(\widehat{\boldsymbol{\vartheta}}\right)} \oslash \boldsymbol{\vartheta} \tag{6.5}$$

as normalized quality measures for general estimators (cf. section 2.2). Variance estimates are typically used to construct confidence intervals rather than reported themselves. The coverage rates of these intervals (confidence interval coverage rates, CI-rates or CIRs) in the simulation are therefore examined to assess the quality of variance estimates. These CI-rates are defined by

$$\mathrm{CIR}\left(\widehat{\vartheta}_{ij}\right) \coloneqq \mathrm{E}\left(\mathbb{I}\left(\mathrm{Abs}\left(\widehat{\vartheta}_{ij} - \vartheta_{ij}\right) \leq \Phi^{-1}\left(1 - \alpha/2\right) \cdot \sqrt{\widehat{\mathrm{V}}\left(\widehat{\vartheta}_{ij}\right)}\right)\right) \tag{6.6}$$

for all estimates $\widehat{\vartheta}_{ij}$, where $\Phi^{-1}\left(1 - \alpha/2\right)$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution. The common significance level of $\alpha = 0.05$ is used in the simulation (cf. e.g. Bethlehem and Biffignandi, 2012, p. 170; Dekking et al., 2005, pp. 383 ff; Efron and Tibshirani, 1986, p. 55; Elliott and Valliant, 2017, p. 261; Rafei, Flannagan and Elliott, 2020, p. 162).

Since repeated sample selection in the simulation is used to approximate the behavior of estimators $\widehat{\boldsymbol{\vartheta}}$ over all possible samples, the expectations used to define equalities 6.4 to 6.6 are all evaluated over the 1 000 estimates obtained from each of the Monte Carlo iterations. The true values (statistics of interest $\boldsymbol{\vartheta}$) to be estimated in the simulation are obtained from the known finite scenario populations. All the specific estimators and statistics of interest that are relevant in the simulation are defined in section 2.2.

An overview of all outlined steps for the simulation study in form of a flowchart is provided in figure 6.4. The results obtained from this study are presented and discussed in the following section 6.3.2.

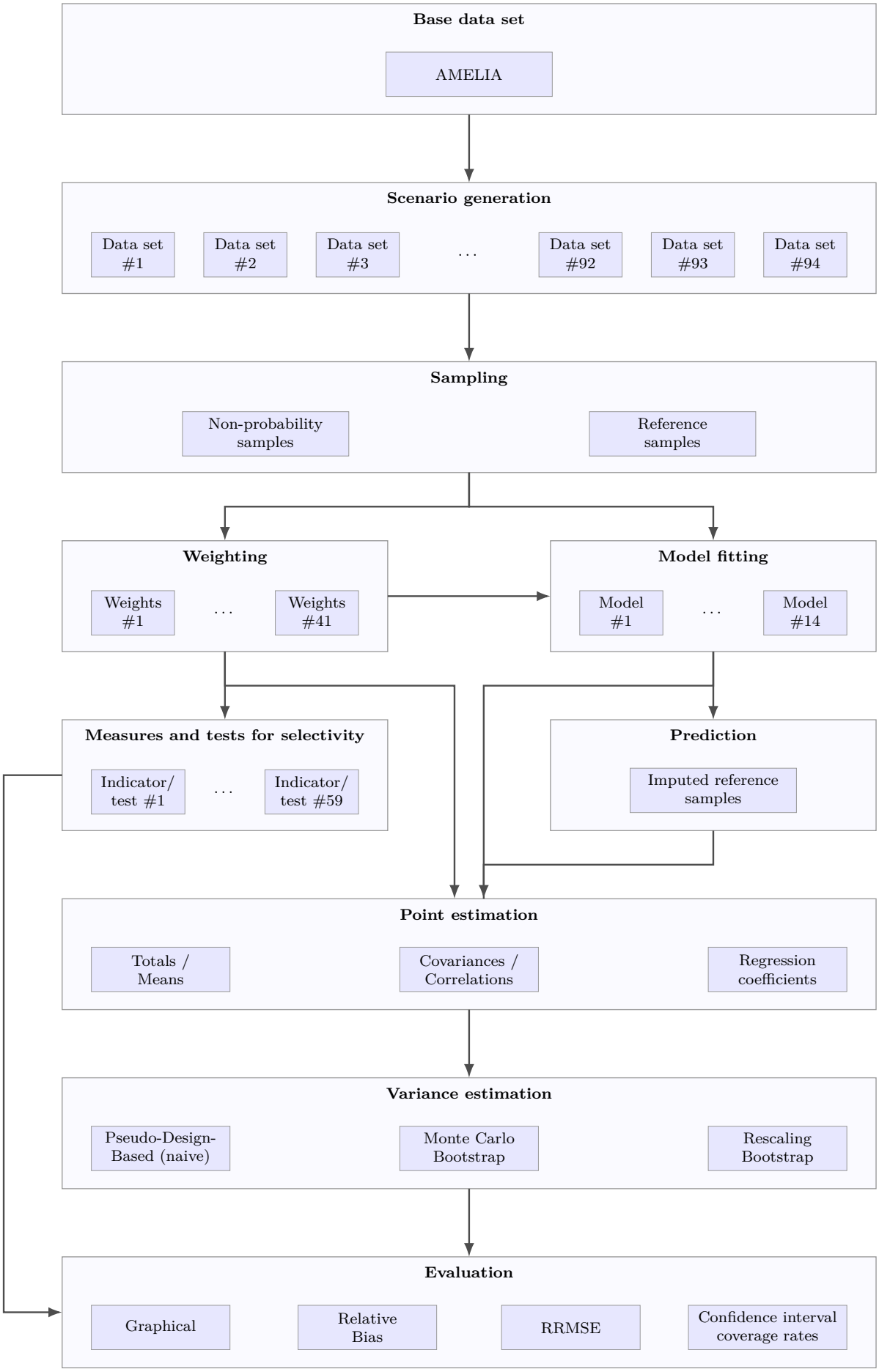


Figure 6.4: Flowchart of the Monte Carlo simulation study

6.3.2 Results of the Simulation Study

Considering the number of scenario populations and approaches for weighting, prediction and selectivity assessment, the simulation study's outcomes are extensive. To provide an overview, the following discussion focuses on a selected subset of results. This subset is chosen to cover important overall findings from the Monte Carlo simulation by illustrating patterns that generalize to other similar cases in the simulation study. Following the order in figure 6.4 and chapters 3 to 5, methods for selectivity assessment are evaluated in section 6.3.2.1. Simulation results for point estimation and inferential approaches follow in sections 6.3.2.2 and 6.3.2.3.

6.3.2.1 Methods for Assessing Selectivity and Bias

Different approaches to examine whether a non-probability sample is selective and biased are discussed throughout chapter 3. To that end, these methods typically make joint use of a non-probability and a reference sample (cf. e.g. equations 3.2 to 3.5). For each of these approaches for selectivity assessment, selected simulation results are successively presented in the following paragraphs. The order corresponds to that in chapter 3, starting with statistical tests for selectivity patterns, followed by matching, R-indicators and MSE-intervals.

Tests for Selectivity

As a first formal approach to assess potential selectivity of non-probability samples, statistical tests are frequently applied (cf. section 3.4). To check whether they are actually fit for this purpose, it is evaluated whether decisions based on such tests can provide some guidance on the magnitude of error occurring in non-probability sample estimates. This *estimation error* is the difference between the estimated and the true statistic of interest (cf. section 2.2). Results are represented in figure 6.5, applying combined difference tests for $\boldsymbol{X}$ and $\boldsymbol{Z}$.

Since most of the following figures follow a highly similar pattern, the general structure is summarized as a first step. As outlined in table 6.2, correlational structure and coverage of the target population are the only factors varied in the simulation. In figure 6.5, twelve distinct scenario populations are considered, representing different types of linear and quadratic relations between auxiliary and target variables $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$. These dependencies are structured and labeled in rows and columns to form a grid, such that each cell in the grid represents one of these twelve scenarios. As discussed above, certain correlations between simulation variables are kept fix across all results in this figure. These are indicated below the plot. As a result of conditional independence, all other correlations are products of the denoted ones. The degree to which the target population is covered by the non-probability samples is also fixed and denoted in the figure's caption. Within the grid cells, each boxplot refers to the distribution of estimates under a certain setting. The mean of the respective estimates is depicted as a purple line within the box while the corresponding values of RBias and RRMSE are denoted in the two columns next to it. The distribution of estimates is represented as deviation from the true values in the respective population, which are represented as red lines at zero deviation.

In the current figure 6.5, each boxplot represents the distribution of unweighted estimates $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ in dependency of the decisions resulting from the various difference tests discussed in section 3.4. A significance level of 5% and non-informative weights $\widetilde{\boldsymbol{w}} = \mathbf{1}_{n^{\text{nps}} \times 1}$ for the non-probability sample are used. The tests are jointly performed for $\boldsymbol{X}$ and $\boldsymbol{Z}$, applying

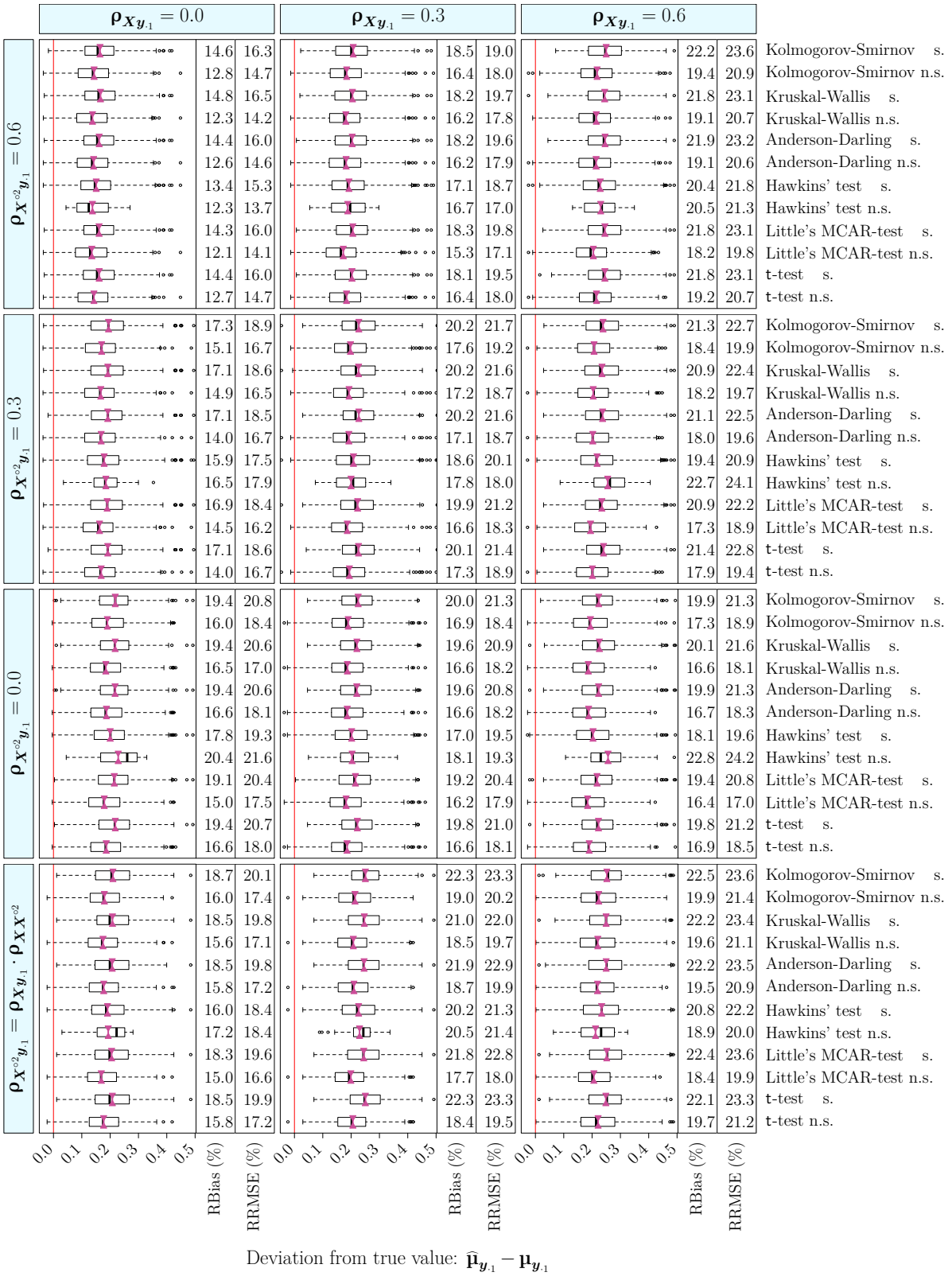


Figure 6.5: Representativity assessment for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$: combined difference tests for $\boldsymbol{X}$ and $\boldsymbol{Z}$, and estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ for 100% coverage – weighting model: unweighted

Rosenblatt's (1952) theorem where the original test does not allow for more than one variable. Correlations $\boldsymbol{\rho}_{\boldsymbol{XZ}}$, $\boldsymbol{\rho}_{\boldsymbol{Zy}_{\cdot 1}}$ and $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathsf{nps}}}$ between $\boldsymbol{Z}$ and respectively $\boldsymbol{X}$, $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{\pi}^{\mathsf{nps}}$ are all fixed to 0.6, while linear and quadratic dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ are varied across the grid. All other relations between the variables are defined by conditional independence (cf. section 6.3.1).

Since difference tests indicate whether there is a significant inequality between non-probability and reference sample, non-significant result are commonly used as evidence for absence of selectivity, even though they formally are an improper tool to accept hypotheses of equality (cf. section 3.4). Results in figure 6.5 show that for each but the Hawkins test, a significant difference in $\boldsymbol{X}$ and $\boldsymbol{Z}$ is indeed associated with a higher bias and MSE in $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$. However, a large amount of bias remains even when the tests are non-significant. This pattern holds over all examined relational structures between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$. Regardless of the resulting test decision, the bias of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ in most cases tends to increase with higher correlation between $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{X}$, i.e. from left to right grid cells. The increase is caused by inclusion probabilities $\boldsymbol{\pi}^{\mathsf{nps}}$ and $\boldsymbol{X}$ both being correlated with $\boldsymbol{Z}$ since the auxiliary variables are not used to compensate for selectivity in this setting.

When comparing the different types of tests, Hawkins' test exhibits comparably poor performance. It is the only test where RBias and RRMSE of the estimated mean are not generally lower if the test outcomes are non-significant. This behavior may be due to the fact that it tests for inequality in variances, which constitute an important component especially for the remaining parametric tests. Differences between most other types of tests are rather small over all scenario populations, but Little's test slightly outperforms its competitors regarding bias and MSE in many cases. From the non-parametric tests alone, the Kruskal-Wallis test seems slightly better than the Anderson-Darling or Kolmogorov-Smirnov test in average over all scenarios.

These findings are largely stable when using weights, i.e. for performing weighted tests to assess weighted estimates. The same holds when estimating correlations or regression coefficients instead of the mean $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ and for the different further scenario populations. For all these cases considered in the simulation, results are highly similar to those presented in figure 6.5. In general, some amount of bias is detected when testing for differences in the auxiliary variables. Therefore, it seems reasonable to assume that such significant differences in auxiliary variables indicate a higher bias in comparison to non-significant ones. However, such tests are typically applied in reality to provide evidence that there is no difference between non-probability and reference sample (cf. section 3.4). The above results show that it is not generally appropriate to assume non-selectivity and unbiasedness when differences in auxiliary variables are not significant. For most cases in the simulation, a considerable amount of bias remains even when the difference tests' results are not significant.

Although it is from a theoretical perspective more appropriate to show coherence of two data sources (cf. section 3.4), the use of equivalence tests leads to quite similar findings. The bias is typically lower when equivalence is significant, but the contrary occurs in various cases as well. A probable reason for this finding is the necessity to choose boundaries for the equivalence interval (cf. inequalities 3.20), which requires more effort and specific knowledge than the application of difference tests. Therefore, the use of equivalence tests for non-probability samples appears more situational and the corresponding results are more ambiguous than for difference tests.

In the simulation, tests for (non-)selectivity therefore permit some degree of bias assessment but do not allow drawing reliable conclusions about unbiasedness of non-probability sample estimates. This limited explanatory power is likely at least partially caused by rather small sampling fractions, resulting in relatively imprecise statistical tests. Although selected by simple random sampling, each reference sample only constitutes a small random subset of the population, from which the distributions of $\boldsymbol{X}$ and $\boldsymbol{Z}$ are used for testing. In addition, selective non-probability samples are subject to some degree of purely random variation as well if the selection process is not fully deterministic. Systematic differences between non-probability and reference sample may, thus, be concealed by the random variation of both data sources. In cases where samples are considerably larger (or more precise due to any other reason; cf. Groves, 1989; Weisberg, 2005), statistical tests might therefore be better indicators for selectivity and bias of non-probability samples. Nevertheless, available data in real applications is often limited and sampling fractions have to be considered fixed in many applications (cf. chapter 2; section 3.2; Bethlehem, 2008b, p. 35). Alternative ways to examine selectivity of non-probability samples may hence be more useful for the settings considered in the simulation study.

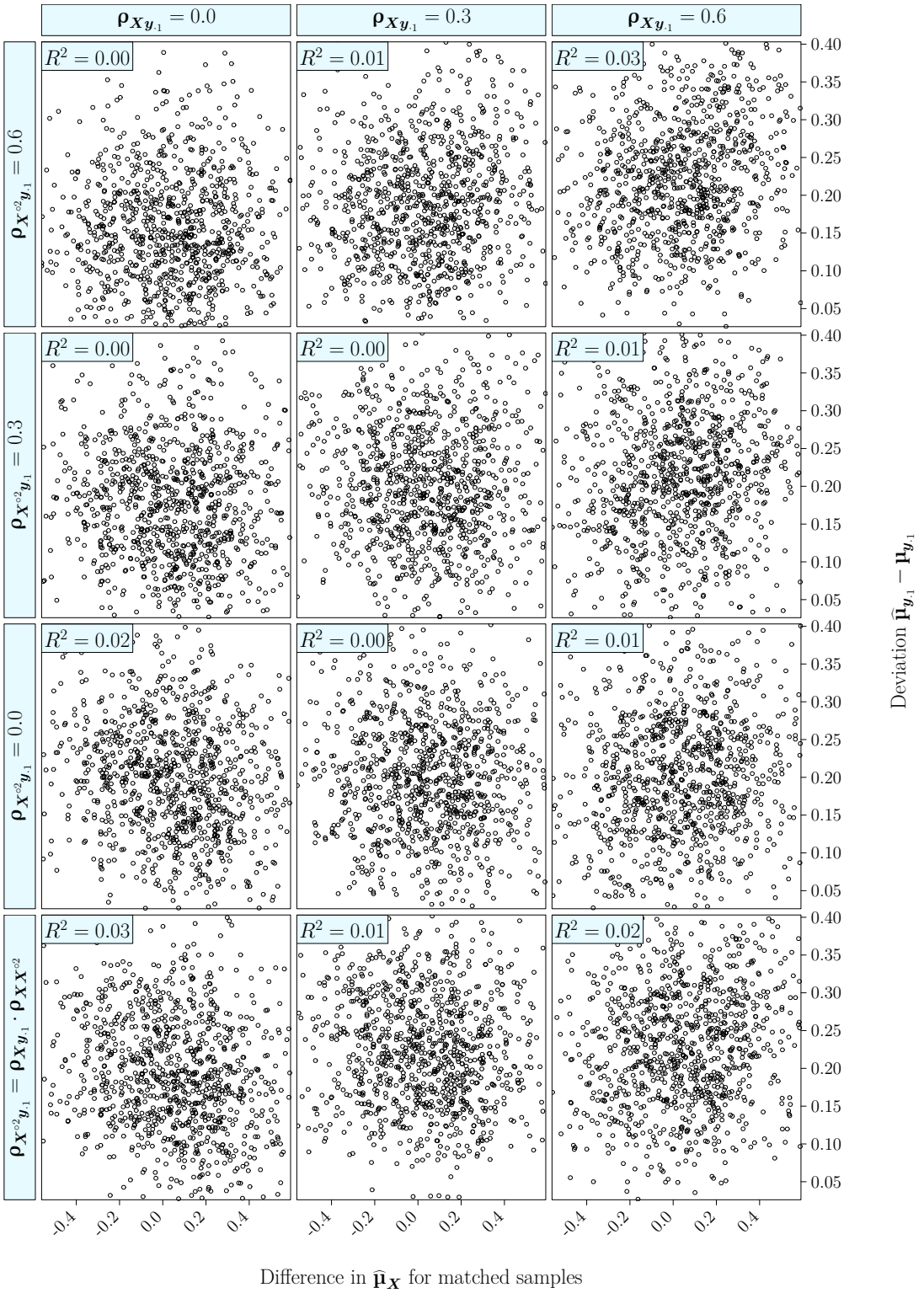
Matching

Matching constitutes an alternative approach for selectivity assessment of non-probability samples. As discussed in section 3.5, the distributions of matched observations in reference and non-probability sample are compared to assess selection bias in the latter, and differences in matched samples are used as an indication about systematic errors in non-probability sample estimates.

Simulation results to examine this usage of matching are shown in figure 6.6, following the same fundamental structure and scenarios as before (cf. figure 6.5). The approximated bias of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$ obtained from matching is used as a proxy for the estimation error of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ (cf. approximation 3.24). The performance of this approach is evaluated by plotting the expected difference in $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$ against actual discrepancies between $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ and $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ in a scatter plot. For each of the twelve scenario populations, these plots represent the dependency between approximate and actual error over all 1 000 samples. Additionally denoted is the coefficient of determination R^2, which in this bivariate case is the squared correlation between both quantities (cf. definition 2.17c). This coefficient expresses the share of variation of the error $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}} - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ that can be explained by a linear function of the approximated bias of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$. As before, unweighted estimates are considered, such that there is no correction for selectivity in the estimates.

It is remarkable that the difference for $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$ in the two matched samples is nearly unrelated to the error made when estimating $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ from the non-probability sample. Even if $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ exhibit considerable correlation, matching results for $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$ in the simulation hardly allow for any valid conclusions about the quality of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$. These results hold regardless of the underlying dependency between both variables, and non-linear dependencies are as well not evident.

This pattern is also found when considering different scenario populations, estimates other than the mean $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ (such as correlation or regression coefficients) or weighted estimates. In general, results for other settings of matching in the simulation are very similar. The general finding is that matching provides useful information about the bias of a non-probability sample only if the actual variable(s) of interest can be compared for the matched samples. To that end, the variables relevant for the estimator (i.e. $\boldsymbol{y}_{\cdot 1}$ in the



R^2: Coefficient of determination
$\rho_{XZ} = 0.6$
$\rho_{Zy_{.1}} = 0.6$
Matching variables: Z
$\rho_{Z\pi^{\text{nps}}} = 0.6$
All other correlations result as products of the stated ones.

Figure 6.6: Representativity assessment for different dependencies between X and $y_{.1}$: difference in $\widehat{\mu}_X$ for matched samples, and estimation of $\mu_{y_{.1}}$ for 100% coverage – weighting model: unweighted

current case) must be part of the auxiliary variables, as they need to be observed in the reference sample. In some real applications of non-probability sampling, this may be the case. However, non-probability samples are more commonly used to measure target variables that are not (yet) observed in existing probability samples. In such cases, evaluation of matched samples with respect to $\boldsymbol{y}_{\cdot 1}$ is infeasible because this variable is unknown for the reference sample (cf. chapter 2).

In the present simulation, differences in auxiliary variables (e.g. in $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$) between matched samples exhibit poor explanatory power for predicting selectivity in target variables (e.g. bias in $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$). As for the statistical tests considered in figure 6.5, this may again be at least partially explained by rather small sampling fractions of non-probability and reference samples. The bias component of the MSE may again be concealed by the random variation of both data sources, such that samples of higher precision might be more suitable for examining selectivity and bias of non-probability samples by means of matching.

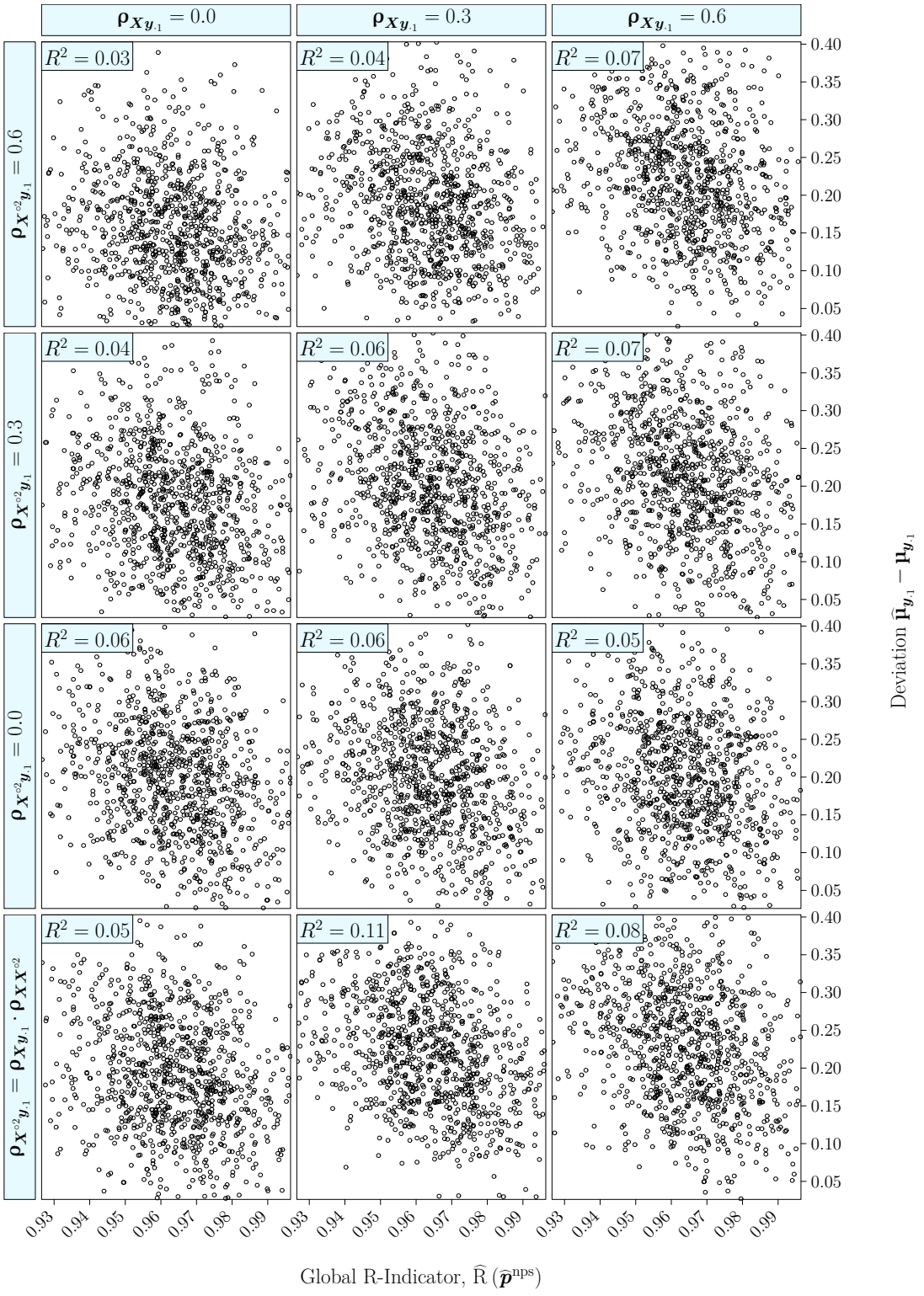
Representativity Indicators

R-indicators constitute the third option considered for selectivity assessment of non-probability samples (cf. section 3.7). These indicators depend on the variability of the estimated response propensities $\widehat{\boldsymbol{p}}^{\text{nps}}$, such that a higher overall R-indicator $\widehat{\text{R}}\left(\widehat{\boldsymbol{p}}^{\text{nps}}\right)$ suggests less dependency between non-probability selection process and independent variables used in the response model. When these auxiliary variables are correlated with the target variables, a higher value of $\widehat{\text{R}}\left(\widehat{\boldsymbol{p}}^{\text{nps}}\right)$ should, hence, indicate a lower systematic difference between estimated and true population statistic of interest. The utility of these indicators for assessing selectivity in the simulation study is evaluated in analogy to that of matching.

Results are presented in figure 6.7, for which response propensities are estimated from a parametric logit GLM. For each scenario population, differences between $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ and $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ are evaluated in dependency of the overall R-indicators calculated from these propensities. The structure and scenarios are the same as for the previous figures, such that $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ is fixed, while (non-)linear dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ are determined by the grid cells. In comparison to matching (cf. figure 6.6), the R-indicator provides slightly more evidence about the error made by an unweighted non-probability sample estimate $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$. The general tendency is a negative correlation between R-indicator and error in $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ for the non-probability sample, which meets the above expectations that higher values in $\widehat{\text{R}}\left(\widehat{\boldsymbol{p}}^{\text{nps}}\right)$ imply lower bias in $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$. Since $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{Z}$ are always correlated, this pattern holds over all examined dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$, although a stronger linear relation $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$ tends to allow slightly better results. Nevertheless, the predictive power of the R-indicator for the error of unweighted non-probability sample estimates $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ is rather small in all considered scenarios. Even in the best case, the coefficient of determination does not exceed 11% and non-linear dependencies are not evident.

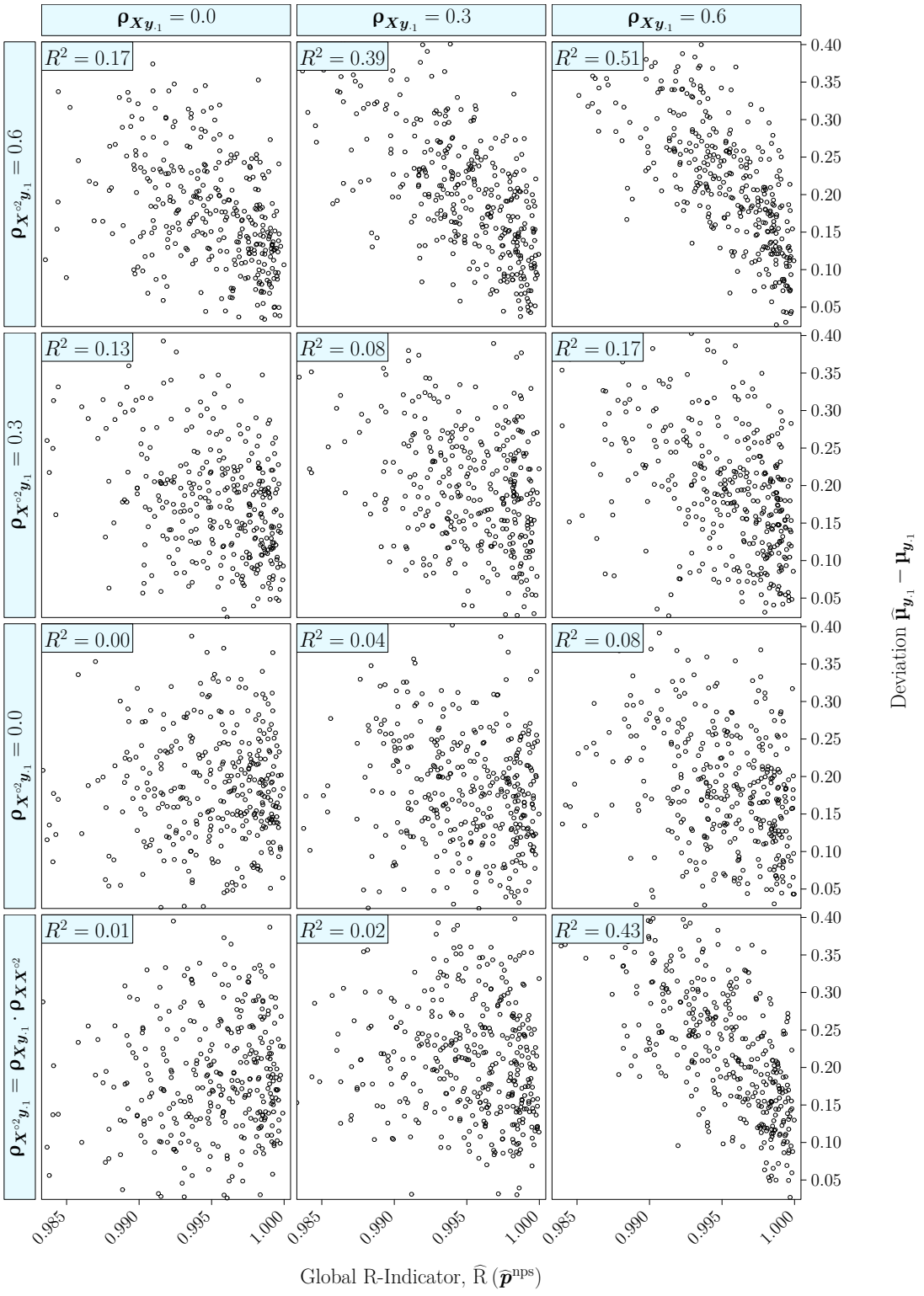
A similar pattern recurs in other settings of the simulation. In most cases, the error of unweighted non-probability sample estimates is largely unrelated to overall, conditional and unconditional representativity indicators. This holds when estimating univariate statistics, such as means and totals, as well as multivariate ones, which include correlation and regression coefficients.

A major improvement is evident only when additional population benchmarking information in form of calibration constraints is incorporated in the response propensity model, e.g. by using calibrated neural networks (cf. section 5.2.3). An example is shown in figure 6.8, considering identical scenarios as above. The propensity model used for computing



R^2: Coefficient of determination
$\rho_{XZ} = 0.6$
$\rho_{Zy_{.1}} = 0.6$
Response model variables: $\boldsymbol{X}, \boldsymbol{Z}$
$\rho_{Z\pi^{\mathrm{nps}}} = 0.6$
All other correlations result as products of the stated ones.

Figure 6.7: Representativity assessment for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$: global R-Indicator $\widehat{\mathrm{R}}\,(\widehat{\boldsymbol{p}}^{\mathrm{nps}})$, and unweighted estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ for 100% coverage – propensity model: logit model (parametric), using a reference sample



R^2: Coefficient of determination

$\boldsymbol{\rho}_{\boldsymbol{XZ}} = 0.6$ $\quad$ $\boldsymbol{\rho}_{\boldsymbol{Z}y_{.1}} = 0.6$ $\quad$ $\boldsymbol{\rho}_{\boldsymbol{Z}\pi^{\mathrm{nps}}} = 0.6$

Calibration variables: $\boldsymbol{X}$

Response model variables: $\boldsymbol{X}, \boldsymbol{Z}$

All other correlations result as products of the stated ones.

Figure 6.8: Representativity assessment for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$: global R-Indicator $\widehat{\mathrm{R}}\,(\widehat{\boldsymbol{p}}^{\mathrm{nps}})$, and unweighted estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ for 100% coverage – propensity model: calibrated ANN (parametric), using a reference sample, total and covariance constraints

the R-indicator is of the same form as in figure 6.7 but additionally incorporates total and variance constraints for the $\boldsymbol{X}$-variable by using soft calibration as described in section 6.3.1. Incorporating such benchmarks in the response model considerably increases the value of all types of R-indicators for measuring selectivity in non-probability samples, such that they are more valuable predictors for the error of the (still unweighted) estimates $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$. Because total and covariance calibration respectively exploit linear and non-linear dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$, this is especially true if $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$ and $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$ are both high. If this is the case, up to 51% of the error's variability can be explained in figure 6.8. A probable reason for this positive effect when calibration constraints are incorporated is the restriction of the feasible parameter region. These restrictions limit the possible values and hence variability of the R-indicators, leading to a stronger relation to the error of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$.

However, further simulation results show that this correlation vanishes once propensities (or any other form of compensating weights) are used for actual weighted estimation. Even with this increased predictive power, representativity indicators seem therefore more appropriate to identify and compensate selectivity during the sampling stage, which corresponds to their original purpose (cf. Schouten et al., 2012, p. 389). For assessing the error of estimates that use compensation methods discussed in chapter 5, R-indicators appear to be less applicable.

When additional auxiliary information in form of calibration benchmarks is included in the propensity model, the value of representativity indicators to determine selectivity of non-probability samples considerably increases in the current simulation. The proposed calibrated neural networks constitute an option for incorporating such information. However, R-indicators are originally based on pure response models that do not perform any calibration. In such cases, use of these indicators for evaluating selectivity of non-probability samples is generally rather limited. As in the case of matching, these results may partially be explained by small sampling fractions of non-probability and reference samples. The sampling variability might simply conceal any systematic differences in the response model underlying the R-indicators. To make representativity indicators more valuable for non-probability samples, it might again be sensible to use larger samples (or otherwise more precise reference samples) where possible.

MSE-intervals

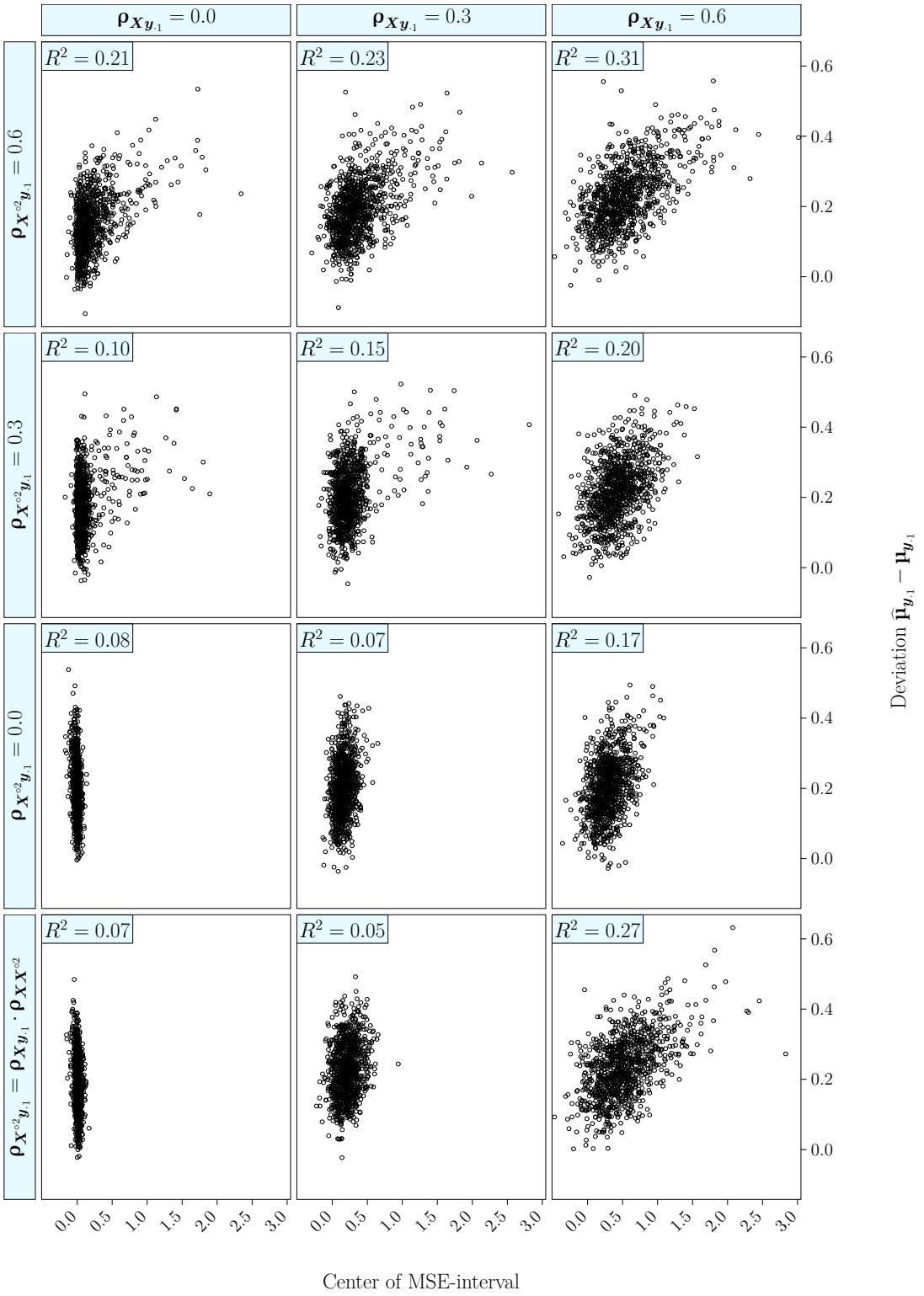
MSE-intervals are the last option considered for detecting selectivity in non-probability samples (cf. section 3.8). The core element here is to obtain an interval for the correlation between target variable and non-probability inclusion indicator, which is based on their respective correlations with a third variable. For a given target variable and sample size, an interval for the MSE of design linear estimators can be directly obtained from this correlation. Although such an interval is rarely sufficiently narrow to be used for actual inference (cf. also sections 5.4 and 6.3.2.3), it may still be applied to examine the potential error of a non-probability sample estimate. Since this interval is applied to quantify the difference between estimated and true statistic, its midpoint should ideally perfectly describe the estimation error in a non-probability sample. Using equalities 3.37 and 3.38 to approximate the estimation error of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}(\widetilde{\boldsymbol{w}})$ in the outlined manner leads to

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}(\widetilde{\boldsymbol{w}}) - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}} \approx \widehat{\boldsymbol{\rho}}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}(\widetilde{\boldsymbol{w}}) \cdot \widehat{\boldsymbol{\rho}}_{\boldsymbol{X}\widetilde{\boldsymbol{r}}^{\mathsf{nps}}}(\boldsymbol{w}^{\mathsf{u}}) \cdot \sqrt{\frac{1-fr^{\mathsf{nps}}}{fr^{\mathsf{nps}}}} \cdot \sqrt{\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_{\cdot 1}}(\widetilde{\boldsymbol{w}})} \cdot \sqrt{1+\frac{\left(\mathrm{CV}(\widetilde{\boldsymbol{w}})\right)^2}{1-fr^{\mathsf{nps}}}} \quad . \tag{6.7}$$

In approximation 6.7, a vector of weights $\widetilde{\boldsymbol{w}}$ is used for the non-probability sample, and $\boldsymbol{w}^{\mathsf{u}}$ denotes the combined vector of weights for the non-probability and the reference sample (cf. equation 3.5). As the third variable required to approximate $\boldsymbol{\rho}_{\widetilde{\boldsymbol{r}}^{\mathsf{nps}}\boldsymbol{y}_{\cdot 1}}$, $\boldsymbol{X}$ is used. When the true population variance and correlations are known and plugged in, this approximation is exact under conditional independence of $\boldsymbol{y}_{\cdot 1}$ and $\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}$ given $\boldsymbol{X}$ (cf. assumption 5.1). Both requirements do not hold in the simulation, and estimates are used for approximation 6.7 to assess the performance of MSE-intervals as estimated measures of selectivity for realized samples. As outlined in section 3.8, $\widehat{\boldsymbol{\rho}}_{\boldsymbol{X}\widetilde{\boldsymbol{r}}^{\mathsf{nps}}}$ is determined from the combined non-probability and reference sample. Estimates $\widehat{\boldsymbol{\rho}}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{X}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_{\cdot 1}}$ as well as the sampling fraction fr^{nps} and the coefficient of variation CV ($\widetilde{\boldsymbol{w}}$) are obtained from the non-probability sample alone.

Simulation results for left- and right-hand side of approximation 6.7 are presented in figure 6.9, considering again the unweighted case ($\widetilde{\boldsymbol{w}} = \mathbf{1}_{n^{\mathsf{nps}}\times 1}$). Structure and scenarios are the same as for the previous figures. As discussed in section 3.8, the quality of approximation 6.7 depends on the correlation of a third variable ($\boldsymbol{X}$) with both the target variable and the weighted response indicator. For the current results, the population's correlation of $\boldsymbol{X}$ and $\boldsymbol{r}^{\mathsf{nps}}$ is fixed through $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{\pi}^{\mathsf{nps}}} = \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} \cdot \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.36$. Consequently, the MSE-intervals' centers provide very limited information about the error of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ in case of small or no dependency between $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{X}$ as well as its squared term $\boldsymbol{X}^{\circ 2}$. But when at least one of these correlations is higher, the linear association between the center of the estimated MSE-intervals and the estimation error of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ improves considerably. Especially when the specified linear and non-linear dependencies are both high, the coefficient of determination increases up to 31%. In all scenarios, the approximated error indeed seems to be linearly associated with the true one. However, the ranges of possible values are considerably different, such that the relation between both still deviates substantially from identity. Furthermore, the amount of unexplained variability remaining in all settings is too high to be neglected. Approximation 6.7 is, thus, still not a precise measure of estimation error.

Nevertheless, these results indicate that MSE-intervals mostly outperform the previously discussed approaches whenever only a reference sample is used for assessing selectivity with regard to design linear estimators (cf. figures 6.5 to 6.8). These findings are similar in the further simulation scenarios and hold whether conditional independence is fulfilled or not. They also extend to weighted estimates of approximation 6.7 and $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$. In particular when using plain propensity weights, MSE-intervals can provide valuable indications for the resulting estimation errors. An example is shown in figure 6.10, where the errors of mean estimates that use propensity weighting are assessed under otherwise coinciding conditions as in figure 6.9. For medium to high population correlations between $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{X}$ as well as its squared term $\boldsymbol{X}^{\circ 2}$, the estimation error of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ is to a large extent (up to 79%) explainable as a linear function of the MSE-interval's midpoints. When $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$ is not additionally controlled, this is also feasible in case of high values for $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$ alone. As before, the relation between approximated and true error appears far from identity since the ranges of the possible values are visibly different. Nevertheless, the linear dependency between approximated and true error is considerable. The scatterplots suggest that this may to a large extent be caused by samples where both sides of approximation 6.7 strongly depart from the majority of samples. In presence of adequately correlated auxiliary variables, approximation 6.7 seems therefore especially worthwhile to assess whether a propensity weighted non-probability sample is particularly prone to large errors in linear statistics.

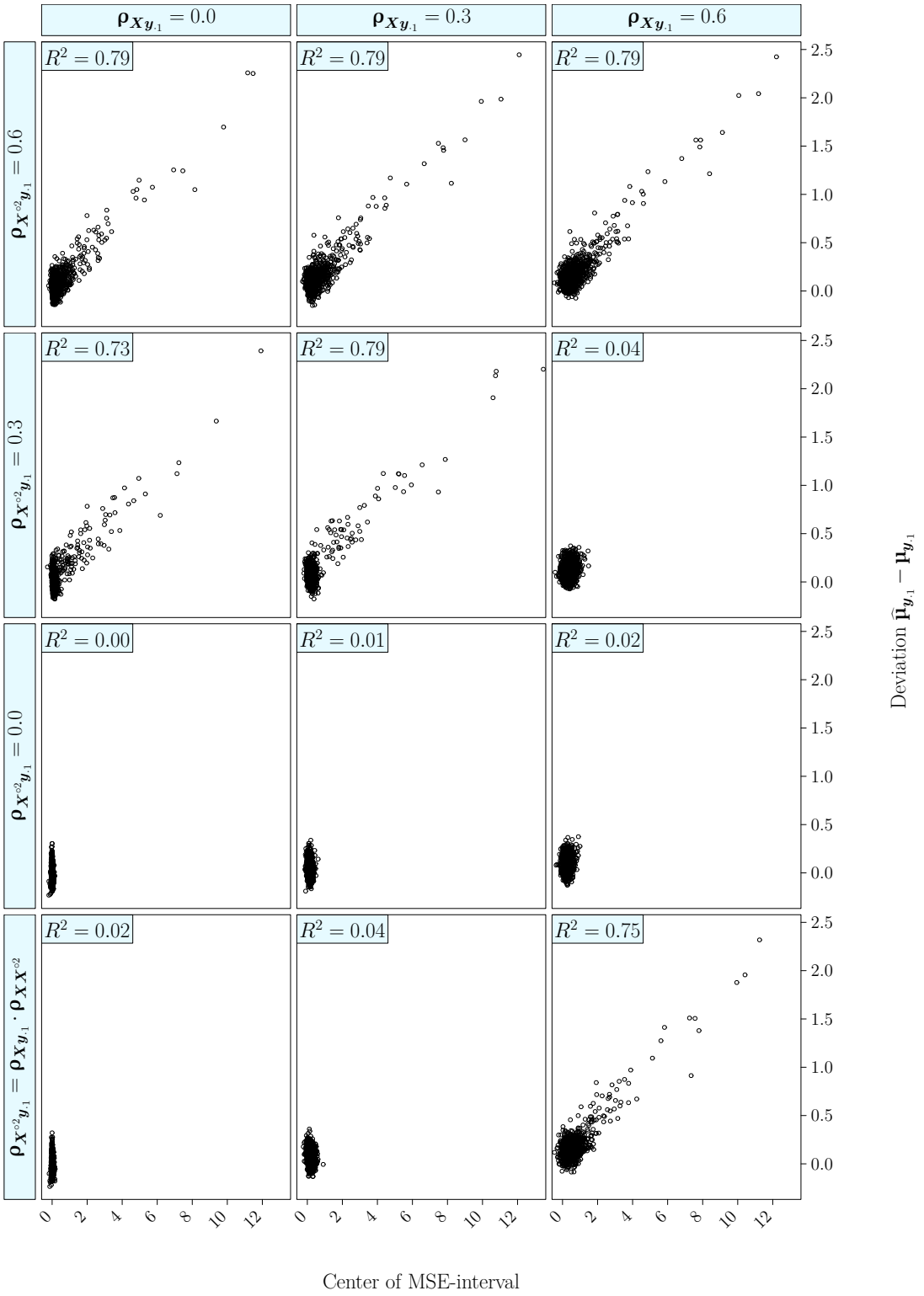


R^2: Coefficient of determination Auxiliary variable for correlation: $\boldsymbol{X}$

$\rho_{\boldsymbol{XZ}} = 0.6$ $\rho_{\boldsymbol{Z}y_{.1}} = 0.6$ $\rho_{\boldsymbol{Z}\pi^{\text{nps}}} = 0.6$

All other correlations result as products of the stated ones.

Figure 6.9: Representativity assessment for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$: MSE-interval based on $\boldsymbol{X}$, and estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ for 100% coverage – weighting model: unweighted



R^2: Coefficient of determination Auxiliary variable for correlation: $\boldsymbol{X}$ Response model variables: $\boldsymbol{X}, \boldsymbol{Z}$

$\rho_{\boldsymbol{XZ}} = 0.6$ $\rho_{\boldsymbol{Z}y_{.1}} = 0.6$ $\rho_{\boldsymbol{Z}\pi^{\text{nps}}} = 0.6$

All other correlations result as products of the stated ones.

Figure 6.10: Representativity assessment for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$: MSE-interval based on $\boldsymbol{X}$, and estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ for 100% coverage – weighting model: logit model (parametric), using a reference sample

However, these results only hold as long as total calibration is not applied for the auxiliary variable that is used for the MSE-intervals. This is caused by the fact that calibration adjusts for biases in linear statistics as far as $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ are correlated. The amount of bias determined by this correlation $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$ is exactly the systematic error that is captured by approximation 6.7, which is the case due to the underlying conditional independence assumption (cf. section 5.2; Bethlehem, 2008b, p. 33). Consequently, MSE-intervals can mainly help to assess non-probability sampling bias when this dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ is not already fully exploited for compensation. For valid estimates of the MSE's variance component to represent the entire estimation error, however, the correlations used in the simulation are simply too small.

Throughout the current section 6.3.2.1, the approaches for assessing selectivity of a non-probability sample discussed in chapter 3 are evaluated and compared. All of these approaches depend on auxiliary variables that are highly correlated with both the target variable and the non-probability sample's inclusion indicator. Auxiliary variables to be used for this purpose should, therefore, be chosen with regard to these correlations (cf. Schouten, 2007, pp. 60 ff). The tremendous importance of such highly correlated variables is underlined by the simulation results, despite none of the considered methods being able to perfectly describe the error of the non-probability sample estimates in any of the simulated conditions. Nevertheless, particular strategies appear more appropriate under certain circumstances. When only a relatively small reference sample is available, the considered statistical tests for selectivity allow rather limited conclusions about the estimation error, and the same holds for matching. Both methods are presumably affected by the high uncertainty of the reference sample estimates, which typically has to be considered as given for realistic scenarios (cf. e.g. Bethlehem, 2008b, p. 35; Isaksson and Lee, 2005, p. 3143). The use of representativity indicators is more feasible in some situations. It is noteworthy mainly for evaluating selectivity and sample composition when suitable auxiliary variables are used for calibrating the underlying response propensities. However, R-indicators seem less useful for determining the errors of estimates that are already based on any of the compensation methods discussed in chapter 5, e.g. for propensity weighted estimates. In such cases, MSE-intervals based on the work of Meng (2018) and Schouten (2007) appear as the most promising approach, at least for their intended use with regard to design linear estimators. The midpoints of these intervals predict the estimation errors relatively well, for both unweighted as well as weighted estimates. However, these results only hold as long as total calibration is not applied for the auxiliary variable that is used for the MSE-intervals. Otherwise, this calibration adjusts exactly for the amount of bias that is measured by these intervals. A detailed investigation of these and other methods that attempt to compensate for selectivity of non-probability samples in point estimation is provided in the following section 6.3.2.2.

6.3.2.2 Methods for Point Estimation

The point estimation methods for non-probability samples are partitioned by the model- and the pseudo-design-based paradigm, with some approaches aiming at a combination of both (cf. sections 5.1 to 5.3). In the current section 6.3.2.2, the objective is to evaluate, compare and discuss the performance of these methods in the simulation study. As in the previous section 6.3.2.1, only selected results can be presented due to the manifold different scenarios considered in the simulation. The following selection is chosen to provide a profound overview, considering different settings of sample selectivity, available auxiliary information and statistics to be estimated. Following the structure in chapter 5, the model-based methods are discussed first, before examining the pseudo-design-based techniques and approaches for integrating both.

Model-based Methods

In section 5.1, the model-based paradigm for estimation from non-probability samples is introduced, which is strongly based on conditional independence assumption 5.1. When this assumption holds, the conditional distribution $f_{\boldsymbol{Y}}(\boldsymbol{y}_{i\cdot} \mid \boldsymbol{x}_{i\cdot})$ of the target variables $\boldsymbol{Y}$ given some auxiliary variables $\boldsymbol{X}$ can be estimated unbiasedly from the non-probability sample. The unconditional distribution of $\boldsymbol{Y}$ is then obtained by using external information about the distribution of $\boldsymbol{X}$, e.g. based on the reference sample (cf. equations 5.7 and 5.8). The model-based paradigm relies on a bandwidth of specific statistical or machine learning models to represent the target variables' conditional distribution (or certain aspects thereof), ranging from matching (cf. section 5.1.1) to support vector machines (cf. section 5.1.10). In the simulation, each of these prediction models is used to obtain estimates for all 1 000 non-probability samples drawn from every scenario population. With regard to these estimates, the respective performance of the different models in point estimation is evaluated and compared. As before, a summary and discussion of important findings is given in the context of the following figures. As a base-line and reference point for comparison, plain non-probability sample estimates ('nps-estimates') that do not use any model are included.

The potential of model-based methods to estimate the mean $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ by using mass-imputation for the reference sample is evaluated in figure 6.11. The simulated scenarios and structure coincide with those in the previous figures 6.5 to 6.10. Different linear and non-linear dependencies between predictors $\boldsymbol{X}$ and target variable $\boldsymbol{y}_{\cdot 1}$ are considered in form of a grid, using fixed correlations $\boldsymbol{\rho}_{\boldsymbol{XZ}} = \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$ (as stated below the figure) to determine the scenario populations. All other relations between these variables are determined by products of the stated ones, corresponding to conditional independence (cf. section 6.3.1). Similar as in figure 6.5, the distributions of the estimates obtained from each of the models over all 1 000 samples are represented as boxplots. RBias and RRMSE are supplementarily reported beside each boxplot.

Examining the reference point of the rather naive non-probability sample estimates that do not use any method to compensate for selectivity, it is evident that the plain (unweighted) non-probability sample mean is severely biased in all represented scenarios. This is the case because selection probabilities $\boldsymbol{\pi}^{\mathsf{nps}}$ and target variable $\boldsymbol{y}_{\cdot 1}$ are both correlated with $\boldsymbol{Z}$. The bias to a large extent determines the MSE of estimators and increases with higher correlations $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$ because $\boldsymbol{X}$ and $\boldsymbol{Z}$ are also correlated.

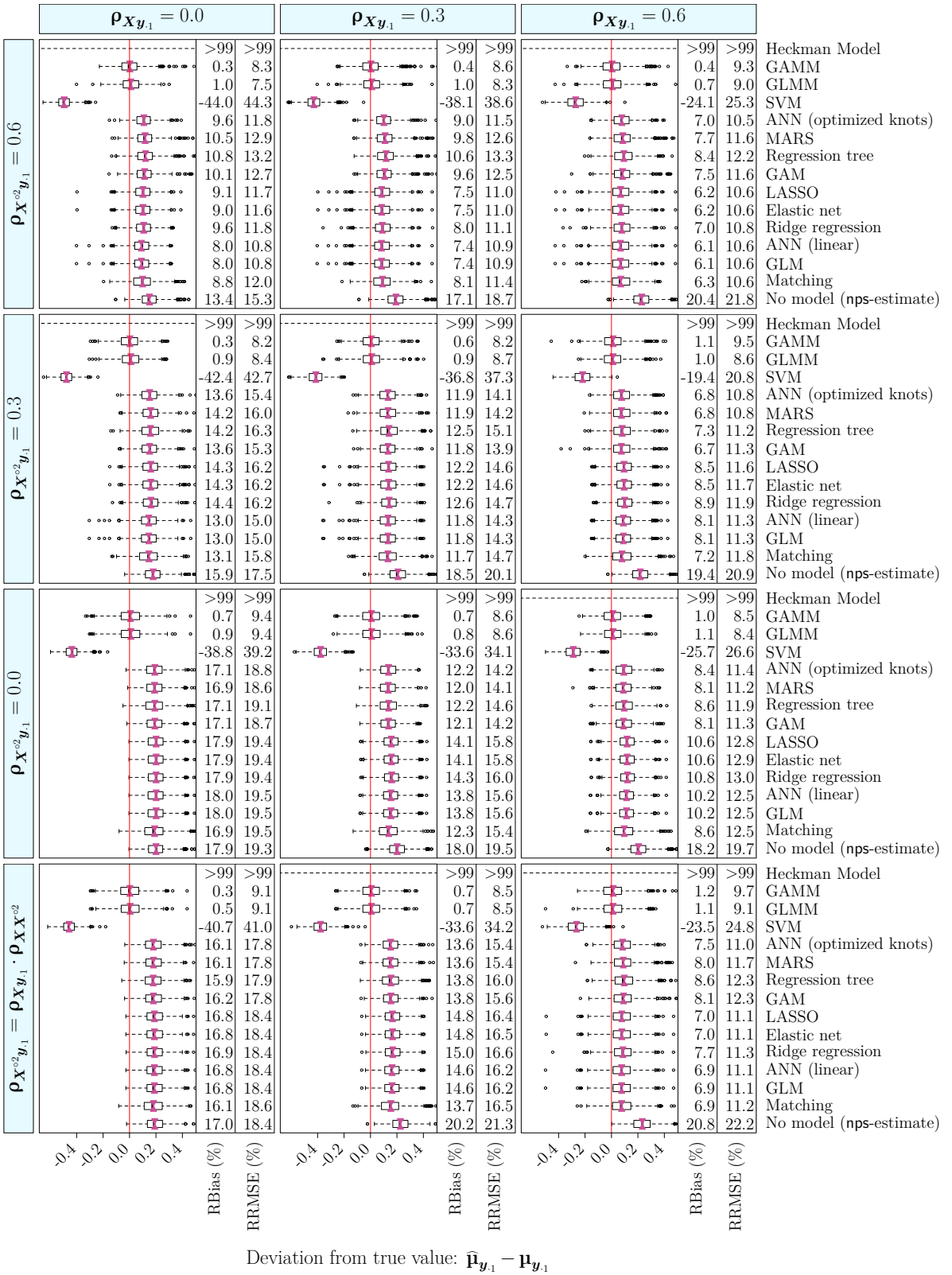


Figure 6.11: Comparison of prediction models for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ for 100% coverage – weighting model: unweighted (estimation from imputed reference sample)

As one would expect, the results in figure 6.11 indicate that the potential of model-based methods to compensate for selection bias and reduce the MSE in comparison to the above naive estimates strongly depends on the relation between auxiliary and target variables. When there is no dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$, estimation by using the imputed reference sample performs similar to the unweighted non-probability sample mean for most of the considered models. But if any such dependency is present, most model-based estimates are able to considerably reduce the bias and MSE. The results show that the amount of bias reduction depends less on the choice of a specific model than on the strength of the linear and non-linear dependencies between target variable and predictor, which are determined by $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$ and $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$. However, prediction models are usually by far simpler to choose, alter and modify than available auxiliary variables in real applications (cf. section 3.2). Therefore, some differences between the considered imputation methods are nevertheless worth discussing.

The importance of conditional independence for model-based estimation discussed in section 5.1 is underlined by the performance of GLMMs and GAMMs. These mixed models make use of $\boldsymbol{Z}$ as additional auxiliary variable. Employing a coarsened version of $\boldsymbol{Z}$ in form of ten classes used to specify the random effects leads to quite good results and nearly eliminates selection bias across all scenarios. The reason is that $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{r}^{\mathsf{nps}}$ are actually conditionally independent given $\boldsymbol{Z}$ in the present context, which corresponds to a MAR pattern for the non-probability sampling mechanism (cf. section 6.3.1). When comparing both types of mixed models, the linear ones typically yield slightly higher biases but lower MSEs than the additive ones. This is an example for the bias-variance trade-off that commonly occurs when higher model complexity allows reducing bias but increases variability (cf. e.g. Hastie, Tibshirani and Friedman, 2008, pp. 37 f). In contrast to all other considered models, which solely rely on $\boldsymbol{X}$ for prediction, the use of an additional explanatory variable leads to considerably better results for these mixed models. The conclusion that this gain in precision is caused by the use of $\boldsymbol{Z}$ is further underpinned in the following discussion (cf. figure 6.18). Despite also employing $\boldsymbol{Z}$ as auxiliary variable for the response model, the Heckman model is generally highly unstable and hence not useful in the present context. This is caused by the fact that this model is highly vulnerable to violated assumptions (cf. section 5.3.1). In particular, availability of auxiliary variables that are solely related to the non-probability selection process but independent from the target variables is required for this model to perform well. This requirement is not fulfilled in the simulation, and the same holds in most real applications (cf. section 5.3.1; Weisberg, 2005, pp. 151 ff). To still be able to visually distinguish the remaining methods, figure 6.11 is set to partially exclude the boxes for Heckman models.

When focusing on models that use only a single auxiliary variable $\boldsymbol{X}$ and are thus subject to MNAR selectivity, most of these models perform somewhat similar to each other and result in larger biases than the mixed models discussed above. When the dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ is of mainly linear nature $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} > \boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}\right)$, matching often results in slightly lower or similar bias in comparison to the other models, but its MSE is typically higher. With increasing non-linear dependencies, however, even purely linear prediction methods are better than matching. Regarding these linear models, least-squares (the plain GLM) seems more suitable for obtaining parameters in the current setting than its penalized versions in form of the LASSO, ridge and elastic net regression (cf. sections 5.1.2 and 5.1.11). Least-squares results in lower bias and MSEs across all scenarios where any dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ is present. The linear ANN is exactly of the same

form as the GLM and hence yields coinciding results. It is included for illustrative and code benchmarking purposes only. Because the `ann`-package with backpropagation is used to fit this model, no Hessian information is required for this purpose. The fact that the results are consistent with those for the GLM indicates that the implementation works as expected and gradient information suffices for estimating this rather simple model. It is somewhat surprising that of all models that use only $\boldsymbol{X}$, these two linear regressions yield the lowest bias in case of a strong correlation between $\boldsymbol{y}_{\cdot 1}$ and squared $\boldsymbol{X}$-variable $\left(\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = 0.6\right)$. Although differences between models are rather small, a strong non-linear dependency at least does not seem to generally reduce the bias for models that do not assume a purely linear influence of $\boldsymbol{X}$ on $\boldsymbol{y}_{\cdot 1}$.

Considering such non-linear prediction models, GAMs are better than GLMs in terms of bias and MSE only when the linear relation between both variables is stronger than the non-linear one $\left(\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} < \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}\right)$, and similar findings occur for the other non-linear models. In comparison to regression trees, which use base functions of zeroth-order to approximate $\boldsymbol{y}_{\cdot 1}$ by a locally constant function, the use of higher order splines in the general MARS model typically improves bias and MSE (cf. section 5.1.7). In particular when $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = 0$, this model outperforms all others that use the same auxiliary information. The performance of regression splines that are based on non-parametric ANNs with optimized knots (cf. section 5.1.9) is typically slightly worse than that of MARS models when the dependencies are mainly linear. Nevertheless, these ANNs are better or at least highly similar in bias and MSE for stronger non-linear dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \leq \boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}\right)$. As for the linear ANN, optimization through backpropagation hence seems to work as expected. Estimates based on predictions from a support vector regression are typically much worse than when using any other (except the Heckman) model, and even unweighted estimates from the non-probability sample perform better. This latter result may be partially explained by the fact that the minimal distance e from the hyperplane as well as the radial kernel transformation are both predetermined across the whole simulation for these SVMs. Methods for making a separate choice for each sample, e.g. by means of cross-validation, could help to overcome this issue (cf. section 5.1.11; Chang and Lin, 2011, p. 24; Drucker et al., 1997, p. 160; Buelens, Burger and van den Brakel, 2015, p. 17).

None of the non-linear models in figure 6.11 appears better than linear ones unless $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$ is smaller than $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$. A possible explanation for this somewhat unexpected and counterintuitive result may again be given by low sampling fractions. Fitting a non-linear model can be advantageous but less stable than a linear one (cf. e.g. Hastie, Tibshirani and Friedman, 2008, p. 22) and therefore requires more information. The sample sizes in the present context might thus be too small to provide sufficient information for gaining efficiency by using models that can represent non-linear dependencies, leading to increasing sampling errors for these models. Linear models do not suffer from this problem because they assume a purely linear relationship, regardless of whether non-linear dependencies exist or not. In contrast, non-linear models can also represent plain linear relationships and offer more flexibility. Higher stability in their estimated parameters may therefore be a reason why accuracy of these models is higher when the actual dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ is mainly linear.

When considering other scenarios in the simulation, the results for estimates of means and totals of $\boldsymbol{y}_{\cdot 1}$ are very similar. Especially the relative performance of prediction models in comparison to each other is largely stable across conditions. In general, model-based

methods are able to (nearly) fully compensate the bias when selectivity is MAR, as in case of the mixed models in figure 6.11. In case of selection mechanisms that follow a MNAR pattern, the bias is typically reduced but not fully compensated through model-based estimation, e.g. for the models that use only $\boldsymbol{X}$ in the above results. However, a general issue with all considered prediction models occurs when the aim is to obtain non-linear estimates, such as for correlation coefficients. Analyzing such measures of association is a common use for non-probability samples (cf. chapter 2), which may be complicated by the model-based paradigm.

An example is given in figure 6.12, evaluating the estimated correlations between $\boldsymbol{y}_{\cdot 1}$ and the second target variable $\boldsymbol{y}_{\cdot 2}$ under the same conditions as in figure 6.11. Although unweighted non-probability sample estimates are biased in all scenarios, they still perform better than any estimator obtained from mass imputation of the reference sample. There are again some dissimilarities between models: as above, GLMMs and GAMMs gain some advantage through additionally using $\boldsymbol{Z}$, and the Heckman model suffers from violated assumptions. All other models under consideration perform more or less similar. In summary, the bias resulting from all model-based methods is severe, such that none of them seems appropriate for estimating $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{y}_{\cdot 2}}$ in the simulation. This bias is due to the fact that these model-based methods for non-probability samples are typically implemented by directly using predictions for mass imputation (cf. e.g. Buelens, Burger and van den Brakel, 2018, p. 330; Kim et al., 2018, p. 7; Yang and Kim, 2018, pp. 4 ff). In the most common case where these predictions correspond to the conditional mean under the model, this strategy can be adequate for estimating statistics like means or totals under conditional independence (cf. equations 5.2 and 5.5). It is, however, of limited use to represent the target variables' actual distribution since any conditional (or residual) covariance is ignored, which induces bias to the model-based estimates in figure 6.12. Similar results are obtained for other measures of association, such as covariances or regression coefficients. For estimating such statistics that incorporate not only means but also (co-)variances or other higher moments of one or multiple variables, it is advisable to rely not only on predictions but to also consider variance components that are not explained by the model. A common strategy to achieve this when compensating for non-response and other forms of missing data in survey research is by means of multiple imputation (cf. e.g. van Buuren, 2018, pp. 63 ff; Little and Rubin, 2019, p. 72; Rubin, 1987, p. 159). It appears sensible to consider this approach for future research on estimation from non-probability samples as well (cf. also Elliott and Valliant, 2017, p. 261).

In general, incorporating auxiliary variables that ensure conditional independence of target variables and sample inclusion is of tremendous importance when using statistical or machine learning models to compensate for selection bias in non-probability samples. When the selectivity pattern is MAR, such prediction methods are able to (almost) completely counterbalance the selection bias for linear statistics. In cases of selectivity that is MNAR, model-based estimation can still reduce but not fully compensate the bias. For estimating non-linear statistics, however, common approaches that use predictions for imputation are of very questionable benefit. Strategies that do not exclusively focus on predictions but also incorporate the unexplained variability, such as common methods for multiple imputation, should therefore be considered for estimation from non-probability samples in future research. An alternative option is the use of weighting methods or their synthesis with prediction models. These are evaluated in the following discussion.

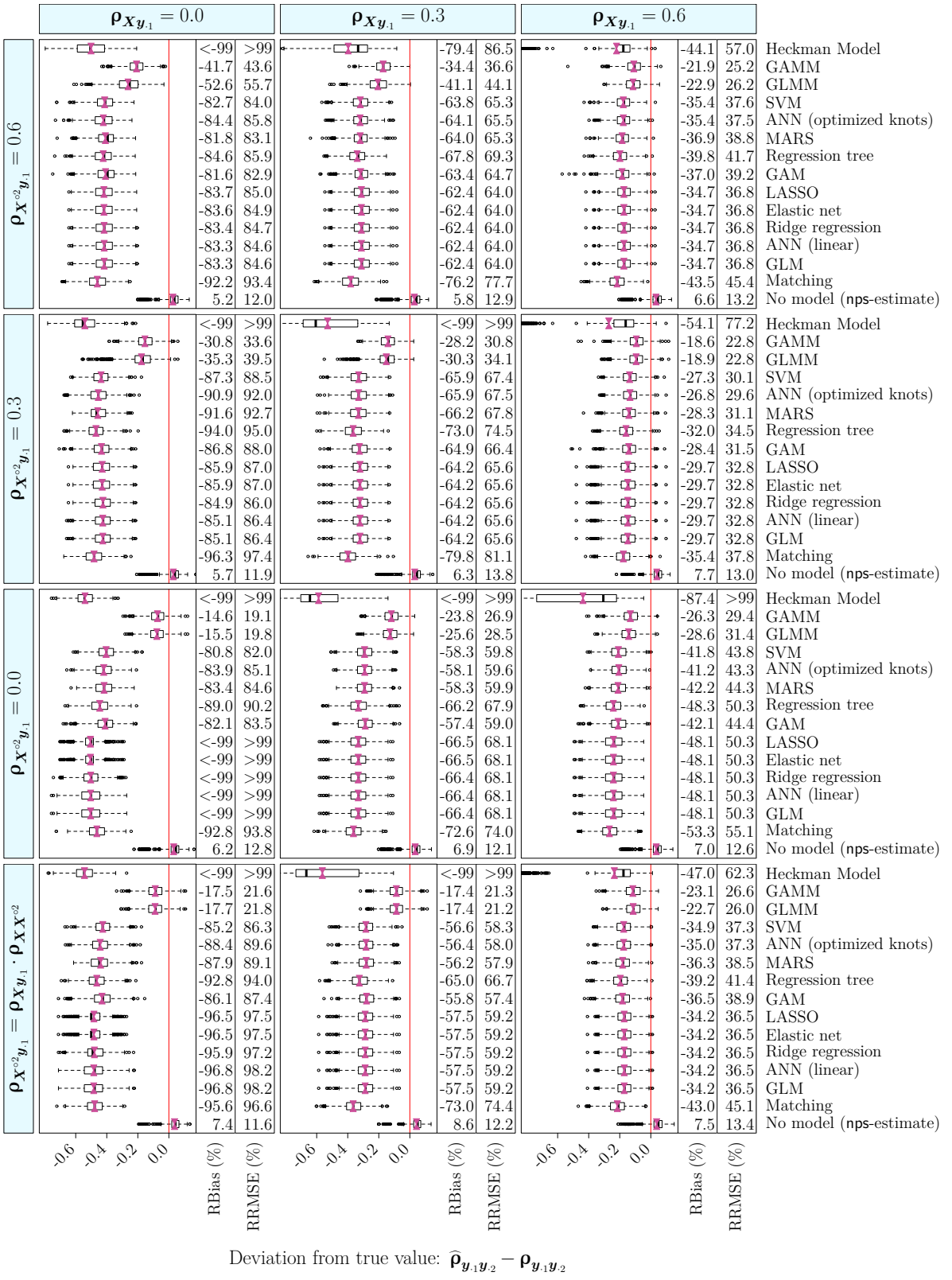


Figure 6.12: Comparison of prediction models for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\rho_{\boldsymbol{y}_{\cdot 1}\boldsymbol{y}_{\cdot 2}}$ for 100% coverage – weighting model: unweighted (estimation from imputed reference sample)

Pseudo-design-based Methods

Although pseudo-design-based methods attempt to use auxiliary information in a different way than model-based approaches, conditional independence assumption 5.1 is the common ground underlying both frameworks. Based on this foundation, different methods to generate pseudo-design weights for non-probability samples are discussed in section 5.2, leading to the proposal of an integrative weighting approach in form of calibrated semi-parametric artificial neural networks. All these pseudo-design-based methods rely on a model for the non-probability sampling process and/or adjustments of the data to external benchmarks (cf. e.g. equations 5.152 to 5.154). Depending on the type and extent of available auxiliary information, various forms of pseudo-design weights for non-probability samples are eligible. As discussed in section 6.3.1, different possible settings of such auxiliary information are considered in the simulation. Results for each of these settings are successively presented in the following paragraphs. Similar as before, a baseline or reference point for comparison is included in each of the figures, which is given by plain non-probability sample estimates that do not use any weights ('unweighted').

In figure 6.13, it is assumed that only a reference sample is available as auxiliary information for modeling response propensities. In such cases, the common weighting strategy is to rely on inverse estimated response propensities, for which the discussion in sections 5.1 and 5.2.1 facilitates different modeling strategies. The most common choice is to rely on inverse predicted probabilities from a GLM with logit link ('logit model'). In the simulation, this model is compared to an ANN of the same structural form (i.e. a single layer with softmax activation and deviance loss function) and 'pseudo-weights' (cf. section 5.2). Following Elliott and Valliant (2017, p. 257), these pseudo-weights are as well based on a GLM, but obtained as outlined in equation 5.137 rather than using the inverse predictions directly. The neural networks that are employed for weighting follow the framework described in section 5.2.3 and hence are generally labeled as calibrated ('cal. ANN'). Although no calibration is applied for the results in figure 6.13, this facilitates coherence with the following figures. For each of these weighting methods, a parametric and a non-parametric model specification is considered. The latter is constituted in form of B-spline regression with five evenly spaced ('fixed') knots, while the former does not use any transformation of the independent variables (cf. section 5.1.6). In case of ANNs, knot optimization as proposed in section 5.1.9 is additionally considered ('optimized knots'). For all of these models, the auxiliary variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ observed in the non-probability and the reference sample are used.

From the results in figure 6.13, it is evident that unweighted estimates from the non-probability sample are biased in all depicted scenarios. This is because target variable $\boldsymbol{y}_{\cdot 1}$ and selection probability $\boldsymbol{\pi}^{\mathsf{nps}}$ are both correlated with $\boldsymbol{Z}$ as in the previous figures 6.11 and 6.12. Selectivity is MAR for all of the considered propensity models since $\boldsymbol{Z}$ is used as auxiliary variable. Consequently, most of the represented pseudo-design-based approaches are able to reduce the bias in comparison to unweighted estimates over all scenarios but to different degrees. In terms of MSE, only some methods perform better than unweighted estimates across all scenarios.

Considering parametric models first, pseudo-weights typically result in the lowest bias in this category if there is any linear or non-linear relationship between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$. Especially for stronger non-linear dependencies, however, parametric ANNs result in lower MSEs, with biases usually lying in between those of logit and pseudo-weights. Propensity weights obtained from the logit model perform worse than both other parametric weighting

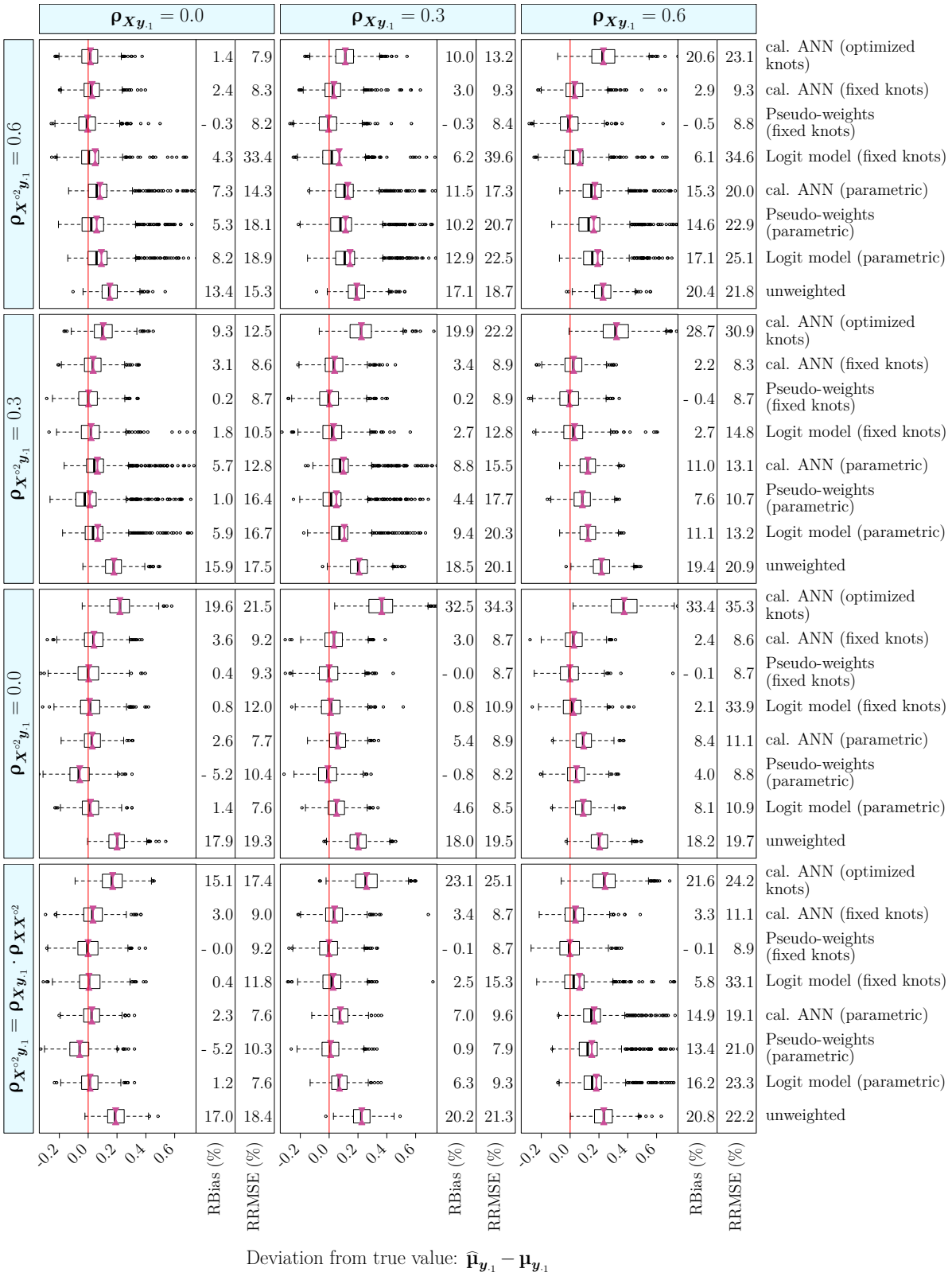


Figure 6.13: Comparison of weighting methods for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ for 100% coverage, using a reference sample

approaches unless the correlation of $\boldsymbol{y}_{\cdot 1}$ with $\boldsymbol{X}$ and its squared values are both zero. Since ANN and logit model are of the same structural form (cf. section 5.1.8), backpropagation seems more suitable than Fisher scoring for propensity models, especially in presence of non-linear dependencies. This is presumably caused by the fact that the use of (expected) Hessian matrices for optimization is prone to finding saddle points rather than global optima (cf. Hagan et al., 1996, p. 281). Instead of using a GLM to calculate pseudo-weights, which is the strategy proposed by Elliott and Valliant (2017, p. 257), it may therefore be worthwhile to use ANNs (or GLMs that use gradient descent) to achieve better stability and lower bias, thereby combining the advantages of both methods.

When applying spline regression with fixed knots rather than parametric propensity models, the gain in bias correction is considerable. This improvement is particularly strong in case of pseudo-weights, where it nearly eliminates the bias across all scenarios. Nevertheless, the magnitude of biases in case of logit model and ANN decreases as well if there is any dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$. In case of the logit model, this bias reduction often comes at the price of a MSE that is higher than for the parametric variant. This drawback hardly occurs for pseudo-weights or ANNs, and only if the non-linear dependency of $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ is negligible. In most cases, the MSEs of these two latter approaches are rather close to each other. As in the parametric case, this suggests that a combination of both may be sensible since pseudo-weights have lower biases but ANNs have lower variances. As before, this is presumably due to the fact that gradient descent is less prone to finding saddle point solutions than Fisher scoring, which is used for calculating pseudo-weights. Propensity weighting that is based on an ANN with optimized knots in general performs rather poor and partially even worse than the unweighted estimator. If at all, it may be considered worthwhile for weighting in case of strong and purely non-linear dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0 \text{ and } \boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = 0.6\right)$, where the resulting MSE is the lowest of all methods. Since ANNs with optimized knots perform reasonably well for purely predictive tasks (cf. figure 6.11), this poor quality in case of propensity weighting is likely caused by over-fitting, such that the particular sample characteristics have more influence on weights than the general selectivity pattern.

For plain propensity weighting, the results when estimating the mean $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ in other settings of the simulation are highly similar, although (nearly) unbiased estimation is only possible when selectivity is based on a MAR pattern. Different degrees of selectivity, coverage and violation of conditional independence (cf. assumption 5.1) affect the extent of bias and MSE, and no method achieves unbiasedness in case of MNAR. Weighting by inverse propensities derived from a logit GLM is typically worse than using any of the other two considered models. Pseudo-weights seem preferable for the simulated scenarios since they typically result in the lowest bias and provide similar MSEs as the ANNs. In general, non-parametric response models appear better than parametric ones, especially in case of strong dependency between variable of interest and sample inclusion. However, this typically increases the MSE for logit models. Overall, the relative performance of propensity weighting methods in comparison to each other is quite stable across different simulation scenarios.

This finding is underlined by the results presented in figure 6.14. For achieving selectivity that violates conditional independence assumption 5.1 and hence corresponds to a MNAR scenario, sample inclusion and target variable $\boldsymbol{y}_{\cdot 1}$ are set to be correlated by $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$. The remaining setting is the same as in the previous figure 6.13. These results demonstrate that none of the propensity weighting methods is able to fully compensate

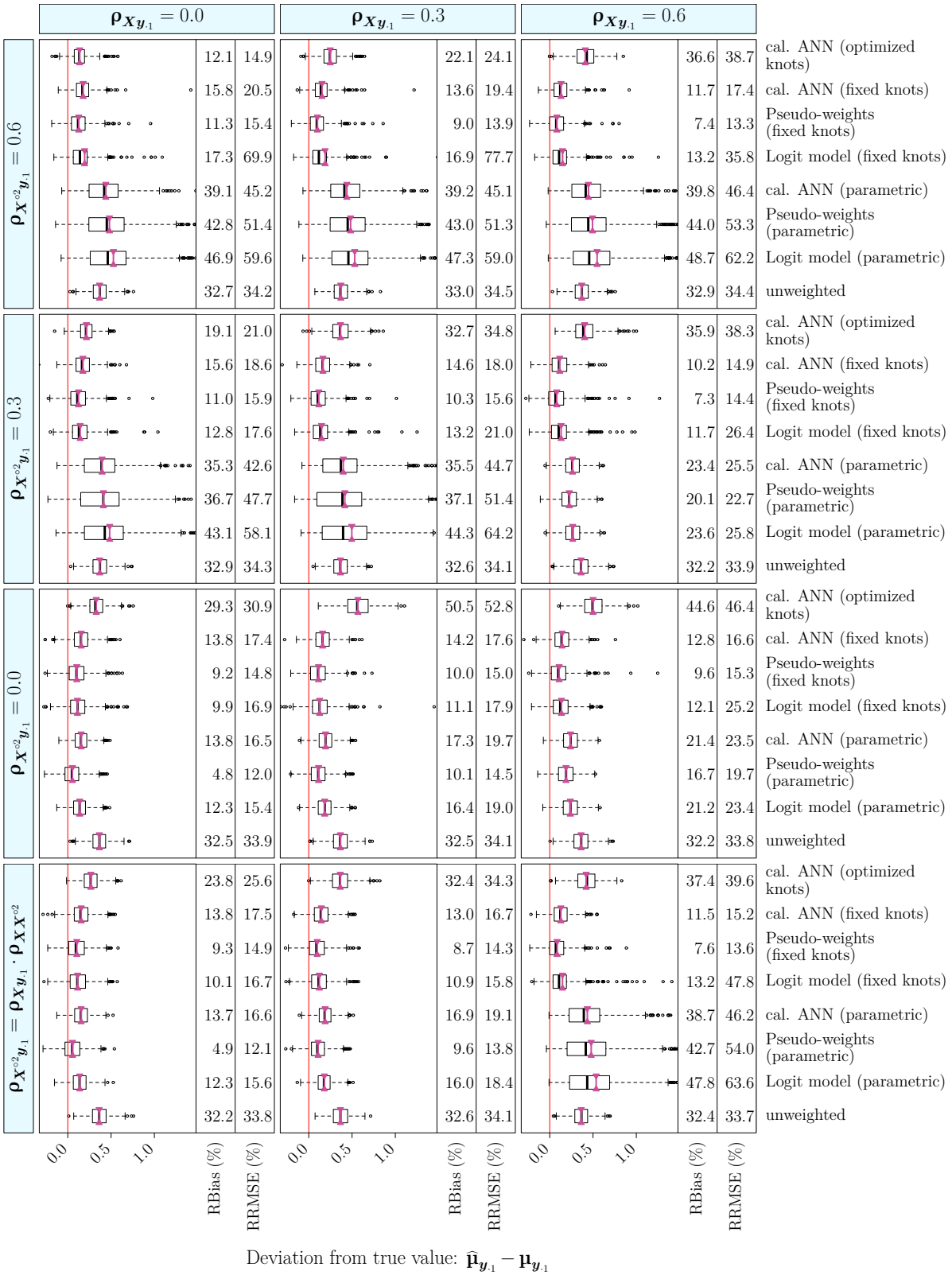


Figure 6.14: Comparison of weighting methods for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ for 100% coverage, using a reference sample

for selection bias in this case of stronger (MNAR) selectivity. Nevertheless, the finding that pseudo-weights outperform all other propensity weights under consideration is clearly the same as for the MAR case in figure 6.13. If anything, this is even more evident from the MNAR scenario in figure 6.14 because all changes in the methods' relative performances in comparison to figure 6.13 are in favor of the pseudo-weights. In real applications, it is typically very difficult or even impossible to detect whether it is reasonable to assume conditional independence and hence MAR rather than MNAR selectivity. Since the latter is by far the more challenging case (cf. Elliott and Valliant, 2017, p. 262; Little and Rubin, 2019, p. 26; Mercer et al., 2017, p. 257; Pfeffermann, 2011, p. 117), the following figures focus on settings where selectivity is MNAR as in figure 6.14.

The general finding that pseudo-weights lead to better estimates than the other considered propensity weights is similar for estimation of multi- or bi- rather than univariate statistics in cases of selectivity patterns that are MAR. However, the general picture is clearly different if selectivity is MNAR. In this case, pseudo-weights seem less appropriate for estimating measures of dependencies, while non-parametric ANNs with knot optimization appear favorable. An example for this finding is presented in figure 6.15, considering estimates for the correlation between $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{y}_{\cdot 2}$. The setting is the same as in figure 6.14. Although unweighted estimates are again considerably biased, all parametric weighting models even increase this bias in all scenarios. In some cases, the respective MSEs are lower for the weighted than for unweighted estimates, but these differences are rather small. As before, non-parametric propensity models with fixed spline knots show some advantages over parametric ones in terms of bias as well as MSE. However, the relative gain is not as large as in figure 6.13. Consequently, there is at best some minor improvement in bias and MSE over using no weights at all, and results are still even partially worsened by applying propensity weighting. Overall, parametric as well as non-parametric response models perform only gradually different in this setting as long as the B-spline knots are predetermined. In contrast, the improvements when using the proposed knot optimization technique in semi-parametric ANNs are considerable. Although unbiasedness is still not achieved, this estimation approach outperforms all others in figure 6.15 in each but one scenario $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.3 \text{ and } \boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = 0\right)$ in terms of bias as well as MSE.

Further results for pure propensity weighting in the simulation are highly similar as in figures 6.13 to 6.15. The relative performance and ranking of the considered methods is stable although the degrees of bias are different, e.g. when considering under-coverage of the target population or estimating regression instead of correlation coefficients. Summarizing these results, pseudo-weights as proposed by Elliott and Valliant (2017) seem a promising approach to estimate linear statistics, in particular when a non-parametric response model is used. To gain this advantage over the considered propensity weights that are based on GLMs or ANNs, pseudo-weights require the sampling design of the reference sample to be (nearly) perfectly described by the auxiliary variables used in the response model. Since reference samples are drawn by simple random sampling, this requirement is perfectly fulfilled in the present simulation. However, it is easily violated in real applications, where complex survey designs are the usual case (cf. Elliott and Valliant, 2017, p. 257). For example, design variables used for the reference sample may be not available in the non-probability sample or even removed from the reference data, e.g. to avoid potential disclosure risks (cf. e.g. Skinner, 2009, p. 383; Willenborg and De Waal, 2001, p. 10). Performance of pseudo-weights under such less than ideal conditions may, hence, be less efficient than in the current simulation study. Some evidence for

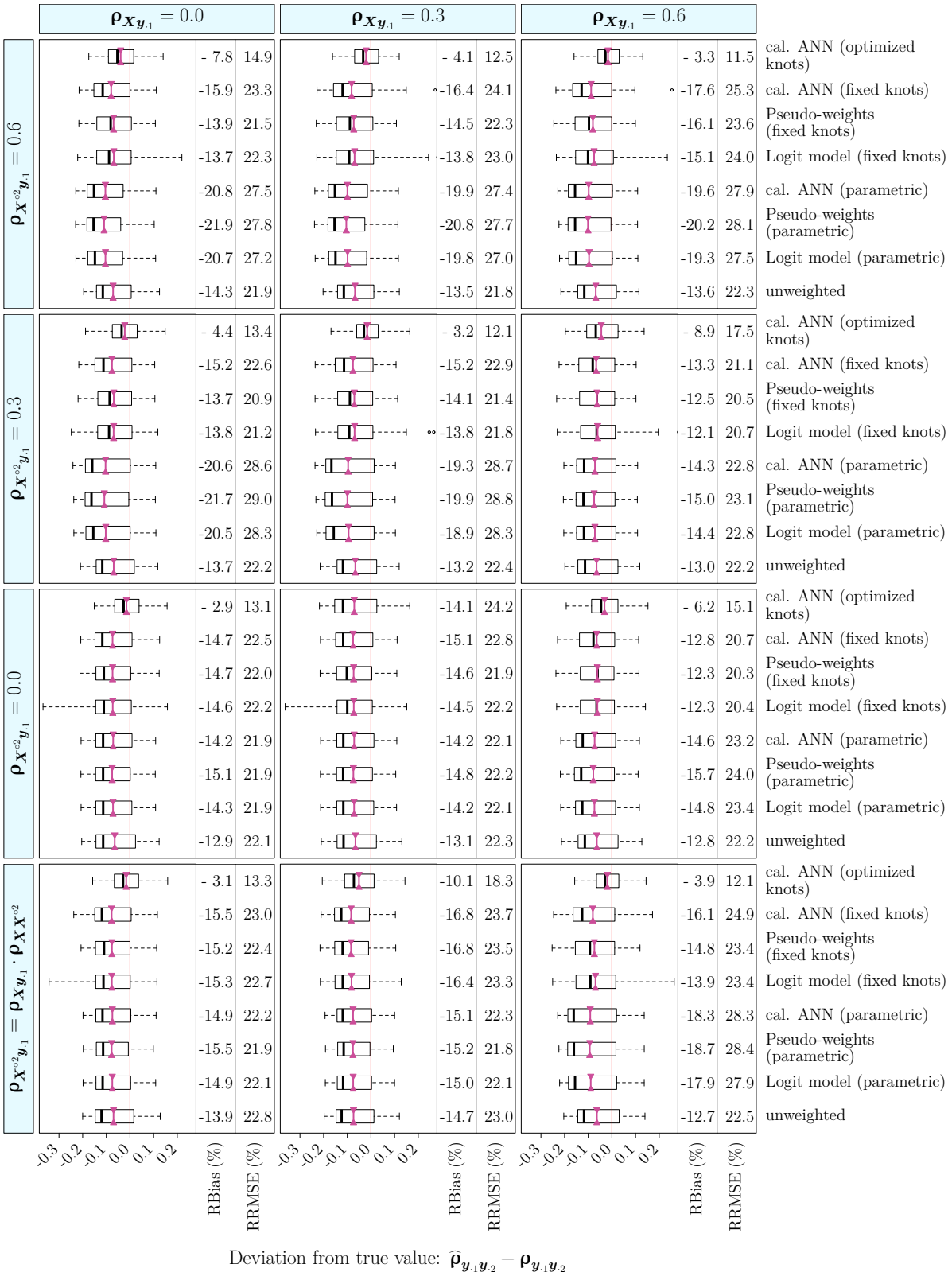


Figure 6.15: Comparison of weighting methods for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\rho_{\boldsymbol{y}_{\cdot 1}\boldsymbol{y}_{\cdot 2}}$ for 100% coverage, using a reference sample

this limitation is provided in case of the application example that is discussed in chapter 7. As is apparent by reference to the considered ANNs, the use of gradient descent rather than Fisher scoring apparently achieves higher stability for propensity weighted estimates. Determining pseudo-weights from such ANNs (or other models that are fit using only first derivatives for optimization) may, hence, be a sensible alternative to using GLMs, especially when non-linear dependencies are present. For non-linear estimates (such as measures of dependencies), the best approach for propensity weighting depends on the sample's selectivity. In MAR scenarios, pseudo-weights yield good results, but when selectivity is more severe (MNAR), semi-parametric artificial neural networks with knot optimization seem particularly promising.

In figures 6.13 and 6.15, the available auxiliary information external to the non-probability sample is limited to a small reference sample, which is exclusively used for response (propensity) modeling. When considering availability of other types of external information, the eligible methods for determining pseudo-design weights are different as well. One such setting is presented in figure 6.16, considering pseudo-design-based estimates for $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ which use total and covariance benchmarks for $\boldsymbol{X}$ as auxiliary information for calibration. The presumably most common weighting approach that can make use of such information is the GREG (cf. e.g. Dever, Rafferty and Valliant, 2008, p. 60; Mercer et al., 2017, p. 264). It is compared to different forms of calibrated semi-parametric artificial neural networks (cf. section 5.2). As discussed in sections 5.2.3 and 6.2, one option to specify these ANNs is in correspondence to the GREG, resembling the approaches of Burgard, Münnich and Rupp (2019), Guggemos and Tillé (2010) and Rupp (2018, p. 126) under additional constraints (cf. problem 5.151). This is achieved by using one parameter per observation, in conjunction with a linear activation function and squared penalty for the parameters. An alternative specification of calibrated ANNs incorporates the functional form of a logit model, considering parametric and non-parametric options as in the previous figures 6.13 and 6.15. Following the functional form approach, these models obtain weights as a function of response model variables $\boldsymbol{X}$ and $\boldsymbol{Z}$, which are observed only in the non-probability sample (cf. section 5.2). Since no reference sample is available in this setting, all weights are determined from calibration constraints alone. The considered ANNs apply soft calibration with box-constraints, allowing a maximum of 2.5% deviation from total and 10% from variance benchmarks (cf. equations 5.152 to 5.155).

Scenarios in figure 6.16 are the same as in figure 6.15, such that sample inclusion and target variable $\boldsymbol{y}_{\cdot 1}$ are again correlated by $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$. Since this implies stronger selectivity in comparison to figure 6.13, unweighted estimates are systematically more biased and have higher MSEs. None of the weighting approaches under consideration completely eliminates this bias because calibration is limited to $\boldsymbol{X}$ and selectivity, thus, MNAR. Nevertheless, all of these methods are able to reduce the bias and the MSE if there is any dependency between auxiliary and target variable or sampling mechanism. For the GREG and calibrated ANN which both use one parameter per observation, this is the case when $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ show any linear or non-linear relationship since these methods determine weights based on $\boldsymbol{X}$ alone. From these two, the GREG is better in almost all scenarios. This is due to the fact that soft calibration is mainly designed for benchmarks that are subject to (sampling) errors or highly abundant (cf. e.g. Burgard, Münnich and Rupp, 2020, p. 12; Deville, Särndal and Sautory, 1993, p. 1015), while a single population total and variance are used for calibration in this context. Therefore, the GREG seems to be the best of these two choices, especially in cases where a strong and mainly linear relationship

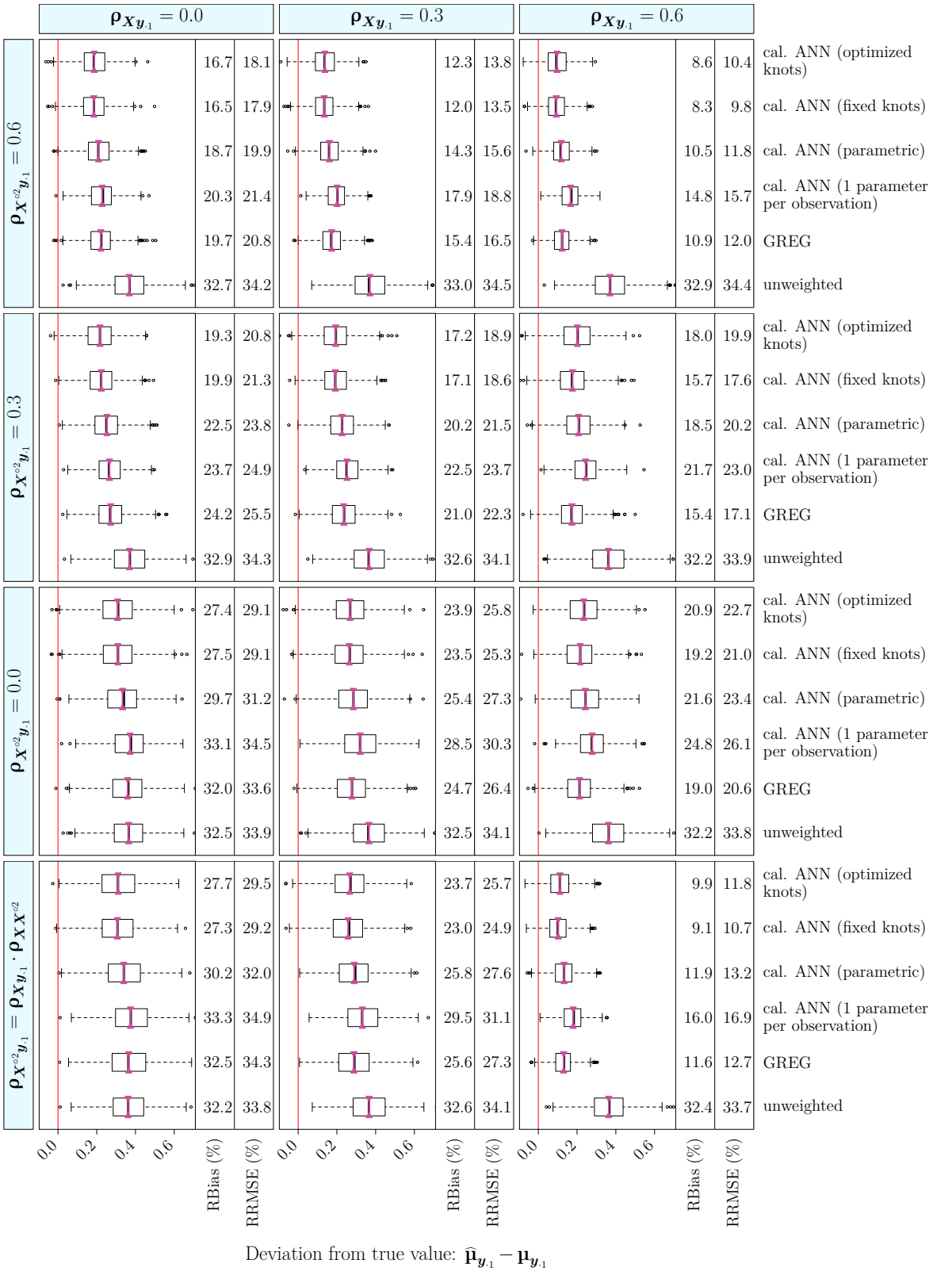


Figure 6.16: Comparison of weighting methods for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$: estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ for 100% coverage, using total and covariance constraints

between calibration and target variable is present $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6 \text{ and } \boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} < 0.6\right)$. However, because no calibration benchmarks are available for $\boldsymbol{Z}$, the fact that this variable is related to both $\boldsymbol{\pi}^{\mathsf{nps}}$ and $\boldsymbol{y}_{\cdot 1}$ cannot be exploited by the GREG. Calibrated ANNs that determine weights as a function of $\boldsymbol{X}$ and $\boldsymbol{Z}$ are therefore quite competitive with the GREG, although selectivity is still MNAR when using $\boldsymbol{Z}$ because $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$. As before, the non-parametric specification with predetermined (evenly spaced) knots performs best for these models. In most scenarios, such a non-parametric calibrated ANN is better than all other approaches under consideration in terms of bias as well as MSE. Even in the few settings where this is not the case, it is only slightly worse than the best option. As before (cf. figure 6.13), knot optimization performs worse than evenly spaced knots, at least for design linear estimates.

Similar results for calibration weighting occur in the further simulation scenarios and also hold for estimates of other statistics, such as for totals, correlation or regression coefficients. While soft calibration is mainly designed for cases where many or not fully reliable benchmarks are used (cf. Burgard, Münnich and Rupp, 2020, p. 12; Deville, Särndal and Sautory, 1993, p. 1015), the GREG is very useful when calibration is done exclusively to relatively few known population totals. Although the GREG can also incorporate covariance or correlation benchmarks by calibrating totals of interaction terms or squared variables (cf. section 5.2.3), the results of calibrated ANNs are generally more convincing when such benchmarks are available. As for other calibration methods that follow the functional form approach, an advantage of such ANNs in comparison to the GREG is that they can make use of the explanatory power of $\boldsymbol{Z}$ for $\boldsymbol{y}_{\cdot 1}$ and/or $\boldsymbol{r}^{\mathsf{nps}}$ even if $\boldsymbol{Z}$ is observed only in the non-probability sample. This is the case when external information is limited to calibration targets of $\boldsymbol{X}$, but $\boldsymbol{Z}$ is additionally measured solely in the non-probability sample. In principle, even $\boldsymbol{Y}^{\mathsf{nps}}$ itself could be used to specify the pseudo-design weights in this setting.

The third scenario of auxiliary information investigated in the simulation therefore assumes joint availability of calibration benchmarks and a reference sample for propensity modeling. In this setting, all weighting methods can make use of auxiliary variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ to the same degree. As for figures 6.13 to 6.15, both auxiliary variables are observed in the non-probability and the reference sample. All other simulated conditions, and in particular the calibration benchmarks obtained from the population, are the same as in the previous figure 6.16. While calibrated ANNs can readily incorporate reference sample and calibration benchmarks as auxiliary information, other types of propensity weights need to be combined with a calibration technique to achieve this. A common approach in that regard is to obtain response propensities from a GLM with logit link and calibrate the resulting propensity weights by means of the GREG in a second step ('logit model and GREG', cf. e.g. section 6.2.2; Enderle, Münnich and Bruch, 2013, p. 94; Lee and Valliant, 2009, p. 335; Valliant and Dever, 2011, p. 109).

Results to compare these strategies are presented in figure 6.17. When all approaches use information about $\boldsymbol{X}$ and $\boldsymbol{Z}$, the proposed calibrated ANNs are outperformed by the combination of logit propensity model and GREG in all scenarios. The latter two-step approach appears generally favorable in the current setting, regardless of whether a parametric or spline regression model is used for the propensities. Similar as in figure 6.14, calibrating the propensities from a parametric logit model almost generally results in the lowest MSEs but yields slightly higher biases than the non-parametric variant for most scenarios where dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ is exclusively non-linear $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0\right)$.

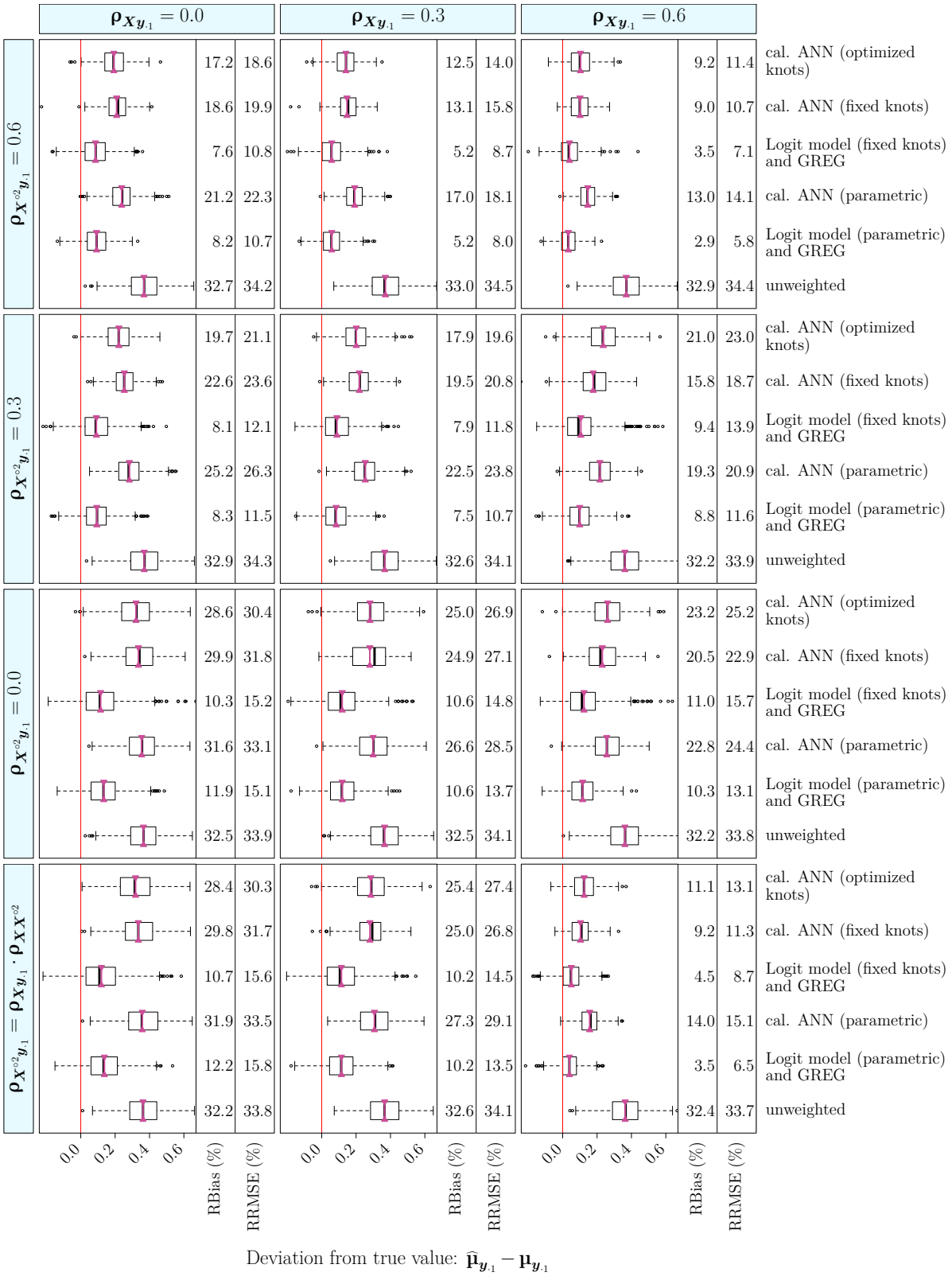


Figure 6.17: Comparison of weighting methods for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$: estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ for 100% coverage, using a reference sample, total and covariance constraints

Further results in the simulation are quite similar, regardless of the degree of selectivity, coverage of the target population and considered estimates (e.g. correlation or regression coefficients). When jointly using calibration benchmarks and a reference sample for modeling response propensities, the combination of GLM and GREG mostly outperforms calibrated ANNs. When comparing figures 6.16 and 6.17, it is evident that the additional information provided by the reference sample is advantageous when supplementing the GREG with a propensity model. In contrast, calibrated ANNs do not gain any advantage from this additional information. Instead, the resulting biases and MSEs even increase for all of these models that are available under both conditions. The main reason is the composition of the distance function that is used for fitting calibrated ANNs. In figure 6.16, these models are fit solely by minimizing the distance between estimates and calibration targets, whereas the deviance (negative binomial log-likelihood) is added as a second component in the context of figure 6.17. On its own, each of these elements is justified and used for different established approaches, such that this combination aims at an integration and trade-off regarding these existing methods (cf. section 5.2). However, the joint use of both distance components is of somewhat limited value in the simulation, mainly because a way to combine them that is useful in all scenarios is difficult to find. A vector of importance weights $\mathbf{v}$ is introduced in equation 5.155, which allows for flexible specifications of such a combination. For the current simulation, the strategy summarized in equalities 6.2 is applied. It is based on choosing $\mathbf{v}$ such that each component (deviance, penalties for total and covariance soft calibration as well as parameter shrinkage) has the same maximum contribution to the overall distance. While this approach performs relatively well for the preliminary simulation in section 6.2, the current results indicate that it may be too simplistic for more complex cases. Therefore, it seems essential to develop more refined strategies to choose a proper mixture of distance metric components when importance weights are not uniquely determined by user-defined priorities. Alternative strategies proposed for finding such weights based on quality measures (such as the estimated variance of a point estimator) are hardly applicable in the present context because these measures are not generally valid for non-probability samples. Therefore, further research is needed to establish a more adequate strategy for choosing the importance weights to combine the multiple considered optimization criteria (cf. sections 5.4 and 6.2; Chang and Kott, 2008, p. 559; Guggemos and Tillé, 2010, p. 3205; Marler and Arora, 2004; 2010).

Nevertheless, the results in figures 6.13 to 6.16 indicate that calibrated ANNs perform quite well for weighting when the distance components are not combined. Consequently, the choice of an adequate weighting method depends on the quality and type of available auxiliary information as well as the quantities to be estimated. When plain propensity weighting is used, pseudo-weights perform superior to the competing approaches for means or total estimates, but semi-parametric ANNs seem better for estimating correlations. A limitation of pseudo-weights is that the sampling design of the reference sample has to be (almost) perfectly described by the available auxiliary variables. This requirement is easily violated in real applications and not needed if the propensity weights are based on ANNs or logit models. Considering pure calibration weights, the GREG performs well when only a few population totals are used as calibration targets. However, it is often too strict in presence of many constraints or when benchmarks are subject to inaccuracies (such as sampling errors; cf. Burgard, Münnich and Rupp, 2020, p. 12; Guggemos and Tillé, 2010, p. 3199). When incorporating not only total but also covariance or correlation

benchmarks for calibration, the GREG is still applicable, but the proposed calibrated ANNs perform better because they can use variables which are measured only in the non-probability sample. For jointly using a reference sample and calibration benchmarks, combining deviance and soft calibration distance components for such calibrated ANNs still requires some refinement (cf. equation 5.155 and section 6.2). Nevertheless, calibrated ANNs seem especially promising when dependencies between variables are of interest, either as calibration benchmark or as the statistic to be estimated from a non-probability sample. This implication gains some further relevance in the following discussion, where joint usages of pseudo-design and model-based paradigms are evaluated.

Synthesis of Model- and Pseudo-design-based Methods

In the context of the previous figures 6.11 to 6.17, the model- and the pseudo-design-based paradigm are considered as mutually exclusive. When selectivity is MAR, i.e. under conditional independence of target variables and sample inclusion given the respective auxiliary variables, both paradigms can provide estimates that are (nearly) unbiased. However, such a strict separation of the two paradigms is not necessarily of major relevance in actual applications. A combination of both may be useful especially in case of stronger (MNAR) selectivity, where neither weighting nor prediction methods can be expected to eliminate selection bias (cf. figures 6.11 and 6.17; Buelens et al., 2012, p. 18; Gelman et al., 2016a, pp. 109 f; Wang et al., 2015). As discussed in section 5.3, possible strategies proposed for integrating the model- and the pseudo-design-based paradigm are

a) to jointly model selection process and target variable,

b) to use weighted loss functions for model fitting, or

c) to apply weighted aggregation of predictions in the non-probability sample.

The Heckman model is the most common example for the first approach. As already indicated in figures 6.11 and 6.12, it is of limited use when the underlying assumptions are violated, which is commonly the case in real applications (cf. Weisberg, 2005, pp. 151 ff). The performance of strategies ***b)*** and ***c)*** is evaluated in the following paragraphs, considering various combinations of weighting and prediction models discussed in sections 5.1 and 5.2. To check whether the Heckman model can be improved by additionally incorporating pseudo-design weights, it is included as well for these evaluations.

When a reference sample is the only available auxiliary information, the results in figure 6.13 show particular advantages of pseudo-weights obtained from a non-parametric logit model for estimating the mean $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$. In figure 6.18, the aim is to evaluate whether a combination of these weights with prediction models can further improve estimation. Considering the same scenarios as in the previous figures 6.15 to 6.17, selectivity is again MNAR due to $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$. When fitting models to survey samples, strategy ***b)*** is the typical case. Consequently, a propensity weighted loss function is applied for fitting the prediction models, which are then used for imputing the reference sample (cf. e.g. Breidt and Opsomer, 2017; Beaumont, 2000; Pfeffermann and Sverchkov, 1999). As a reference point for comparison, the purely pseudo-design-based estimates from the non-probability sample ('nps-estimates') are included as in figures 6.11 and 6.12.

The results in figure 6.18 illustrate that when $\boldsymbol{X}$ is a valuable predictor for $\boldsymbol{y}_{\cdot 1}$, the use of mass imputation from a weighted model typically reduces bias in relation to pure propensity weighted estimates even if $\boldsymbol{X}$ is already used for the propensity model. As in the unweighted case, different prediction models again perform partially dissimilar.

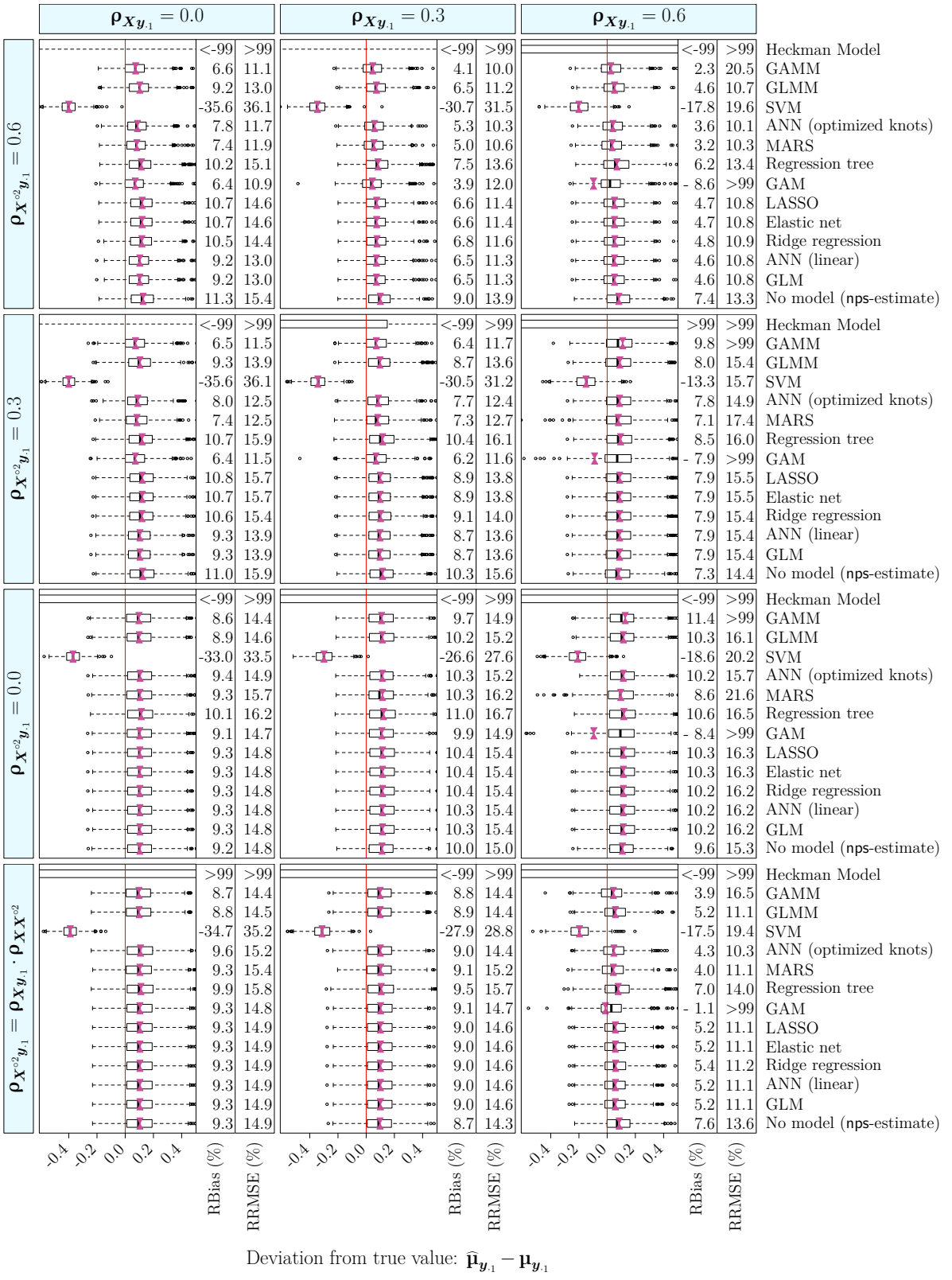


Figure 6.18: Comparison of prediction models for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ for 100% coverage – weighting model: pseudo-weights (fixed knots), using a reference sample (estimation from imputed reference sample)

Presumably for the same reasons as discussed with regard to figure 6.11, support vector machines yield rather bad estimates. Even worse results are obtained from the Heckman model, for which estimates are again often far off the scale due to violated assumptions. These findings are similar in the following figures. The distributions of estimates obtained from all other models are at least rather close to each other. Although mixed models are still useful in this context, their strong advantage for the unweighted case in figure 6.11 is caused by the use of $\boldsymbol{Z}$ to specify the random effects. A lot of this advantage is actually compensated when including $\boldsymbol{Z}$ in the response model that is used as foundation for fitting propensity weighted prediction models. Considering purely linear models, it seems again favorable to choose a least-squares fit, i.e. the plain GLM, rather than its penalized versions. However, non-linear models are slightly better than linear ones in all scenarios, which is different from figure 6.11. For a strong dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6\right)$, regression splines with optimized knots seem particularly suitable. In this setting, MARS models perform comparatively well in terms of bias and MSE, but ANNs with knot optimization are better when considering only the MSE. Additive (mixed) models can lead to slightly lower biases in this setting but typically result in much worse MSEs, while linear (mixed) models mostly lead to higher biases. In scenarios where the linear dependency is less pronounced $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} < 0.6\right)$, GAMs seem more favorable when $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} \geq 0.3$. In all other scenarios that are not explicitly considered above, models provide at best a minor benefit over purely pseudo-design-based estimation.

The simulation results furthermore indicate that biases for weighted prediction models are also lower than for the unweighted ones. For example, this is evident when comparing figures 6.18 and 6.11. Although selectivity is more severe in the former, most of the presented estimates are still better than in the latter. Mixed models are an exception because selectivity for those is MAR in figure 6.11 but MNAR in figure 6.18. The results are similar for other forms of propensity weighted estimates of the mean $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$, although non-parametric propensity models often perform slightly worse than parametric ones in this context. These general patterns extend to further scenarios considered in the simulation as well. In summary, the better $\boldsymbol{y}_{\cdot 1}$ is predicted by $\boldsymbol{X}$, the more gain in efficiency for mean or total estimates can be achieved under strong (MNAR) selectivity by using an adequate weighted prediction model rather than plain pseudo-design-based estimation. The same argument applies vice versa: the better the selection processes can be modeled, the more accuracy is gained from using propensity weighted instead of unweighted loss-functions for prediction models in mass imputation. However, these results hold for propensity weighting only when estimating univariate statistics like means or totals. As discussed with regard to figure 6.12, the common use of predictions as imputed values often leads to considerable bias for non-linear (e.g. bi- or multivariate) estimates because the residual (co-)variance is ignored, a limitation that also occurs for weighted prediction models. Furthermore, the outlined advantages of combining propensity weights with model-based approaches only apply when the weighted models are actually used for imputation in the reference sample (strategy ***b)***). Weighted aggregation of prediction as outlined in strategy ***c)*** does not improve efficiency over plain pseudo-design-based estimates when using inverse propensity scores for weighting.

When a reference sample is used as the only auxiliary information for estimating means or totals, purely model- or pseudo-design-based methods are only preferable if conditional independence (cf. assumption 5.1) holds. If this is the case, combinations of propensity and prediction models can be less accurate than one of the approaches on its own. In all

other settings, their joint usage in form of propensity weighted mass imputation models appears more suitable but requires identification of adequate specifications for both the weighting and the prediction model. Since a reference sample is anyhow required for fitting a propensity model, the only additional requirement (e.g. in comparison to figure 6.13) is that relevant predictors for the target variables are measured in non-probability as well as reference sample.

The setting is different when calibration benchmarks are the only available auxiliary information, in which case imputation of the reference sample is infeasible. Under such conditions, weighted aggregation of predictions in the non-probability sample, which corresponds to the above strategy ***c)***, is the main approach to combine the model- and the pseudo-design-based paradigm. This procedure applies pseudo-design weighted estimation to model predictions rather than to the observed values in the non-probability sample but does not impute for any reference sample. Results for this technique are presented in figure 6.19, considering the same scenarios as before. Note that in order to exactly meet the typical specification of MRP as the most common realization of these methods (cf. section 5.3.2), calibration and random effects variables coincide just for this single example.

Based on these results, it is evident that weighted aggregation of predictions only rarely reduces bias or MSE in comparison to calibration (i.e. post-stratification) alone. Such an improvement only occurs when $\boldsymbol{y}_{.1}$ is strongly correlated with at least one of either $\boldsymbol{X}$ or $\boldsymbol{X}^{\circ 2}$, and the gain in efficiency is rather small even in these cases. When $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{.1}} = 0.6$ and $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{.1}} = 0.3$, SVMs yield the lowest absolute bias and MSE. However, this may be the case because estimates resulting from SVMs are systematically lower under these quite specific conditions of positive bias for the pseudo-design-based estimates and, therefore, not necessarily generalizable. The only case of a systematic improvement is given by GLMMs. In case of a strong non-linear dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$ $\left(\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{.1}} = 0.6\right)$ or a strong and purely linear one $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{.1}} = 0.6 \text{ and } \boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{.1}} = \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{.1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}\right)$, these models generally result in the lowest absolute biases and MSEs. Although this improvement provides at least some justification for MRP, it is still only a minor one in comparison to the remaining inaccuracy.

The results are highly similar when using the original calibration variable $\boldsymbol{X}$ rather than (as in figure 6.19) classes of $\boldsymbol{Z}$. This holds when using classes of $\boldsymbol{X}$ for post-stratification as well as when only a total benchmark for the actual $\boldsymbol{X}$-variable is available. Therefore, it seems that the benefits of strategy ***c)*** are generally rather limited in the current simulation if only totals are available as auxiliary information. A possible explanation for the only minor advantages especially of MRP may be the fact that $\boldsymbol{Z}$ is divided into only ten categories to constitute the random effects and post-stratification cells. This quantity might be simply too low for the combination of mixed models and calibration to perform well, especially when considering the thousand or more possible cross-combinations that are used in real applications. In more realistic cases, the stabilizing effect obtained from prediction models may therefore help to achieve better estimates (cf. e.g. Gelman et al., 2016b, p. 90; Wang et al., 2015, p. 981).

When such cross-combinations are used, post-stratification basically includes not only totals but also covariances of indicator variables (cf. appendix B.5.2; Lenau and Münnich, 2017, pp. 62 f). Indeed, the simulation results are somewhat different when not solely applying total but also covariance calibration. An example is given in figure 6.20, considering the same setting as in figure 6.19. Weights for the non-probability sample are

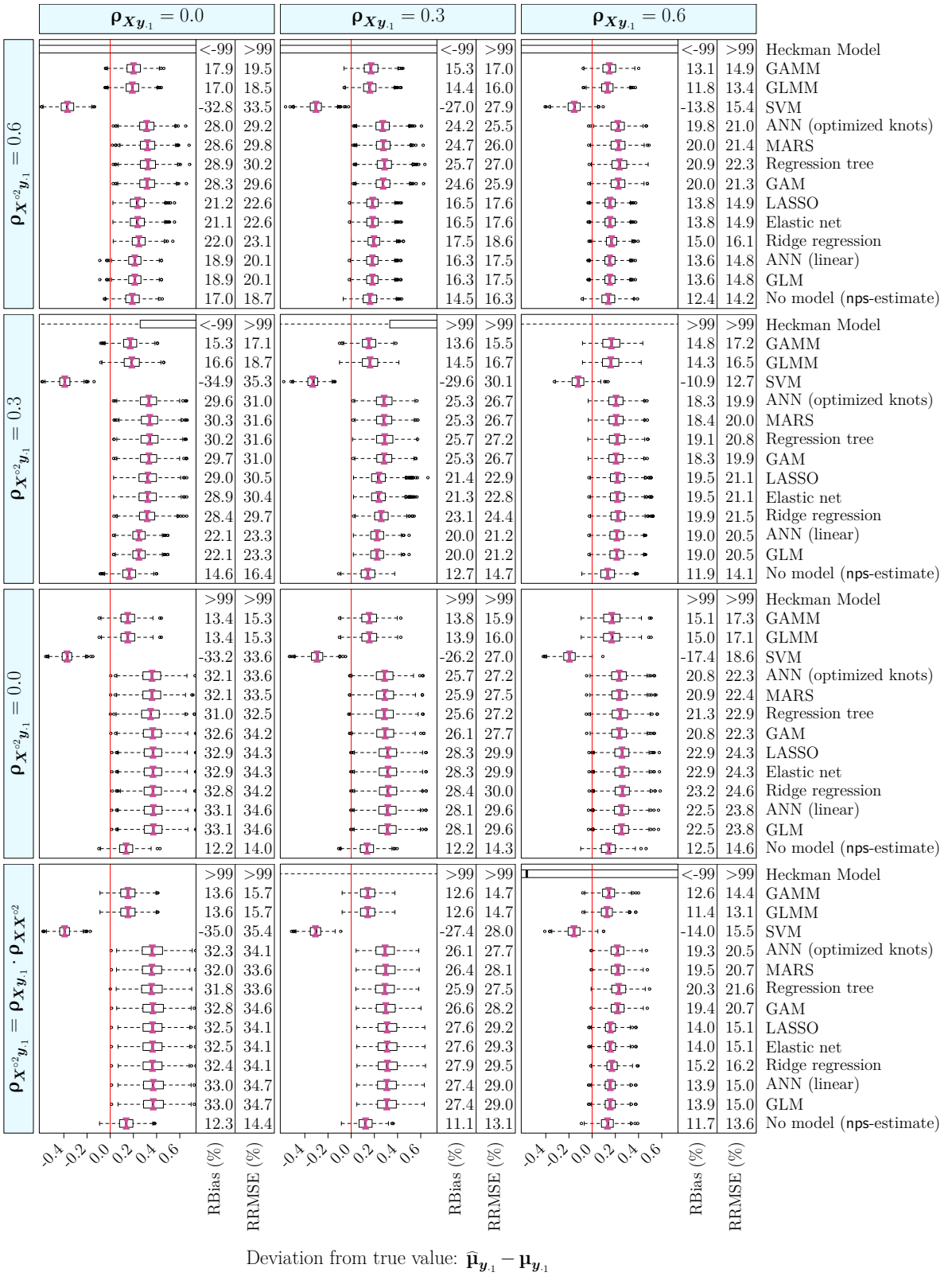


Figure 6.19: Comparison of prediction models for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$: estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ for 100% coverage – weighting model: post-stratification, using total constraints (estimation by weighted aggregation of predictions)

determined as a function of $\boldsymbol{X}$ and $\boldsymbol{Z}$ by using a calibrated neural network that has the structure of a parametric logit model (i.e. employs a single layer with softmax activation function). Since no reference sample is available, the weighting model is exclusively fit to meet the benchmarks for total and variance of $\boldsymbol{X}$, as in figure 6.16.

From the results in figure 6.20, it is evident that the performance of purely pseudo-design weighted estimates is worse than in figure 6.19 for most of the twelve scenario populations. The only exception occurs in presence of a strong non-linear and at least a medium linear dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ $\left(\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = 0.6 \text{ and } \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \geq 0.3\right)$, in which case calibrating total and variance of $\boldsymbol{X}$ leads to a lower bias and MSE than post-stratification with regard to classes of $\boldsymbol{Z}$. This is because population level auxiliary information about $\boldsymbol{Z}$ is valuable for modeling the selection process across all scenarios in figure 6.19. In contrast, $\boldsymbol{X}$ is not directly related to the sample selection mechanism. The potential to reduce the bias of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot 1}}$ by means of the corresponding calibration benchmarks in figure 6.20 is therefore determined by the use of $\boldsymbol{X}$ for predicting the target variable $\boldsymbol{y}_{\cdot 1}$. Considering the different prediction methods that are used in conjunction with weighted estimation in this latter figure, the bias for SVMs is again in opposite direction when compared to all other estimators. Similar as before, these models lead to the lowest absolute bias and MSE in case of a mainly linear dependency $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \geq 0.3 \text{ and } \boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} < \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}\right)$. Because the results for SVMs are worse than for any other method under consideration across all other scenarios, it seems that this is caused by a specific dependency of variables, coupled with a positive bias of the pseudo-design-based estimates. Although this pattern appears to be a bit more systematic than in figure 6.19, these conditions may be hard to identify in real applications. In all other cases where a dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ is present, ordinary and mixed linear models yield better results than any other method. Even in cases of no dependency between the two variables, the resulting estimates when using GLMMs are in the worst case only slightly less accurate than plain pseudo-design-based estimates. In comparison to MRP in figure 6.19, it furthermore becomes evident that there is no necessity for limiting calibration to the variables that explain the response process and constitute the random effects in the GLMMs. For these and other models, calibration for further auxiliary variables can yield even better results, in particular when these variables are strongly related to the variable of interest.

For the further conditions used in the simulation, coinciding results occur when estimating $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ using calibration of total and variance of $\boldsymbol{X}$, regardless of whether calibrated neural networks or the GREG is applied. While the potential of support vector machines to reduce the bias that remains after pseudo-design-based estimation is highly situational, weighted aggregation of GLMM predictions is typically at least as good as pseudo-design weighting applied to the observed target variables in the non-probability sample. Therefore, it appears that for stabilizing estimators, predictions from linear mixed models can indeed be a sensible replacement even for values of $\boldsymbol{y}_{\cdot 1}$ that are actually observed in the non-probability sample. Nevertheless, it should be kept in mind that this stabilization typically requires random effects that are strongly related to the selection mechanism in order to be as close to conditional independence as possible. In addition, only a relatively fine-grained weighting method actually suggests such stabilization (cf. figures 6.19 and 6.20; Gelman et al., 2016b, p. 90; Wang et al., 2015, p. 981). Furthermore, this does again not extend to bi- or multivariate statistics because the conditional mean is used as imputed value, which ignores the residual (co-)variance for estimation (cf. figure 6.12).

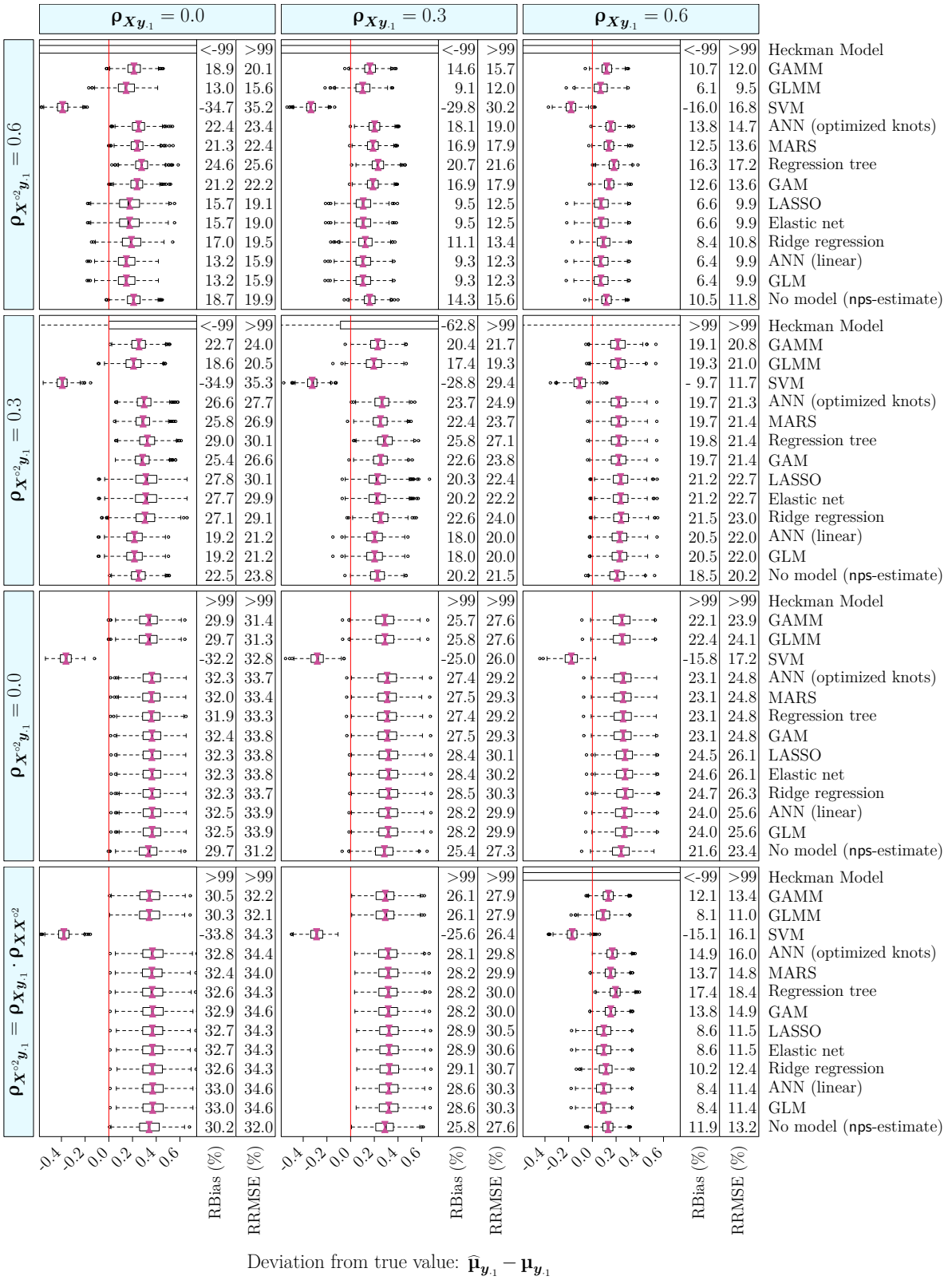


Figure 6.20: Comparison of prediction models for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$: estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ for 100% coverage – weighting model: calibrated ANN (parametric), using total and covariance constraints (estimation by weighted aggregation of predictions)

When considering joint availability of a reference sample and calibration constraints as auxiliary information, it is in principle possible to apply either one of the above strategies ***b)*** and ***c)***. However, the former (using weighted loss functions for fitting imputation models) performs better in nearly all simulation scenarios and is, therefore, presented in the following example. In this setting of available auxiliary information, the combination of parametric logit model and GREG performs well for pseudo-design-based estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ (cf. figure 6.17). The use of this weighting method in conjunction with weighted imputation models is therefore examined in figure 6.21. All other simulated conditions coincide with those discussed above.

In comparison to using plain propensity weighting in figure 6.18, the supplementary use of total and variance calibration improves the efficiency of pseudo-design weighted nps-estimates in most cases where the squared $\boldsymbol{X}$-variable is related to $\boldsymbol{y}_{\cdot 1}$ $\left(\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} > 0\right)$. This largely cancels out the gain in accuracy that can be achieved by additionally using prediction models in case of propensity weighting because calibration leads to a stronger utilization of the dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ already in the pseudo-design weights. When comparing the different prediction methods, the findings are quite similar as in figure 6.18. The only considered imputation approach that yields biases which are at least similar or even better than for the plain pseudo-design weighted non-probability sample estimates is based on additive mixed models. However, the potential improvements in terms of the bias are again rather small and counterbalanced by an increase in variance and also MSE.

The results in other scenarios where a response propensity model is used in conjunction with total benchmarks to estimate means or totals are quite similar. This holds regardless of whether covariance benchmarks are included or not. Based on the simulation results, it therefore seems sensible to rely on purely pseudo-design-based methods when jointly using a reference sample and total calibration. However, when propensities are calibrated to only meet the variance, bias correction through calibration does not make use of any linear dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$. In this case, weighted prediction models can typically reduce the bias in comparison to pseudo-design-based estimates of $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$.

These findings indicate a possible limitation for the generalizability of the results presented in figures 6.18 to 6.21. It is induced by using very similar auxiliary information for prediction and weighting methods since both make use of the auxiliary variable $\boldsymbol{X}$. Joint usage of model- and pseudo-design-based methods may hence perform different when the available information for both is more distinct. However, since both of these approaches ideally require auxiliary variables which are highly related to selection mechanism and target variables to achieve conditional independence, it appears rather plausible that similar auxiliary variables are used for weighting as well as prediction (cf. e.g. Baker et al., 2010, p. 47; Isaksson and Lee, 2005, p. 3148; Kreuter et al., 2010, p. 404; Steinmetz et al., 2014, p. 286). Adaptation of these considerations for the simulation is also required to limit the computational burden of the already large-scale study (cf. section 6.3.1).

In summary, the benefits of combining model- and pseudo-design-based methods are most evident when a reference sample is the only available auxiliary information. In that case, using pseudo-weights to fit weighted prediction models is particularly beneficial in case of MNAR selectivity, i.e. when conditional independence is not given. For this purpose, however, it must be possible to specify adequate response as well as prediction models based on the available auxiliary variables. When only calibration benchmarks are available

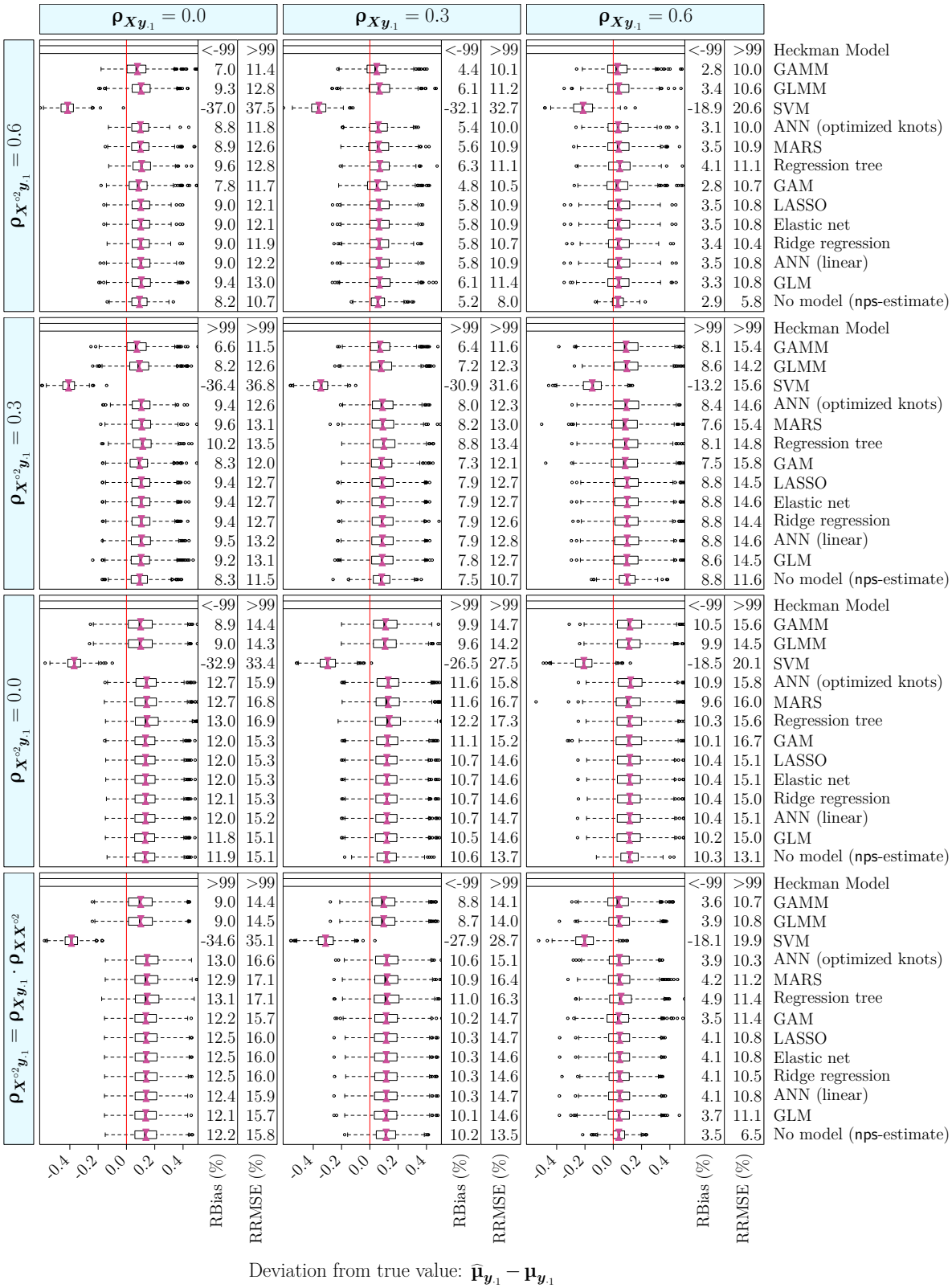


Figure 6.21: Comparison of prediction models for different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$: estimation of $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ for 100% coverage – weighting model: logit model (parametric) and GREG, using a reference sample, total and covariance constraints (estimation from imputed reference sample)

as auxiliary information, it depends on the calibration problem's complexity whether prediction models can be additionally used to improve pseudo-design weighted estimates. In case of calibration constraints that are manifold or hard to meet, weighted aggregation of predictions can help to stabilize the resulting estimates, e.g. when using MRP. In cases where calibration weights themselves are already relatively stable, it is typically better to apply them directly to the observed values rather than predictions for estimation. Furthermore, MRP and similar methods may often be inadequate to estimate bi- or multivariate statistics because the predictions typically represent the target variable's conditional mean rather than the actual conditional distribution.

The empirical findings discussed in the current section 6.3.2.2 demonstrate that full compensation of the selection bias in non-probability samples is possible when auxiliary variables facilitate conditional independence of sample selection and variable(s) of interest. In such MAR scenarios, the model- and the pseudo-design-based paradigm both provide a valid framework for unbiased estimation, underlining the theoretical considerations in chapter 5. However, in cases where selectivity is MNAR, none of the methods is able to completely adjust for selection bias. Unless all available auxiliary information is already fully exploitable by one of the methods alone, valuable improvements can be achieved by combining both paradigms in such cases. Even though the remaining bias in these settings is typically considerably different from zero, it often is still lower than when relying on weighting or prediction alone. This is an important feature when considering inferential methods for non-probability samples in the following section 6.3.2.3, because the approaches that are proposed for this purpose typically rely on variance estimation and assume that the (remaining) bias can be ignored (cf. section 5.4).

6.3.2.3 Methods for Inference

As discussed in section 5.4, typical approaches for inference from non-probability samples assume that the bias due to sample selection can be compensated by model- or pseudo-design-based point estimation methods. However, it is evident from the previous section 6.3.2.2 that this is highly dependent on the selectivity as well as available auxiliary information and, therefore, not necessarily true. Nonetheless, inferential methods are usually limited to the same set of assumptions and auxiliary information that is used for point estimation. In this case, any assessment of bias for the purpose of inference could be used to compensate for the bias already during point estimation. Therefore, the remaining bias can usually not be estimated from a single non-probability sample unless some information or assumption is deliberately dropped for point estimation but not for inference (cf. e.g. Pfeffermann, 2015, pp. 441 ff; Schouten, 2007). In the typical case where all available knowledge and assumptions are already used for point estimation, the only strategy that is proposed for quantifying an estimator's error is variance estimation. For this purpose, estimated within and between variance components given the prediction and/or weighting model are used either separately or jointly in different publications (cf. equality 5.188; Buelens, Burger and van den Brakel, 2018, p. 330; Yang and Kim, 2018, p. 5; Rafei, Flannagan and Elliott, 2020, p. 160). The following simulation results highlight possibilities and limitations of these inferential approaches, considering the naive as well as the resampling variance estimates described in sections 5.4 and 6.3.1. To provide an overview and summary of findings in the simulation, the discussion focuses on inference for selected point estimators which occurred in the previous discussion. As in section 6.3.2.2, model-based estimates are considered first, followed by pseudo-design-based estimation and approaches for combining both paradigms.

Inference for Model-based Methods

For the model-based estimates, the first example in table 6.3 represents CI-rates of 95% confidence intervals for selected approaches to estimate $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ presented in figure 6.11. As before, naive estimates from the non-probability sample are included as a reference point. The selected estimation approaches (naive nps-estimates, GLM and GLMM) are denoted in boxes beside the table rows. CI-rates are presented for varying (non-)linear dependencies $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$ and $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$ between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$, which are denoted in the first two columns of the table and correspond to the grid cells in the previous plots. In the current setting, correlations $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = \boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$ are all fixed and specified below the table. All other relations between the considered variables result in conditional independence and are therefore determined by products of the stated ones. RBias and RRMSE of the point estimates are represented in the third and fourth column, and correspond to the values in figure 6.11 in this case. The remaining columns five to eleven show the CI-rates that result from the different variance estimation strategies described in sections 5.4 and 6.3.1. The 'naive' approach treats the pseudo-design weighted non-probability or imputed reference sample as if it was a probability sample with known design weights and variables of interest and applies classical design-based variance estimates to these samples. For the Monte Carlo and rescaling bootstrap, variance components $\widehat{\mathbf{V}}_{\mathbf{b}}$, $\widehat{\mathbf{V}}_{\mathbf{w}}$ and their combination $\widehat{\mathbf{V}}_{\mathbf{t}}$ are considered, which are defined in equations 5.189 (cf. also section 2.2). All values except for the correlations are provided in percent.

As discussed with regard to the point estimates, plain nps-estimates do not use any compensation for selectivity and are biased in all scenarios due to the correlations of selection probabilities $\boldsymbol{\pi}^{\mathsf{nps}}$ and target variable $\boldsymbol{y}_{\cdot 1}$ with $\boldsymbol{Z}$. As $\boldsymbol{X}$ and $\boldsymbol{Z}$ are also correlated, this bias increases with higher correlations $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$. If such biases are not accounted for, it is evident from table 6.3 that inference can hardly perform well. For the naive variance estimates, CI-rates are not even close to the 95% nominal coverage in any of the cases. Higher relative bias strongly determines lower CI-rates, indicating that this is due to the bias component of the MSE which is ignored by variance estimates. The results are similar but even worse for the between variance estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ obtained from the Monte Carlo or rescaling bootstrap. In case of both resampling techniques, this estimated variance component leads to confidence intervals that cover the true value even less than the naive confidence intervals, and the rescaling bootstrap leads to 0% CI-rates in all cases. Underlining this finding, the estimated within variance $\widehat{\mathbf{V}}_{\mathbf{w}}$, calculated as the average of the naive variance estimates obtained from all resamples, seems to be more effective. For both types of resampling, this component outperforms the confidence intervals which are based on $\widehat{\mathbf{V}}_{\mathbf{b}}$ across all scenarios. Furthermore, its estimation when using the Monte Carlo bootstrap performs clearly better than the naive variance estimates in all scenarios. However, CI-rates are still considerably lower than 95% for all considered scenarios and worse in case of the rescaling bootstrap. Using the combination $\widehat{\mathbf{V}}_{\mathbf{t}}$ of both variance components discussed so far considerably improves the results. Especially the Monte Carlo bootstrap yields confidence intervals which achieve CI-rates of up to 93% in this case. Nevertheless, its CI-rates still decrease with higher biases and are therefore still far below the nominal coverage in almost all scenarios. The rescaling bootstrap does not improve visibly from adding $\widehat{\mathbf{V}}_{\mathbf{b}}$ to $\widehat{\mathbf{V}}_{\mathbf{w}}$ because its estimates of the between component are generally very low. In summary, estimating the total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$ by means of the Monte Carlo bootstrap seems the most favorable of the considered approaches to make inference for unweighted non-probability sample estimates.

Table 6.3: Confidence interval coverage rates for selected prediction models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\mathrm{V}\left(\widehat{\bar{\mu}}_{\boldsymbol{y}_{\cdot 1}}\right)$ for 100% coverage – weighting model: unweighted (estimation from imputed reference sample)

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\rho_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\rho_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
No model (nps-estimate)	0.0	$\rho_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \rho_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	17.0	18.4	33	23	58	79	0	42	42
		0.0	17.9	19.3	30	16	45	76	0	32	32
		0.3	15.9	17.5	42	21	59	86	0	40	40
		0.6	13.4	15.3	59	32	78	93	0	65	65
	0.3	$\rho_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \rho_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	20.2	21.3	18	12	47	59	0	22	22
		0.0	18.0	19.5	30	21	53	70	0	37	37
		0.3	18.5	20.1	29	10	52	70	0	26	26
		0.6	17.1	18.7	38	15	62	83	0	30	30
	0.6	$\rho_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \rho_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	20.8	22.2	20	13	42	61	0	28	28
		0.0	18.2	19.7	29	16	59	76	0	32	32
		0.3	19.4	20.9	23	13	42	58	0	25	25
		0.6	20.4	21.8	22	10	43	62	0	23	23
GLM	0.0	$\rho_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \rho_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	16.8	18.4	1	26	0	28	0	0	0
		0.0	18.0	19.5	0	16	3	19	0	2	2
		0.3	13.0	15.0	8	33	11	39	0	9	9
		0.6	8.0	10.8	28	60	31	71	0	31	31
	0.3	$\rho_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \rho_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	14.6	16.2	11	34	13	46	0	13	13
		0.0	13.8	15.6	8	35	13	44	0	13	13
		0.3	11.8	14.3	21	34	17	47	0	17	17
		0.6	7.4	10.9	52	57	46	72	0	45	45
	0.6	$\rho_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \rho_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	6.9	11.1	77	52	79	90	0	79	79
		0.0	10.2	12.5	39	43	36	76	0	36	36
		0.3	8.1	11.3	69	39	76	88	0	74	74
		0.6	6.1	10.6	82	46	81	93	0	82	82
GLMM	0.0	$\rho_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \rho_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	0.5	9.1	87	53	82	89	0	83	83
		0.0	0.9	9.4	90	53	88	90	0	87	87
		0.3	0.9	8.4	84	63	82	91	0	80	80
		0.6	1.0	7.5	78	79	74	91	0	75	75
	0.3	$\rho_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \rho_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	0.7	8.5	89	47	86	88	0	86	86
		0.0	0.8	8.6	87	61	85	90	0	85	85
		0.3	0.9	8.7	86	62	86	94	0	86	86
		0.6	1.0	8.3	84	71	85	95	0	85	85
	0.6	$\rho_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \rho_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	1.1	9.1	90	49	88	92	0	88	88
		0.0	1.1	8.4	84	60	78	91	0	79	79
		0.3	1.0	8.6	89	55	84	91	0	85	85
		0.6	0.7	9.0	91	53	88	91	0	88	88

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables
$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate
Prediction model variables: $\boldsymbol{X}$ Random effects variables: $\boldsymbol{Z}$-classes
Correlations determining the simulation scenario (all others result as products of the stated ones):
$\rho_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\rho_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\rho_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$
All numbers except for the correlations are in (rounded) percentage points.

The advantages of this inferential approach are similarly evident for model-based estimates. Considering mass imputation from a GLM, confidence interval coverage is typically lower than for nps-estimates in cases where the magnitude of bias is similar. However, the point estimates obtained from GLM predictions allow for a considerable reduction of bias and MSE when $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ are related, and CI-rates are again strongly influenced by the remaining bias. The dependency between auxiliary and target variable thus determines whether inference works better for estimates obtained from the unweighted non-probability or the imputed reference sample. In particular when a strong linear dependency between the two variables $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6\right)$ allows a large amount of bias reduction, confidence intervals for the GLM-based estimates exhibit better coverage than those for the unweighted non-probability sample. The resulting CI-rates for Monte Carlo bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{t}}$ range from 76 to 93%. In this case, point estimation as well as inference can be improved by using a linear prediction model, although selectivity is still MNAR for these models because they do not incorporate $\boldsymbol{Z}$ as a predictor. For the other (linear as well as non-linear) models that only use $\boldsymbol{X}$ as independent variable, highly similar results are obtained.

Under MAR selectivity, point estimation and confidence interval coverage can be improved further. In the present context, this is the case for the mixed models because the random effects variable $\boldsymbol{Z}$ guarantees conditional independence of $\boldsymbol{y}_{\cdot 1}$ and $\boldsymbol{r}^{\mathsf{nps}}$. The non-probability selection bias is, thus, almost fully compensated, and MSEs are considerably lower than for other estimation methods. The results for the GLMM in table 6.3 indicate that naive variance estimation performs relatively well in this case and leads to CI-rates between 78 and 91% in all scenarios. Similar as for the plain non-probability sample estimates, solely estimating the variance components $\widehat{\mathbf{V}}_{\mathbf{w}}$ or $\widehat{\mathbf{V}}_{\mathbf{b}}$ by means of resampling does not provide a general improvement over using naive variance estimates. But when applying the Monte Carlo bootstrap, joint usage of these components in form of $\widehat{\mathbf{V}}_{\mathbf{t}}$ leads to the best CI-rate in almost all scenarios (except when $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.3$ and $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$, where it is 0.1% worse than the naive estimate). CI-rates between 88 and 95% can be achieved with this approach. Although these results are still not perfect in view of the nominal 95% coverage that is reached in only one scenario, they nevertheless seem rather convincing and reasonable. The simplifying assumption that non-probability sample selection can be approximated by simple random sampling for bootstrapping clearly introduces some imperfections to inference. Note that the estimator $\widehat{\mathbf{V}}_{\mathbf{t}}$ relies on some important concepts of multiple imputation and seems clearly favorable for inference in the present context (cf. section 5.4). This provides a further indication to apply multiple imputation in the context of model-based estimation for non-probability samples, underlining the results in figure 6.12. For additive mixed models, which allow a similar degree of bias reduction as GLMMs, the resulting CI-rates are almost the same.

When comparing the different variance estimation methods, the Monte Carlo bootstrap seems preferable over the rescaling bootstrap in all cases. As the point estimator's bias is not accounted for in variance estimation and leads to CI-rates that are typically too low, this is mainly because the latter results in tremendously lower estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$. The finite population correction, which is included in the rescaling but not in the Monte Carlo bootstrap, may be a partial explanation for this difference. However, the large discrepancies between both resampling techniques suggest that the above-mentioned simplifying assumptions pose stronger drawbacks for the rescaling than for the Monte Carlo bootstrap. Although better performance of the rescaling bootstrap can be found for pseudo-design-

Table 6.4: Confidence interval coverage rates for selected prediction models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\mathrm{V}\left(\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{y}_{\cdot 2}}\right)$ for 100% coverage – weighting model: unweighted (estimation from imputed reference sample)

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
No model (nps-estimate)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	7.4	11.6	66	44	85	94	0	65	65
		0.0	6.2	12.8	64	55	85	93	0	73	73
		0.3	5.7	11.9	68	59	84	93	0	69	69
		0.6	5.2	12.0	70	56	78	93	0	65	65
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	8.6	12.2	62	48	82	90	0	68	68
		0.0	6.9	12.1	65	51	82	90	0	67	67
		0.3	6.3	13.8	60	45	74	84	0	63	63
		0.6	5.8	12.9	64	51	76	91	0	61	61
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	7.5	13.4	57	39	72	88	0	50	50
		0.0	7.0	12.6	62	52	81	92	0	64	64
		0.3	7.7	13.0	61	46	71	86	0	63	63
		0.6	6.6	13.2	61	48	77	92	0	61	61
GLMM	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	- 17.7	21.8	69	4	71	72	0	71	71
		0.0	- 15.5	19.8	76	7	79	80	0	79	79
		0.3	- 35.3	39.5	24	3	26	33	0	26	26
		0.6	- 52.6	55.7	4	0	3	5	0	3	3
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	- 17.4	21.2	71	7	73	75	0	73	73
		0.0	- 25.6	28.5	45	2	46	49	0	46	46
		0.3	- 30.3	34.1	36	2	46	47	0	45	45
		0.6	- 41.1	44.1	14	2	16	20	0	16	16
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	- 22.7	26.0	55	0	51	54	0	51	51
		0.0	- 28.6	31.4	37	2	39	40	0	38	38
		0.3	- 18.9	22.8	67	10	72	74	0	72	72
		0.6	- 22.9	26.2	54	6	50	52	0	50	50

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables
$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate
Prediction model variables: $\boldsymbol{X}$ Random effects variables: $\boldsymbol{Z}$-classes
Correlations determining the simulation scenario (all others result as products of the stated ones):
$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$
All numbers except for the correlations are in (rounded) percentage points.

based estimates (cf. e.g. table 6.5), the Monte Carlo bootstrap seems generally more appropriate for model-based estimation. Especially when it is used for estimating the total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$, results are generally at least as good as those obtained from any other inferential technique under consideration. Despite relying on simplifying assumptions for variance estimation, nominal CI-rates are nearly reached when the bias of point estimators is close to zero. These results can be as well confirmed for the other considered prediction models. As for the ones shown in table 6.3, confidence interval coverage typically depends on the degree of bias that is not compensated by the model. Mainly by reducing the bias in point estimation, adequate prediction models can therefore lead to better inference than using unweighted estimates from non-probability samples.

As presented in figure 6.12, estimating the correlation $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{y}_{\cdot 2}}$ through mass-imputation is rather inadequate when using predictions as imputed values. Because such predictions typically represent the target variable's conditional means rather than its actual conditional distribution, residual (co-)variances are ignored. The bias for non-linear estimates that are based on such predictions is therefore usually even higher than for unweighted non-probability sample estimates and mostly prevents valid inference. This is illustrated in table 6.4, where confidence intervals for some of the point estimation methods for $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{y}_{\cdot 2}}$ presented in figure 6.12 are evaluated.

In case of the unweighted non-probability sample estimates that do not use any compensation for selectivity, the relative point estimation biases range from 5.2 to 8.6%. For these estimates, all variance estimation approaches result in CI-rates below 95% in every scenario. Nevertheless, CI-rates between 84 and 94% can be achieved by using Monte Carlo bootstrap estimates of the total variance $\left(\widehat{\mathbf{V}}_{\mathbf{t}}\right)$, which lead to the most conservative intervals. This is clearly not perfect for a nominal CI-rate of 95%, but still way better than what can be achieved for any of the model-based estimation methods.

In figure 6.12, mixed models perform better than any other prediction model under consideration but still exhibit a larger magnitude of bias than the unweighted nps-estimates considered above. As a consequence, inference based on variance estimation is much less reliable for these model-based estimates. Considering GLMMs, the most conservative approach to inference occurs again for the Monte Carlo bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{t}}$. But even in this case, the resulting CI-rates that range from 5 to 80% are much too low to be considered adequate. Results when using GAMMs for prediction are very similar as in case of GLMMs, while all other considered prediction models yield even higher biases and lower CI-rates than these two types of mixed models.

When considering model-based point estimation by means of mass-imputation, Monte Carlo bootstrap estimates for the total variance are typically the best option for making inference. As is evident from tables 6.3 and 6.4, a comparatively lower bias in point estimates generally helps to come closer to the nominal CI-rates because all considered inferential approaches are based on variance rather than MSE estimation. Nevertheless, perfectly unbiased point estimates are neither a necessary nor sufficient condition for adequate inference since variance estimation in this context is a simplifying approximation rather than an unbiased estimate for the repeated sampling variance (cf. sections 5.4 and 6.3.1). These findings are also underlined in the following tables 6.5 to 6.13. Furthermore, point estimation and inference for measures of dependency are generally rather inaccurate when using model-based methods in the current manner. As discussed with regard to figure 6.12, the reason is that imputation of model predictions does not account for variability that is not explained by the model. To estimate statistics that incorporate not only means but also covariances or other higher moments in the context of non-probability samples, it therefore seems essential to apply more refined forms of (multiple) imputation (cf. e.g. van Buuren, 2018, pp. 63 ff; Little and Rubin, 2019, p. 72; Rubin, 1987, p. 159). Unless this is the case, it seems generally more reasonable to rely on pseudo-design-based approaches for this purpose. Inference for such pseudo-design-based methods is considered in the following paragraphs.

Inference for Pseudo-design-based Methods

Considering the same scenarios as in table 6.3, confidence interval coverage rates for selected propensity weighted estimates of $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ are presented in table 6.5, which follows the same structure as the previously discussed ones. These results correspond to a subset of point estimates shown in figure 6.13, which use a reference sample for the response model as sole auxiliary information. Note that the unweighted non-probability sample estimates coincide with the previous table 6.3 and are therefore not repeated.

While the Monte Carlo bootstrap estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$ appears generally preferable for the model-based approaches shown in table 6.3, there is no unique best choice for inference in case of pseudo-design-based approaches. From the weighting methods represented in figure 6.13, pseudo-weights obtained from a non-parametric logit model perform particularly well in all scenarios and almost completely eliminate selection bias for point estimation. Considering inference for these estimates, naive variance estimates are not too bad but generally fall short in achieving the nominal 95% CI-rate, even though a slight increase in CI-rates for stronger non-linear dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ occurs. Estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ for the between variance obtained from the rescaling bootstrap again result in confidence intervals that are generally too narrow to cover the true value. Monte Carlo bootstrap estimates of this component perform better and are typically similar or slightly superior in comparison to naive variance estimates. Even higher coverage can be achieved when using within variance estimates $\widehat{\mathbf{V}}_{\mathbf{w}}$ for both resampling methods. However, while coverage in case of the rescaling bootstrap is still generally below 95%, the Monte Carlo bootstrap results in confidence intervals that are too wide in most cases and lead to CI-rates of up to 99%. Even though wider intervals usually seem preferable over too narrow ones because they lead to more conservative inference, the ideal would be to exactly meet the nominal coverage (cf. Särndal, Swensson and Wretman, 1992, p. 83; Wolter, 2007, p. 26). The same argument applies for the total estimated variance $\widehat{\mathbf{V}}_{\mathbf{t}}$, for which coverage in case of the Monte Carlo bootstrap is way too high. For the rescaling bootstrap, this combination of $\widehat{\mathbf{V}}_{\mathbf{b}}$ and $\widehat{\mathbf{V}}_{\mathbf{w}}$ behaves exactly as the latter, simply because $\widehat{\mathbf{V}}_{\mathbf{b}}$ is negligibly small. Therefore, none of the considered methods performs perfect for these pseudo-weighted point estimates. The most appropriate choice seems to be based on estimates $\widehat{\mathbf{V}}_{\mathbf{w}}$ using the Monte Carlo bootstrap, even though confidence intervals are somewhat too conservative in that case.

In contrast and despite being slightly more biased, propensity weighted point estimates when using a non-parametric neural network as response model (with fixed B-spline knots as for the pseudo-weights) result in higher CI-rates for each variance estimate and across all scenarios. While naive variance estimates still fall short in reaching the nominal CI-rate of 95%, Monte Carlo bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ are mostly close to or a bit higher than 95%. For the rescaling bootstrap, this estimated component again leads to intervals which are insufficiently narrow, even though coverage is still tremendously higher than for the pseudo-weights. The estimated within variance $\widehat{\mathbf{V}}_{\mathbf{w}}$ is again clearly larger than the between component in all cases. This leads to confidence intervals that are too wide in case of both resampling methods, which is particularly severe for the Monte Carlo bootstrap. The resulting excess length of confidence intervals is carried over and magnified for the total estimated variance $\widehat{\mathbf{V}}_{\mathbf{t}}$ because the latter includes the within component. As a consequence, the Monte Carlo bootstrap estimate of $\widehat{\mathbf{V}}_{\mathbf{b}}$ seems the most adequate method to obtain confidence intervals for point estimates that use plain propensity weights

Table 6.5: Confidence interval coverage rates for selected weighting models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\mathrm{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}\right)$ for 100% coverage, using a reference sample

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
Pseudo-Weights (fixed knots)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	0.0	9.2	84	84	96	98	0	76	76
		0.0	0.4	9.3	83	84	96	99	0	85	85
		0.3	0.2	8.7	87	91	98	99	0	87	87
		0.6	- 0.3	8.2	91	90	99	100	0	88	88
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	- 0.1	8.7	85	86	97	99	0	81	81
		0.0	0.0	8.7	87	88	98	99	0	84	84
		0.3	0.2	8.9	87	90	96	100	0	83	83
		0.6	- 0.3	8.4	89	93	98	99	0	90	90
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	- 0.1	8.9	86	84	97	98	0	81	81
		0.0	- 0.1	8.7	86	89	97	100	0	84	84
		0.3	- 0.4	8.7	85	83	92	98	0	80	80
		0.6	- 0.5	8.8	86	85	96	98	0	82	82
calibrated ANN (fixed knots)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	3.0	9.0	88	94	99	100	74	98	98
		0.0	3.6	9.2	88	95	100	100	84	98	99
		0.3	3.1	8.6	90	98	99	100	92	98	99
		0.6	2.4	8.3	94	98	100	100	92	99	100
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	3.4	8.7	90	94	99	100	76	97	98
		0.0	3.0	8.7	90	98	100	100	82	99	100
		0.3	3.4	8.9	91	95	99	100	80	97	99
		0.6	3.0	9.3	93	95	100	100	94	99	99
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	3.3	11.1	92	94	99	100	85	98	98
		0.0	2.4	8.6	89	97	100	100	78	99	100
		0.3	2.2	8.3	88	90	100	100	80	94	99
		0.6	2.9	9.3	91	95	100	100	83	97	100

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables
$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate
Response model variables: $\boldsymbol{X}$,$\boldsymbol{Z}$
Correlations determining the simulation scenario (all others result as products of the stated ones):
$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$
All numbers except for the correlations are in (rounded) percentage points.

obtained from a non-parametric artificial neural network. A more conservative alternative would be to estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$ from the rescaling bootstrap. The results for the non-parametric logit model are highly similar to those for the ANN, which is why it is not considered separately.

For propensity weighted estimates of the mean $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$, estimating a single variance component rather than the total variance performs better in the current setting. However, there appears to be some interplay between the applied propensity model and the best method to choose for inference. Although Monte Carlo bootstrap variance estimates seem preferable for both types of propensity weights considered in table 6.5, it is the estimated within variance $\widehat{\mathbf{V}}_{\mathbf{w}}$ that performs best in case of pseudo-weights, while $\widehat{\mathbf{V}}_{\mathbf{b}}$ appears to

be better for the non-parametric ANN as propensity model. Unlike for the model-based estimates (cf. table 6.3), an approach for variance estimation that is preferable regardless of the propensity model is therefore not identifiable.

In cases of MAR selectivity patterns as in table 6.5, propensity weighting can largely compensate selection bias. However, it is typically unknown in real applications whether sample selection actually follows such a pattern. Since the bias is hardly fully compensable in cases where selectivity is MNAR, inference needs to be evaluated for such cases as well, and in particular when estimating the variance rather than the MSE, as is the case in the present context (cf. chapter 5; Elliott and Valliant, 2017, p. 262; Mercer et al., 2017, p. 257). MNAR selectivity is therefore considered in table 6.6, representing CI-rates for selected point estimates depicted in figure 6.14. Apart from applying a different selection mechanism through $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$, the setting is the same as in the previous table 6.5.

In comparison to the results in table 6.5, the unweighted point estimates are severely more biased in the context of table 6.6 due to selectivity that is MNAR. Since this bias is completely ignored when inference is solely based on variance estimation, nominal CI-rates are achieved with none of the considered inferential approaches. As before, estimates $\widehat{\mathbf{V}}_{\mathbf{t}}$ obtained from the Monte Carlo bootstrap as the most conservative inferential approach seem preferable to construct confidence intervals for unweighted estimates. Nevertheless, the resulting CI-rates between 56 and 64% are still far from the nominal 95% in all cases, a finding that is even more severe for all other inferential approaches under consideration.

Since bias and MSE for propensity weighted estimates are likewise higher when selectivity is MNAR rather than MAR, inference becomes correspondingly more difficult. Nevertheless, the CI-rates do not necessarily decline. As before, point estimates which derive pseudo-weights from a non-parametric GLM perform comparably well in figure 6.14. Considering inference for these estimates, CI-rates in case of naive variance estimation are again below 95% for all scenarios. Between variance estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ obtained by the Monte Carlo bootstrap perform better than in the previous table 6.5 and yield CI-rates between 89 and 97%. Estimating the same variance component with the rescaling bootstrap again leads to 0% coverage in all cases. Higher CI-rates are achieved when using the estimated within variance $\widehat{\mathbf{V}}_{\mathbf{w}}$. For the Monte Carlo bootstrap, CI-rates ranging from 98 to 100% indicate that confidence intervals are too wide, while the opposite is true for the rescaling bootstrap variant, where CI-rates between 80 and 92% are achieved. Similarly, the estimated total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$ overshoots nominal coverage when using the Monte Carlo bootstrap and falls short in achieving it in case of the rescaling bootstrap. Therefore, the Monte Carlo bootstrap seems more reasonable for inference in case of pseudo-weights. However, the resulting estimates $\widehat{\mathbf{V}}_{\mathbf{w}}$ of the within variance are clearly too conservative for inference, while confidence intervals based on the estimated between component $\widehat{\mathbf{V}}_{\mathbf{b}}$ are too narrow in most cases. Therefore, both components have their flaws and none of them is clearly preferable for all cases.

Despite higher biases and MSEs in point estimates, inference seems again slightly easier when applying a non-parametric ANN with fixed B-spline knots as response model. Again, the confidence intervals which are based on naive variance estimation are too narrow to achieve 95% CI-rates. While the same holds for the estimated between variance $\widehat{\mathbf{V}}_{\mathbf{b}}$ obtained from the rescaling bootstrap, the Monte Carlo bootstrap estimate of this variance component leads to CI-rates between 92 and 99%. Since within variance estimates $\widehat{\mathbf{V}}_{\mathbf{w}}$ are again on average larger than $\widehat{\mathbf{V}}_{\mathbf{b}}$, the resulting CI-rates range from 95 to 100% in case

Table 6.6: Confidence interval coverage rates for selected weighting models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\mathrm{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}\right)$ for 100% coverage, using a reference sample

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
unweighted estimates	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	32.2	33.8	10	3	37	63	0	12	12
		0.0	32.5	33.9	8	4	27	58	0	9	9
		0.3	32.9	34.3	8	4	33	59	0	10	10
		0.6	32.7	34.2	11	4	34	64	0	14	14
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	32.6	34.1	10	2	42	63	0	9	9
		0.0	32.5	34.1	8	6	35	56	0	14	14
		0.3	32.6	34.1	8	1	28	61	0	5	5
		0.6	33.0	34.5	8	0	32	61	0	9	9
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	32.4	33.7	9	4	44	63	0	16	16
		0.0	32.2	33.8	10	3	33	60	0	15	15
		0.3	32.2	33.9	10	2	34	57	0	10	10
		0.6	32.9	34.4	9	5	33	58	0	12	12
Pseudo-Weights (fixed knots)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	9.3	14.9	84	93	100	100	0	86	86
		0.0	9.2	14.8	84	91	98	100	0	87	87
		0.3	11.0	15.9	84	90	98	100	0	84	84
		0.6	11.3	15.4	85	95	100	100	0	89	89
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	8.7	14.3	87	90	100	100	0	80	80
		0.0	10.0	15.0	83	89	100	100	0	83	83
		0.3	10.3	15.6	84	93	99	100	0	84	84
		0.6	9.0	13.9	90	97	99	100	0	92	92
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	7.6	13.6	90	91	100	100	0	84	84
		0.0	9.6	15.3	85	90	100	100	0	84	84
		0.3	7.3	14.4	87	90	98	100	0	84	84
		0.6	7.4	13.3	90	96	100	100	0	92	92
calibrated ANN (fixed knots)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	13.8	17.5	77	92	99	100	72	96	99
		0.0	13.8	17.4	77	93	100	100	70	96	99
		0.3	15.6	18.6	75	96	99	100	72	97	99
		0.6	15.8	20.5	78	97	100	100	78	97	99
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	13.0	16.7	82	96	100	100	73	98	99
		0.0	14.2	17.6	77	92	100	100	70	96	100
		0.3	14.6	18.0	78	95	99	100	66	95	100
		0.6	13.6	19.4	85	99	100	100	88	99	100
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	11.5	15.2	88	97	100	100	83	97	99
		0.0	12.8	16.6	79	93	100	100	70	97	100
		0.3	10.2	14.9	84	97	100	100	76	97	100
		0.6	11.7	17.4	89	99	100	100	86	100	100

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables

$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate

Response model variables: $\boldsymbol{X}$,$\boldsymbol{Z}$

Correlations determining the simulation scenario (all others result as products of the stated ones):
$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$

All numbers except for the correlations are in (rounded) percentage points.

of the rescaling and from 99 to 100% in case of the Monte Carlo bootstrap, such that especially the latter is too conservative when using this component. Consequently, and as in table 6.5, the estimated total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$ likewise generally overshoots the nominal CI-rate for both resampling methods. In case of propensity weights that are obtained from such a non-parametric ANN, Monte Carlo bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ of the between variance therefore again seem to be the most appropriate approach for constructing confidence intervals that achieve CI-rates of roughly 95%. As a more conservative alternative, the rescaling bootstrap estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$ could be used in this case as well.

The most adequate method for inference in the present context again depends on the non-probability sampling mechanism, applied weighting model and available auxiliary information. Monte Carlo bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ in general work well when using a non-parametric ANN as propensity model. Because this finding coincides with the MAR case (cf. table 6.5), it indicates a clearly favorable approach for inference when using such a propensity model. This can be considered an advantage since selection mechanism and population structure, which define the scenarios in the simulation, are typically not perfectly known in reality. As described above, the picture is less clear when using pseudo-weights, for which Monte Carlo bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ or $\widehat{\mathbf{V}}_{\mathbf{w}}$ may be used. Nevertheless, both of these options have their flaws, such that inference is more straightforward when using response propensities which are predicted from an ANN.

As is evident from figure 6.15, performance and ranking of propensity weighting approaches are different when estimating $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{y}_{\cdot 2}}$ of other multi- or bi- rather than univariate statistics in cases of selectivity patterns that are MNAR. To evaluate methods for inference in this setting, CI-rates for a subset of point estimates considered in that figure are presented in table 6.7.

In case of unweighted estimates, the most conservative approach to construct confidence intervals seems preferable, just as in the previous tables 6.5 and 6.6. It is based on estimates $\widehat{\mathbf{V}}_{\mathbf{t}}$ obtained from the Monte Carlo bootstrap and achieves CI-rates between 94 and 98%. These seem quite good, considering that selectivity is MNAR and that simplifying assumptions are made for point as well as variance estimation (cf. section 5.4). All other considered methods for variance estimation lead to CI-rates that are way below the nominal 95% in this case.

Although pseudo-weights that rely on a non-parametric GLM decrease point estimation bias and MSE in at least most of the simulated scenarios, the relative benefit is much smaller than for estimates of the mean presented in table 6.6. Nevertheless, CI-rates are generally higher for pseudo- than for unweighted estimates. Naive and rescaling bootstrap variance estimates still fall short of reaching the nominal CI-rate, but the Monte Carlo bootstrap seems to be more adequate for this purpose. When using between variance estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$, the resulting CI-rates range from 91 to 98%. The estimated within variance component $\widehat{\mathbf{V}}_{\mathbf{w}}$ only yields 36 to 57%, while the combination of both in form of $\widehat{\mathbf{V}}_{\mathbf{t}}$ is typically too large and results in CI-rates of 99 to 100%. Unless such highly conservative confidence intervals are required, Monte Carlo bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ seem to be the most reasonable approach for this setting.

The best propensity weighted estimates for the correlation in figure 6.15 are based on non-parametric ANNs which apply the proposed knot optimization technique. The biases and MSEs in point estimation are considerably lower than when using pseudo- or no weights in almost all scenarios. The resulting CI-rates are often but not generally higher

Table 6.7: Confidence interval coverage rates for selected weighting models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of V $\left(\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{y}_{\cdot 2}}\right)$ for 100% coverage, using a reference sample

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
unweighted estimates	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	-13.9	22.8	32	73	37	96	0	29	29
		0.0	-12.9	22.1	34	84	42	96	0	36	36
		0.3	-13.7	22.2	34	80	39	97	0	36	36
		0.6	-14.3	21.9	36	84	38	98	0	33	33
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	-14.7	23.0	31	77	29	94	0	23	23
		0.0	-13.1	22.3	32	80	36	97	0	32	32
		0.3	-13.2	22.4	32	80	34	95	0	30	30
		0.6	-13.5	21.8	34	79	34	94	0	29	29
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	-12.7	22.5	31	75	36	97	0	28	28
		0.0	-12.8	22.2	35	77	44	96	0	40	40
		0.3	-13.0	22.2	34	75	41	95	0	36	36
		0.6	-13.6	22.3	34	76	36	97	0	32	32
Pseudo-Weights (fixed knots)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	-15.2	22.4	38	95	45	100	0	41	41
		0.0	-14.7	22.0	40	92	44	100	0	41	41
		0.3	-13.7	20.9	42	97	48	100	0	40	40
		0.6	-13.9	21.5	44	98	48	99	0	44	44
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	-16.8	23.5	36	91	36	99	0	31	31
		0.0	-14.6	21.9	39	94	42	99	0	39	39
		0.3	-14.1	21.4	42	93	44	99	0	39	39
		0.6	-14.5	22.3	42	91	43	100	0	37	37
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	-14.8	23.4	39	93	43	99	0	38	38
		0.0	-12.3	20.3	49	93	57	99	0	52	52
		0.3	-12.5	20.5	48	92	53	100	0	45	45
		0.6	-16.1	23.6	40	92	43	99	0	39	39
calibrated ANN (optimized knots)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	- 3.1	13.3	63	92	73	98	83	56	95
		0.0	- 2.9	13.1	64	92	86	100	83	69	94
		0.3	- 4.4	13.4	60	94	77	100	87	53	95
		0.6	- 7.8	14.9	54	100	73	100	95	57	99
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	-10.1	18.3	38	92	41	99	83	27	91
		0.0	-14.1	24.2	29	95	35	100	83	30	95
		0.3	- 3.2	12.1	70	91	85	98	86	72	95
		0.6	- 4.1	12.5	66	96	81	100	89	65	97
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	- 3.9	12.1	68	96	80	98	90	68	96
		0.0	- 6.2	15.1	52	96	63	100	90	49	98
		0.3	- 8.9	17.5	41	94	58	99	86	42	97
		0.6	- 3.3	11.5	73	96	82	100	90	66	95

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables

$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate

Response model variables: $\boldsymbol{X}$,$\boldsymbol{Z}$

Correlations determining the simulation scenario (all others result as products of the stated ones):
$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\text{nps}}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\text{nps}}} = 0.6$

All numbers except for the correlations are in (rounded) percentage points.

than in case of the pseudo-weights summarized above. Again, naive variance estimates are too small to achieve CI-rates of 95% or more. The same holds for estimates $\widehat{\mathbf{V}}_{\mathbf{w}}$ of the within variance for both considered resampling methods and for $\widehat{\mathbf{V}}_{\mathbf{b}}$ obtained from the rescaling bootstrap. When the between variance is estimated by means of the Monte Carlo bootstrap, CI-rates between 91 and 100% are more convincing, while estimating the total variance through Monte Carlo bootstrapping again exceeds the nominal CI-rate in all scenarios. The most adequate results are achieved when using the rescaling bootstrap estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$ of the total variance, where CI-rates range from 91 to 99% and are closer to 95% than for Monte Carlo bootstrap estimates of the between variance in most cases.

Therefore, the best method for inference again depends on the propensity model applied for point estimation of correlation coefficients. The results for the considered pseudo-weights are best when relying on Monte Carlo bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$. When using propensity weights obtained from a non-parametric neural network with knot optimization, it is better to rely on the rescaling bootstrap and the estimated total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$.

Considering cases where calibration benchmarks are used as the only available auxiliary information, inference for a subset of pseudo-design weighted estimators of the mean $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ depicted in figure 6.16 is evaluated in table 6.8. The unweighted point estimates for this setting coincide with those in the previous table 6.6 and are therefore not shown again. In comparison to these unweighted estimates, a considerable amount of selection bias can be compensated through calibration weighting if there is any relation between calibration and target variables $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$, especially if this dependency is linear. However, since selectivity is again MNAR, some bias remains even after applying calibration weights.

From the calibration methods that use one parameter per observation in the non-probability sample, the GREG performs best for estimating $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ from the non-probability sample (cf. figure 6.16). When evaluating inference for these estimates in table 6.8, there is no unique approach to obtain confidence intervals that works best in all scenarios. Intervals based on naive variance estimates as well as $\widehat{\mathbf{V}}_{\mathbf{b}}$ obtained from both resampling methods are clearly too short in all cases. CI-rates are consequently below 95%, although coverage increases with higher correlation $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$ due to lower remaining biases. Performance of inference using either $\widehat{\mathbf{V}}_{\mathbf{w}}$ or $\widehat{\mathbf{V}}_{\mathbf{t}}$ likewise depends on the relation of $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$. In case of no to medium linear association $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} < 0.6\right)$, the estimated within variance yields CI-rates that are generally below 95%. In such a setting, the use of $\widehat{\mathbf{V}}_{\mathbf{t}}$ seems to be the most adequate choice. Because CI-rates tend to be below 95%, the rescaling bootstrap seems preferable due to its slightly more conservative results, even though CI-rates are still too low in some and too high in other scenarios. In contrast, CI-rates of 99 to 100% are clearly too high when using $\widehat{\mathbf{V}}_{\mathbf{t}}$ in cases of stronger linear dependency and bias reduction $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6\right)$. The estimated within variance $\widehat{\mathbf{V}}_{\mathbf{w}}$ is more useful in this context, for which again the rescaling bootstrap yields slightly better CI-rates, ranging from 95 to 100%. This is again not perfect, but appears adequate when considering that selectivity is MNAR and simplifying assumptions are made for inference (cf. section 5.4).

Being based on the functional form approach, the proposed non-parametric neural networks determine weights as a function of $\boldsymbol{X}$ and $\boldsymbol{Z}$, but do not require information about $\boldsymbol{Z}$ outside the non-probability sample (cf. section 5.2.3). In comparison to the GREG, this strategy leads to lower point estimation biases for most scenario populations in the current table 6.8, and the corresponding CI-rates are higher in all scenarios where this is the case. Nevertheless, and as in table 6.6, CI-rates for naive variance estimation are all

Table 6.8: Confidence interval coverage rates for selected weighting models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\mathrm{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}\right)$ for 100% coverage, using total and covariance constraints

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
GREG	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	32.5	34.3	17	24	65	92	20	64	91
		0.0	32.0	33.6	13	14	57	87	14	58	88
		0.3	24.2	25.5	11	11	42	67	9	40	68
		0.6	19.7	20.8	19	21	51	80	16	51	83
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	25.6	27.3	36	46	85	95	36	82	97
		0.0	24.7	26.4	31	33	81	94	29	85	96
		0.3	21.0	22.3	21	14	61	86	11	55	87
		0.6	15.4	16.5	37	23	69	91	21	70	92
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	11.6	12.7	65	46	96	100	44	96	100
		0.0	19.0	20.6	52	50	95	99	53	95	99
		0.3	15.4	17.1	73	60	100	100	63	100	100
		0.6	10.9	12.0	68	48	94	100	41	95	100
calibrated ANN (fixed knots)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	27.3	29.2	26	79	94	100	31	75	97
		0.0	27.5	29.1	23	74	96	100	23	67	95
		0.3	19.9	21.3	38	69	89	100	27	63	87
		0.6	16.5	17.9	52	75	94	100	41	79	94
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	23.0	24.9	46	87	98	100	44	90	99
		0.0	23.5	25.3	37	86	99	100	31	89	98
		0.3	17.1	18.6	51	80	97	100	42	77	96
		0.6	12.0	13.5	72	91	100	100	62	93	100
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	9.1	10.7	84	98	100	100	77	99	99
		0.0	19.2	21.0	49	88	99	100	47	91	98
		0.3	15.7	17.6	70	95	100	100	60	97	100
		0.6	8.3	9.8	88	95	100	100	74	97	99

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables
$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate
Response model variables: $\boldsymbol{X}$,$\boldsymbol{Z}$ Calibration variables: $\boldsymbol{X}$
Correlations determining the simulation scenario (all others result as products of the stated ones):
$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$
All numbers except for the correlations are in (rounded) percentage points.

below 95%. The same holds when $\widehat{\mathbf{V}}_{\mathbf{b}}$ is estimated from a rescaling bootstrap, but Monte Carlo bootstrap estimates for this variance component seem to be the most reasonable choice for inference when $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6$. For lower linear dependencies $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} < 0.6\right)$, better CI-rates are achieved by using estimates $\widehat{\mathbf{V}}_{\mathbf{w}}$. This holds regardless of the resampling method, but the rescaling bootstrap appears clearly less favorable due to intervals which never achieve 95% CI-rates. Estimates $\widehat{\mathbf{V}}_{\mathbf{t}}$ lead to CI-rates of 100% across all scenarios for the Monte Carlo bootstrap. Although the rescaling bootstrap seems more reasonable in case of this total variance estimate, confidence intervals are still too wide in most scenarios. Nevertheless, it may be considered as a slightly more conservative alternative to estimating $\widehat{\mathbf{V}}_{\mathbf{w}}$ by means of the Monte Carlo bootstrap when $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} < 0.6$.

As in case of propensity weights (cf. tables 6.5 and 6.6), there is no variance estimation technique that is unambiguously favorable for all types of calibration weights and scenarios. When using the GREG, it depends on the population correlation $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$ whether rescaling bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{t}}$ or $\widehat{\mathbf{V}}_{\mathbf{w}}$ perform better, and inference is mainly reliable when this correlation is non-negligible. This dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ likewise determines whether Monte Carlo bootstrap variance estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ or $\widehat{\mathbf{V}}_{\mathbf{w}}$ are more suitable for constructing confidence intervals when weights are obtained from a calibrated non-parametric ANN, although inference in this case is a bit more reliable than for the GREG. An adequate choice of an inferential method therefore depends again on the underlying selectivity, structural form of the weighting model and available auxiliary information.

The case of pseudo-design-based estimates for $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ under joint availability of a reference sample and calibration benchmarks is considered in table 6.9. These results correspond to selected point estimates presented in figure 6.17. All other conditions, and thus the unweighted point estimates for this setting, are the same as before in table 6.6 and therefore not recapitulated.

Results in figure 6.17 indicate that to jointly use reference sample and calibration constraints for estimating $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$, the preferable approach is to fit a GLM for predicting response propensities and subsequently calibrate the propensity weights by means of the GREG. Considering confidence intervals for this weighting method, higher linear dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ lead to increasing CI-rates due to reduction of the point estimates' bias, which is similar as for the GREG on its own (cf. table 6.8). Regardless of whether a parametric or a non-parametric GLM is used for the outlined two-step approach, confidence intervals when using naive variance estimation reach or exceed 95% only if $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6$, in which case they perform quite well. For less pronounced linear dependencies $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} < 0.6\right)$, the estimated between variance $\widehat{\mathbf{V}}_{\mathbf{b}}$ obtained from a Monte Carlo bootstrap is more adequate for achieving the nominal CI-rate. This approach may also be sensible for $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6$ when using the non-parametric propensity model, although coverage in this case is higher than 95% in all scenarios. However, this excess is too big when applying parametric GLMs, such that the naive approach to variance estimation seems clearly favorable in this setting. For both types of GLMs to combine with the GREG, the estimated within variance leads to confidence intervals that are too wide in almost all cases. Consequently, confidence intervals that use $\widehat{\mathbf{V}}_{\mathbf{w}}$ alone or as a component of $\widehat{\mathbf{V}}_{\mathbf{t}}$ are usually too conservative for adequate inference.

In contrast, weights obtained from a calibrated non-parametric neural network (with fixed B-spline knots) in this setting lead to considerably higher biases and MSEs in point estimation. When using these weights, it is difficult to identify a valid approach for inference in most scenarios. Naive variance estimation generally results in CI-rates that are below 95%. For bootstrap variance estimates it depends on the scenario whether CI-rates are below or above this nominal CI-rate. For $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.0$, the Monte Carlo bootstrap estimates of $\widehat{\mathbf{V}}_{\mathbf{w}}$ are at least close to 95%. Similar but slightly worse is the performance of confidence intervals that are based on $\widehat{\mathbf{V}}_{\mathbf{t}}$ obtained from a rescaling bootstrap. The remaining options lead to CI-rates that are either way too low or too high in all scenarios. In case of higher correlations $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$, there is rarely any choice for variance estimation that achieves the nominal CI-rate. For $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.3$ and $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = 0.6$, Monte Carlo bootstrap estimates of the between variance $\widehat{\mathbf{V}}_{\mathbf{b}}$ or total variance estimates $\widehat{\mathbf{V}}_{\mathbf{t}}$ obtained

Table 6.9: Confidence interval coverage rates for selected weighting models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\mathrm{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}\right)$ for 100% coverage, using a reference sample, total and covariance constraints

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
Logit model (parametric) & GREG	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	12.2	15.8	83	95	99	100	73	98	99
		0.0	11.9	15.1	84	99	100	100	67	99	100
		0.3	8.3	11.5	81	95	99	100	65	98	99
		0.6	8.2	10.7	80	94	99	100	69	97	99
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	10.2	13.5	91	98	100	100	83	100	100
		0.0	10.6	13.7	88	100	100	100	79	100	100
		0.3	7.5	10.7	86	97	100	100	73	98	100
		0.6	5.2	8.0	91	98	100	100	83	99	99
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	3.5	6.5	97	100	100	100	84	100	100
		0.0	10.3	13.1	93	99	100	100	88	100	100
		0.3	8.8	11.6	97	100	100	100	88	100	100
		0.6	2.9	5.8	98	100	100	100	84	100	100
Logit model (fixed knots) & GREG	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	10.7	15.6	84	95	100	100	81	99	100
		0.0	10.3	15.2	86	95	100	100	84	99	100
		0.3	8.1	12.1	80	91	98	100	76	94	95
		0.6	7.6	10.8	82	96	99	100	82	96	98
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	10.2	14.5	89	97	100	100	85	100	100
		0.0	10.6	14.8	86	96	100	100	85	99	100
		0.3	7.9	11.8	83	93	99	100	75	95	99
		0.6	5.2	8.7	92	97	99	100	81	97	99
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	4.5	8.7	97	97	100	100	84	99	100
		0.0	11.0	15.7	90	98	100	100	87	99	100
		0.3	9.4	13.9	95	98	100	100	91	99	100
		0.6	3.5	7.1	97	100	100	100	88	100	100
calibrated ANN (fixed knots)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	29.8	31.7	18	82	93	100	32	71	93
		0.0	29.9	31.8	14	68	96	100	20	60	96
		0.3	22.6	23.6	23	61	87	100	16	52	84
		0.6	18.6	19.9	38	63	93	100	26	63	93
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	25.0	26.8	34	90	100	100	35	85	100
		0.0	24.9	27.1	28	88	100	100	44	75	100
		0.3	19.5	20.8	43	77	91	100	27	73	86
		0.6	13.1	15.8	61	95	100	100	45	91	95
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	9.2	11.3	83	100	100	100	72	100	100
		0.0	20.5	22.9	40	89	100	100	33	78	100
		0.3	15.8	18.7	58	100	100	100	40	100	100
		0.6	9.0	10.7	83	95	100	100	68	95	100

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables
$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate
Response model variables: $\boldsymbol{X}$,$\boldsymbol{Z}$ Calibration variables: $\boldsymbol{X}$
Correlations determining the simulation scenario (all others result as products of the stated ones):
$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$
All numbers except for the correlations are in (rounded) percentage points.

from the rescaling bootstrap yield 95% coverage. If $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.3$ and $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} < 0.6$, all CI-rates are either considerably below or above 95%. When $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = \boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = 0.6$, the Monte Carlo bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ as well as the rescaling bootstrap estimates $\widehat{\mathbf{V}}_{\mathbf{w}}$ both achieve 95% coverage. When the non-linear dependency is less pronounced $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6 \text{ and } \boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} < 0.6\right)$, the aim of achieving 95% CI-rates is usually missed as before.

In summary, there is no unique solution for inference that performs best in the context of all pseudo-design weighted point estimates. Even for a specific weighting model, the choice of an inferential method has to be made with respect to the actual selectivity pattern and the potential of available auxiliary information to account for the selectivity. This makes inference for pseudo-design-based estimation clearly more situational than for model-based strategies. It also has an effect on inference for estimates that jointly use pseudo-design- and model-based methods, which is considered in the following paragraphs.

Inference for the Synthesis of Model- and Pseudo-design-based Methods

In a final step, the following discussion focuses on inference when integrating the model- and the pseudo-design-based paradigm. Results for weighted prediction models that are fit using pseudo-weights obtained from a non-parametric model are presented in table 6.10. This is a subset of point estimates shown in figure 6.18, where MARS and additive mixed models yield comparatively good results in relation to the other prediction models under consideration. Inferential methods for these two models are evaluated and again contrasted with the reference point where no weighted prediction model but only the pseudo-design weights are used for estimation (nps-estimates).

As $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$ implies a MNAR scenario for selectivity, these nps-estimates are biased across all scenarios. As a consequence, naive variance estimation generally results in CI-rates below 95%. Estimated between variances $\widehat{\mathbf{V}}_{\mathbf{b}}$ when using the Monte Carlo bootstrap lead to better CI-rates, ranging from 89 to 97%. When using the corresponding rescaling bootstrap estimates, however, the coverages drop to 0% across all scenarios. The within component $\widehat{\mathbf{V}}_{\mathbf{w}}$ provides higher CI-rates for both resampling techniques. The results are CI-rates that are generally above 95% for the Monte Carlo but still below 95% for the rescaling bootstrap, with the same holding for the combination $\widehat{\mathbf{V}}_{\mathbf{t}}$ of both components. Although not perfect, estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ obtained through Monte Carlo bootstrapping therefore come closest to the nominal CI-rate. A considerably more conservative approach is obtained when using estimates for the within instead of the between component calculated from the Monte Carlo bootstrap.

In comparison to the reference point of weighted non-probability sample estimates, using a weighted MARS model for imputation results in point estimates that have similar or lower biases in almost all scenarios. As for the purely model-based estimates presented in table 6.3, however, CI-rates are considerably lower for mass-imputation than for plain pseudo-design weighted estimates when the magnitude of bias is similar. In combination with the uncertainty and hence variation that is introduced by additionally applying prediction models, it is more cumbersome to achieve adequate inference for the joint usage of model- and pseudo-design-based methods. As a consequence, none of the considered variance estimates actually achieves the nominal 95% CI-rate in any of the scenarios. Since all empirical CI-rates fall below this target, it is the most conservative approach that comes closest to it. This is the case for the total variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$ obtained through Monte

Table 6.10: Confidence interval coverage rates for selected prediction models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\mathrm{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}\right)$ for 100% coverage – weighting model: pseudo-weights (fixed knots), using a reference sample (estimation from imputed reference sample)

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
No model (nps-estimate)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	9.3	14.9	84	93	100	100	0	86	86
		0.0	9.2	14.8	84	91	98	100	0	87	87
		0.3	11.0	15.9	84	90	98	100	0	84	84
		0.6	11.3	15.4	85	95	100	100	0	89	89
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	8.7	14.3	87	90	100	100	0	80	80
		0.0	10.0	15.0	83	89	100	100	0	83	83
		0.3	10.3	15.6	84	93	99	100	0	84	84
		0.6	9.0	13.9	90	97	99	100	0	92	92
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	7.6	13.6	90	91	100	100	0	84	84
		0.0	9.6	15.3	85	90	100	100	0	84	84
		0.3	7.3	14.4	87	90	98	100	0	84	84
		0.6	7.4	13.3	90	96	100	100	0	92	92
MARS	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	9.3	15.4	16	65	33	76	0	23	23
		0.0	9.3	15.7	20	71	25	78	0	21	21
		0.3	7.4	12.5	26	68	35	78	0	26	26
		0.6	7.4	11.9	39	76	46	86	0	42	42
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	9.1	15.2	49	66	51	82	0	47	47
		0.0	10.3	16.2	48	71	51	88	0	45	45
		0.3	7.3	12.7	54	64	51	77	0	42	42
		0.6	5.0	10.6	70	70	71	90	0	70	70
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	4.0	11.1	85	68	77	92	0	75	75
		0.0	8.6	21.6	70	66	72	89	0	71	71
		0.3	7.1	17.4	83	68	82	94	0	81	81
		0.6	3.2	10.3	86	67	85	93	0	83	83
GAMM	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	8.7	14.4	79	58	76	91	0	76	76
		0.0	8.6	14.4	80	58	83	92	0	84	84
		0.3	6.5	11.5	79	59	79	87	0	79	79
		0.6	6.6	11.1	73	63	75	90	0	78	78
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	8.8	14.4	75	56	72	88	0	73	73
		0.0	9.7	14.9	69	58	69	86	0	69	69
		0.3	6.4	11.7	77	52	71	86	0	72	72
		0.6	4.1	10.0	85	62	82	93	0	83	83
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	3.9	16.5	90	55	83	89	0	81	81
		0.0	11.4	139.6	71	61	75	92	0	72	72
		0.3	9.8	132.1	84	62	89	92	0	79	79
		0.6	2.3	20.5	90	52	89	96	0	91	91

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables
$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate
Prediction model variables: $\boldsymbol{X}$ Random effects variables: $\boldsymbol{Z}$-classes Response model variables: $\boldsymbol{X}$,$\boldsymbol{Z}$
Correlations determining the simulation scenario (all others result as products of the stated ones):
$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$
All numbers except for the correlations are in (rounded) percentage points.

Carlo bootstrapping. At least for $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{.1}} = 0.6$, CI-rates of 89 to 94% can be obtained with this technique. Although such coverage clearly is not perfect, it nevertheless seems reasonable when considering that selectivity is MNAR. However, these rates drop down to 76% for some of the lower linear dependencies $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{.1}} < 0.6\right)$.

Especially for these less pronounced linear relationships $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{.1}} < 0.6\right)$, additive mixed models yield slightly lower biases and MSEs than MARS models. Correspondingly, inference for these former models appears somewhat less situational and more stable across the different simulation scenarios than for the latter. Still, naive and rescaling bootstrap variance estimates generally fall short in achieving nominal CI-rates. The same holds for the separate use of variance components $\widehat{\mathbf{V}}_{\mathbf{b}}$ and $\widehat{\mathbf{V}}_{\mathbf{w}}$ estimated by means of the Monte Carlo bootstrap, but their joint usage in form of $\widehat{\mathbf{V}}_{\mathbf{t}}$ leads to CI-rates between 86 and 96%. As for the MARS model, these results are clearly less than perfect, yet seem reasonable for selectivity that is MNAR.

In correspondence to inference for purely model-based estimates examined in table 6.3, Monte Carlo bootstrap estimates for the total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$ are preferable even for propensity weighted imputation models. This pattern holds quite generally when using mass-imputation for the reference sample, regardless of the chosen prediction model. Nevertheless, confidence interval coverage depends still critically on the magnitude of the remaining bias, such that mass imputation improves inference for non-probability samples mainly when it considerably reduces point estimation bias.

In cases where calibration benchmarks are the only available auxiliary information, imputation of the reference sample is infeasible. In this context, MRP is the most commonly discussed technique to combine the model- and the pseudo-design-based paradigm by weighted aggregation of predictions in the non-probability sample. Indeed, the linear mixed model is the only prediction method that provides a clear improvement over purely pseudo-design-based estimates of the mean $\boldsymbol{\mu}_{\boldsymbol{y}_{.1}}$ in at least some of the scenarios presented in figure 6.19. Inferential approaches for these two strategies which apply weighted estimation to either the observed values ('no model') or the predictions from such a GLMM are evaluated in table 6.11 for simulation scenarios corresponding to the previous table 6.10. Note that in order to exactly meet the specification of MRP as in figure 6.19, calibration and random effects variables are both determined by classes of $\boldsymbol{Z}$ for this single example.

In case of the plain nps-estimates that use the observed target variable, naive variance estimation already performs relatively well when the non-linear dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{.1}$ is rather small $\left(\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{.1}} < 0.3\right)$. CI-rates based on these variance estimates range from 90 to 96% in these cases. When $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{.1}} \geq 0.3$, however, the corresponding CI-rates drop down to 70% in the worst case, and seem to be still reasonable only when $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{.1}} = 0.6$. As before, the estimated between variance component $\widehat{\mathbf{V}}_{\mathbf{b}}$ is too small to achieve 95% coverage in any of the scenarios, regardless of the employed resampling method. Better results are obtained by using the within variance. When estimated by means of the Monte Carlo bootstrap, confidence intervals tend to be too conservative and lead to CI-rates of 92 to 100%. Yet, this seems to be the best possible setting especially for $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{.1}} < 0.6$. Although the rescaling bootstrap usually results in CI-rates below 95%, it seems to be an adequate alternative when $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{.1}} = 0.6$, for which the resulting estimates $\widehat{\mathbf{V}}_{\mathbf{w}}$ or $\widehat{\mathbf{V}}_{\mathbf{t}}$ yield CI-rates between 92 and 94%.

Table 6.11: Confidence interval coverage rates for selected prediction models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\mathrm{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}\right)$ for 100% coverage – weighting model: post-stratification, using total constraints (estimation by weighted aggregation of predictions)

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
No model (nps-estimate)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	12.3	14.4	91	65	99	100	0	92	92
		0.0	12.2	14.0	93	48	100	100	0	93	93
		0.3	14.6	16.4	83	39	99	100	0	84	84
		0.6	17.0	18.7	70	39	92	99	0	74	74
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	11.1	13.1	96	66	100	100	0	98	98
		0.0	12.2	14.3	92	61	99	99	0	92	92
		0.3	12.7	14.7	91	47	100	100	0	90	90
		0.6	14.5	16.3	83	48	97	99	0	83	83
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	11.7	13.6	94	64	100	100	0	94	94
		0.0	12.5	14.6	90	54	98	99	0	92	92
		0.3	11.9	14.1	92	63	97	100	0	94	94
		0.6	12.4	14.2	93	60	98	100	0	92	92
GLMM	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	13.6	15.7	37	64	53	91	0	39	39
		0.0	13.4	15.3	35	48	57	93	0	33	33
		0.3	16.6	18.7	22	25	45	80	0	25	25
		0.6	17.0	18.5	24	33	45	72	0	27	27
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	12.6	14.7	34	66	49	93	0	34	34
		0.0	13.9	16.0	26	63	40	88	0	31	31
		0.3	14.5	16.7	36	33	60	91	0	36	36
		0.6	14.4	16.0	45	37	70	87	0	49	49
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	11.4	13.1	73	56	94	100	0	76	76
		0.0	15.0	17.1	22	56	28	83	0	24	24
		0.3	14.3	16.5	29	55	44	89	0	29	29
		0.6	11.8	13.4	76	49	93	98	0	81	81

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables
$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate
Prediction model variables: $\boldsymbol{X}$ Random effects variables: $\boldsymbol{Z}$-classes Calibration variables: $\boldsymbol{X}$
Correlations determining the simulation scenario (all others result as products of the stated ones):
$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathrm{nps}}} = 0.6$
All numbers except for the correlations are in (rounded) percentage points.

As in tables 6.3 and 6.10, the CI-rates when a prediction model is additionally incorporated are way lower than for pure pseudo-design-based estimates that exhibit similar magnitudes of bias. Even in cases where the GLMM's predictions allow for some degree of bias reduction in comparison to the plain weighted estimates, CI-rates are lower than 95% for almost all variance estimates. Therefore, the most conservative variance estimator, which is $\widehat{\mathbf{V}}_{\mathbf{t}}$ obtained from the Monte Carlo bootstrap, provides the most adequate results in most of the considered scenarios. Nevertheless, this approach still misses nominal coverage in many cases, yielding CI-rates between 72 and 100%. As a consequence, inference for this example of MRP is less reliable than for post-stratified estimation that does not

additionally use a GLMM. Furthermore, the amount of bias reduction obtained from using the GLMM is relatively small in comparison to the remaining bias, such that it may be better to rely on pseudo-design weighting alone in the present context.

As discussed with regard to figure 6.20, the stabilizing effect obtained by using mixed model predictions in place of actually observed target variables $\boldsymbol{y}_{\cdot 1}$ in the non-probability sample is more advantageous in case of a fine-grained weighting method that leads to stronger variability in pseudo-design weights. Results for the corresponding confidence intervals when estimating $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ are presented in table 6.12. Calibration is used for the total and variance of $\boldsymbol{X}$. The weighting model is a calibrated neural network that has the structure of a parametric logit model but is fit exclusively with regard to the calibration benchmarks. The remaining settings and considered estimation approaches are kept as in the previous table 6.11.

For estimates that use pseudo-design weights and observed target variable (nps-estimates), naive variance estimation once again results in confidence intervals that are generally too short to reach the nominal 95% coverage. Nevertheless, CI-rates increase with higher correlations $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$, which help to reduce biases in point estimation. As in table 6.10, CI-rates are higher for almost all bootstrap variance estimates. For the estimated between variance $\widehat{\mathbf{V}}_{\mathbf{b}}$, these rates are still mostly below 95% unless the Monte Carlo bootstrap is used in scenarios where $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6$. CI-rates when using within variance estimates $\widehat{\mathbf{V}}_{\mathbf{w}}$ are generally higher than for the between component. When estimating $\widehat{\mathbf{V}}_{\mathbf{w}}$, Monte Carlo bootstrapping yields quite adequate CI-rates when $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.3$ but results in confidence intervals which are mostly too narrow in case of $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.0$ and too wide for $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6$. For the latter case, the rescaling bootstrap performs better in most cases, yielding CI-rates between 87 and 99%. The estimated total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$ leads to confidence intervals that are clearly too wide for the Monte Carlo bootstrap in all scenarios. When using the rescaling bootstrap, however, CI-rates ranging from 85 to 98% seem at least fairly acceptable for $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.0$.

Predictions from GLMMs are particularly suitable for reducing the bias that remains after pseudo-design weighting in certain scenarios considered in figure 6.20. Nevertheless, and underlining the results in tables 6.3 and 6.10, the corresponding CI-rates are considerably lower when using model predictions rather than the observed target variable in cases where the magnitudes of bias are similar. Therefore, inference seems to be quite situational for weighted aggregation of GLMM predictions, and CI-rates that are close to the nominal 95% are achieved only for specific combinations of variance estimation methods and correlations determining the simulation scenarios. CI-rates obtained through naive variance estimation are generally too low, and the same holds for almost all cases where the estimated between variance $\widehat{\mathbf{V}}_{\mathbf{b}}$ is used. Results for within variance estimates $\widehat{\mathbf{V}}_{\mathbf{w}}$ are occasionally either worse or better. However, even the estimated total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$ is mostly too small to achieve the nominal CI-rate, such that its most conservative estimate, which is obtained by means of the Monte Carlo bootstrap, yields the best CI-rates in most scenarios. Exceptions occur only for $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = \boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = 0.6$, where $\widehat{\mathbf{V}}_{\mathbf{w}}$ obtained from the Monte Carlo bootstrap is a better choice, and for $\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6$ and $\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}} = \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$, where the rescaling bootstrap estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$ or the Monte Carlo bootstrap estimate $\widehat{\mathbf{V}}_{\mathbf{b}}$ yield better results.

Table 6.12: Confidence interval coverage rates for selected prediction models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of V $\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}\right)$ for 100% coverage – weighting model: calibrated ANN (parametric), using total and covariance constraints (estimation by weighted aggregation of predictions)

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
No model (nps-estimate)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	30.2	32.0	17	85	96	100	35	71	98
		0.0	29.7	31.2	15	80	90	100	25	63	97
		0.3	22.5	23.8	17	70	79	100	27	41	85
		0.6	18.7	19.9	25	61	86	100	30	54	88
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	25.8	27.6	31	88	98	100	55	83	98
		0.0	25.4	27.3	27	87	96	100	47	79	98
		0.3	20.2	21.5	26	79	93	100	19	64	93
		0.6	14.3	15.6	46	86	98	100	46	82	100
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	11.9	13.2	65	96	100	100	63	99	100
		0.0	21.6	23.4	38	87	100	100	51	87	100
		0.3	18.5	20.2	53	91	98	100	43	96	98
		0.6	10.5	11.8	70	97	100	100	62	97	100
GLMM	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	30.3	32.1	2	33	11	75	7	7	35
		0.0	29.7	31.3	2	25	14	75	12	7	24
		0.3	18.6	20.5	16	50	55	82	21	29	51
		0.6	13.0	15.6	38	65	70	77	40	59	66
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	26.1	27.9	5	38	14	81	10	5	26
		0.0	25.8	27.6	4	38	15	77	13	13	28
		0.3	17.4	19.3	22	59	67	90	22	44	62
		0.6	9.1	12.0	59	81	81	96	57	74	81
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	8.1	11.0	71	96	100	100	79	93	96
		0.0	22.4	24.1	4	36	20	73	11	13	24
		0.3	19.3	21.0	10	40	38	96	6	17	26
		0.6	6.1	9.5	79	92	96	100	78	90	94

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables
$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate
Prediction model variables: $\boldsymbol{X}$ Random effects variables: $\boldsymbol{Z}$-classes
Response model variables: $\boldsymbol{X}$,$\boldsymbol{Z}$ Calibration variables: $\boldsymbol{X}$
Correlations determining the simulation scenario (all others result as products of the stated ones):
$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\text{nps}}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\text{nps}}} = 0.6$
All numbers except for the correlations are in (rounded) percentage points.

In comparison to pseudo-design weighted point estimates, MRP and other forms of weighted aggregation of predictions can have a positive effect on the bias in point estimation. Nevertheless, the use of prediction models complicates generalization to the target population, such that there is no single method for inference that works best for all simulated scenarios. As discussed with regard to table 6.11, it hence has to be considered carefully whether bias reduction through using predicted rather than observed values is (presumably) strong enough to counterbalance these inferential difficulties. Inference is typically more reliable for the plain pseudo-design weighted estimates.

After considering separate availability of either a reference sample or calibration constraints in tables 6.10 to 6.12, the final setting of possible auxiliary information is determined by the joint availability of both. As summarized in the context of figure 6.21, using a weighted loss function for fitting imputation models seems to be the best strategy to combine model- and pseudo-design-based approaches for point estimation in such a setting. Some of the corresponding results for inference when estimating $\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}$ are depicted in table 6.13. The simulated conditions are the same as before in table 6.12, apart from the calculation of the employed pseudo-design weights: the reference sample is used to fit a logit GLM as response propensity model, and the resulting propensity weights are calibrated by means of the GREG to meet the calibration benchmarks.

As discussed with regard to figure 6.21, the results in table 6.13 show that the use of calibrated propensity weights increases the potential to reduce bias in comparison to plain propensity weighting. In that way, calibration constraints to a large extent absorb the capability of prediction models to reduce the bias that remains after weighting in table 6.10. Pseudo-design-based point estimates (nps-estimates) are therefore considerably less biased than in the previous table 6.12, which leads to higher CI-rates in almost all scenarios. Consequently, the naive approach for variance estimation already allows quite reasonable inference for these point estimates when $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ are strongly related $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} = 0.6\right)$. The resulting CI-rates between 93 and 98% are closer to the nominal 95% than for all considered resampling variance estimates. When the linear dependency between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$ is less pronounced $\left(\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} < 0.6\right)$, inference based on between variance estimates $\widehat{\mathbf{V}}_{\mathbf{b}}$ obtained by the Monte Carlo bootstrap seems more adequate since it leads to CI-rates between 94 and 100%. In contrast, confidence intervals based on rescaling bootstrap estimates of the between variance are generally too short to achieve the nominal coverage. For both resampling methods, the CI-rates when using the estimated within variance $\widehat{\mathbf{V}}_{\mathbf{w}}$ and thus also the total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$ estimates are generally too large across all scenarios, resulting in CI-rates which are always above 95%.

As is evident from figure 6.21, the sole case where imputation based on a weighted prediction model reliably yields similar or lower biases than the plain pseudo-design weighted non-probability sample estimates occurs when using a GAMM. However, the potential improvements in bias are rather small and CI-rates in case of weighted imputation models are again lower than for pure pseudo-design-based estimates when a similar degree of bias remains. Therefore, the separate usage of estimated variance components $\widehat{\mathbf{V}}_{\mathbf{b}}$ and $\widehat{\mathbf{V}}_{\mathbf{w}}$ as well as naive variance estimation lead to CI-rates that are generally lower than 95%. Consequently, and as in table 6.10, the estimated total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$ leads to the most adequate CI-rates. For obtaining these estimates, the rescaling bootstrap performs slightly better than the Monte Carlo variant, yielding CI-rates between 92 and 97%.

As for the plain pseudo-design-based estimates, there is no unique best method for inference when it comes to combinations of model- and pseudo-design-based methods. Once again, the best strategy for inference depends on the sample's selectivity, available auxiliary information and point estimation method. Whenever a propensity model is used for weighting, the estimated total variance $\widehat{\mathbf{V}}_{\mathbf{t}}$ seems to be the best choice for achieving the nominal CI-rate. However, it depends on the additional utilization of calibration benchmarks whether the Monte Carlo or the rescaling bootstrap are a better choice for estimating this total variance. When considering weighted aggregation of predictions,

Table 6.13: Confidence interval coverage rates for selected prediction models under different dependencies between $\boldsymbol{X}$ and $\boldsymbol{y}_{\cdot 1}$: estimation of $\mathrm{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot 1}}\right)$ for 100% coverage – weighting model: logit model (parametric) and GREG, using a reference sample, total and covariance constraints (estimation from imputed reference sample)

	Simulation scenario		Quality of point estimates		Confidence interval coverage rates						
					Naive	Monte Carlo bootstrap			Rescaling bootstrap		
	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}}$	$\boldsymbol{\rho}_{\boldsymbol{X}^{\circ 2}\boldsymbol{y}_{\cdot 1}}$	RBias	RRMSE		$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$	$\widehat{\mathbf{V}}_{\mathbf{b}}$	$\widehat{\mathbf{V}}_{\mathbf{w}}$	$\widehat{\mathbf{V}}_{\mathbf{t}}$
No model (nps-estimate)	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	12.2	15.8	83	95	99	100	73	98	99
		0.0	11.9	15.1	84	99	100	100	67	99	100
		0.3	8.3	11.5	81	95	99	100	65	98	99
		0.6	8.2	10.7	80	94	99	100	69	97	99
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	10.2	13.5	91	98	100	100	83	100	100
		0.0	10.6	13.7	88	100	100	100	79	100	100
		0.3	7.5	10.7	86	97	100	100	73	98	100
		0.6	5.2	8.0	91	98	100	100	83	99	99
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	3.5	6.5	97	100	100	100	84	100	100
		0.0	10.3	13.1	93	99	100	100	88	100	100
		0.3	8.8	11.6	97	100	100	100	88	100	100
		0.6	2.9	5.8	98	100	100	100	84	100	100
GAMM	0.0	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	9.0	14.4	79	72	77	96	72	79	96
		0.0	8.9	14.4	81	74	84	93	75	89	93
		0.3	6.6	11.5	79	68	79	91	72	81	92
		0.6	7.0	11.4	74	74	78	94	75	80	93
	0.3	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	8.8	14.1	77	68	75	97	70	76	95
		0.0	9.9	14.7	70	75	68	91	76	71	93
		0.3	6.4	11.6	78	63	72	93	65	77	93
		0.6	4.4	10.1	85	72	83	94	71	85	95
	0.6	$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{y}_{\cdot 1}} \cdot \boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{X}^{\circ 2}}$	3.6	10.7	90	61	82	91	63	84	93
		0.0	10.5	15.6	68	77	77	97	84	84	97
		0.3	8.1	15.4	82	76	87	97	80	92	97
		0.6	2.8	10.0	90	61	88	96	63	89	97

Naive: Naive variance estimate, assuming a probability sample with observed weights and target variables

$\widehat{\mathbf{V}}_{\mathbf{b}}$: Between variance estimate $\widehat{\mathbf{V}}_{\mathbf{w}}$: Within variance estimate $\widehat{\mathbf{V}}_{\mathbf{t}}$: Total variance estimate

Prediction model variables: $\boldsymbol{X}$ Random effects variables: $\boldsymbol{Z}$-classes

Response model variables: $\boldsymbol{X}$,$\boldsymbol{Z}$ Calibration variables: $\boldsymbol{X}$

Correlations determining the simulation scenario (all others result as products of the stated ones):

$\boldsymbol{\rho}_{\boldsymbol{X}\boldsymbol{Z}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{y}_{\cdot 1}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{Z}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$ $\boldsymbol{\rho}_{\boldsymbol{y}_{\cdot 1}\boldsymbol{\pi}^{\mathsf{nps}}} = 0.6$

All numbers except for the correlations are in (rounded) percentage points.

generalization to the target population is generally more situational and less reliable. The results do not allow identification of a single best method for inference in this case.

In the current section 6.3.2.3, different inferential methods are examined with regard to their suitability for non-probability samples. As discussed in sections 5.4 and 6.3.1, all of these methods are based on variance estimates, which are simplifying approximations rather than unbiased estimates for the repeated sampling variances in non-probability samples. In this regard, a point estimator's bias is not considered for inference. Adequate CI-rates are therefore easier to achieve when this bias is relatively low, although this is neither a necessary nor sufficient condition. As a consequence, the choice of the most appropriate method for inference depends on a non-probability sample's selectivity, available auxiliary information and applied point estimation method. In case of mass-imputation for the reference sample, Monte Carlo bootstrap estimates for the total variance generally perform quite well, even though the rescaling bootstrap may occasionally yield slightly better results. This holds regardless of whether the prediction model is fit with or without pseudo-design weights. When using purely pseudo-design-based estimates or weighted aggregation of predictions, the findings are far less clear. In such cases, no single method for inference that performs best for all cases can be identified.

All results presented in the current chapter 6 are obtained by means of Monte Carlo simulations, considering synthetic populations and simulated samples. To evaluate and compare methods proposed for dealing with the challenges of non-probability samples, such simulation studies are a common and usually more appropriate way than the application to a single real data set that is obtained by non-probability sampling (cf. e.g. section 6.3.1; Buelens, Burger and van den Brakel, 2018, p. 327; Enderle, Münnich and Bruch, 2013, p. 95). Nevertheless, such an application to a real non-probability sample can provide additional insights into particular benefits and pitfalls that are not considered in the simulations. Moreover, it allows assessing the practical relevance of the simulation results. To these ends, an application example for data collected in the WageIndicator volunteer web survey is presented in the following chapter 7.

7 Application to the WageIndicator Web Survey

In the previous chapter 6, methods proposed for selectivity assessment, point estimation and inference in the context of non-probability sampling are evaluated and compared by means of Monte Carlo simulation studies. A crucial advantage when using such simulations for this purpose is that the true population and underlying sample selection mechanism are both known (cf. chapter 2; Enderle, Münnich and Bruch, 2013, p. 95). Furthermore, any influence of other aspects of data quality and accuracy, e.g. due to dissimilar survey modes, questionnaire designs and degrees of response error when comparing different samples, can be prevented. This is typically not the case when considering real non-probability samples (cf. e.g. Bethlehem and Biffignandi, 2012, pp. 103 ff; Bianchi, Biffignandi and Lynn, 2017, p. 387; Japec et al., 2015, p. 853; Münnich and Lenau, 2019). To nevertheless illustrate the use of all the methods introduced in chapters 3 and 5 for a real non-probability sample and to highlight additional pitfalls not considered in the simulation, an application to data from the WageIndicator web survey (WI) is presented in the current chapter 7.

The WI is a project implemented by an international non-profit organization, carrying out online surveys for 143 countries in 56 languages (as of February 2020; cf. Tijdens, 2020, p. 1). The international online portal is accessible via `https://wageindicator.org`. Since the major goal of the WI and its conductors is to improve labor market transparency by publishing wage-related information, the web survey's target population is the full labor force in the surveyed countries. This does not only include workers in formal dependent employment, but also groups like informal or self-employed workers as well as retired persons and pupils having side jobs (cf. Tijdens, 2008, p. 92; Tijdens et al., 2010, p. 14). With regard to this population and research interest, data collected in the WI is used for various analyses, e.g. for quantifying mean income and income inequalities between groups or for comparing working conditions as well as job and life satisfaction within and between occupations (cf. e.g. Pedraza, Guzi and Tijdens, 2020; Tijdens et al., 2014; Visintin et al., 2015). The WI is a continuous and voluntary web survey, implying that its online questionnaires are permanently available to anyone that is willing to participate. Sample inclusion is therefore determined by the opportunity, awareness and active decision of potential respondents to visit the homepage and fill in the questionnaire (cf. Tijdens, 2008, p. 98). Recruitment strategies for the WI rely on different advertising channels to increase awareness and willingness to participate in the survey, e.g. via newspapers and online banners. Furthermore, there are lottery incentives for participation in the survey (cf. Pedraza et al., 2010, pp. 112 ff; Smyk, Tyrowicz and Van der Velde, 2021, p. 436; Steinmetz et al., 2014, p. 277; Tijdens, 2014, p. 27). Since the sample selection process is merely influenced but neither fully controlled nor known even by the researchers who are gathering the data, the WI is a non-probability sample (cf. chapter 2). A comprehensive overview of the motivation and methods of the WI is given by Tijdens (2008) and Tijdens et al. (2010). To have a clearly defined target population for which high-quality auxiliary information is available, the following example is based on the German WI sample from the year 2012.

All non-trivial methods for assessing and compensating selection bias in non-probability samples require some kind auxiliary information (cf. chapters 3 and 5). In the current application, the German Microcensus from the corresponding year (2012; cf. e.g.

Statistisches Bundesamt, 2013; 2017) is used as a source of auxiliary information.[10] It is considered to be a high-quality reference sample for modeling sample selection processes due to certain advantages (cf. Enderle, Münnich and Bruch, 2013, p. 92). Most importantly, the Microcensus is a 1% probability sample of the German resident population conducted by the German Statistical Offices, with mandatory participation and duty of disclosure for a large variety of variables established by law. Therefore, it contains a large number of observations and only minimal unit or item non-response (cf. Statistisches Bundesamt, 2013, p. 7). The Microcensus' target population is the whole resident population in Germany, from which the sample is drawn using a stratified single-stage cluster sampling design. Primary sampling units (PSUs) are the so-called 'Auswahlbezirke', which represent clusters of neighboring dwellings that are geographically close to each other. These dwellings can be located within either a single larger or several smaller buildings, with a benchmark of roughly twelve dwellings per PSU. All secondary sampling units (households and persons) within the selected primary sampling units are surveyed. For stratification of these PSUs, a cross-combination of two variables is used.[11] On the one hand, 201 regional strata are used to represent geographical areas with on average 350 000 inhabitants, where metropolitan areas constitute separate strata. On the other hand, PSUs are grouped in five strata that classify the size of the building(s) containing the sampling units, with two additional strata for communal accommodations and new buildings. Once the sampling units are selected, computer-assisted personal interviewing (CAPI) is carried out for most of the respondents, using paper questionnaires sent via mail for cases where attempts for personal contact fail multiple times. A description of the Microcensus' sampling design and data collection procedures in greater detail is given by Rendtel and Schimpl-Neimanns (2001), Schimpl-Neimanns (2011, pp. 21 ff) and the German Federal Statistical Office (cf. Statistisches Bundesamt, 2013, pp. 5 ff). Classical design-based estimation for such probability samples is summarized in section 2.2, with a more in-depth theoretical discussion being e.g. provided by Cochran (1977, pp. 300 ff), Fuller (2009, pp. 29 ff) or Särndal, Swensson and Wretman (1992, pp. 124 ff).

To provide a meaningful but concise overview, the following discussion focuses on four selected core variables. These are income, gender, age and education, which are chosen for the following reasons: as described above, wages and related information are pivotal for the WageIndicator web survey, which even is directly evident from its name (cf. Tijdens, 2008, p. 92; Tijdens et al., 2010, p. 14). Income earned from work is, thus, a major topic and typical variable of interest for this survey. In the current example, the monthly net earned income is used as the target variable. Usually, there is a strong influence of a survey's topic on voluntary participation, especially when advertising channels relating to this topic are used, which is the case for the WI (cf. section 2.1; Baker et al., 2010, p. 38; Faas and Schoen, 2006, p. 180; Pedraza et al., 2010, p. 116). Publications that attempt to compensate for this kind of selectivity in the WI typically focus on the three socio-demographic variables that are additionally chosen (cf. e.g. Pedraza et al., 2010, p. 112;

[10] The analysis is not based on the Microcensus scientific use file but on a larger panel data set accessed at the Economic and Social Statistics Department of Trier University.

[11] Strictly speaking, the use of systematic sampling leads to another implicit level of stratification, where 100 PSUs at a time are randomly assigned to constitute one stratum, from which exactly one is drawn to constitute the 1% sample. However, this stratification is typically disregarded since it cannot be considered in unbiased inference and PSUs are assigned randomly to these strata (cf. Rendtel and Schimpl-Neimanns, 2001, pp. 89 f; Wolter, 2007, pp. 298 ff).

Smyk, Tyrowicz and Van der Velde, 2021, pp. 447 ff; Steinmetz et al., 2014, p. 296). Smyk, Tyrowicz and Van der Velde (2021, pp. 442 ff) adduce substantial as well as pragmatical reasons for this strategy. On the one hand, the authors argue that gender, age and education are important indicators for human capital and, therefore, inherently related to wages as the core survey topic. On the other hand, some sort of external auxiliary information about these three variables is commonly available for the countries represented in the WI data. Considering the large number of countries for which this is the case, a weighting approach that is applicable across all countries is facilitated when using such a common set of variables. A larger set of auxiliary variables would often require carrying out probability reference surveys specifically for this purpose in all covered countries. Furthermore, the authors state that the proportion of missing values in these variables is comparably low across different countries. Therefore, additional difficulties in dealing with non-response are reduced when using only the three socio-demographic variables mentioned above (cf. also Steinmetz, Tijdens and Pedraza, 2009, pp. 45 ff; Steinmetz et al., 2014; Tijdens and Steinmetz, 2016). Although other options and further variables for weighting are evaluated as well (cf. e.g. Steinmetz and Tijdens, 2009; Steinmetz et al., 2014), the only pseudo-design weights that are published with the WI data are propensity weights which are exclusively based on gender and age (cf. WageIndicator Foundation, 2011). For comparability and correspondence with the referred publications, the following discussion focuses on gender, age and education as auxiliary variables. Nevertheless, this rather small set of variables may clearly result in over-simplification when assessing and compensating selection bias, a limitation which is discussed at the end of the current chapter 7. In the remaining part of this chapter, findings regarding the selectivity and potential biases in the WI are presented in section 7.1. Results for point estimation are considered in section 7.2, while a summarizing discussion of the results and limitations of this application study is provided in section 7.3.

7.1 Assessment of Selectivity and Potential Biases

Assessment of selectivity as introduced in chapter 3 is commonly a first step and toolbox for studying the properties of a non-probability sampling process. Although selectivity can be understood with respect to either a sampling process or a single sample, it is typically examined focusing on a single realized sample. This is due to the fact that the non-probability selection process is usually at least partially unknown, such that the realized sample is the only available information. An examination of the potential selectivity can provide indications on the variables that characterize this unknown sampling process and on the potential bias that it may induce.

As described in sections 3.3 and 3.4, a common starting point for this purpose is to look for (dis-)similarities in the auxiliary variables between non-probability and reference sample. The three socio-demographic auxiliary variables are measured consistently in the Microcensus and the WI, such that a comparison is straightforward. However, monthly net earned income in the Microcensus is only available for workers in formal dependent employment (cf. Statistisches Bundesamt, 2017, p. 260), which only constitute a subgroup of the WI's target population (cf. Tijdens, 2008, p. 92; Tijdens et al., 2010, p. 14). Following Smyk, Tyrowicz and Van der Velde (2021, pp. 453 f), the monthly net total income in the Microcensus is therefore used as a proxy measure to still provide an indication of selectivity for the target variable without excluding relevant subgroups. This strategy appears reasonable because the total income heavily depends on the wage

Education: ISCED levels five and six are joined into one category for this representation due to very small case numbers in the highest category.

Income: For the WI, monthly net earned income is the target variable. Since it is not measured in the Microcensus, the total net income is used as a proxy benchmark in this representation.

Figure 7.1: Comparison of the German WageIndicator web survey and Microcensus 2012

(cf. e.g. Checchi and García-Peñalosa, 2010; Lerman and Yitzhaki, 1985) and is available for the whole Microcensus. Monthly net total income in the Microcensus is also used for benchmarking the estimates in the subsequent discussion, but not for estimation itself. It is measured as an interval-censored variable, as indicated in figure 7.1 (cf. Statistisches Bundesamt, 2017, p. 257).

In figure 7.1, the estimated distributions obtained from the WI (purple) and the Microcensus (green) are compared for the four variables outlined above. Design weights are used for the Microcensus to calculate population estimates as described in section 2.2. Due to the high data quality of the Microcensus summarized above, the estimates obtained from this sample are used as surrogates for the true population distributions and, thus, considered as benchmarks. Since inclusion probabilities (and thus design weights) are unknown for the WI, all observations in this non-probability sample are given the same weight, leading to unweighted estimates in case of the proportions in figure 7.1.

Considering the gender distribution first, the Microcensus estimates suggest approximately 49% male and 51% female persons in the population. Both proportions are much more dissimilar in case of the WI, where roughly 65% men and 35% women are observed. This is a strong indication that men are over-represented in this non-probability sample.

The age distributions estimated from both data sources are likewise quite different. In the Microcensus estimates, respondents being between 40 and 49 years old constitute the modal cohort, which contains roughly 17% of all observations. The proportions decrease when moving towards the distribution's tails, showing a relatively steep decrease at the very upper tail. The largest cohort in the WI data is likewise that between the ages of 40 and 49. However, with about 30% of the respondents falling into this category, the proportion is much higher than estimated from the Microcensus. Similar but slightly less pronounced findings hold for the remaining age groups near this modal category, while all other cohorts (aged below 20 or above 59) are much less frequent than in the Microcensus estimates. This can be partially explained by the different target populations for both surveys mentioned above, especially when considering the lower proportions for ages below 20 in the WI. Nevertheless, this does not seem to fully justify the differing age distributions because the WI's target population includes pupils as well as retired workers (cf. Tijdens, 2008, p. 92). Therefore, these results suggest that respondents close to the age distribution's center are over-represented in the web survey, presumably because these persons are more typical for the population of internet users and the labor force. The steeper decline in proportions towards the tails of the distribution furthermore implies that the relation between age and sample inclusion is non-linear.

The respondent's highest educational qualification is measured by the International Standard Classification of Education (ISCED) in its 1997 version (cf. Schroedter, Lechert and Lüttinger, 2006; UNESCO, 2006; 2012). Comparing the estimated proportions, it is evident that the ISCED levels one, two, four and six are much more frequent in the WI than estimated from the Microcensus, implying that respondents falling in these categories are likely over-represented in the web survey. The opposite holds for the remaining levels three and five, which are the most frequent ones according to the Microcensus estimates but way less common in the WI. Since the differences between both data sources do not uniformly in- or decrease with increasing ISCED level, the relation between educational level and sample inclusion appears non-linear as well.

In case of monthly net earned income, the pattern is somewhat more straightforward. All benchmark proportions for incomes below 900 € estimated from the Microcensus are higher than those in the WI, while the proportions of all but one income classes above 900 € are clearly higher in the WI than suggested by the benchmark distribution. The only exception from this pattern is the share of respondents between 7 500 and 10 000 €, where the benchmark proportion (0.24%) is higher than that in the non-probability sample (0.15%), but this difference appears negligible. These results indicate an under-representation of lower and over-representation of higher incomes in the WI. This finding appears even more severe when considering that earned income from in the WI is compared to total income measured in the Microcensus. Since the earned income is only a part of the total income, the earned income (WI) figures should on average be lower than those for the total income (Microcensus), yet the opposite is the case.

For all considered variables, the graphical evaluation in figure 7.1 indicates that the WI deviates systematically from the benchmarks. Different approaches to provide a more

Table 7.1: Selectivity measures and tests for the German WageIndicator web survey 2012

Measure of selectivity	**Variable(s)**				
	Gender	**ISCED levels**	**Age**	**Income classes**	**ISCED, Age, Income classes**
Tests for selectivity: p-values for differences between non-probability and reference sample					
Fisher's exact test	< 1e-10	–	–	–	–
t-test	–	< 1e-10	< 1e-10	< 1e-10	–
Kruskal-Wallis test	–	< 1e-10	< 1e-10	< 1e-10	–
Kolmogorov-Smirnov test	–	< 1e-10	< 1e-10	< 1e-10	–
Anderson-Darling test	–	1.4e-09	1.2e-09	1.3e-09	–
Little's MCAR-test	–	–	–	–	< 1e-10
Hawkins' test	–	–	–	–	< 1e-10
Matching: mean difference in samples before and after matching on auxiliary variables					
Mahalanobis matching	0.00	0.00	0.00	2.15	–
Propensity score matching	0.03	0.01	-1.59	2.09	–
Coarsened exact matching	0.00	0.00	0.01	2.23	–
Exact matching	0.00	0.00	0.00	2.17	–
Representativity indicators					
R-indicators for propensity model *without* income classes					
Global	0.62	0.62	0.62	–	–
Unconditional	0.05	0.18	0.01	–	–
Conditional	0.05	0.18	0.04	–	–
R-indicators for propensity model *with* income classes					
Global	0.77	0.77	0.77	0.77	–
Unconditional	0.03	0.10	0.00	0.05	–
Conditional	0.01	0.10	0.02	0.05	–
Estimated error for mean income based on MSE-interval					
Lower bound	-145136	-152165	-152640	–	–
Midpoint	2227	-2424	-830	–	–
Upper bound	149590	147318	150981	–	–

Coding of variables: *Gender* is represented as a dummy variable with male=0 and female=1. For *ISCED levels* and *income classes*, the ranks of the levels/classes as presented in figure 7.1 are used as numeric values. *Age* is used as continuous variable.

Matching: Nearest neighbor/exact matching is applied with replacement, using the five closest observations from the reference sample. Coarsened exact matching is based on age classes depicted in figure 7.1 in place of the continuous age variable but leaves ISCED levels and gender unchanged.

Propensity models: A GLM with logit link is used as propensity model for computing R-indicators. The model with and without income classes as auxiliary variable are represented for illustrative purposes.

Global R-indicator: The global R-indicator is not variable-specific and, therefore, the same for all auxiliary variables under a given propensity model.

formal evaluation of selectivity are described in sections 3.4 to 3.8. The results obtained from applying these methods in the current example are presented in table 7.1. For this purpose, the ranks of the levels or classes are used as numeric values in case of ISCED as well as income, and gender is represented by a dummy variable that indicates whether a respondent is female.

From the statistical tests discussed in section 3.4, `t`-, Kruskal-Wallis, Kolmogorov-Smirnov and Anderson-Darling tests are designed for single variables. Therefore, each of them is applied to ISCED levels, age and income classes individually. Little's and Hawkins' tests assume multiple variables and are applied to these three variables jointly, but the resulting test decisions are not altered when any of the three variables is excluded. Note that for gender as a binary variable, none of these tests is reasonable. To nevertheless apply a statistical difference test for gender proportions, Fisher's (1922a) exact test is used for the binary variable although this approach is not explicitly proposed for the context of non-probability samples (cf. e.g. Witting, 1985, pp. 379 ff). Each of these tests reveals significant differences between non-probability and reference sample estimates for all respective variables it is applicable to. This underlines the differences indicated in figure 7.1.

Matching as a measure for differences which are (not) explainable by the auxiliary variables is introduced in section 3.5. In the present context, the three socio-demographic auxiliary variables are used for matching, but income as the target variable is not. Each observation in the WI is matched with five respondents in the Microcensus, considering the different methods for matching described in sections 3.5 and 3.6, i.e. Mahalanobis, propensity score, coarsened and fully exact matching. The matches are chosen with replacement. By calculating differences in the expected values of the matched observations, a measure for the bias of the conditional mean given the matching variables can be obtained from approximation 3.24. By definition, it holds that the expected difference in the auxiliary variables used for matching is zero in case of exact matches. Despite using the age classes depicted in figure 7.1 in place of the continuous age variable, coarsened exact matching achieves quite similar results. Only a negligible difference of 0.01 years in the mean age of respondents remains. Although nearest neighbor matching does not need to be exact in any of the auxiliary variables, Mahalanobis matching likewise eliminates all differences in the auxiliary variables identified in figure 7.1. In contrast, the expected differences between non-probability and reference sample are different from zero for all socio-demographic variables in case of propensity score matching. Although the differences between matched observations in terms of gender and education are too small to appear meaningful, observations in the non-probability sample are on average 1.59 years younger than their matched counterparts in the Microcensus. Therefore, propensity score matching does not fully compensate the over-representation of younger respondents in the WI. As a measure for potential selection bias in the income variable, matching indicates that respondents in the WI on average report an income class that is more than two ranks higher than that of the respective matched observations in the Microcensus. The expected difference between the WI and the Microcensus benchmark varies between 2.09 in case of propensity score and 2.23 in case of coarsened exact matching, suggesting that higher incomes are severely over-represented in the web survey even after controlling for the effects of gender, age and education. This is a confirmation and reinforcement of the findings in figure 7.1 and an indication that selectivity of the WI is MNAR when considering only the three socio-demographic auxiliary variables.

R-indicators introduced in section 3.7 are based on the variance of estimated response propensities. To calculate these indicators, two nested logit propensity models are considered in table 7.1 for comparison. The first model is based on the socio-demographic variables only, the second additionally uses income as independent variable. The global R-indicator is not specific to a single variable and, therefore, constant across all variables included in the model. Although its possible values range from below zero to one (cf. equation 3.29 and the related discussion), it is difficult to judge the severity of potential selectivity by using only this indicator, especially when it cannot be compared for different data sets. In comparison to the values obtained in the simulation, global R-indicators of 0.62 in the model without and 0.77 in the model with income indicate dependencies between the considered variables and sample selection (cf. figures 6.7 and 6.8). Yet, both values suggest different degrees of selectivity with regard to the considered variables since incorporating income as independent variable in the propensity model leads to a higher indicated representativity. Such an increase by using an auxiliary variable that is, as discussed above, most likely subject to selection bias does not seem desirable. This emphasizes the heavy dependency of R-indicators on the underlying propensity model and the caution that is required when interpreting them. For the unconditional and conditional R-indicators, a higher value indicates less representativeness. Although this makes interpretation slightly counter-intuitive in comparison to the global R-indicator (where it is the other way round), this canonical definition is used here as well (cf. section 3.7; Schouten, Cobben and Bethlehem, 2009, pp. 104 ff; Schouten, Shlomo and Skinner, 2011, pp. 236 f). For each variable, unconditional and conditional partial R-indicator are rather close to each other, and even differences between the two models are in most cases fairly small. The results indicate that education has the strongest impact on non-representativeness of the WI as measured by the underlying propensity model. In comparison, gender has a much smaller influence, followed by age in the last place of the three socio-demographic variables. When additionally used in the propensity model, income appears to be of less importance than education but more relevant than the other socio-demographic variables for the non-representativeness expressed by the model. Nevertheless, part of the selectivity that is attributed to income in this case is ascribed to each of the socio-demographic variables when income is not considered in the propensity model. Once again, this underlines the dependence of R-indicators on a specific response model and urges cautious interpretation.

An approach to estimate an interval for the error of design linear estimators in non-probability samples is introduced in section 3.8, utilizing the correlations of response indicator and target variable with a single auxiliary variable. Using each of the three socio-demographic variables for this purpose, the last part of table 7.1 represents the resulting midpoints and boundaries of these intervals for the estimated mean income. To obtain intervals that are sufficiently narrow to be of actual value for inference, very high correlations between auxiliary variable and selection mechanism as well as target variable would be required. Therefore, it is not surprising that the interval's width is quite large for all considered auxiliary variables. Since all intervals enclose zero, the possibility of no selection bias cannot be ruled out with regard to these results. In the simulation, the intervals' midpoints perform relatively well as measures of estimation error (cf. figures 6.9 and 6.10). In the present context, however, these midpoints are of considerably different magnitudes and even signs, depending on the auxiliary variable that is used for their calculation. While gender results in a midpoint of 2 227, the corresponding value of

−2 424 when using education is even larger in absolute terms but negatively signed. Using age, one obtains a value of −830 that lies somewhere in between the others but is still negative. Since the findings of the graphical evaluation as well as matching indicate a clearly positive bias for the mean income in the WI, the only midpoint that really matches with this result is obtained from gender. Such ambiguous results can occur because the multivariate dependencies between selection mechanism, auxiliary and target variables are ignored for these intervals, which are exclusively based on bivariate correlations. These findings illustrate the issues of directly interpreting the interval's midpoint. It is only an exact measure of estimation error when selection mechanism and target variable are conditionally independent given the auxiliary variable and when variances and correlations are known (cf. equations 3.38 and 6.7 as well as the related discussion).

Although some of the results in figure 7.1 and table 7.1 have to be interpreted with caution, they altogether provide an indication that the WI deviates from the Microcensus estimates with regard to each of the four considered variables. These findings are backed by the results in various publications focusing on the WI in different countries and time periods. Specifically for Germany, Steinmetz, Tijdens and Pedraza (2009, pp. 25 ff) find differences which are mostly substantial for the four variables considered above when comparing the distributions in the WI with benchmark data. Similar results are obtained by Steinmetz et al. (2014, pp. 279 ff) for the Netherlands and by Pedraza, Tijdens and Bustillo (2007, pp. 14 ff) for Spain. Considering a larger variety of countries from different continents, Tijdens and Steinmetz (2016) come to the same conclusion for the socio-demographic variables in ten developing countries, while the results for income are underpinned by the findings of Smyk, Tyrowicz and Van der Velde (2021, pp. 446 f) for 17 industrialized and developing countries. Therefore, the considerable differences between the distribution of respondents and the respective target population found in the present example are regularly emphasized in other analyses of the WI as well.

Such differences clearly are an indication for selectivity and often straightforwardly referred to as selection bias due to non-probability sampling (cf. e.g. Schillewaert and Meulemeester, 2005, p. 177; Steinmetz, Tijdens and Pedraza, 2009, pp. 25 f). Although it is usually justified to trust benchmarks obtained from a probability sample more than estimates from a non-probability sample, such interpretations often ignore alternative potential explanations. Even perfectly unbiased estimates can be far off from some known benchmarks solely due to sampling variance (cf. e.g. Särndal, Swensson and Wretman, 1992, p. 41). Therefore, such random variation alone may to some extent explain deviations between the non-probability sample estimates and the benchmarks. However, the systematic patterns and magnitudes of these differences in figure 7.1 and table 7.1 seem too considerable to be attributed to random variation alone (cf. Pasek, 2016, p. 283; Steinmetz et al., 2014, p. 288; Yeager et al., 2011, p. 27). But even systematic deviations between WI and Microcensus are not necessarily exclusively attributable to bias that is due to non-probability sampling. The discrepancies may also be caused by other sources of error, such as effects of questionnaire design, satisficing or social desirability. This is especially true since both samples are obtained by means of different survey modes, for which magnitudes and directions of the various sources of estimation error are typically not the same. Yet, it is usually not possible to differentiate between these potential explanations when only a single non-probability and reference sample are available (cf. Bethlehem and Biffignandi, 2012, pp. 97 ff, 242; Buelens, Burger and van den Brakel, 2018, p. 327; Groves, 1989, pp. 295 ff; Weisberg, 2005). This potential coincidence of multiple sources of error is the reason why

simulations are considered preferable for the methodological evaluation and comparison in chapter 6. As selection bias is the most serious concern in the general context of non-probability samples and for the WI in particular, it thus has to be assumed that differences between the WI and the Microcensus are an adequate measure for this bias in the current example. This is only the case if the contamination by other sources of error is negligible (cf. Bethlehem and Biffignandi, 2012, pp. 303 ff; Buelens, Burger and van den Brakel, 2018, p. 327; Smyk, Tyrowicz and Van der Velde, 2021, p. 435; Yeager et al., 2011, p. 27).

7.2 Point Estimation

A variety of approaches for estimation from non-probability samples in presence of such selectivity is introduced in chapter 5. The following discussion provides an overview of the results obtained by applying these methods to the German WageIndicator web survey data from 2012. As outlined above, monthly net earned income is the target variable, and the three auxiliary variables used for estimation (i.e. prediction and weighting) are gender, age and ISCED levels. As a measure of estimation accuracy, the estimated distribution of the target variable (income) is compared with the benchmark distribution, which is based on design-based estimates for total income measured in the Microcensus as above (cf. section 2.2). As in figure 7.1, the following analysis is focused on income classes because these are the only information on income that is available from the Microcensus. To facilitate a fair and realistic comparison of the considered methods, this auxiliary information on income is not used for estimation but solely for measuring accuracy, although one could e.g. use calibration to exactly meet this distribution (cf. section 5.2).

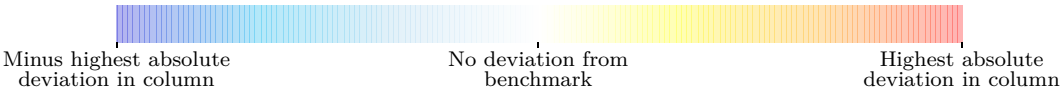
The results for the purely model-based approaches are presented in table 7.2. The considered prediction models are those introduced in sections 5.1 and 5.3, using the same specifications as in section 6.3. All these models are fit to the WI data and used for mass imputation of the monthly net earned income in the Microcensus. The estimates obtained from each of these models are compared with the benchmark frequencies for each of the income classes. To ease analysis and interpretation, the benchmarks are represented at the very top of the table. Denoted below are the deviations of the model-based estimates from this benchmark distribution in rounded percentage points for all 13 prediction models. Negative deviations, which correspond to underestimating the respective frequency, are shaded in blue. Positive deviations correspond to overestimation of the class frequency and are highlighted in red. For each income class, the highest magnitude of deviation from the benchmark frequency is indicated by the darkest color, ranging from dark blue for minus the highest absolute deviation to dark red for the highest absolute deviation. Lighter colors represent values in between these limits, with white indicating a perfect congruence of estimates and benchmark. The last column represents the mean absolute error over all income class frequencies.

It is noteworthy that all of the models except matching (which is based on the Mahalanobis distance in this case) can result in predicted income values below zero. Such values do not exist in the actual Microcensus (and also not in the WI) data and are, thus, implausible as imputed values. If not specifically accounted for, this can be a serious drawback of the model-based approaches, but fortunately, most of the considered models yield only few or no predictions below zero in the current example. However, a substantial number of such implausible values occurs when applying the GAMM or the Heckman model, where about 81 and 62% of all predictions are smaller than zero, respectively. Therefore, these models

Table 7.2: Results for model-based estimation in the German WageIndicator web survey 2012

Deviation from benchmark distribution, in percentage points (rounded)

unweighted WI	0.0	-20	-2.7	-4.3	-4.9	-3.0	0.2	2.0	4.5	5.0	6.9	5.7	3.1	1.9	1.5	1.4	1.0	0.7	0.1	0.1	0.0	0.0	-0.1	0.0	0.7	2.8
Matching	0.0	-21	-3.3	-6.8	-5.8	-3.7	-3.3	-0.9	2.6	5.6	10.4	7.1	2.4	11.4	3.1	0.6	-0.5	-0.7	-0.4	-0.3	0.0	0.5	0.2	1.5	1.0	3.7
GLM (OLS)	0.1	-21	-2.5	-5.4	-6.8	-6.5	-5.2	-3.7	-3.8	-2.5	2.9	4.8	8.3	9.8	7.3	8.3	7.4	5.1	2.6	1.5	0.0	-0.3	-0.2	-0.2	-0.1	4.6
GLM (Ridge)	0.0	-21	-3.0	-5.0	-6.8	-6.6	-5.2	-3.6	-3.7	-2.6	3.1	5.2	8.1	9.9	7.4	8.6	7.2	4.8	2.7	1.1	-0.1	-0.3	-0.2	-0.2	-0.1	4.7
GLM (LASSO)	0.1	-21	-2.5	-5.4	-6.8	-6.5	-5.2	-3.7	-3.7	-2.6	3.1	4.7	8.3	9.8	7.4	8.5	7.2	5.1	2.6	1.5	0.0	-0.3	-0.2	-0.2	-0.1	4.7
GLM (Elastic net)	0.1	-21	-2.5	-5.4	-6.8	-6.5	-5.2	-3.7	-3.7	-2.6	3.1	4.7	8.3	9.8	7.4	8.5	7.2	5.1	2.6	1.5	0.0	-0.3	-0.2	-0.2	-0.1	4.7
GAM (fix)	0.1	-21	-2.5	-5.4	-6.8	-6.5	-5.2	-3.7	-3.8	-2.5	2.9	4.8	8.3	9.8	7.3	8.3	7.4	5.1	2.6	1.5	0.0	-0.3	-0.2	-0.2	-0.1	4.6
Regression tree	0.0	-21	-3.3	-6.8	-7.3	-7.8	-8.0	-8.1	34.1	-6.2	-2.8	-5.2	45.1	-2.0	-1.9	-1.5	-0.9	-0.9	-0.5	-0.3	-0.2	-0.3	-0.2	6.0	-0.1	6.8
MARS	0.1	-21	-2.3	-5.6	-5.9	-5.5	-5.9	-4.9	-4.4	36.6	-4.6	-2.3	34.5	-2.0	-1.9	-1.5	-0.9	-0.9	-0.5	-0.3	-0.2	-0.3	-0.2	-0.2	-0.1	5.7
ANN (opt.)	0.4	-20	-2.4	-3.0	-6.7	-3.1	-4.7	-4.1	-3.7	-1.4	6.4	2.4	3.7	10.3	6.5	6.7	5.1	6.1	1.1	1.7	-0.2	-0.3	-0.2	-0.2	-0.1	4.0
SVM	0.0	-21	-3.3	-6.8	-6.9	-4.4	-4.1	-1.1	7.6	5.5	14.6	10.4	13.1	3.3	-1.9	-1.5	-0.9	-0.9	-0.5	-0.3	-0.2	-0.3	-0.2	-0.2	-0.1	4.4
GLMM	0.1	-21	-2.5	-5.4	-6.8	-6.5	-5.2	-3.7	-3.8	-2.5	2.9	4.8	8.3	9.8	7.3	8.3	7.4	5.1	2.6	1.5	0.0	-0.3	-0.2	-0.2	-0.1	4.6
GAMM (fix)	80.7	- 2	-3.3	-6.8	-7.3	-7.8	-8.0	-8.1	-7.3	-6.2	-6.8	-5.2	-3.5	-2.0	-1.9	-1.5	-0.9	-0.9	-0.5	-0.3	-0.2	-0.3	-0.2	-0.2	-0.1	6.5
Heckman model	62.2	-15	5.1	0.4	-2.5	-4.2	-5.7	-6.9	-6.4	-5.0	-5.4	-4.5	-3.3	-1.9	-1.9	-1.5	-0.9	-0.9	-0.5	-0.3	-0.2	-0.3	-0.2	-0.2	-0.1	5.4



Mean abs. err.: Mean absolute error over all income classes
unweighted WI: Unweighted (observed) frequencies in the WI
(fix): non-parametric, fixed knots
(opt.): non-parametric, optimized knots

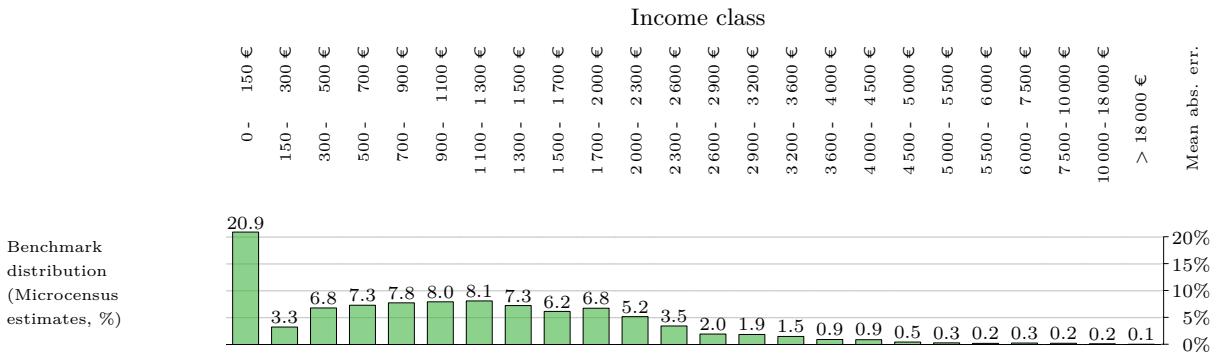
are rather unreliable in the present context, which is also indicated by their mean absolute error being considerably higher than for most other models. Except for this undesirable peak in imputed incomes being smaller than zero, only few predictions in the income range below 1 300€ are obtained from any of the models. This is a direct consequence of higher incomes being over-represented in the WI (cf. figure 7.1 and table 7.1) and leads all model-based methods to tremendously underestimate the share of persons with plausible lower incomes. Extending the view to higher incomes as well, similar problems occur for the upper tail of the distribution. Briefly speaking, nearly all prediction models tend to neglect the tails and attribute more mass towards the center of the distribution in comparison to the benchmarks. This is again caused by the fact that unexplained variation is ignored for prediction (cf. e.g. figure 6.12) and constitutes an example for the 'regression towards the mean' effect (cf. e.g. Galton, 1886; Samuels, 1991). As a consequence, none of the model-based methods actually improves the mean absolute estimation error in comparison to using the unweighted frequencies observed in the WI, which are presented in figure 7.1 and considered in the first row ('unweighted WI') of table 7.2 for comparison.

Some differences between the prediction models are nevertheless worth mentioning. Matching is the model-based approach that suffers least from the regression towards the mean effect and the only model that results in systematic over-estimation of frequencies at the upper tail of the income distribution. Estimates obtained by matching are therefore relatively close to the benchmark frequencies especially for incomes between 500 and 6 000€, yielding the lowest mean absolute error of all prediction models. A model that works comparatively well especially for the very lower tail of the distribution is the semi-parametric ANN incorporating knot optimization, which outperforms all other considered methods for income classes below 500€. The proposed ANN therefore seems to be well-suited for predicting low incomes but, as most other prediction models, suffers from the regression towards the mean effect. Nevertheless, the resulting mean absolute error over all income classes is the second lowest of all prediction models. Between all considered variants of linear (mixed) models, there are only minor differences. All these models, and also the additive model, strongly exhibit the regression towards the mean effect outlined above and therefore heavily over-estimate the frequencies of all income classes between 1 700 and 5 500€. This over-estimation is strongly linked to the over-representation of the medium to high income classes in the WI (cf. figure 7.1) and leads to mean absolute prediction errors which are medium in comparison to the other models. Slightly distinct from this pattern are the results obtained from regression trees and MARS, which both result in a large number of predictions falling into only two income classes. As in the simulation study (cf. e.g. figure 6.11), the use of higher order splines in the general MARS model results in improvements over regression trees, which use base functions of zeroth-order (cf. section 5.1.7). Nevertheless, both of these models respectively result in the third largest and largest mean absolute error. This is mostly caused by the fact that all auxiliary variables except age are categorical, while MARS and regression trees are designed in particular for continuous independent variables. Therefore, they seem to be of limited applicability for imputing the monthly net earned income in the present context. Support vector regression likewise results in severe over-estimation of frequencies in certain income ranges, but the predictions are less concentrated within single classes than in case of regression trees and MARS. Considering the mean absolute error, this SVM yields results that are slightly better than for linear and additive (mixed) models, but worse than for matching and the ANN.

In summary, none of the model-based methods is able to reliably reduce selection bias when estimating the income distribution from the WI. Therefore, even the unweighted frequencies from this non-probability sample are in most cases closer to the benchmark distribution than mass imputation estimates. An important reason for this pitfall of model-based approaches is the regression towards the mean effect. When predictions are used as imputed values, which is typically the case for non-probability samples (cf. e.g. Buelens, Burger and van den Brakel, 2018, p. 330; Kim et al., 2018, p. 7), the conditional (or residual) variance of the dependent variable given the auxiliaries is ignored. As discussed in section 6.3, it is therefore advisable for mass imputation to also consider variance components that are not explained by the model (cf. Elliott and Valliant, 2017, p. 261), for example by means of multiple imputation methods (cf. e.g. van Buuren, 2018, pp. 63 ff; Little and Rubin, 2019, p. 72; Rubin, 1987, p. 159).

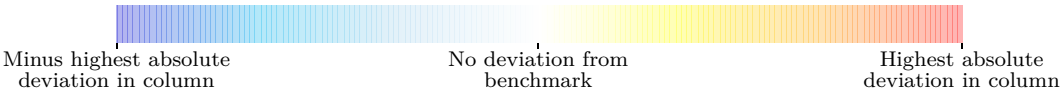
The results for pseudo-design-based estimation are presented below, following the same structure as for the model-based approaches in table 7.2. As introduced in section 5.2, the proposed methods to determine pseudo-design weights can use different types of auxiliary

Table 7.3: Results for pseudo-design-based estimation in the German WageIndicator web survey (weighting methods without response propensity model)



Deviation from benchmark distribution, in percentage points (rounded)

Calibration benchmarks: None																								
unweighted WI	-20	-2.7	-4.3	-4.9	-3.0	0.2	2.0	4.5	5.0	6.9	5.7	3.1	1.9	1.5	1.4	1.0	0.7	0.1	0.1	0.0	0.0	-0.1	0.0	0.7	2.8
Calibration benchmarks: Totals																								
GREG	-20	-3.3	-5.2	-5.0	-2.4	0.4	-1.0	2.1	4.9	4.8	6.8	4.6	3.8	2.1	2.6	1.8	2.0	0.2	0.2	0.0	0.1	-0.1	0.0	0.7	3.0
cal. ANN (1 par./obs.)	-20	-2.9	-4.9	-5.4	-3.3	-0.2	2.2	5.3	5.4	7.7	6.4	3.5	1.7	1.5	1.2	1.0	0.6	0.0	0.0	-0.1	-0.1	-0.1	0.0	0.6	3.0
cal. ANN (par.)	-20	-3.1	-4.8	-5.2	-2.9	0.0	0.5	3.9	5.8	6.6	6.9	4.2	2.8	1.5	1.7	0.8	0.8	-0.1	-0.1	-0.1	0.1	-0.1	-0.1	0.7	2.9
cal. ANN (fix)	-20	-3.1	-4.8	-5.1	-2.7	0.1	0.5	3.7	5.5	6.4	6.8	4.2	2.8	1.7	1.7	0.9	0.9	-0.1	0.0	-0.1	0.1	-0.1	-0.1	0.7	2.9
cal. ANN (opt.)	-20	-3.0	-5.2	-5.6	-3.6	-0.7	0.9	3.8	4.9	7.1	6.6	4.1	2.4	2.3	1.8	1.8	1.5	0.3	0.2	0.0	-0.1	0.0	0.0	0.7	3.1
Calibration benchmarks: Covariances																								
GREG	-20	-2.9	-5.0	-5.3	-3.2	-0.5	1.7	4.1	5.4	7.4	6.3	3.7	2.1	1.7	1.4	1.3	1.0	0.2	0.1	0.0	-0.1	0.0	0.0	0.7	3.0
cal. ANN (1 par./obs.)	-20	-2.9	-4.5	-5.2	-3.2	-0.4	1.5	4.8	5.2	7.5	6.8	3.5	2.1	1.6	1.3	0.9	0.6	0.0	0.0	-0.1	-0.1	-0.1	0.0	0.7	2.9
cal. ANN (par.)	-20	-3.1	-5.5	-6.5	-5.6	-1.8	-0.4	4.2	6.9	8.2	8.7	5.6	3.6	1.4	2.1	1.0	0.8	-0.1	-0.1	-0.2	0.0	-0.1	-0.1	0.8	3.5
cal. ANN (fix)	- 6	-1.5	5.6	2.4	-5.4	-4.3	-3.0	9.7	-2.1	12.7	-1.7	-1.5	-0.9	-1.0	-0.8	-0.5	-0.5	-0.3	-0.2	-0.2	-0.2	-0.2	-0.1	0.1	2.4
cal. ANN (opt.)	-20	-3.0	-5.6	-5.6	-3.7	-1.4	0.9	3.1	5.4	7.9	7.2	5.0	2.7	2.4	2.9	1.2	0.4	-0.1	0.1	-0.1	0.0	-0.2	0.0	0.6	3.2
Calibration benchmarks: Totals and covariances																								
GREG	-20	-3.3	-6.8	-6.4	-4.0	-1.3	-2.1	0.4	5.1	6.0	8.4	6.9	3.8	2.9	3.5	3.0	2.5	0.5	0.2	0.1	-0.1	0.0	-0.1	0.8	3.5
cal. ANN (1 par./obs.)	-20	-3.3	-7.9	-7.1	-4.7	-3.1	-0.8	1.9	6.2	8.9	9.0	7.5	2.7	3.2	3.1	2.3	1.3	0.2	-0.1	0.0	-0.1	0.0	-0.2	1.0	3.8
cal. ANN (par.)	-20	-2.7	-2.6	-1.4	3.9	5.7	5.2	5.5	5.3	3.7	2.2	-0.6	-0.9	-0.8	-0.9	-0.4	-0.8	-0.3	-0.3	-0.1	-0.3	-0.2	0.0	0.2	2.5
cal. ANN (fix)	-20	-2.7	-4.3	-4.9	-2.9	-0.6	-0.8	1.3	4.4	5.3	7.1	4.9	4.4	3.2	1.9	1.1	1.6	-0.1	0.1	-0.1	0.4	-0.1	-0.1	0.9	2.9
cal. ANN (opt.)	-20	-3.0	-5.6	-5.6	-3.6	-1.3	0.7	3.4	5.1	7.8	7.4	5.0	2.5	2.6	2.7	1.2	0.5	-0.1	0.2	-0.1	0.0	-0.2	0.0	0.7	3.2



Mean abs. err.: Mean absolute error over all income classes
unweighted WI: Unweighted (observed) frequencies in the WI
cal. ANN: calibrated ANN
(1 par./obs.): one parameter per observation (as for the GREG)
(par.): parametric
(fix): non-parametric, fixed knots
(opt.): non-parametric, optimized knots

information, i.e. benchmarks for calibration of totals and covariances (or correlations) as well as the reference sample for modeling the response process. The corresponding results for the WI are therefore grouped with respect to the type of auxiliary information that is used for weighting. As outlined above, the three socio-demographic auxiliary variables (gender, age and education) are used as independent variables for the response model as well as for calibration, and calibration benchmarks correspond to population estimates obtained from the Microcensus.

In table 7.3, the results for pseudo-design-based estimates that solely use calibration to incorporate auxiliary information are considered. The blocks in this table are used to group estimates with regard to the type of calibration benchmarks that are applied, where the first block contains unweighted WI estimates, which do not use any auxiliary information. As already discussed above, the results in this unweighted case indicate that higher incomes are over-represented in the WI (cf. figure 7.1). The consequence is an under-estimation of frequencies for all income classes below and an over-estimation for nearly all classes above 900€, a pattern which carries over to most pseudo-design-based strategies presented in table 7.3. The second block encompasses weighting methods that exclusively use total calibration, all of which tend to over-estimate the proportions of medium to high income classes. In general, the magnitude of this bias is slightly higher than for the unweighted estimates but far less severe than for the model-based estimates (cf. table 7.2). From all methods using total calibration, the calibrated ANN performs best in terms of the mean absolute error over all income classes when using a parametric or fixed-knot specification and worst when optimized knots are used. However, there are only rather small differences between these two settings. Lying somewhere in between these two cases, the GREG as well as the calibrated ANN resembling it by using one parameter per observation (cf. sections 6.2 and 5.2) exhibit a highly similar pattern of estimated frequencies. However, none of the weighting methods that use only total calibration seems clearly preferable over unweighted estimation from the non-probability sample. These findings are quite similar for most of the pseudo-design-based estimates that exclusively use covariance calibration, which are presented in the third block. However, the calibrated semi-parametric ANN with fixed B-spline knots is an exception because the pseudo-design weights obtained from it largely counterbalance the positive bias for frequencies of medium to high income classes. The corresponding deviations from the benchmark frequencies are therefore in most cases considerably smaller than for almost all other estimation approaches. The same holds for the mean absolute error over all income classes, which also outperforms that of the unweighted estimates. When jointly using total and covariance calibration, as presented in the fourth block, the results for the different weighting methods are a bit more diverse than when using either type of the benchmarks on its own. For the GREG, overestimation of frequencies for medium to high incomes is more severe than in the first blocks, and this also holds for the mean absolute error. This pattern is even more pronounced for the calibrated ANN that uses one parameter per observation and is, therefore, closely linked to the GREG (cf. section 6.2.2). In comparison, estimation errors for the parametric specification of such an ANN are mostly smaller than for all other methods that use the same auxiliary information. The resulting mean absolute error is, thus, the second lowest of all approaches considered in table 7.3. When using a calibrated semi-parametric ANN with fixed or optimized knot positioning, under- and over-estimation of class frequencies is more severe than for the parametric specification but less serious than when using one parameter per observation or the GREG.

Comparing the three types of calibration, covariance benchmarks are of limited use for the GREG. However, these benchmarks can increase efficiency for calibrated ANNs, which then outperform the GREG when specified correctly. In summary, exclusively relying on covariance benchmarks for fitting the semi-parametric calibrated ANN with fixed knots performs best in terms of the mean absolute error over all income classes. Only slightly worse is the parametric ANN incorporating total as well as covariance constraints. Since the semi-parametric specification is less rigid than the parametric one, this is an indication that stronger structural assumptions for calibrated ANNs can help to counterbalance overfitting and instability of weights in presence of more calibration constraints that are harder to meet (cf. also section 5.1.11).

Complementary to table 7.3, pseudo-design-based methods that additionally or alternatively incorporate an explicit propensity model for non-probability sample membership are considered in table 7.4. These approaches additionally use the individual observations in the Microcensus rather than only aggregated calibration benchmarks. The results are structured in the same manner as before. The first block encompasses pure propensity models that do not apply any calibration constraints. Unweighted frequencies observed in the WI are included for comparison as before, exhibiting an over-representation of medium to high incomes. Again, this selectivity pattern carries over to all pseudo-design-based methods considered in table 7.4, such that frequencies of higher incomes are over-estimated in all cases. Indeed, the weighting approaches contained in the first block perform mostly similar to the plain unweighted estimates obtained from the non-probability sample and are not clearly better than this rather naive reference approach. Inverse response propensities as well as pseudo-weights are obtained from a generalized linear logit model (cf. section 5.2.1). Differences between both as well as between parametric or non-parametric model specifications are rather small. Consequently, the resulting mean absolute error is almost the same for all of the four cross-combinations resulting from these choices. This finding underlines the importance of the additional assumptions that are required for pseudo-weights to outperform propensity weights. Following Elliott and Valliant (2017, p. 257), beta regression is used to model the probabilities of persons in the WI to be in the Microcensus, which are required for computing the pseudo-weights (cf. section 5.2.1; Ferrari and Cribari-Neto, 2004). This is necessary because the Microcensus' sampling design is not adequately described by the auxiliary variables used in the response model or any other overlapping variables between both data sets (cf. Statistisches Bundesamt, 2017; Tijdens et al., 2010). It is presumably due to the additional imprecision introduced by this strategy that the strong advantages of pseudo-weights found in the simulation (cf. section 6.3.2.2) are not apparent for the current application example. Although closely related, all ANNs that are used as plain propensity models in this first block perform slightly worse than the logit or pseudo-weights but similar to each other in terms of the mean absolute error. In contrast to the results in figure 6.13, it seems better to actually use second derivatives when fitting the propensity model in this case. When incorporating total calibration in the second block, calibrated ANNs perform fairly poor when using a parametric specification or a non-parametric one with fixed B-spline knots. The use of optimized rather than fixed knots results in improvements, but a combination of the (especially semi-parametric) logit model and the GREG still appears to be more adequate. However, none of the results is clearly better than the reference case of unweighted frequencies obtained from the WI, and the additional use of total benchmarks does not clearly improve performance in comparison to pure propensity weighting in the

Table 7.4: Results for pseudo-design-based estimation in the German WageIndicator web survey 2012 (weighting methods with response propensity model)

Income class

	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1 100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €	Mean abs. err.
Benchmark distribution (Microcensus estimates, %)	20.9	3.3	6.8	7.3	7.8	8.0	8.1	7.3	6.2	6.8	5.2	3.5	2.0	1.9	1.5	0.9	0.9	0.5	0.3	0.2	0.3	0.2	0.2	0.1	

Deviation from benchmark distribution, in percentage points (rounded)

Calibration benchmarks: None

unweighted WI	-20	-2.7	-4.3	-4.9	-3.0	0.2	2.0	4.5	5.0	6.9	5.7	3.1	1.9	1.5	1.4	1.0	0.7	0.1	0.1	0.0	0.0	-0.1	0.0	0.7	2.8
Logit (par.)	-20	-3.1	-4.4	-4.6	-2.3	0.9	1.2	4.8	5.6	6.3	6.3	3.5	2.3	1.1	1.4	0.6	0.5	-0.2	-0.1	-0.1	0.0	-0.2	-0.1	0.6	2.8
Pseudo-Weights (par.)	-20	-2.9	-4.3	-4.6	-2.3	1.2	2.2	5.0	4.9	6.4	5.4	3.1	1.9	1.2	1.1	0.7	0.6	0.0	0.0	-0.1	0.0	-0.2	0.0	0.6	2.7
cal. ANN (par.)	-20	-2.9	-5.0	-5.5	-3.5	-0.6	1.5	4.5	4.9	7.5	6.7	3.9	2.1	2.0	1.6	1.3	0.9	0.1	0.1	0.0	-0.1	-0.1	0.0	0.7	3.0
Logit (fix)	-20	-2.9	-4.1	-4.7	-3.2	0.5	0.7	4.7	5.5	6.4	6.6	4.2	2.4	0.9	1.4	0.6	0.5	-0.2	-0.1	-0.1	0.0	-0.2	-0.1	0.6	2.8
Pseudo-Weights (fix)	-20	-2.9	-4.2	-4.6	-2.3	1.8	3.2	6.2	5.1	6.4	4.9	2.6	1.4	0.7	0.6	0.4	0.2	-0.1	-0.1	-0.1	-0.1	-0.2	0.0	0.6	2.7
cal. ANN (fix)	-20	-2.9	-5.0	-5.5	-3.5	-0.6	1.5	4.5	4.9	7.5	6.7	3.9	2.1	2.0	1.6	1.3	0.9	0.1	0.1	0.0	-0.1	-0.1	0.0	0.7	3.0
cal. ANN (opt.)	-20	-2.9	-5.0	-5.5	-3.5	-0.6	1.5	4.5	4.9	7.5	6.7	3.9	2.1	2.0	1.6	1.3	0.9	0.1	0.1	0.0	-0.1	-0.1	0.0	0.7	3.0

Calibration benchmarks: Totals

Logit (par.) and GREG	-20	-3.2	-5.1	-5.1	-2.3	-0.3	-0.5	2.5	5.0	5.9	7.3	4.8	3.4	2.4	2.3	1.1	1.3	0.0	0.0	0.0	0.2	-0.2	-0.1	0.8	3.0
cal. ANN (par.)	-21	-2.0	-5.1	-7.3	-3.7	-3.9	-6.9	-6.6	7.2	8.0	9.6	2.4	1.6	4.6	7.2	1.7	1.7	2.7	3.0	1.2	1.2	-0.2	-0.2	4.6	4.5
Logit (fix) and GREG	-20	-3.1	-4.5	-4.8	-2.5	0.5	0.8	4.1	4.8	5.9	6.3	5.1	2.4	1.4	1.5	0.8	0.8	-0.1	0.0	-0.1	0.1	-0.2	-0.1	0.7	2.8
cal. ANN (fix)	-21	-2.2	-5.2	-7.3	-3.2	-3.6	-6.8	-5.9	8.4	8.1	10.1	2.2	1.7	4.5	6.3	1.4	1.4	2.4	2.7	1.0	1.0	-0.2	-0.2	4.3	4.4
cal. ANN (opt.)	-20	-2.9	-5.0	-5.5	-3.5	-0.6	1.5	4.4	4.9	7.5	6.7	3.9	2.1	2.0	1.6	1.3	0.9	0.1	0.1	0.0	-0.1	-0.1	0.0	0.7	3.0

Calibration benchmarks: Covariances

Logit (par.) and GREG	-20	-2.9	-4.8	-5.2	-3.1	0.0	2.0	4.5	4.9	7.0	6.1	3.3	1.9	1.7	1.3	1.3	1.0	0.2	0.2	-0.1	-0.1	0.0	0.0	0.6	2.9
cal. ANN (par.)	-20	-2.7	-4.5	-5.8	-4.6	-2.1	-0.4	3.5	4.2	9.0	8.7	4.7	2.8	2.6	1.9	1.2	0.8	0.2	0.1	0.0	-0.1	-0.2	0.0	0.8	3.2
Logit (fix) and GREG	-20	-2.8	-4.3	-4.8	-3.0	0.1	1.9	4.4	5.0	6.8	6.1	3.4	2.0	1.4	1.3	1.1	0.7	0.1	0.1	-0.1	-0.1	-0.1	0.0	0.6	2.8
cal. ANN (fix)	-20	-2.9	-4.6	-4.7	-1.4	0.9	2.4	4.3	6.6	7.1	5.6	3.0	1.4	0.7	0.4	0.7	0.2	0.0	-0.1	-0.1	-0.2	0.0	-0.1	0.6	2.7
cal. ANN (opt.)	-20	-3.0	-5.2	-5.6	-3.6	-0.9	1.1	4.1	5.0	7.8	7.4	4.2	2.3	2.1	1.8	1.1	0.7	0.0	0.1	0.0	-0.1	-0.1	0.0	0.7	3.1

Calibration benchmarks: Totals and covariances

Logit (par.) and GREG	-21	-3.3	-5.7	-4.5	-1.4	1.0	-2.8	1.3	3.6	3.9	6.8	4.5	5.7	2.5	3.8	2.4	2.3	0.2	0.0	0.0	0.0	-0.1	0.0	0.7	3.1
cal. ANN (par.)	-21	-2.0	-5.1	-7.3	-3.7	-3.9	-6.9	-6.6	7.2	8.0	9.6	2.4	1.6	4.6	7.2	1.7	1.7	2.7	3.0	1.2	1.2	-0.2	-0.2	4.6	4.5
Logit (fix) and GREG	-20	-3.0	-4.6	-5.1	-3.0	0.5	0.1	3.8	5.2	6.1	7.4	3.0	3.7	1.2	2.1	1.2	1.1	0.0	0.0	-0.2	0.1	-0.2	-0.1	0.7	2.9
cal. ANN (fix)	-21	-2.2	-5.2	-7.3	-3.2	-3.6	-6.8	-5.9	8.4	8.1	10.1	2.2	1.7	4.5	6.3	1.4	1.4	2.4	2.7	1.0	1.0	-0.2	-0.2	4.3	4.4
cal. ANN (opt.)	-20	-3.0	-5.3	-5.6	-3.6	-1.0	1.0	4.0	5.1	7.9	7.5	4.4	2.3	2.2	1.9	1.1	0.6	0.0	0.1	0.0	-0.1	-0.1	0.0	0.8	3.1

Minus highest absolute deviation in column — No deviation from benchmark — Highest absolute deviation in column

Mean abs. err.: Mean absolute error over all income classes
unweighted WI: Unweighted (observed) frequencies in the WI
cal. ANN: calibrated ANN
Logit: Weights from GLM with logit link
Logit and GREG: Weights from GLM with logit link, calibrated using the GREG

(par.): parametric
(fix): non-parametric, fixed knots
(opt.): non-parametric, optimized knots

first block as well. The results are mostly similar when considering covariance calibration in the third block. For the combination of logit model and GREG, there are only minor differences in comparison to total calibration. While the calibrated semi-parametric ANN with optimized B-spline knots is slightly worse than in case of total benchmarks, the remaining two variants of such ANNs perform better for covariance calibration. Similar as in table 7.3, the results are particularly good for the semi-parametric ANN using fixed B-spline knots, which outperforms all other approaches in table 7.4 when considering the mean absolute error over all income classes. As before, combining total and covariance calibration in the fourth block does not improve estimates in comparison to using either total or covariance calibration on its own. The joint usage of logit model and GREG is again preferable to calibrated ANNs, but none of the methods is better than plain propensity or even unweighted estimates. Overall, the calibrated semi-parametric ANN that uses fixed B-spline knots and covariance constraints is the only method in table 7.4 that provides estimates which are better than using unweighted frequencies in terms of the mean absolute error. However, the improvement over using no weights is still only minor.

A comparison of the results in tables 7.3 and 7.4 furthermore indicates that it is better to exclusively rely on calibration constraints rather than to additionally incorporate a response propensity model for weighting. A similar finding particularly for calibrated ANNs is encountered in figures 6.16 and 6.17. A different combination of the distance function's components (cf. equation 5.155) used for fitting these models may therefore help to improve the resulting pseudo-design weights (cf. section 6.3.2). However, not only the proposed calibrated ANN but also the combination of logit model and GREG fails to gain efficiency from additionally incorporating the response propensity model. An explanation for this finding may be that the same auxiliary variables are used for calibration and propensity models. In comparison to their separate use, a combination of both weighting approaches is therefore prone to increasing the weights' variability without adding any meaningful information to compensate for selection bias.

7.3 Summary and Limitations

In summary, it is evident that there is some selectivity in the WageIndicator web survey with regard to all considered variables, and the consequences in terms of biased estimates are not easily compensated. From all the model- and pseudo-design-based methods represented in tables 7.2 to 7.4, only a few actually perform better than the unweighted non-probability sample estimates. In terms of the mean absolute error over all income class frequencies, the best of all considered estimation methods is to apply pseudo-design weighting based on the calibrated semi-parametric artificial neural network with fixed B-spline knots that only uses covariance benchmarks for fitting. This approach also outperforms all applicable combinations of model- and pseudo-design-based methods introduced in section 5.3. Since a large number (949) of such combinations is possible, the corresponding results are quite extensive, and a detailed overview is deferred to appendix D. Nevertheless, none of the examined methods seems capable of fully eliminating selection bias in the WI by using only the three socio-demographic auxiliary variables.

This finding is in line with expectations when considering the results in section 6.3 and table 7.1 because selectivity is presumably MNAR in the present example. A major reason to suspect violation of conditional independence assumption 5.1 is that gender,

age and education are the only auxiliary variables used throughout the current chapter 7. As outlined above, the main motivation for this strategy is to achieve compliance and comparability with relevant publications and the actual pseudo-design weights that are included in the WI data set. Nevertheless, exclusively relying on three socio-demographic variables presumably leads to over-simplification when assessing and compensating selectivity, which is why the bias can be reduced by some of the methods but not fully compensated. Similar results are obtained by Steinmetz et al. (2014, pp. 284 ff). An obvious alternative that may help to overcome this limitation is to use more variables to compensate for selectivity. However, available auxiliary information as well as selectivity patterns cannot be considered being equal between (groups of) countries in a worldwide web survey. Therefore, such a less simplifying approach for compensating the non-probability selection of the WI may require applying strategies that are way more country-specific and, thus, not applicable to the web survey as a whole (cf. Steinmetz et al., 2014, p. 287). Moreover, one has to watch out for potential over-fitting of weighting or prediction models in case of too many auxiliary variables (cf. also section 5.1.11).

Further limitations arise because neither population nor non-probability sampling process are fully known for the WI. To still evaluate the considered estimation methods for a single realized non-probability sample, some sort of benchmarks for the quantities to be estimated are required because an estimator's quality is typically defined with respect to such a target statistic (cf. e.g. equations 2.10 to 2.12 and 6.4 to 6.6). Since the common strategy in this case is to compare a non-probability sample with a probability sample of high quality, the Microcensus appears well-suited as a reference point for comparison (cf. Enderle, Münnich and Bruch, 2013, p. 92). As outlined above, however, the target populations of both samples do not perfectly coincide since the WI is intended to cover the German labor force, while the Microcensus covers the resident population. In addition, monthly net earned income as the target variable is only measured in the WI, such that Microcensus estimates for the monthly net total income have to be used as surrogate. In conjunction with the Microcensus estimates being subject to sampling error, the use of two single realized samples does, hence, not allow distinguishing between different potential explanations for inconsistencies between the non-probability and the reference sample (cf. Bethlehem and Biffignandi, 2012, pp. 97 ff, 242; Biffignandi and Bethlehem, 2012, p. 370; Groves, 1989, pp. 295 ff; Pasek, 2016, p. 283; Weisberg, 2005). As a consequence, it has to be assumed that differences between the WI and the Microcensus are an adequate measure for the WI's selection bias.

The problem is even more evident when considering methods for inference. While it is possible to use benchmarks obtained from external sources for point estimates, there is no similar strategy to evaluate inferential approaches in the present context. To make this possible, it would be necessary to assume that a model is able to fully deduce the true sampling process from the realized samples. Such an assumption appears implausible in the present context, in particular when considering the results in tables 7.2 to 7.4. Furthermore, assuming a model for the true selection mechanism is prone to favoring exactly this model for estimation, simply by design of the evaluation procedure (cf. e.g. Kim et al., 2018, pp. 12 ff; Setoguchi et al., 2008, p. 548; Stürmer et al., 2007, p. 1111). As a consequence, the methods for inference introduced in section 5.4 may be adequate or not in case of the WI, but the data itself does not allow assessing this issue in a reliable manner. Corresponding results for variance estimation are therefore deferred to appendix D.

In general, simulations are often more suitable than application studies to evaluate estimation methods. This is the reason why the Monte Carlo simulation studies discussed in the previous chapter 6 constitute the main focus in terms of evaluating and comparing the performance of methods in this thesis. Among other advantages, this strategy allows for a more profound assessment of inferential approaches. Nevertheless, the exemplary application to the WageIndicator web survey in the current chapter 7 emphasizes additional issues that occur in the context of real non-probability samples, ranging from the lack of perfectly adequate auxiliary information to potential sources of error being indistinguishable.

8 Conclusion and Outlook

Prevalence and relevance of non-probability samples in various areas are increasing, especially due the growing abundance and availability of new digital data sources throughout the recent years. Because an adequate use of such samples poses substantial statistical challenges, the overarching aims of this thesis are to discuss, expand and evaluate methods for dealing with the specific issues of non-probability samples in a common framework.

By pursuing these goals, the current thesis provides a unifying overview as well as new theoretical developments and empirical findings on methods for non-probability samples. Important opportunities and challenges posed by such samples are discussed in the formal theoretical context of survey sampling as well as with regard to practical applications. Various pre-existing methods to tackle these challenges are identified and reviewed. In addition, new methodological approaches are developed for this purpose. Complementing the theoretical discussion, particular attention is paid to the mathematical and algorithmic aspects required to implement the respective methods. To evaluate and compare the considered methods with regard to their performance in the context of non-probability sampling, simulation and application studies are conducted. A summary of the core theoretical and empirical findings made in this thesis is provided in the following paragraphs.

The first major challenge when using non-probability samples is to operationalize and quantify the selectivity (or synonymously non-representativity) of the respective sampling mechanisms. To tackle this issue, suitable auxiliary information is used to examine the dependency between sample inclusion and variables of interest. The basic ideas are to check for (non-)compliance of non-probability samples with relevant external information or to describe the non-probability sample inclusion and its effects using auxiliary variables.

In the simulation study, statistical tests and matching provide only little information about the quality of a non-probability sample and the corresponding estimates obtained from it. The same holds for representativity indicators in their original specification, but their performance can be considerably improved by incorporating calibration constraints in the underlying propensity model. Even more conclusive results in case of design linear estimators can be obtained by means of MSE-intervals. In the application study, manual comparisons, statistical tests and matching provide agreeing and plausible conclusions. In contrast, the results of representativity indicators and MSE-intervals are highly dependent on the choice of auxiliary variables required for their computation and difficult to interpret in absence of other non-probability samples that can be used for comparison. Therefore, these findings indicate that distinct measures of selectivity serve for different purposes. Manual comparisons, statistical tests and matching provide straightforward interpretability for a single realized non-probability sample, while representativity indicators and MSE-intervals allow better comparisons of different samples. In any case, the effect of a particular non-probability selection mechanism depends on the actual variables and estimates of interest. Unlike probability sampling, which allows for a variety of unbiased estimators with regard to arbitrary target variables, the question of quality when using a non-probability sample is therefore whether it is "fit for purpose" (Baker et al., 2013b, p. 102) concerning specific research interests.

The second fundamental challenge when using non-probability samples is to account for their potential selectivity and resulting biases in point estimation and inference. Two broader paradigms that address this challenge can be distinguished. The *model-based*

paradigm focuses on predicting information about the variables of interest outside the non-probability sample. Various statistical and machine learning methods relevant for this purpose are considered. In contrast, the *pseudo-design-based paradigm* constructs weights that are meant to mimic the design weights in probability sampling, for which different methods are discussed. Although the model- and the pseudo-design-based framework are often considered as mutually exclusive, a bandwidth of strategies to achieve a synthesis of both paradigms is presented as well. Inferential methods for non-probability samples are furthermore examined, which again refer to the ideas of these two paradigms.

The two major methodological novelties proposed in the context of this thesis contribute to either of the two paradigms outlined above. Semi-parametric artificial neural networks integrate B-spline layers and optimal knot positioning in the general structure and fitting procedure of artificial neural networks, thereby combining and extending existing prediction methods. Calibrated semi-parametric artificial neural networks constitute response propensity models of adaptable complexity that allow incorporating soft and exact calibration of totals, covariances and correlations. This is an integration and extension of fundamental pseudo-design-based concepts. Complementing their theoretical foundation, custom-made computational implementations for fitting (calibrated) semi-parametric artificial neural networks by means of (stochastic) gradient descent, `BFGS` and sequential quadratic programming algorithms are developed in `C++` as part of three `R`-packages.

The empirical results of the simulation study show that most of the discussed model-based methods allow reducing selection bias for linear estimators, e.g. for estimated means and totals. For all considered prediction models, the degree of bias reduction is heavily determined by the dependencies between the available auxiliary and target variables as well as the selection mechanism. The proposed semi-parametric ANNs perform adequately for prediction and yield results that are similar to those when using MARS models of comparable complexity and flexibility. However, simpler methods like statistical matching or linear regression mostly appear more suitable for the same purpose in the simulation. Although some of the more complex and non-linear statistical or machine learning models can perform slightly better under certain circumstances, these conditions may be hard to identify in real applications. An issue with all of the model-based methods for non-probability samples is that they are commonly implemented by imputing the conditional mean under the model (cf. e.g. Buelens, Burger and van den Brakel, 2018, p. 330; Kim et al., 2018, p. 7). Because residual (co-)variances that are not explained by the model are ignored by this strategy, it performs rather poor in case of non-linear estimators, e.g. for correlations, regression coefficients or a variable's distribution. This finding is underlined by the results of the application study, where none of the model-based approaches reliably reduces selection bias in the income class frequencies estimated from the WageIndicator web survey. As discussed below, future research may help to overcome this limitation.

Considering the simulation results for pseudo-design-based methods, non-parametric response models usually outperform parametric ones for propensity weighting. Despite being commonly used, inverse predictions from generalized linear logit models are less reliable than other approaches for propensity weighting. Especially in case of design linear estimates, pseudo-weights proposed by Elliott (2009, pp. 2 f) as well as Elliott and Valliant (2017, pp. 256 f) appear particularly valuable for this purpose. With regard to calibration weighting, the generalized regression estimator is mainly useful when calibrating to only a few known population totals. Yet, calibrated semi-parametric ANNs can yield more reliable estimates in most scenarios because the functional form approach

allows for additional incorporation of variables for which no calibration benchmarks are available. However, specifying an adequate combination of the multiple optimization criteria used for fitting such ANNs still provokes some difficulties. Therefore, a joint application of propensity model and GREG is usually more adequate than calibrated ANNs when considering a combination of propensity and calibration weighting. In the simulation, such ANNs perform best when either response or calibration weighting are used on their own. This finding is underlined by the results of the application study. For estimating income class frequencies from the WageIndicator web survey, pseudo-design weights obtained from a calibrated ANN that is fit solely using covariance benchmarks outperforms all other considered methods in terms of compensating selection bias.

As is evident from the theoretical and empirical results, the model- and the pseudo-design-based paradigm both allow (almost) fully counterbalancing selection bias when selectivity corresponds to a MAR pattern. This underlines the relevance of the conditional independence assumption discussed in chapter 5. When this assumption is violated, none of the considered methods allows for unbiased estimation. Nevertheless, the simulation results indicate that a combination of model- and pseudo-design-based strategies can still improve efficiency in such cases of MNAR selectivity. Especially when a reference sample is the only available auxiliary information, propensity weighted imputation models perform better than either propensity weighting or unweighted imputation methods on their own. In case of calibration weighting, MRP and other forms of weighted aggregation of predictions can be used to stabilize estimation. Yet, this improves performance only in presence of relatively volatile weights, e.g. if a large number of calibration benchmarks is incorporated. Considering the joint use of propensity and calibration weighting, combining model- and pseudo-design-based methods results in a loss rather than a gain in efficiency, such that a synthesis of the two paradigms does not seem to be advisable in this case. For estimation of income class frequencies from the WageIndicator web survey in the application study, the selectivity pattern is presumably again MNAR. Nevertheless, none of the combinations of model- and pseudo-design-based methods outperforms the pseudo-design weighted estimates when using the calibrated ANN that solely relies on covariance benchmarks.

With regard to inference, the findings in this thesis do not suggest a unique best approach. In the simulation study, valid inference is often easier to achieve when a point estimator's bias is comparably small in magnitude, although this is neither a necessary nor sufficient condition. For estimation that is based on mass-imputation, Monte Carlo bootstrap estimates of the total variance generally perform comparatively well, regardless of whether a weighted or an unweighted prediction model is applied. The results are far more ambiguous in case of pseudo-design-based estimates or weighted aggregation of predictions, such that the best approach for inference still depends on the actual sampling mechanism, available auxiliary information and specific point estimation method. Because the underlying non-probability selection process is typically unknown in real applications, there is no reliable strategy to evaluate inferential methods in the application study without using additional assumptions and/or simulations.

This highlights an aspect of the Monte Carlo simulation in section 6.3 that may be considered as either an advantage or drawback. This central simulation study is designed to cover a bandwidth of prototype scenarios that can occur when using non-probability samples, considering a mixture and trade-off between settings that are used in other relevant publications. A crucial advantage of this strategy is that it provides a general overview that can be interpreted with regard to a variety of types and applications of non-

probability sampling and highlights important conditions under which specific methods may be more or less suitable. A drawback of this setup is that it may be too general to allow for a fully detailed evaluation of methods with regard to a specific real non-probability sample. To precisely identify the best estimation strategy for such a particular sample, more in-depth evaluations and simulation studies considering the distinct non-probability selection mechanism and relevant variables may be required. Moreover, the reference sample's sampling design as well as the sizes of the non-probability and reference sample may have an impact on the relative advantages and drawbacks of certain methods. Further important aspects of both samples should also be taken into account, such as effects of survey modes, questionnaire designs, non-response and data editing. Such characteristics are not directly related to the issues of non-probability sampling, but can provoke additional difficulties when assessing and compensating selectivity. In general, there is no method that fits all types of non-probability samples and purposes, such that the most adequate strategy for quality assessment, point estimation and inference has to be determined for each and every particular non-probability sample, research interest and available auxiliary information.

The broad definition of non-probability sampling encompasses manifold selection processes. Due to the heterogeneous nature and evolution of these processes, research on methods for non-probability samples presumably needs to be a steadily ongoing process. This may be facilitated by pursuing the following important topics for future research.

The proposed calibrated semi-parametric ANNs perform well in certain settings. However, the weighted-sum approach used for multi-criteria optimization still poses some difficulties in specifying a general rule to define the required importance weights. Different strategies can be used for this purpose, but typical approaches in the context of weighting rely on design-based variance estimates for probability sampling and are therefore not straightforwardly applicable to non-probability samples. Hence, further research is needed to extend the promising results of calibrated neural networks to more general circumstances. Moreover, it may be beneficial to integrate pseudo-weights and calibrated ANNs in further research, as both ideas appear valuable in different settings.

In addition, various more general areas for further research on non-probability samples in the field of survey statistics exist. For example, future research could address potential dependencies between observations in certain non-probability samples that lead to violation of the i.i.d. assumption underlying most of the considered prediction or weighting models. This issue is typically covered in publications on respondent-driven sampling (cf. e.g. Frank and Snijders, 1994; Heckathorn, 2002) but may be extended towards more general methods for non-probability samples. Generalized linear and additive mixed models allow for dependencies between observations, and there are approaches for also using random effects in support vector machines (cf. Cho, 2010; Luts et al., 2012) and artificial neural networks (cf. Tran et al., 2020; Xiong, Kim and Singh, 2019). To account for such dependencies between observations in the context of pseudo-design-based methods, it may be sensible to explicitly consider the joint conditional probability of two elements being simultaneously part of the non-probability sample in the weighting model. This would lead to a pseudo-design-based analogy to second order inclusion probabilities.

To better incorporate the uncertainty of prediction or weighting models in point estimation and inference is a further important topic for future research. As outlined above, the straightforward and common use of conditional means expressed by a statistical or

machine learning model for mass-imputation in model-based estimation ignores important (co-)variance components that are not explained by the model. Similarly, treating pseudo-design weights like known design weights can lead to severe bias especially in inference for pseudo-design-based point estimates. Considering the challenges of non-probability samples as a missing data problem (cf. figure 5.1), the use of multiple imputation (cf. e.g. van Buuren, 2018; Little and Rubin, 2019; Rubin, 1987) may help to improve point estimation and inference for non-probability samples (cf. sections 5.4 and 6.3.2.3; Elliott and Valliant, 2017, p. 261; Rafei, Flannagan and Elliott, 2020, p. 160). A different but related extension to the inferential methods considered in this thesis could be based on a double bootstrap strategy. A first stage of resampling can be used to estimate the uncertainty of the prediction or weighting model parameters while a second stage may be used to represent the point estimator's variability conditional on the parameters determined in the first stage (cf. also section 5.4; Kuk, 1989; Hinkley and Shi, 1989).

Moreover, there are alternative uses and potential extensions for measures of selectivity. For example, Chambers, Dorfman and Wehrly (1993, p. 270) derive an expression for the bias in model-based estimates of finite population totals when using certain non-parametric prediction models. Assuming a particular joint distribution for non-probability sample inclusion and variable of interest, Andridge et al. (2019), Little et al. (2020) and West et al. (2021) introduce indicators for the selectivity of a non-probability sample in a Bayesian framework. These and similar approaches apply model-based assumptions already for assessments of selectivity. Their development and evaluation under realistic conditions is a further field for current and future research. In addition, a continuous non-probability sampling process often allows calculating intermediate measures of a sample's selectivity already during data collection. Future research could focus on the use of such intermediate measures for implementing adaptive non-probability sampling procedures, potentially even in real-time. For example, sub-groups that are under-represented in the current data can be specifically addressed to increase their participation, or additional relevant auxiliary variables can be identified and measured to better account for selectivity in the estimation stage (cf. Biffignandi and Pratesi, 2002, p. 65; Mercer et al., 2017, p. 266). These considerations highlight the general relevance of variable and model selection in all stages of conducting and analyzing non-probability samples and the potential use of selectivity measures for this purpose that may be enhanced in future research.

Last but not least, there are further potential applications for the methods introduced in this thesis. Because the development of calibrated (semi-parametric) ANNs is focused on the particular challenges of non-probability samples, the specific features of these models constitute an extension to well-established methods mainly in closely related research areas. For example, quasi-experimental and other observational studies (cf. e.g. Rosenbaum, 2010) as well as non-response (cf. e.g. Särndal and Lundström, 2005) can be seen as special cases of non-probability sampling for which calibrated ANNs may be applied. Even more versatile is the proposed optimization of B-spline knots in semi-parametric ANNs. Although motivated with regard to non-probability sampling in this thesis, these models can as well be used for estimation in classical probability samples (cf. e.g. Breidt and Opsomer, 2009). Furthermore, there are potential applications beyond the scope of survey statistics, for example in engineering or computer sciences (cf. e.g. Folgheraiter, 2016; Piazza et al., 1997; Wang and Lei, 2001). Application and evaluation of the newly proposed methods in such diverse contexts is therefore a further topic for future research within or beyond the context of non-probability sampling.

Statistical and Machine Learning Methods for Handling Selectivity in Non-Probability Samples

Appendix and Bibliography

to the

Doctoral Thesis

approved by the Department IV of the University of Trier
in partial fulfillment of the requirements for the degree of

Dr. rer. pol.

by

Simon Jonas Lenau, M.Sc.

Trier

Date of submission: June 07, 2021
Date of defence: December 02, 2022
Published: February 2023

Supervisors:
Prof. Dr. Ralf Münnich (Trier University)
Prof. Dr. Silvia Biffignandi (University of Bergamo)

Appendix A Mathematical Background of the BFGS-update

As described in section 4.2.3, the BFGS-algorithm is based on approximating the inverse Hessian matrix, such that $\mathbf{H}^{-1} \approx \widetilde{\mathbf{B}}$. In each iteration, the weighted Frobenius-Norm between old and new approximation is minimized, subject to the secant and symmetry conditions 4.21 and 4.22.

As in equations 4.19 and 4.20, the changes in parameters and gradient are given by

$$\begin{aligned} \boldsymbol{s} &:= \boldsymbol{\Theta}^{(a+1)} - \boldsymbol{\Theta}^{(a)} \\ \boldsymbol{y} &:= \mathbf{J}_{\mathrm{L}}\left(\boldsymbol{\Theta}^{(a+1)}\right) - \mathbf{J}_{\mathrm{L}}\left(\boldsymbol{\Theta}^{(a)}\right) \quad . \end{aligned} \tag{A.1}$$

Additionally denoting the update for the current inverse Hessian approximation used to determine the new one by

$$\widetilde{\mathbf{B}}^{\Delta} := \widetilde{\mathbf{B}}^{(a+1)} - \widetilde{\mathbf{B}}^{(a)} \quad , \tag{A.2}$$

the optimization problem is

$$\begin{aligned} \widetilde{\mathbf{B}}^{(a+1)} &= \underset{\widetilde{\mathbf{B}}^{(a+1)}}{\operatorname{argmin}} \left(\left\| \widetilde{\mathbf{B}}^{\Delta} \right\|_{\mathrm{F}_{\boldsymbol{W}}}^{2} \right) \\ \text{s. t.} \quad & \widetilde{\mathbf{B}}^{(a+1)} \boldsymbol{y} = \boldsymbol{s} \\ & \widetilde{\mathbf{B}}^{(a+1)} = \left(\widetilde{\mathbf{B}}^{(a+1)} \right)^{\mathsf{T}} \quad , \end{aligned} \tag{A.3}$$

where

$$\left\| \widetilde{\mathbf{B}}^{\Delta} \right\|_{\mathrm{F}_{\boldsymbol{W}}}^{2} = \operatorname{tr}\left(\left(\widetilde{\mathbf{B}}^{\Delta} \right)^{\mathsf{T}} \boldsymbol{W}^{\mathsf{T}} \widetilde{\mathbf{B}}^{\Delta} \boldsymbol{W} \right) \tag{A.4}$$

(cf. equations 4.21 to 4.23; Gill, Murray and Wright, 1981, pp. 121 ff; Jarre and Stoer, 2004, pp. 173 ff; Nocedal and Wright, 1999, pp. 194 ff).

The Lagrange function of optimization problem A.3 resulting from equation 4.11 is given by

$$\begin{aligned} \mathrm{L}\left(\widetilde{\mathbf{B}}^{(a+1)} \right) &= \operatorname{tr}\left(\left(\widetilde{\mathbf{B}}^{\Delta} \right)^{\mathsf{T}} \boldsymbol{W}^{\mathsf{T}} \widetilde{\mathbf{B}}^{\Delta} \boldsymbol{W} \right) + \boldsymbol{\lambda}^{\mathsf{T}} \left(\widetilde{\mathbf{B}} \boldsymbol{y} - \boldsymbol{s} \right) + \sum_{i=2}^{h} \sum_{j=1}^{i-1} \alpha_{(i/2\cdot(i-1)+j)} \left(\widetilde{\mathrm{b}}_{ij}^{a+1} - \widetilde{\mathrm{b}}_{ji}^{a+1} \right) \\ &= \operatorname{tr}\left(\left(\widetilde{\mathbf{B}}^{\Delta} \right)^{\mathsf{T}} \boldsymbol{W}^{\mathsf{T}} \widetilde{\mathbf{B}}^{\Delta} \boldsymbol{W} \right) + \boldsymbol{\lambda}^{\mathsf{T}} \left(\widetilde{\mathbf{B}} \boldsymbol{y} - \boldsymbol{s} \right) + \sum_{i=1}^{h} \sum_{j=1}^{h} a_{ij} \widetilde{\mathrm{b}}_{ij}^{a+1} \quad , \end{aligned} \tag{A.5}$$

where $\boldsymbol{W}$ is an arbitrary symmetric weighting matrix that fulfills equation 4.24, $\widetilde{\mathrm{b}}_{ij}$ is an entry of $\widetilde{\mathbf{B}}$, and

$$a_{ij} := \begin{cases} 0 & , \text{if} \quad i = j \\ \alpha_{(i/2\cdot(i-1)+j)} & , \text{if} \quad i > j \\ -\alpha_{(i/2\cdot(i-1)+j)} & , \text{if} \quad i < j \end{cases} \tag{A.6}$$

are entries of a matrix $\boldsymbol{A}$ containing the Lagrange multipliers $\boldsymbol{\alpha}$ for the symmetry condition with property

$$\boldsymbol{A}^{\mathsf{T}} = -\boldsymbol{A} \quad . \tag{A.7}$$

Consequently, the derivative follows as

$$\begin{aligned}\frac{\partial\left(\mathrm{L}\left(\widetilde{\mathbf{B}}^{(a+1)}\right)\right)}{\partial\left(\widetilde{\mathbf{B}}^{(a+1)}\right)} &= \frac{\partial\left(\left(\widetilde{\mathbf{B}}^{\Delta}\right)^{\mathsf{T}}\boldsymbol{W}^{\mathsf{T}}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W}\right)}{\partial\left(\widetilde{\mathbf{B}}^{(a+1)}\right)} + \boldsymbol{\lambda}\boldsymbol{y}^{\mathsf{T}} + \boldsymbol{A} \\ &= \frac{\partial\left(\widetilde{\mathbf{B}}^{\Delta}\right)}{\partial\left(\widetilde{\mathbf{B}}^{(a+1)}\right)}\left(\boldsymbol{W}^{\mathsf{T}}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W}\right) + \frac{\partial\left(\boldsymbol{W}^{\mathsf{T}}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W}\right)}{\partial\left(\widetilde{\mathbf{B}}^{(a+1)}\right)}\widetilde{\mathbf{B}}^{\Delta} + \boldsymbol{\lambda}\boldsymbol{y}^{\mathsf{T}} + \boldsymbol{A} \\ &= \mathbf{I}_h\left(\boldsymbol{W}^{\mathsf{T}}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W}\right) + \boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}\widetilde{\mathbf{B}}^{\Delta} + \boldsymbol{\lambda}\boldsymbol{y}^{\mathsf{T}} + \boldsymbol{A} \\ &= 2\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W} + \boldsymbol{\lambda}\boldsymbol{y}^{\mathsf{T}} + \boldsymbol{A} \quad .\end{aligned} \tag{A.8}$$

The KKT-conditions resulting from equation 4.12 are given by

$$\begin{aligned}\frac{\partial\left(\mathrm{L}\left(\widetilde{\mathbf{B}}^{(a+1)}\right)\right)}{\partial\left(\widetilde{\mathbf{B}}^{(a+1)}\right)} &= \mathbf{0}_{h\times h} \\ \widetilde{\mathbf{B}}^{(a+1)}\boldsymbol{y} &= \boldsymbol{s} \\ \widetilde{\mathbf{B}}^{(a+1)} &= \left(\widetilde{\mathbf{B}}^{(a+1)}\right)^{\mathsf{T}} \quad .\end{aligned} \tag{A.9}$$

By symmetry of $\widetilde{\mathbf{B}}^{\Delta}$ and $\boldsymbol{W}$ as well as asymmetry of $\boldsymbol{A}$, it additionally holds that

$$\left(2\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W} + \boldsymbol{\lambda}\boldsymbol{y}^{\mathsf{T}} + \boldsymbol{A}\right)^{\mathsf{T}} = 2\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W} + \boldsymbol{y}\boldsymbol{\lambda}^{\mathsf{T}} - \boldsymbol{A} \quad , \tag{A.10}$$

which can be used to eliminate $\boldsymbol{A}$ from the first KKT-condition by

$$\begin{aligned}\mathbf{0}_{h\times h} &= 2\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W} + \boldsymbol{\lambda}\boldsymbol{y}^{\mathsf{T}} + \boldsymbol{A} + 2\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W} + \boldsymbol{y}\boldsymbol{\lambda}^{\mathsf{T}} - \boldsymbol{A} \\ &= 4\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W} + \boldsymbol{\lambda}\boldsymbol{y}^{\mathsf{T}} + \boldsymbol{y}\boldsymbol{\lambda}^{\mathsf{T}} \quad .\end{aligned} \tag{A.11}$$

By definition of $\widetilde{\mathbf{B}}^{(a+1)}$ and $\boldsymbol{W}$, the equalities

$$\begin{aligned}\boldsymbol{W}\boldsymbol{s} &= \boldsymbol{y} \\ \boldsymbol{s}^{\mathsf{T}}\boldsymbol{W} &= \boldsymbol{y}^{\mathsf{T}} \\ \boldsymbol{W}^{-1}\boldsymbol{y} &= \boldsymbol{s} \\ \boldsymbol{y}^{\mathsf{T}}\boldsymbol{W}^{-1} &= \boldsymbol{s}^{\mathsf{T}} \\ \widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y} &= \left(\widetilde{\mathbf{B}}^{(a+1)} - \widetilde{\mathbf{B}}^{(a)}\right)\boldsymbol{y} = \left(\boldsymbol{s} - \widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\right)\end{aligned} \tag{A.12}$$

hold. From equalities A.11 and A.12, it follows that

$$\begin{aligned}&\left(\left(4\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W} + \boldsymbol{\lambda}\boldsymbol{y}^{\mathsf{T}} + \boldsymbol{y}\boldsymbol{\lambda}^{\mathsf{T}}\right)\boldsymbol{s}\right)^{\mathsf{T}}\boldsymbol{s} \\ &\quad= 4\left(\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W}\boldsymbol{s}\right)^{\mathsf{T}}\boldsymbol{s} + \left(\boldsymbol{\lambda}\boldsymbol{y}^{\mathsf{T}}\boldsymbol{s}\right)^{\mathsf{T}}\boldsymbol{s} + \left(\boldsymbol{y}\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{s}\right)^{\mathsf{T}}\boldsymbol{s} \\ &\quad= 4\left(\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W}\boldsymbol{s}\right)^{\mathsf{T}}\boldsymbol{s} + \boldsymbol{s}^{\mathsf{T}}\boldsymbol{y}\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{s} + \boldsymbol{s}^{\mathsf{T}}\boldsymbol{y}\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{s} \\ &\quad= 4\left(\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W}\boldsymbol{s}\right)^{\mathsf{T}}\boldsymbol{s} + 2\boldsymbol{s}^{\mathsf{T}}\boldsymbol{y}\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{s} \\ &\quad= 0 \quad ,\end{aligned} \tag{A.13}$$

and thus

$$\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{s} = -2 \cdot \frac{\left(\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}\right)^\mathsf{T}\boldsymbol{s}}{\boldsymbol{s}^\mathsf{T}\boldsymbol{y}} \quad . \tag{A.14}$$

Post-multiplying equation A.11 by $\boldsymbol{s}$ and plugging in equality A.14 yields

$$\begin{aligned} \mathbf{0}_{h\times h} &= 4\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W}\boldsymbol{s} + \boldsymbol{\lambda}\boldsymbol{y}^\mathsf{T}\boldsymbol{s} + \boldsymbol{y}\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{s} \\ &= 4\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W}\boldsymbol{s} + \boldsymbol{\lambda}\boldsymbol{y}^\mathsf{T}\boldsymbol{s} - 2\cdot\boldsymbol{y}\frac{\left(\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}\right)^\mathsf{T}\boldsymbol{s}}{\boldsymbol{s}^\mathsf{T}\boldsymbol{y}} \quad , \end{aligned} \tag{A.15}$$

which can be solved for $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = 2\boldsymbol{y}\frac{\left(\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}\right)^\mathsf{T}\boldsymbol{s}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - 4\frac{\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} = 2\boldsymbol{y}\frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - 4\frac{\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \quad . \tag{A.16}$$

Inserting equality A.16 into equation A.11 as well as pre- and post-multiplication with $\boldsymbol{W}^{-1}$ results in

$$\begin{aligned} \mathbf{0}_{h\times h} &= 4\cdot\boldsymbol{W}^{-1}\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{W}\boldsymbol{W}^{-1} + \boldsymbol{W}^{-1}\boldsymbol{\lambda}\boldsymbol{y}^\mathsf{T}\boldsymbol{W}^{-1} + \boldsymbol{W}^{-1}\boldsymbol{y}\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{W}^{-1} \\ &= 4\cdot\widetilde{\mathbf{B}}^{\Delta} + \boldsymbol{W}^{-1}\boldsymbol{\lambda}\boldsymbol{s}^\mathsf{T} + \boldsymbol{s}\left(\boldsymbol{W}^{-1}\boldsymbol{\lambda}\right)^\mathsf{T} \\ &= 4\cdot\widetilde{\mathbf{B}}^{\Delta} + \left(2\boldsymbol{W}^{-1}\boldsymbol{y}\frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - 4\frac{\boldsymbol{W}^{-1}\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\boldsymbol{s}^\mathsf{T} \\ &\quad + \boldsymbol{s}\left(2\boldsymbol{W}^{-1}\boldsymbol{y}\frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - 4\frac{\boldsymbol{W}^{-1}\boldsymbol{W}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)^\mathsf{T} \\ &= 4\cdot\widetilde{\mathbf{B}}^{\Delta} + \left(2\boldsymbol{s}\frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - 4\frac{\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\boldsymbol{s}^\mathsf{T} + \boldsymbol{s}\left(2\boldsymbol{s}\frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - 4\frac{\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)^\mathsf{T} \\ &= 4\cdot\widetilde{\mathbf{B}}^{\Delta} + 2\boldsymbol{s}\frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2}\boldsymbol{s}^\mathsf{T} - 4\frac{\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\boldsymbol{s}^\mathsf{T} + 2\boldsymbol{s}\frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2}\boldsymbol{s}^\mathsf{T} - 4\boldsymbol{s}\frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{\Delta}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \\ &= \widetilde{\mathbf{B}}^{\Delta} + \boldsymbol{s}\frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2}\boldsymbol{s}^\mathsf{T} - \frac{\widetilde{\mathbf{B}}^{\Delta}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\boldsymbol{s}^\mathsf{T} - \boldsymbol{s}\frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{\Delta}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \\ &= \widetilde{\mathbf{B}}^{\Delta} + \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\left(\boldsymbol{s}-\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\right)\boldsymbol{s}^\mathsf{T}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - \frac{\left(\boldsymbol{s}-\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\right)\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} - \frac{\boldsymbol{s}\left(\boldsymbol{s}-\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\right)^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \\ &= \widetilde{\mathbf{B}}^{\Delta} + \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - \frac{\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} + \frac{\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} - \frac{\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} + \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \\ &= \widetilde{\mathbf{B}}^{\Delta} + \frac{\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} - \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - \frac{\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} + \frac{\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} - \frac{\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} + \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \\ &= \widetilde{\mathbf{B}}^{(a+1)} - \widetilde{\mathbf{B}}^{(a)} + \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} + \frac{\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} - \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} - \frac{\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \\ &= \widetilde{\mathbf{B}}^{(a+1)} - \left(\widetilde{\mathbf{B}}^{(a)} - \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\left(\mathbf{I}_h - \frac{\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right) - \frac{\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \\ &= \widetilde{\mathbf{B}}^{(a+1)} - \left(\mathbf{I}_h - \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\widetilde{\mathbf{B}}^{(a)}\left(\mathbf{I}_h - \frac{\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right) - \frac{\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \quad . \end{aligned} \tag{A.17}$$

The update for $\widetilde{\mathbf{B}}$ is hence determined by

$$\widetilde{\mathbf{B}}^{(a+1)} = \left(\mathbf{I}_h - \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\widetilde{\mathbf{B}}^{(a)}\left(\mathbf{I}_h - \frac{\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right) + \frac{\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \quad , \tag{A.18}$$

which corresponds to equation 4.25. Equality A.18 can be equivalently written as

$$\begin{aligned}\widetilde{\mathbf{B}}^{(a+1)} &= \widetilde{\mathbf{B}}^{(a)} - \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} - \frac{\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} + \frac{\boldsymbol{s}\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\boldsymbol{s}^\mathsf{T}}{\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^2} + \frac{\boldsymbol{s}\boldsymbol{s}^\mathsf{T}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} \\ &= \widetilde{\mathbf{B}}^{(a)} - \boldsymbol{U}\boldsymbol{V}^\mathsf{T} \quad ,\end{aligned} \tag{A.19}$$

where

$$h \coloneqq \left(\boldsymbol{y}^\mathsf{T}\boldsymbol{s}\right)^{-1} \quad , \tag{A.20a}$$

$$\boldsymbol{U} \coloneqq \left[\boldsymbol{s} \quad \widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\right] \cdot h \tag{A.20b}$$

and

$$\boldsymbol{V}^\mathsf{T} \coloneqq \begin{bmatrix} 1 + h\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y} & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{s}^\mathsf{T} \\ \boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)} \end{bmatrix} \tag{A.20c}$$

(cf. Nocedal and Wright, 1999, pp. 197, 541). Using the Woodbury formula (cf. Nocedal and Wright, 1999, p. 605), it follows that

$$\widetilde{\mathbf{H}}^{(a+1)} = \widetilde{\mathbf{H}}^{(a)} - \widetilde{\mathbf{H}}^{(a)}\boldsymbol{U}\left(\mathbf{I}_2 + \boldsymbol{V}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{U}\right)^{-1}\boldsymbol{V}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)} \quad . \tag{A.21}$$

The components of equation A.21 are obtained by

$$\widetilde{\mathbf{H}}^{(a)}\boldsymbol{U} = \widetilde{\mathbf{H}}^{(a)}\left[\boldsymbol{s} \quad \widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\right] \cdot h = \left[\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} \quad \boldsymbol{y}\right] \cdot h \quad , \tag{A.22a}$$

$$\boldsymbol{V}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)} = \begin{bmatrix} 1 + \dfrac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)} \\ \boldsymbol{y}^\mathsf{T} \end{bmatrix} \quad , \tag{A.22b}$$

as well as

$$\begin{aligned}&\boldsymbol{V}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{U} \\ &= \begin{bmatrix} 1 + \dfrac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)} \\ \boldsymbol{y}^\mathsf{T} \end{bmatrix}\left[\boldsymbol{s} \quad \widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\right] \cdot h \\ &= \begin{bmatrix} 1 + \dfrac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} & \boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y} \\ \boldsymbol{y}^\mathsf{T}\boldsymbol{s} & \boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y} \end{bmatrix} \cdot h \\ &= \begin{bmatrix} 1 + \dfrac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}} & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} & \boldsymbol{y}^\mathsf{T}\boldsymbol{s} \\ \boldsymbol{y}^\mathsf{T}\boldsymbol{s} & \boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y} \end{bmatrix} \cdot h \\ &= \begin{bmatrix} \left(1 + \dfrac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} - \boldsymbol{y}^\mathsf{T}\boldsymbol{s} & \left(1 + \dfrac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\boldsymbol{y}^\mathsf{T}\boldsymbol{s} - \boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y} \\ -\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} & -\boldsymbol{y}^\mathsf{T}\boldsymbol{s} \end{bmatrix} \cdot h \\ &= \begin{bmatrix} \left(1 + \dfrac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} \cdot h - 1 & 1 \\ -\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} \cdot h & -1 \end{bmatrix} \quad .\end{aligned} \tag{A.22c}$$

Using equalities A.22, the matrix to be inverted in equality A.21 is thus given by

$$\left(\mathbf{I}_2 + \boldsymbol{V}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{U}\right) = \begin{bmatrix} \left(1 + \dfrac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\cdot h & 1 \\ -\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\cdot h & 0 \end{bmatrix} \quad . \tag{A.23}$$

Its inverse is defined as

$$\boldsymbol{C} \coloneqq \left(\mathbf{I}_2 + \boldsymbol{V}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{U}\right)^{-1} \quad , \tag{A.24}$$

such that

$$\boldsymbol{C}^{-1}\boldsymbol{C} = \begin{bmatrix} \left(1 + \dfrac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\cdot h & 1 \\ -\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\cdot h & 0 \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad . \tag{A.25}$$

This is a triangular matrix, for which the strategies discussed in section 4.1.1 can be used. The system of equations is

$$c_{11} \cdot \left(1 + \frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\cdot h + c_{21}\cdot 1 = 1 \tag{A.26a}$$

$$c_{12} \cdot \left(1 + \frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\cdot h + c_{22}\cdot 1 = 0 \tag{A.26b}$$

$$-c_{11}\cdot\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\cdot h + c_{21}\cdot 0 = 0 \tag{A.26c}$$

$$-c_{12}\cdot\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\cdot h + c_{22}\cdot 0 = 1 \quad . \tag{A.26d}$$

Equalities A.26c and A.26d directly yield

$$c_{11} = 0 \tag{A.27a}$$

and

$$c_{12} = -\left(h\cdot\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\right)^{-1} \quad , \tag{A.27b}$$

such that equations A.26a and A.26b are solved by

$$c_{21} = 1 \tag{A.27c}$$

and

$$c_{22} = \left(h\cdot\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\right)^{-1}\cdot\left(1 + \frac{\boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}}{\boldsymbol{y}^\mathsf{T}\boldsymbol{s}}\right)\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\cdot h = 1 + \boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\cdot h \quad . \tag{A.27d}$$

Therefore, $\boldsymbol{C}$ is determined by

$$\boldsymbol{C} = \begin{bmatrix} 0 & -\left(h\cdot\boldsymbol{s}^\mathsf{T}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\right)^{-1} \\ 1 & 1 + \boldsymbol{y}^\mathsf{T}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\cdot h \end{bmatrix} \quad . \tag{A.28}$$

Plugging equality A.28 into equation A.21 results in

$$\begin{aligned}
\widetilde{\mathbf{H}}^{(a+1)} &= \widetilde{\mathbf{H}}^{(a)} - \widetilde{\mathbf{H}}^{(a)}\boldsymbol{UCV}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)} \\
&= \widetilde{\mathbf{H}}^{(a)} - \begin{bmatrix}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} & \boldsymbol{y}\end{bmatrix}\begin{bmatrix}0 & -\left(h\cdot\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\right)^{-1} \\ 1 & 1+\boldsymbol{y}^{\mathsf{T}}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\cdot h\end{bmatrix}\begin{bmatrix}1+\boldsymbol{y}^{\mathsf{T}}\widetilde{\mathbf{B}}^{(a)}\boldsymbol{y}\cdot h & -1 \\ -1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)} \\ \boldsymbol{y}^{\mathsf{T}}\end{bmatrix}\cdot h \\
&= \widetilde{\mathbf{H}}^{(a)} - \begin{bmatrix}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} & \boldsymbol{y}\end{bmatrix}\begin{bmatrix}\left(h\cdot\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\right)^{-1} & 0 \\ 0 & -1\end{bmatrix}\begin{bmatrix}\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)} \\ \boldsymbol{y}^{\mathsf{T}}\end{bmatrix}\cdot h \\
&= \widetilde{\mathbf{H}}^{(a)} - \begin{bmatrix}\left(h\cdot\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\right)^{-1}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s} & -\boldsymbol{y}\end{bmatrix}\begin{bmatrix}\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)} \\ \boldsymbol{y}^{\mathsf{T}}\end{bmatrix}\cdot h \\
&= \widetilde{\mathbf{H}}^{(a)} - \left(\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\right)^{-1}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)} + \boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}\cdot h \\
&= \widetilde{\mathbf{H}}^{(a)} + \frac{\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}}{\boldsymbol{y}^{\mathsf{T}}\boldsymbol{s}} - \frac{\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)}}{\boldsymbol{s}^{\mathsf{T}}\widetilde{\mathbf{H}}^{(a)}\boldsymbol{s}} \qquad ,
\end{aligned} \tag{A.29}$$

which constitutes equality 4.26.

Appendix B Mathematical Background of Weighting and Prediction Methods

B.1 Bias of the Maximum Likelihood Covariance Estimator

As described in section 2.2, the ML covariance estimator $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}_{\cdot\mathrm{k}},\boldsymbol{y}_{\cdot\mathrm{l}}}$ is biased for a finite population covariance $\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot\mathrm{k}},\boldsymbol{y}_{\cdot\mathrm{l}}}$ because in general, $\mathrm{E}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\cdot\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)=\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\cdot\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}$ does not hold. Assuming unbiased estimates $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{s}}\right)$ and $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\left(\boldsymbol{w}^{\mathsf{s}}\right)$, this can be shown by

$$
\begin{aligned}
&\mathrm{E}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\cdot\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)\\
&= \mathrm{E}\left(\frac{1}{\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)^{2}}\left(\sum_{i\in\mathcal{S}^{\mathsf{s}}}w_i^{\mathsf{s}}\cdot y_{ik}\right)\cdot\left(\sum_{j\in\mathcal{S}^{\mathsf{s}}}w_j^{\mathsf{s}}\cdot y_{jl}\right)\right)\\
&= \mathrm{E}\left(\frac{1}{\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)^{2}}\cdot\left(\left(\sum_{i\in\mathcal{S}^{\mathsf{s}}}w_i^{\mathsf{s}}\cdot\left(y_{ik}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\right)\right)+\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\cdot\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\right)\right.\\
&\qquad\left.\cdot\left(\sum_{j\in\mathcal{S}^{\mathsf{s}}}\left(w_j^{\mathsf{s}}\cdot\left(y_{jl}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\right)\right)+\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\cdot\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\right)\right)\\
&= \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\cdot\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\\
&\quad+\mathrm{E}\left(\frac{1}{\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)^{2}}\cdot\left(\sum_{i\in\mathcal{S}^{\mathsf{s}}}\sum_{j\in\mathcal{S}^{\mathsf{s}}}w_i^{\mathsf{s}}w_j^{\mathsf{s}}\cdot\left(y_{ik}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\right)\cdot\left(y_{jl}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\right)\right)\right)\\
&\quad+\mathrm{E}\left(\frac{1}{\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)^{2}}\cdot\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\cdot\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\cdot\left(\sum_{i\in\mathcal{S}^{\mathsf{s}}}\left(w_i^{\mathsf{s}}\cdot\left(y_{ik}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\right)\right)\right)\right)\\
&\quad+\mathrm{E}\left(\frac{1}{\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)^{2}}\cdot\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\cdot\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\cdot\left(\sum_{j\in\mathcal{S}^{\mathsf{s}}}\left(w_j^{\mathsf{s}}\cdot\left(y_{jl}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\right)\right)\right)\right)\\
&= \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\cdot\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}+\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\cdot\mathrm{E}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\left(\boldsymbol{w}^{\mathsf{s}}\right)-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\right)+\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\cdot\mathrm{E}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{s}}\right)-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\right)\\
&\quad+\mathrm{E}\left(\frac{1}{\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)^{2}}\cdot\left(\sum_{i\in\mathcal{S}^{\mathsf{s}}}\left(w_i^{\mathsf{s}}\right)^{2}\cdot\left(y_{ik}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\right)\cdot\left(y_{il}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\right)\right)\right)\\
&\quad+\mathrm{E}\left(\frac{1}{\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}}\right)\right)^{2}}\cdot\sum_{i,j\in\mathcal{S}^{\mathsf{s}}}\sum_{j\neq i}w_i^{\mathsf{s}}w_j^{\mathsf{s}}\cdot\left(y_{ik}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\right)\cdot\left(y_{jl}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\right)\right)
\end{aligned}
\tag{B.1a}
$$

$$\approx \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}} \cdot \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}} + \left(1 + \frac{\widehat{\mathrm{N}}\left((\boldsymbol{w}^{\mathsf{s}})^{\circ 2} \right)}{\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}} \right) \right)^2} \right) \cdot \boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot\mathrm{k}} \boldsymbol{y}_{\cdot\mathrm{l}}} \quad . \tag{B.1b}$$

Equations B.1a can alternatively be written as

$$\begin{aligned}
&\mathrm{E}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{s}} \right) \cdot \widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{s}} \right) \right) \\
&= \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}} \cdot \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}} + \\
&\quad \mathrm{E}\left(\frac{1}{\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}} \right) \right)^2} \cdot \left(\sum_{i \in \mathcal{S}^{\mathsf{s}}} w_i^{\mathsf{s}} \cdot y_{ik} - \widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}} \right) \cdot \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}} \right) \cdot \left(\sum_{j \in \mathcal{S}^{\mathsf{s}}} w_j^{\mathsf{s}} \cdot y_{jl} - \widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}} \right) \cdot \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}} \right) \right) \\
&= \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}} \cdot \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}} + \mathrm{E}\left(\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{s}} \right) - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}} \right) \cdot \left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\left(\boldsymbol{w}^{\mathsf{s}} \right) - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}} \right) \right) \\
&= \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{k}}} \cdot \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot\mathrm{l}}} + \boldsymbol{\Sigma}_{\left[\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}(\boldsymbol{w}^{\mathsf{s}}) \;\; \widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{l}}}(\boldsymbol{w}^{\mathsf{s}}) \right]} \quad .
\end{aligned} \tag{B.1c}$$

Hence, the maximum likelihood estimator's bias is equal to the covariance of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{s}} \right)$ and $\widehat{\boldsymbol{\mu}}_{\boldsymbol{y}_{\cdot\mathrm{l}}}\left(\boldsymbol{w}^{\mathsf{s}} \right)$. However, using approximation B.1b to obtain

$$\nu\left(\boldsymbol{w}^{\mathsf{s}} \right) \coloneqq \frac{\widehat{\mathrm{N}}\left((\boldsymbol{w}^{\mathsf{s}})^{\circ 2} \right)}{\left(\widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{s}} \right) \right)^2} \quad , \tag{B.2}$$

the correction

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}} \right) \coloneqq \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}\left(\boldsymbol{w}^{\mathsf{s}} \right) \cdot \left(1 - \nu\left(\boldsymbol{w}^{\mathsf{s}} \right) \right)^{-1} \tag{B.3}$$

for ML covariance estimators is commonly used for estimating weighted sample covariances. In case of an unweighted sample ($w_i^{\mathsf{s}} = 1$) for all i, it reduces to Bessel's correction (using $n-1$ instead of n in the denominator of the sample variance) and is therefore exact (cf. e.g. Galassi et al., 2009, p. 266; Lumley, 2004; R Core Team, 2018; Särndal, Swensson and Wretman, 1992, pp. 186 f).

B.2 Use of Design Weights for Estimation of Conditional Distributions

Pfeffermann and Sverchkov (1999, p. 185) show that for probability sampling designs, it holds that

$$\begin{aligned}
\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1 \right) &= \frac{\mathrm{P}\left(r_i^{\mathsf{ps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot} \right)}{\mathrm{P}\left(r_i^{\mathsf{ps}} = 1 \,|\, \boldsymbol{z}_{i\cdot} \right)} \cdot \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot} \right) \\
&= \frac{\mathrm{E}\left(\mathrm{P}\left(r_i^{\mathsf{ps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}, \pi_i \right) \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot} \right)}{\mathrm{E}\left(\mathrm{P}\left(r_i^{\mathsf{ps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}, \pi_i \right) \,|\, \boldsymbol{z}_{i\cdot} \right)} \cdot \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot} \right) \\
&= \frac{\mathrm{E}\left(\pi_i \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot} \right)}{\mathrm{E}\left(\pi_i \,|\, \boldsymbol{z}_{i\cdot} \right)} \cdot \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot} \right) \quad ,
\end{aligned} \tag{B.4a}$$

$$\begin{aligned}
\mathrm{f}_{\boldsymbol{w}}\left(w_i \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1 \right) &= \frac{\mathrm{E}\left(\pi_i \,|\, w_i, \boldsymbol{z}_{i\cdot} \right)}{\mathrm{E}\left(\pi_i \,|\, \boldsymbol{z}_{i\cdot} \right)} \cdot \mathrm{f}_{\boldsymbol{w}}\left(w_i \,|\, \boldsymbol{z}_{i\cdot} \right) \\
&= \frac{\mathrm{f}_{\boldsymbol{w}}\left(w_i \,|\, \boldsymbol{z}_{i\cdot} \right)}{w_i \cdot \mathrm{E}\left(\pi_i \,|\, \boldsymbol{z}_{i\cdot} \right)}
\end{aligned} \tag{B.4b}$$

and

$$\begin{aligned} \mathrm{E}\left(w_i \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right) &= \int w_i \cdot \frac{\mathrm{f}_{\boldsymbol{w}}\left(w_i \,|\, \boldsymbol{z}_{i\cdot}\right)}{w_i \cdot \mathrm{E}\left(\pi_i \,|\, \boldsymbol{z}_{i\cdot}\right)} \, d\, w_i = \int \frac{\mathrm{f}_{\boldsymbol{w}}\left(w_i \,|\, \boldsymbol{z}_{i\cdot}\right)}{\mathrm{E}\left(\pi_i \,|\, \boldsymbol{z}_{i\cdot}\right)} \, d\, w_i \\ &= \frac{1}{\mathrm{E}\left(\pi_i \,|\, \boldsymbol{z}_{i\cdot}\right)} \quad . \end{aligned} \tag{B.4c}$$

Using equalities B.4, the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ is

$$\mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot}\right) = \frac{\mathrm{E}\left(w_i \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right)}{\mathrm{E}\left(w_i \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right)} \cdot \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right) \quad . \tag{B.5}$$

The conditional expectation can therefore be estimated using

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot}\right) &= \int \boldsymbol{y}_{i\cdot} \cdot \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right) \cdot \frac{\mathrm{E}\left(w_i \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right)}{\mathrm{E}\left(w_i \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right)} \, d\, \boldsymbol{y}_{j\cdot} \\ &= \frac{\mathrm{E}\left(\boldsymbol{y}_{i\cdot} \cdot \mathrm{E}\left(w_i \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right) \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right)}{\mathrm{E}\left(w_i \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right)} \\ &= \frac{\mathrm{E}\left(w_i \cdot \boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right)}{\mathrm{E}\left(w_i \,|\, \boldsymbol{z}_{i\cdot}, r_i^{\mathsf{ps}} = 1\right)} \quad . \end{aligned} \tag{B.6}$$

B.3 General Motivation of Model- and Pseudo-design-based Methods for Non-probability Samples

The arguments in appendix B.2 apply for probability samples with known design weights and inclusion probabilities. Propensity weights based on a set of variables $\boldsymbol{Z}$ that provide conditional independence of $\boldsymbol{Y}$ and $\boldsymbol{r}^{\mathsf{nps}}$ are motivated in a quite similar manner, by replacing $\boldsymbol{\pi}^{\mathsf{ps}}$ with $\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}\right)$ and, assuming that this conditional probability is positive, $\boldsymbol{w}^{\mathsf{ps}}$ with $1/\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}\right)$.

In general, it holds that

$$\mathrm{f}_{\boldsymbol{Y}, \boldsymbol{r}^{\mathsf{nps}}}\left(\boldsymbol{y}_{i\cdot}, r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right) = \mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}\right) \cdot \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot}\right) \quad . \tag{B.7a}$$

In case of conditional independence

$$\left(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{r}^{\mathsf{nps}}\right) | \, \boldsymbol{Z} \quad , \tag{B.7b}$$

equality B.7a can be rewritten as

$$\mathrm{f}_{\boldsymbol{Y}, \boldsymbol{r}^{\mathsf{nps}}}\left(\boldsymbol{y}_{i\cdot}, r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right) = \mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right) \cdot \mathrm{f}_{\boldsymbol{Y}}\left(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{i\cdot}\right) \quad . \tag{B.7c}$$

From equations B.7, it follows that

$$\mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}\right) = \mathrm{P}\left(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}\right) \tag{B.8}$$

in this case (cf. Dawid, 1979, p. 3). Together, equations B.4a and B.8 imply that the conditional distribution can be estimated unbiasedly from the sample under conditional independence assumption B.7b. When $\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{j\cdot}\right)$ is considered known, estimation and

inference for the target population can then be based on $\mathrm{f}_{\boldsymbol{Z}}(\boldsymbol{z}_{j\cdot})$ since

$$\mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y}_{i\cdot}) = \int \cdots \int \mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y}_{i\cdot} \,|\, \boldsymbol{z}_{j\cdot}) \cdot \mathrm{f}_{\boldsymbol{Z}}(\boldsymbol{z}_{j\cdot})\, d\, z_{i1} \cdots d\, z_{iq} \quad . \tag{B.9}$$

This is the main rationale for using prediction models for $\boldsymbol{Y}$ to obtain estimates from a non-probability sample (cf. section 5.1; Pfeffermann, 2011, pp. 120 f).

Since equality B.8 holds under conditional independence assumption B.7b, it follows that

$$\begin{aligned}
\mathrm{E}\left(\sum_{i \in \mathcal{S}^{\mathrm{P}}} \frac{r_i^{\mathsf{nps}} \cdot \boldsymbol{y}_{i\cdot}}{\mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot})}\right) &= \sum_{i \in \mathcal{S}^{\mathrm{P}}} \mathrm{E}\left(\frac{r_i^{\mathsf{nps}} \cdot \boldsymbol{y}_{i\cdot}}{\mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot})}\right) \\
&= \sum_{i \in \mathcal{S}^{\mathrm{P}}} \mathrm{E}\left(\mathrm{E}\left(\frac{r_i^{\mathsf{nps}} \cdot \boldsymbol{y}_{i\cdot}}{\mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot})} \,\middle|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot}\right)\right) \\
&= \sum_{i \in \mathcal{S}^{\mathrm{P}}} \mathrm{E}\left(\boldsymbol{y}_{i\cdot} \cdot \frac{\mathrm{E}(r_i^{\mathsf{nps}} \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot})}{\mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot})}\right) \\
&= \mathrm{E}\left(\sum_{i \in \mathcal{S}^{\mathrm{P}}} \boldsymbol{y}_{i\cdot}\right)
\end{aligned} \tag{B.10}$$

(cf. Horvitz and Thompson, 1952, pp. 667 ff; Imbens, 2000, p. 708; Lunceford and Davidian, 2004, p. 2941). In this case, the *true* conditional probability $\mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot})$ can therefore cancel out bias when used for weighting of linear statistics in analogy to the Horvitz-Thompson estimator, as long as the conditional probabilities are all positive. This is the main motivation of pseudo-design-based usages of response propensities (cf. section 5.2.1).

Matching on response propensities (cf. section 3.6) is motivated by the fact that

$$\begin{aligned}
&\mathrm{f}_{\boldsymbol{Y}, r^{\mathsf{nps}}}(\boldsymbol{y}_{i\cdot}, r_i^{\mathsf{nps}} = 1 \,|\, \mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot})) \\
&= \mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot})) \cdot \mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y}_{i\cdot} \,|\, \mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{y}_{i\cdot}, \boldsymbol{z}_{i\cdot})) \\
&= \mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot}) \cdot \mathrm{f}_{\boldsymbol{Y}}(\boldsymbol{y}_{i\cdot} \,|\, \mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot})) \quad ,
\end{aligned} \tag{B.11}$$

which implies that $\boldsymbol{r}^{\mathsf{nps}}$ and $\boldsymbol{Y}$ are conditionally independent given $\mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot})$, just as when conditioning on $\boldsymbol{z}_{i\cdot}$ itself. Nevertheless, $\mathrm{P}(r_i^{\mathsf{nps}} = 1 \,|\, \boldsymbol{z}_{i\cdot})$ usually needs to be estimated from a model. Most commonly, this is done by means of a generalized linear model, but other options discussed in section 5.1 are used as well (cf. Rosenbaum and Rubin, 1983, pp. 44 ff).

In their alternative strategy for estimating propensities and obtaining pseudo-weights, Elliott and Valliant (2017, pp. 256 f) assume that

a) the observed values outside the non-probability sample come from a reference sample res, with a corresponding response indicator variable $\boldsymbol{r}^{\mathsf{res}}$, and

b) the sampling design generating $\boldsymbol{r}^{\mathsf{res}}$ is considered known and either perfectly or at least very well describable by variables $\boldsymbol{Z}$, such that $\mathrm{P}(r_i^{\mathsf{res}} = 1 \,|\, \boldsymbol{z}_{i\cdot})$ can be obtained with high precision for all $i \in \mathcal{S}^{\mathrm{P}}$.

By Bayes' theorem, it holds that

$$\mathrm{f}_{\boldsymbol{Z}}(\boldsymbol{z}_{i\cdot}) = \frac{\mathrm{f}_{\boldsymbol{Z}}(\boldsymbol{z}_{i\cdot} \,|r_i^{\mathsf{s}} = 1) \cdot \mathrm{P}(r_i^{\mathsf{s}} = 1)}{\mathrm{P}(r_i^{\mathsf{s}} = 1 \,|\boldsymbol{z}_{i\cdot})} \tag{B.12a}$$

and

$$\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|r_i^{\mathsf{s}}=1\right) \;=\; \frac{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\right)\cdot \mathrm{P}\left(r_i^{\mathsf{s}}=1\,|\boldsymbol{z}_{i\cdot}\right)}{\mathrm{P}\left(r_i^{\mathsf{s}}=1\right)} \tag{B.12b}$$

for $\mathsf{s}\in\{\mathsf{nps};\mathsf{res}\}$, i.e. non-probability and reference sample, such that

$$\frac{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|r_i^{\mathsf{nps}}=1\right)}{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|r_i^{\mathsf{res}}=1\right)} = \frac{\mathrm{P}\left(r_i^{\mathsf{nps}}=1\,|\boldsymbol{z}_{i\cdot}\right)}{\mathrm{P}\left(r_i^{\mathsf{res}}=1\,|\boldsymbol{z}_{i\cdot}\right)}\cdot\frac{\mathrm{P}\left(r_i^{\mathsf{res}}=1\right)}{\mathrm{P}\left(r_i^{\mathsf{nps}}=1\right)} \propto \frac{\mathrm{P}\left(r_i^{\mathsf{nps}}=1\,|\boldsymbol{z}_{i\cdot}\right)}{\mathrm{P}\left(r_i^{\mathsf{res}}=1\,|\boldsymbol{z}_{i\cdot}\right)} \quad . \tag{B.12c}$$

Based on equalities B.12, the true response propensity may be written as

$$\begin{aligned}
\mathrm{P}\left(r_i^{\mathsf{nps}}=1\,|\boldsymbol{z}_{i\cdot}\right) \;&=\; \frac{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|r_i^{\mathsf{nps}}=1\right)\cdot\mathrm{P}\left(r_i^{\mathsf{nps}}=1\right)}{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\right)} \\
&= \frac{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|r_i^{\mathsf{nps}}=1\right)}{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|r_i^{\mathsf{res}}=1\right)}\cdot\frac{\mathrm{P}\left(r_i^{\mathsf{nps}}=1\right)}{\mathrm{P}\left(r_i^{\mathsf{res}}=1\right)}\cdot\mathrm{P}\left(r_i^{\mathsf{res}}=1\,|\boldsymbol{z}_{i\cdot}\right) \\
&\propto \frac{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|r_i^{\mathsf{nps}}=1\right)}{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|r_i^{\mathsf{res}}=1\right)}\cdot\mathrm{P}\left(r_i^{\mathsf{res}}=1\,|\boldsymbol{z}_{i\cdot}\right) \qquad ,
\end{aligned} \tag{B.13}$$

considering $\mathrm{P}\left(r_i^{\mathsf{nps}}=1\right)/\mathrm{P}\left(r_i^{\mathsf{res}}=1\right)$ as a normalizing constant. As in section 5.2.1, denote by $r_i^{*}\coloneqq\mathbb{I}\left(i\in(\mathcal{S}^{\mathsf{nps}}\cup\mathcal{S}^{\mathsf{res}})\right)\in\{0;1\}$ for all $i\in\mathcal{S}^{\mathrm{P}}$ the indicator for whether an element i is part of the non-probability and/or the reference sample. The approximation for equality B.13 proposed by Elliott and Valliant (2017, p. 256) is then given by

$$\begin{aligned}
\frac{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|\,r_i^{\mathsf{nps}}=1\right)}{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|\,r_i^{\mathsf{res}}=1\right)} \;&=\; \frac{\mathrm{P}\left(\boldsymbol{Z}=\boldsymbol{z}_{i\cdot}\cap r_i^{\mathsf{nps}}=1\right)/\mathrm{P}\left(r_i^{\mathsf{nps}}=1\right)}{\mathrm{P}\left(\boldsymbol{Z}=\boldsymbol{z}_{i\cdot}\cap r_i^{\mathsf{res}}=1\right)/\mathrm{P}\left(r_i^{\mathsf{res}}=1\right)} \\
&\approx \frac{\mathrm{P}\left(\boldsymbol{Z}=\boldsymbol{z}_{i\cdot}\cap r_i^{\mathsf{nps}}=1\cap r_i^{*}=1\right)/\mathrm{P}\left(r_i^{\mathsf{nps}}=1\cap r_i^{*}=1\right)}{\mathrm{P}\left(\boldsymbol{Z}=\boldsymbol{z}_{i\cdot}\cap r_i^{\mathsf{nps}}=0\cap r_i^{*}=1\right)/\mathrm{P}\left(r_i^{\mathsf{nps}}=0\cap r_i^{*}=1\right)} \\
&= \frac{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|\,r_i^{\mathsf{nps}}=1,r_i^{*}=1\right)}{\mathrm{f}_{\boldsymbol{Z}}\left(\boldsymbol{z}_{i\cdot}\,|\,r_i^{\mathsf{nps}}=0,r_i^{*}=1\right)} \qquad .
\end{aligned} \tag{B.14}$$

For this approximation, it is assumed that the overlap $(\mathcal{S}^{\mathsf{nps}}\cap\mathcal{S}^{\mathsf{res}})$ is ignorable, such that $\boldsymbol{r}^{*}\approx\boldsymbol{r}^{\mathsf{nps}}+\boldsymbol{r}^{\mathsf{res}}$. Note that Elliott and Valliant (2017, p. 256) justify this approximation by negligible sampling fractions for both samples, such that $\mathrm{P}\left(r_i^{\mathsf{res}}=1\cap r_i^{\mathsf{nps}}=0\right)\approx\mathrm{P}\left(r_i^{\mathsf{res}}=1\right)$ and $\mathrm{P}\left(r_i^{\mathsf{nps}}=1\cap r_i^{\mathsf{res}}=0\right)\approx\mathrm{P}\left(r_i^{\mathsf{nps}}=1\right)$. As becomes evident from equalities B.14, however, the overlap needs to be ignorable for all values of $\boldsymbol{Z}$. Therefore, the stronger assumptions $\mathrm{P}\left(\boldsymbol{Z}=\boldsymbol{z}_{i\cdot}\cap r_i^{\mathsf{nps}}=1\cap r_i^{\mathsf{res}}=0\right)\approx\mathrm{P}\left(\boldsymbol{Z}=\boldsymbol{z}_{i\cdot}\cap r_i^{\mathsf{nps}}=1\right)$ and $\mathrm{P}\left(\boldsymbol{Z}=\boldsymbol{z}_{i\cdot}\cap r_i^{\mathsf{res}}=1\cap r_i^{\mathsf{nps}}=0\right)\approx\mathrm{P}\left(\boldsymbol{Z}=\boldsymbol{z}_{i\cdot}\cap r_i^{\mathsf{res}}=1\right)$ are required for approximation 5.136 to be valid. As discussed in section 5.2.1, $\mathrm{P}\left(r_i^{\mathsf{res}}=1\,|\,\boldsymbol{z}_{i\cdot}\right)$ is then obtained from the reference sample's design or a different model.

B.4 Mathematical Background of Prediction Models

B.4.1 Derivation of the Linear Model

The least squares loss function of the linear regression discussed in section 5.1.2 is

$$\begin{aligned} \delta\left(\boldsymbol{\Theta}\right) &= \left(\boldsymbol{Y}^{\mathsf{s}} - \boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta}\right)^{\mathsf{T}} \boldsymbol{W} \left(\boldsymbol{Y}^{\mathsf{s}} - \boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta}\right) \\ &= \left(\boldsymbol{Y}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{Y}^{\mathsf{s}} - 2 \cdot \left(\boldsymbol{Y}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta} + \left(\boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta} \quad , \end{aligned} \tag{B.15}$$

where $\boldsymbol{W} = \mathbf{diag}\left(\boldsymbol{w}\right)$. The corresponding Jacobian matrix is given by

$$\begin{aligned} \mathbf{J}_{\delta}\left(\boldsymbol{\Theta}\right) &= \left(\frac{\partial(\delta\left(\boldsymbol{\Theta}\right))}{\partial\left(\boldsymbol{\Theta}\right)}\right)^{\mathsf{T}} = 2 \cdot \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta} - \left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{Y}^{\mathsf{s}}\right)^{\mathsf{T}} \\ &= 2 \cdot \left(\boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta} - \boldsymbol{Y}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}} \quad . \end{aligned} \tag{B.16}$$

The minimum of $\delta\left(\boldsymbol{\Theta}\right)$ can be found analytically, by setting the Jacobian matrix to zero:

$$\begin{aligned} & 2 \cdot \left(\boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta} - \boldsymbol{Y}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}} &&= \mathbf{0} \\ \Leftrightarrow \quad & \left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta} &&= \left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{Y}^{\mathsf{s}} \\ \Leftrightarrow \quad & \boldsymbol{\Theta} &&= \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\right)^{-1} \left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{Y}^{\mathsf{s}} \quad . \end{aligned} \tag{B.17}$$

This is equivalent to a single iteration of the Newton-Raphson algorithm 5 since it follows from equation B.16 that the Hessian of the loss function is given by

$$\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}\right) = \left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}} \quad , \tag{B.18}$$

such that for iteration a, the update is

$$\begin{aligned} \boldsymbol{\Theta}^{(a)} &= \boldsymbol{\Theta}^{(a-1)} - \left(\mathbf{H}_{\delta}\left(\boldsymbol{\Theta}^{(a-1)}\right)\right)^{-1} \left(\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}^{(a-1)}\right)\right)^{\mathsf{T}} \\ &= \boldsymbol{\Theta}^{(a-1)} - \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\right)^{-1} 2 \cdot \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\boldsymbol{\Theta}^{(a-1)} - \left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{Y}^{\mathsf{s}}\right) \\ &= -\boldsymbol{\Theta}^{(a-1)} + 2 \cdot \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\right)^{-1} \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{Y}^{\mathsf{s}}\right) \quad . \end{aligned} \tag{B.19}$$

The optimum is found when $\boldsymbol{\Theta}^{(a-1)} = \boldsymbol{\Theta}^{(a)}$, such that

$$\begin{aligned} \boldsymbol{\Theta}^{(a)} &= -\boldsymbol{\Theta}^{(a)} + 2 \cdot \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\right)^{-1} \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{Y}^{\mathsf{s}}\right) \\ \Leftrightarrow \quad \boldsymbol{\Theta}^{(a)} &= \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{X}^{\mathsf{s}}\right)^{-1} \left(\left(\boldsymbol{X}^{\mathsf{s}}\right)^{\mathsf{T}} \boldsymbol{W}\boldsymbol{Y}^{\mathsf{s}}\right) \quad . \end{aligned} \tag{B.20}$$

B.4.2 Additional Newton-Raphson and Fisher Scoring Update Rules for Generalized Linear and Additive Models

To simplify notation, derivation of generalized linear and additive models is considered for the whole population. Incorporating of weights to obtain HT-estimates from a sample is discussed in sections 5.1.3 and 5.1.4.

Denoting the vector of optimization parameters by

$$\boldsymbol{\Theta} \coloneqq \boldsymbol{\kappa} \quad , \tag{B.21}$$

the systematic component is

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\Theta}) \coloneqq \boldsymbol{t}(\boldsymbol{X}, \boldsymbol{\kappa}) \quad . \tag{B.22}$$

Assuming a probability density function belonging to the exponential family (cf. equation 5.16) results in the Log-Likelihood

$$\mathscr{L}(\boldsymbol{\Theta}) = (\boldsymbol{y}_{\cdot l} \circ \boldsymbol{\theta} - \mathbf{b}(\boldsymbol{\theta}))^{\mathsf{T}} (\mathbf{a}(\boldsymbol{\phi}))^{-1} + \mathbf{c}(\boldsymbol{y}_{\cdot l}, \boldsymbol{\phi}) \tag{B.23}$$

for $\boldsymbol{\Theta}$ (cf. Nelder and Wedderburn, 1972, p. 371). The first derivative of equation B.23 with respect to $\boldsymbol{\Theta}$ is

$$\begin{aligned}
\frac{\partial(\mathscr{L}(\boldsymbol{\Theta}))}{\partial(\boldsymbol{\Theta})} &= \frac{\partial(\mathscr{L}(\boldsymbol{\eta}))}{\partial(\boldsymbol{\theta})} \frac{\partial(\boldsymbol{\theta})}{\partial\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)} \frac{\partial\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)}{\partial(\boldsymbol{\eta})} \frac{\partial(\boldsymbol{\eta})}{\partial(\boldsymbol{\Theta})} \\
&= \left(\boldsymbol{y}_{\cdot l} - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}} (\mathbf{a}(\boldsymbol{\phi}))^{-1} \left(\mathbf{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)\right)^{-1} \frac{\partial\left(\boldsymbol{l}^{-1}(\boldsymbol{\eta})\right)}{\partial(\boldsymbol{\eta})} \frac{\partial(\boldsymbol{\eta})}{\partial(\boldsymbol{\Theta})} \\
&= \left(\boldsymbol{y}_{\cdot l} - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}} \left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1} \frac{\partial\left(\boldsymbol{l}^{-1}(\boldsymbol{\eta})\right)}{\partial(\boldsymbol{\eta})} \frac{\partial(\boldsymbol{\eta})}{\partial(\boldsymbol{\Theta})} \quad .
\end{aligned} \tag{B.24}$$

Equalities B.24 result from the components

$$\frac{\partial(\mathscr{L}(\boldsymbol{\eta}))}{\partial(\boldsymbol{\theta})} = \left(\boldsymbol{y}_{\cdot l} - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}} (\mathbf{a}(\boldsymbol{\phi}))^{-1} \quad , \tag{B.25a}$$

$$\frac{\partial(\boldsymbol{\theta})}{\partial\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)} = \frac{\partial(\boldsymbol{\theta})}{\partial\left(\frac{\partial(\mathbf{b}(\boldsymbol{\theta}))}{\partial(\boldsymbol{\theta})}\right)} = \left(\frac{\partial^2 \mathbf{b}(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})}\right)^{-1} = \left(\mathbf{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)\right)^{-1} \quad , \tag{B.25b}$$

and

$$\frac{\partial\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)}{\partial(\boldsymbol{\eta})} = \frac{\partial\left(\boldsymbol{l}^{-1}(\boldsymbol{\eta})\right)}{\partial(\boldsymbol{\eta})} \quad , \tag{B.25c}$$

which are based on equalities 5.17 (cf. Hastie and Tibshirani, 1986, p. 302; McCullagh and Nelder, 1989, pp. 29–43; Nelder and Wedderburn, 1972, pp. 371 ff).

The ML parameter estimates are obtained by setting the score function B.24 to zero. Optimization is typically performed by the Fisher scoring or Newton-Raphson method (cf. algorithm 5), which require the actual or expected Hessian matrix

$$\begin{aligned}\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{\eta}\right)\right)}{\partial^2\left(\boldsymbol{\Theta}\right)} &= \frac{\partial}{\partial\left(\boldsymbol{\Theta}\right)}\left(\left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)}\right)\\ &= \left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\left(\frac{\partial}{\partial\left(\boldsymbol{\Theta}\right)}\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)}\right)\right)\\ &\quad-\left(\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)} \quad .\end{aligned} \tag{B.26}$$

Equations B.25 are used again to obtain equalities B.26.

Since $\mathrm{E}\left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)=\mathrm{E}\left(\boldsymbol{y}_{\cdot l}-\mathrm{E}\left(\boldsymbol{y}_{\cdot l}\right)\right)=\mathbf{0}$, the first product in equation B.26 vanishes in expectation, hence

$$\mathrm{E}\left(\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{\eta}\right)\right)}{\partial^2\left(\boldsymbol{\Theta}\right)}\right) = -\left(\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)} \tag{B.27}$$

(cf. Hastie and Tibshirani, 1986, p. 302; McCullagh and Nelder, 1989, pp. 29–43; Nelder and Wedderburn, 1972, pp. 371 ff). This coincides with the negative Fisher information matrix, which can be shown by

$$\begin{aligned}-\mathbf{V}\left(\frac{\partial(\mathscr{L}\left(\boldsymbol{\Theta}\right))}{\partial\left(\boldsymbol{\Theta}\right)}\right) &= \left(\mathrm{E}\left(\frac{\partial(\mathscr{L}\left(\boldsymbol{\Theta}\right))}{\partial\left(\boldsymbol{\Theta}\right)}\right)\right)\left(\mathrm{E}\left(\frac{\partial(\mathscr{L}\left(\boldsymbol{\Theta}\right))}{\partial\left(\boldsymbol{\Theta}\right)}\right)\right)^{\mathsf{T}}\\ &\quad-\mathrm{E}\left(\left(\frac{\partial(\mathscr{L}\left(\boldsymbol{\Theta}\right))}{\partial\left(\boldsymbol{\Theta}\right)}\right)^{\mathsf{T}}\left(\frac{\partial(\mathscr{L}\left(\boldsymbol{\Theta}\right))}{\partial\left(\boldsymbol{\Theta}\right)}\right)\right)\\ &= -\mathrm{E}\left(\left(\left(\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)\right)\right.\\ &\qquad\left.\left(\left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)}\right)\right)\\ &= -\left(\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)} \quad .\end{aligned} \tag{B.28}$$

The expected and actual value of the Hessian matrix (equations B.27 and B.26) coincide in case of a canonical link function when $\boldsymbol{t}$ is linear in $\boldsymbol{\kappa}$, in which case it holds that

$$\boldsymbol{\eta} = \boldsymbol{\theta} \quad , \tag{B.29}$$

and

$$\frac{\partial^2\left(\boldsymbol{\eta}\right)}{\partial\left(\boldsymbol{\Theta}\right)\partial\left(\boldsymbol{\Theta}\right)} = \mathbf{0} \quad . \tag{B.30}$$

It then follows that

$$\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)} = \frac{\partial\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)}{\partial\left(\boldsymbol{\theta}\right)} = \left(\frac{\partial\left(\boldsymbol{\theta}\right)}{\partial\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)}\right)^{-1} = \mathbf{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right) \quad , \tag{B.31}$$

hence

$$\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)} = \left(\mathbf{a}\left(\boldsymbol{\phi}\right)\right)^{-1}\left(\mathbf{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)\right)^{-1}\mathbf{V}\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right) = \left(\mathbf{a}\left(\boldsymbol{\phi}\right)\right)^{-1} \tag{B.32}$$

and

$$\begin{aligned}\frac{\partial}{\partial\left(\boldsymbol{\Theta}\right)}&\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)}\right)\\ &= \frac{\partial\left(\left(\mathbf{a}\left(\boldsymbol{\phi}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\Theta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\Theta}\right)} + \left(\mathbf{a}\left(\boldsymbol{\phi}\right)\right)^{-1}\frac{\partial^2\left(\boldsymbol{\eta}\right)}{\partial\left(\boldsymbol{\Theta}\right)\partial\left(\boldsymbol{\Theta}\right)}\\ &= \mathbf{0}\end{aligned} \tag{B.33}$$

since $\mathbf{a}\left(\boldsymbol{\phi}\right)$ does not depend on $\boldsymbol{\Theta}$. When equality B.33 holds, the first factor in equation B.26 is always zero, such that Newton-Raphson and Fisher scoring algorithm coincide (cf. Breslow and Clayton, 1993, p. 10; Hastie and Tibshirani, 1986, p. 316; McCullagh and Nelder, 1989, p. 43).

For the generalized *linear* models, the following additional identities hold:

$$\begin{aligned}\boldsymbol{\Theta} &= \boldsymbol{\beta}\\ \frac{\partial\left(\boldsymbol{\eta}\right)}{\partial\left(\boldsymbol{\beta}\right)} &= \boldsymbol{X} \quad .\end{aligned} \tag{B.34}$$

Consequently, the Jacobian and Hessian matrix are determined by

$$\frac{\partial(\mathscr{L}\left(\boldsymbol{\Theta}\right))}{\partial\left(\boldsymbol{\beta}\right)} = \frac{\partial(\mathscr{L}\left(\boldsymbol{\eta}\right))}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{X} \tag{B.35}$$

and

$$\begin{aligned}\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{\Theta}\right)\right)}{\partial^2\left(\boldsymbol{\beta}\right)} = &\quad \left(\boldsymbol{y}_{\cdot l} - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\left(\frac{\partial}{\partial\left(\boldsymbol{\beta}\right)}\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial\left(\boldsymbol{\eta}\right)}{\partial\left(\boldsymbol{\beta}\right)}\right)\right)\\ &- \left(\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial\left(\boldsymbol{\eta}\right)}{\partial\left(\boldsymbol{\beta}\right)}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial\left(\boldsymbol{\eta}\right)}{\partial\left(\boldsymbol{\beta}\right)}\\ = &- \left(\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{X}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{X}\end{aligned} \tag{B.36}$$

(cf. McCullagh and Nelder, 1989, pp. 29–43; Nelder and Wedderburn, 1972, pp. 371 ff).

B.4.3 Additional Newton-Raphson and Fisher Scoring Update Rules for Generalized Linear and Additive Mixed Models

As in appendix B.4.2, derivation of generalized linear and additive mixed models is considered for the whole population to simplify notation. The use of weights to obtain HT-estimates from a sample is discussed in section 5.1.5.

As in section 5.1.5, the model is defined by

$$\begin{aligned} \boldsymbol{\Theta} &\coloneqq \begin{bmatrix} \boldsymbol{\kappa}^\mathsf{T} & \boldsymbol{u}^\mathsf{T} & \boldsymbol{\phi}^\mathsf{T} \end{bmatrix}^\mathsf{T} \\ \boldsymbol{\eta} &= \boldsymbol{\eta}\left(\boldsymbol{\Theta}\right) \coloneqq \boldsymbol{t}\left(\boldsymbol{X}, \boldsymbol{\kappa}\right) + \boldsymbol{D}\boldsymbol{u} \\ \boldsymbol{\phi} &\coloneqq \begin{bmatrix} \left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)^\mathsf{T} & \left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)^\mathsf{T} \end{bmatrix}^\mathsf{T} \\ \boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi) &\coloneqq \boldsymbol{\Sigma}_{\boldsymbol{e}_{\cdot l}}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right) + \boldsymbol{D}\boldsymbol{\Sigma}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\boldsymbol{D}^\mathsf{T} \quad . \end{aligned} \tag{B.37}$$

The component and combined likelihoods are given by

$$\begin{aligned} \mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}} \middle| \boldsymbol{u}\right) &\coloneqq \left\| \left(\mathbf{a}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)\right)^{-1} \left(\boldsymbol{y}_{\cdot l} \circ \boldsymbol{\theta}_{\boldsymbol{y}_{\cdot l}} - \mathbf{b}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\theta}_{\boldsymbol{y}_{\cdot l}}\right)\right) + \mathbf{c}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{y}_{\cdot l}, \boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right) \right\|_1 \\ \mathscr{L}\left(\boldsymbol{u}\right) &= \left\| \left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)^{-1} \left(\boldsymbol{u} \circ \boldsymbol{\theta}_{\boldsymbol{u}} - \mathbf{b}_{\boldsymbol{u}}\left(\boldsymbol{\theta}_{\boldsymbol{u}}\right)\right) + \mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u}, \boldsymbol{\phi}^{(\boldsymbol{u})}\right) \right\|_1 \\ \mathscr{L}\left(\boldsymbol{\Theta}\right) &= \mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}} \middle| \boldsymbol{u}\right) + \mathscr{L}\left(\boldsymbol{u}\right) \end{aligned} \tag{B.38}$$

(cf. Bates, 2018, p. 5; Lee and Nelder, 1996, pp. 620 f).

In analogy to equation B.25, it follows that

$$\begin{aligned} \frac{\partial\left(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}} \middle| \boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\kappa}\right)} &= \frac{\partial\left(\mathscr{L}\left(\boldsymbol{\eta} \middle| \boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\theta}\right)} \frac{\partial\left(\boldsymbol{\theta}\right)}{\partial\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)} \frac{\partial\left(\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)}{\partial\left(\boldsymbol{\eta}\right)} \frac{\partial\left(\boldsymbol{\eta}\right)}{\partial\left(\boldsymbol{\kappa}\right)} \\ &= \left(\boldsymbol{y}_{\cdot l} - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^\mathsf{T} \left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1} \frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)} \frac{\partial\left(\boldsymbol{\eta}\right)}{\partial\left(\boldsymbol{\kappa}\right)} \quad , \end{aligned} \tag{B.39}$$

$$\frac{\partial\left(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}} \middle| \boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)} = \left(\boldsymbol{y}_{\cdot l} - \boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^\mathsf{T} \left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1} \frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)} \boldsymbol{D} \quad , \tag{B.40}$$

and

$$\begin{aligned} \frac{\partial\left(\mathscr{L}\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)} &= \frac{\partial}{\partial\left(\boldsymbol{u}\right)} \left(\left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)^{-1} \left(\boldsymbol{u} \circ \boldsymbol{\theta}_{\boldsymbol{u}} - \mathbf{b}_{\boldsymbol{u}}\left(\boldsymbol{\theta}_{\boldsymbol{u}}\right)\right) + \mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u}, \boldsymbol{\phi}^{(\boldsymbol{u})}\right) \right) \\ &= \left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)^{-1} \boldsymbol{\theta}_{\boldsymbol{u}} + \frac{\partial\left(\mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u}, \boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)}{\partial\left(\boldsymbol{u}\right)} \end{aligned} \tag{B.41}$$

(cf. Hastie and Tibshirani, 1986, p. 302; Lee and Nelder, 1996, pp. 631 f; McCullagh and Nelder, 1989, pp. 29–43; Nelder and Wedderburn, 1972, pp. 371 ff).

Differentiation of the likelihoods w.r.t. the dispersion parameters $\boldsymbol{\phi}$ shows that

$$\begin{aligned}\frac{\partial(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)} &= \frac{\partial\Big(\left(\mathbf{a}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)\right)^{-1}\Big)}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)}\left(\boldsymbol{y}_{\cdot l}\circ\boldsymbol{\theta}_{\boldsymbol{y}_{\cdot l}}-\mathbf{b}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\theta}_{\boldsymbol{y}_{\cdot l}}\right)\right)\\ &\quad+\frac{\partial\Big(\mathbf{c}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{y}_{\cdot l},\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)\Big)}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)}\quad,\end{aligned}\tag{B.42}$$

$$\frac{\partial(\mathscr{L}\left(\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)}=\frac{\partial\Big(\left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)^{-1}\Big)}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)}\left(\boldsymbol{u}\circ\boldsymbol{\theta}_{\boldsymbol{u}}-\mathbf{b}_{\boldsymbol{u}}\left(\boldsymbol{\theta}_{\boldsymbol{u}}\right)\right)+\frac{\partial\Big(\mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u},\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\Big)}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)}\quad,\tag{B.43}$$

and

$$\frac{\partial(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)}=\mathbf{0}\quad.\tag{B.44}$$

For equations B.42 and B.43, it holds for $k\in\{\boldsymbol{y}_{\cdot l},\boldsymbol{u}\}$ that

$$\frac{\partial\Big(\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\Big)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}=-\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\frac{\partial\Big(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\Big)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\tag{B.45}$$

is the derivative of the inverse matrix (cf. Rao, 2003, pp. 100 ff; Searle, Casella and McCulloch, 2006, pp. 235 ff, 456).

The second derivatives are given by

$$\begin{aligned}\frac{\partial^2(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\kappa}\right)\partial\left(\boldsymbol{\kappa}\right)} &= \left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\left(\frac{\partial}{\partial\left(\boldsymbol{\kappa}\right)}\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\kappa}\right)}\right)\right)\\ &\quad-\left(\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\kappa}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\kappa}\right)}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\kappa}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\kappa}\right)}\end{aligned}\tag{B.46}$$

$$\begin{aligned}\frac{\partial^2(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\kappa}\right)\partial\left(\boldsymbol{u}\right)} &= \left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\left(\frac{\partial}{\partial\left(\boldsymbol{u}\right)}\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\kappa}\right)}\right)\right)\\ &\quad-\left(\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{D}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\kappa}\right)}\end{aligned}\tag{B.47}$$

$$\frac{\partial^2(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\kappa}\right)\partial\left(\boldsymbol{\phi}\right)}=\left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\frac{\partial\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}\right)}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\kappa}\right)}\tag{B.48}$$

for cases where the first derivatives are taken with respect to $\boldsymbol{\kappa}$.

In the cases where they are w.r.t. $\boldsymbol{u}$, one obtains

$$\begin{aligned}\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}|\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{u}\right)} &= \left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\left(\frac{\partial}{\partial\left(\boldsymbol{u}\right)}\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{D}\right)\right)\\ &\quad-\left(\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{D}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{D}\end{aligned} \tag{B.49}$$

$$\begin{aligned}\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}|\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{\kappa}\right)} &= \left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\left(\frac{\partial}{\partial\left(\boldsymbol{\kappa}\right)}\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{D}\right)\right)\\ &\quad-\left(\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\kappa}\right)}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{D}\end{aligned} \tag{B.50}$$

$$\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}|\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{\phi}\right)} = \left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\frac{\partial\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}\right)}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{D} \tag{B.51}$$

for the conditional Log-Likelihood of $\boldsymbol{y}_{\cdot l}^{\mathsf{s}}$, as well as

$$\begin{aligned}\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{u}\right)} &= \frac{\partial}{\partial\left(\boldsymbol{u}\right)}\left(\left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)^{-1}\boldsymbol{\theta}_{\boldsymbol{u}}+\frac{\partial\left(\mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u},\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)}{\partial\left(\boldsymbol{u}\right)}\right)\\ &= \frac{\partial^2\left(\mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u},\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{u}\right)}\end{aligned} \tag{B.52}$$

and

$$\begin{aligned}\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{\phi}\right)} &= \frac{\partial}{\partial\left(\boldsymbol{\phi}\right)}\left(\left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)^{-1}\boldsymbol{\theta}_{\boldsymbol{u}}+\frac{\partial\left(\mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u},\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)}{\partial\left(\boldsymbol{u}\right)}\right)\\ &= \frac{\partial\left(\left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}\right)}\boldsymbol{\theta}_{\boldsymbol{u}}+\frac{\partial^2\left(\mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u},\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{\phi}\right)}\end{aligned} \tag{B.53}$$

for the Log-Likelihood of $\boldsymbol{u}$. The second derivatives for the cases where the first derivatives are taken w.r.t. to the variance components $\boldsymbol{\phi}$ are given by

$$\begin{aligned}\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}|\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)\partial\left(\boldsymbol{\kappa}\right)} &= \frac{\partial\left(\left(\mathbf{a}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{T}}\frac{\partial\left(\boldsymbol{\theta}_{\boldsymbol{y}_{\cdot l}}\right)}{\partial\left(\boldsymbol{\eta}\right)}-\frac{\partial\left(\mathbf{b}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\theta}_{\boldsymbol{y}_{\cdot l}}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\right)\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\kappa}\right)}\\ &= \left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\frac{\partial\left(\left(\mathbf{a}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\frac{\partial(\boldsymbol{\eta})}{\partial\left(\boldsymbol{\kappa}\right)}\end{aligned} \tag{B.54}$$

and

$$\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}|\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)\partial\left(\boldsymbol{u}\right)} = \left(\boldsymbol{y}_{\cdot l}-\boldsymbol{\mu}_{\boldsymbol{y}_{\cdot l}}\right)^{\mathsf{T}}\frac{\partial\left(\left(\mathbf{a}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}_{\cdot l}}(\phi)\right)^{-1}\frac{\partial\left(\boldsymbol{l}_{\boldsymbol{y}_{\cdot 1}}^{-1}\left(\boldsymbol{\eta}\right)\right)}{\partial\left(\boldsymbol{\eta}\right)}\boldsymbol{D} \tag{B.55}$$

based on equalities 5.17 and B.25, as well as

$$\frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\partial\left(\boldsymbol{u}\right)} = \frac{\partial^2\left(\mathscr{L}\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{u}\right)\partial\left(\boldsymbol{\phi}\right)} \quad , \tag{B.56}$$

$$\frac{\partial\left(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}\,|\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\phi}\right)\partial\left(\boldsymbol{\phi}\right)} = \frac{\partial^2\left(\left(\mathbf{a}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}\right)\partial\left(\boldsymbol{\phi}\right)}\left(\boldsymbol{y}_{\cdot l}\circ\boldsymbol{\theta}_{\boldsymbol{y}_{\cdot l}}-\mathbf{b}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{\theta}_{\boldsymbol{y}_{\cdot l}}\right)\right)+\frac{\partial\left(\mathbf{c}_{\boldsymbol{y}_{\cdot 1}}\left(\boldsymbol{y}_{\cdot l},\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)\right)}{\partial\left(\boldsymbol{\phi}\right)\partial\left(\boldsymbol{\phi}\right)} \tag{B.57}$$

and, in analogy,

$$\frac{\partial\left(\mathscr{L}\left(\boldsymbol{u}\right)\right)}{\partial\left(\boldsymbol{\phi}\right)\partial\left(\boldsymbol{\phi}\right)} = \frac{\partial^2\left(\left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}\right)\partial\left(\boldsymbol{\phi}\right)}\left(\boldsymbol{u}\circ\boldsymbol{\theta}_{\boldsymbol{u}}-\mathbf{b}_{\boldsymbol{u}}\left(\boldsymbol{\theta}_{\boldsymbol{u}}\right)\right)+\frac{\partial\left(\mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{y}_{\cdot l},\boldsymbol{\phi}^{(\boldsymbol{u})}\right)\right)}{\partial\left(\boldsymbol{\phi}\right)\partial\left(\boldsymbol{\phi}\right)} \quad . \tag{B.58}$$

For equalities B.57 and B.58 ($k\in\{\boldsymbol{\eta},\boldsymbol{u}\}$), it holds that

$$\begin{aligned}
&\frac{\partial^2\left(\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)\partial\left(\boldsymbol{\phi}^{(k)}\right)}\\
&=-\frac{\partial\left(\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}\frac{\partial\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\\
&\quad-\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\left(\frac{\partial\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}\frac{\partial\left(\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}+\frac{\partial^2\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)\partial\left(\boldsymbol{\phi}^{(k)}\right)}\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\right)\\
&=\quad\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\frac{\partial\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\frac{\partial\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\\
&\quad+\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\frac{\partial\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\frac{\partial\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\\
&\quad-\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\frac{\partial^2\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)\partial\left(\boldsymbol{\phi}^{(k)}\right)}\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\\
&=\quad\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\left(\begin{array}{l}2\cdot\frac{\partial\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}\frac{\partial\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)}\\ \qquad-\frac{\partial^2\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(k)}\right)\partial\left(\boldsymbol{\phi}^{(k)}\right)}\end{array}\right)\left(\mathbf{a}_k\left(\boldsymbol{\phi}^{(k)}\right)\right)^{-1}
\end{aligned} \tag{B.59}$$

is the second derivative of the matrix inverse (cf. e.g. Rao, 2003, pp. 100 ff; Searle, Casella and McCulloch, 2006, pp. 235 ff, 456). These components constitute the Jacobian

$$\begin{aligned}
&\mathbf{J}_{\mathscr{L}}\left(\boldsymbol{\Theta}\right) =\\
&\left[\left(\frac{\partial(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\kappa}\right)}\right)^{\mathsf{T}}\left(\frac{\partial(\ell\left(\boldsymbol{u}\right)+\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{u}\right)}\right)^{\mathsf{T}}\left(\frac{\partial(\mathscr{L}\left(\boldsymbol{y}_{\cdot l}^{\mathsf{s}}\,|\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{y}_{\cdot l})}\right)}\right)^{\mathsf{T}}\left(\frac{\partial(\ell\left(\boldsymbol{u}\right))}{\partial\left(\boldsymbol{\phi}^{(\boldsymbol{u})}\right)}\right)^{\mathsf{T}}\right]
\end{aligned} \tag{B.60}$$

and Hessian matrix

$$\mathbf{H}_{\mathscr{L}}(\boldsymbol{\Theta}) =$$

$$\begin{bmatrix} \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial(\boldsymbol{\kappa})\partial(\boldsymbol{\kappa})} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial(\boldsymbol{\kappa})\partial(\boldsymbol{u})} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial(\boldsymbol{\kappa})\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial(\boldsymbol{\kappa})\partial\left(\boldsymbol{\phi}^{(u)}\right)} \\ \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial(\boldsymbol{u})\partial(\boldsymbol{\kappa})} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u})+\mathscr{L}(\boldsymbol{u}))}{\partial(\boldsymbol{u})\partial(\boldsymbol{u})} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial(\boldsymbol{u})\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u})+\mathscr{L}(\boldsymbol{u}))}{\partial(\boldsymbol{u})\partial\left(\boldsymbol{\phi}^{(u)}\right)} \\ \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)\partial(\boldsymbol{\kappa})} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)\partial(\boldsymbol{u})} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)} & \mathbf{0} \\ \mathbf{0} & \frac{\partial^2(\mathscr{L}(\boldsymbol{u}))}{\partial\left(\boldsymbol{\phi}^{(u)}\right)\partial(\boldsymbol{u})} & \mathbf{0} & \frac{\partial^2(\mathscr{L}(\boldsymbol{u}))}{\partial\left(\boldsymbol{\phi}^{(u)}\right)\partial\left(\boldsymbol{\phi}^{(u)}\right)} \end{bmatrix}. \tag{B.61}$$

The expected value of the Hessian used for Fisher scoring algorithm is given by

$$\mathrm{E}\left(\mathbf{H}_{\mathscr{L}}(\boldsymbol{\Theta})\right) = \mathrm{E}\left(\begin{bmatrix} \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial(\boldsymbol{\kappa})\partial(\boldsymbol{\kappa})} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial(\boldsymbol{\kappa})\partial(\boldsymbol{u})} & \mathbf{0} & \mathbf{0} \\ \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial(\boldsymbol{u})\partial(\boldsymbol{\kappa})} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u})+\mathscr{L}(\boldsymbol{u}))}{\partial(\boldsymbol{u})\partial(\boldsymbol{u})} & \mathbf{0} & \frac{\partial^2(\mathscr{L}(\boldsymbol{u}))}{\partial(\boldsymbol{u})\partial\left(\boldsymbol{\phi}^{(u)}\right)} \\ \mathbf{0} & \mathbf{0} & \frac{\partial^2(\mathscr{L}(\boldsymbol{y}^{\mathsf{s}}_{\cdot l}|\boldsymbol{u}))}{\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)\partial\left(\boldsymbol{\phi}^{(y_{\cdot l})}\right)} & \mathbf{0} \\ \mathbf{0} & \frac{\partial^2(\mathscr{L}(\boldsymbol{u}))}{\partial\left(\boldsymbol{\phi}^{(u)}\right)\partial(\boldsymbol{u})} & \mathbf{0} & \frac{\partial^2(\mathscr{L}(\boldsymbol{u}))}{\partial\left(\boldsymbol{\phi}^{(u)}\right)\partial\left(\boldsymbol{\phi}^{(u)}\right)} \end{bmatrix}\right). \tag{B.62}$$

A useful characteristic when assuming $\boldsymbol{u}$ to be normally distributed as

$$\boldsymbol{u} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(u)}\right)\right) \tag{B.63}$$

is that by using equalities 5.18, this results in

$$\begin{aligned} \mathrm{E}\left(\frac{\partial^2(\mathscr{L}(\boldsymbol{u}))}{\partial\left(\boldsymbol{\phi}^{(u)}\right)\partial(\boldsymbol{u})}\right) &= \mathrm{E}\left(\frac{\partial^2(\mathscr{L}(\boldsymbol{u}))}{\partial(\boldsymbol{u})\partial\left(\boldsymbol{\phi}^{(u)}\right)}\right) \\ &= \mathrm{E}\left(\frac{\partial\left(\left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(u)}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}^{(u)}\right)}\boldsymbol{\theta}_{\boldsymbol{u}}\right) + \mathrm{E}\left(\frac{\partial^2\left(\mathbf{c}_{\boldsymbol{u}}\left(\boldsymbol{u},\boldsymbol{\phi}^{(u)}\right)\right)}{\partial\left(\boldsymbol{\phi}^{(u)}\right)\partial(\boldsymbol{u})}\right) \\ &= \mathrm{E}\left(\frac{1}{2}\cdot\frac{\partial\left(\left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(u)}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}^{(u)}\right)}\boldsymbol{\mu}_{\boldsymbol{u}}^{\circ 2}\right) - \mathrm{E}\left(\frac{\partial\left(\left(\mathbf{a}_{\boldsymbol{u}}\left(\boldsymbol{\phi}^{(u)}\right)\right)^{-1}\right)}{\partial\left(\boldsymbol{\phi}^{(u)}\right)}\boldsymbol{u}\right) \\ &= \mathbf{0} \quad , \end{aligned} \tag{B.64}$$

thus rendering the expected Hessian matrix defined in equation B.62 as block-diagonal. In that case, the updating rules can be split in a block-wise manner, as discussed in section 5.1.5 (cf. Rao, 2003, pp. 100 ff; Searle, Casella and McCulloch, 2006, pp. 235 ff).

B.4.4 Derivatives of B-splines

B.4.4.1 Derivatives of B-splines With Respect to the Knots

For B-splines defined in equations 5.58, the derivative with respect to the knots can be found as

$$\frac{\partial\left(\mathbf{B}_k^l\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right)}{\partial\left(K_m^{\boldsymbol{x}_{\cdot j}}\right)} = \lim_{\epsilon \to 0}\left(\frac{\mathbf{B}_k^l\left(x_{ij}, \boldsymbol{I}^{\boldsymbol{x}_{\cdot j}}\right) - \mathbf{B}_k^l\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{\epsilon}\right) \quad , \tag{B.65}$$

where

$$I_k^{\boldsymbol{x}_{\cdot j}} := \begin{cases} K_k^{\boldsymbol{x}_{\cdot j}} + \epsilon & \text{if} \quad k = m \\ K_k^{\boldsymbol{x}_{\cdot j}} & \text{else} \end{cases} \tag{B.66}$$

is the original knot vector with $K_m^{\boldsymbol{x}_{\cdot j}}$ increased by ϵ. Equation B.65 can be computed by inserting $K_m^{\boldsymbol{x}_{\cdot j}}$ into $\boldsymbol{I}^{\boldsymbol{x}_{\cdot j}}$ and $I_m^{\boldsymbol{x}_{\cdot j}}$ into $\boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}$, such that

$$\widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}} := \left[K_1^{\boldsymbol{x}_{\cdot j}} \quad \cdots \quad K_m^{\boldsymbol{x}_{\cdot j}} \quad I_m^{\boldsymbol{x}_{\cdot j}} \quad \cdots \quad K_u^{\boldsymbol{x}_{\cdot j}}\right] \tag{B.67a}$$

or, equivalently,

$$\widetilde{K}_k^{\boldsymbol{x}_{\cdot j}} = \begin{cases} K_k^{\boldsymbol{x}_{\cdot j}} & \text{if} \quad k \leq m \\ I_k^{\boldsymbol{x}_{\cdot j}} & \text{if} \quad k = m+1 \\ K_{k+1}^{\boldsymbol{x}_{\cdot j}} & \text{if} \quad k > m+1 \end{cases} \tag{B.67b}$$

is used for both functions. Using the knot inserting formulae for B-splines derived by Böhm (1980; cf. also Böhm, Farin and Kahmann, 1984, p. 18; Eck and Hadenfeld, 1995, p. 260), the two B-spline base functions in equality B.65 can be reformulated using the common knot vector $\widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}$:

$$\mathbf{B}_k^l\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) = \begin{cases} \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \text{if} \quad 0 < k < m-l \\ \dfrac{I_m^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \\ \quad + \dfrac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - I_m^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \text{if} \quad m-l \leq k \leq m \\ \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \text{if} \quad m < k < u \end{cases} \tag{B.68a}$$

and

$$\mathbf{B}_k^l\left(x_{ij}, \boldsymbol{I}^{\boldsymbol{x}_{\cdot j}}\right) = \begin{cases} \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \text{if} \quad 0 < k < m-l-1 \\ \dfrac{K_m^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \\ \quad + \dfrac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - K_m^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \text{if} \quad m-l-1 \leq k < m \\ \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \text{if} \quad m \leq k < u \end{cases} \tag{B.68b}$$

(cf. Piegl and Tiller, 1998, p. 931).

Combining equalities B.68, the numerator of equality B.65 is

$$\mathbf{B}_k^l\left(x_{ij}, \boldsymbol{I}^{\boldsymbol{x}_{\cdot j}}\right) - \mathbf{B}_k^l\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \;=$$

$$\begin{cases} 0 & \text{if} \quad 0 < k < m-l-1 \\ \\ \dfrac{K_m^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \\ + \dfrac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - K_m^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \text{if} \quad k = m-l-1 \\ - \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \\ \\ \dfrac{K_m^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \\ + \dfrac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - K_m^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \\ - \dfrac{I_m^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \text{if} \quad m-l \leq k \leq m-1 \\ - \dfrac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - I_m^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \\ \\ \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \\ - \dfrac{I_m^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \text{if} \quad k = m \\ - \dfrac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - I_m^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) & \\ \\ 0 & \text{if} \quad m < k < u \quad . \end{cases} \tag{B.69}$$

The different cases in equality B.69 can be simplified. In the first non-zero case ($k = m-l-1$), one obtains

$$\begin{aligned} & \frac{K_m^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) + \frac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - K_m^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) - \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) \\ = & \frac{\widetilde{K}_m^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_m^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) + \frac{\widetilde{K}_{m+1}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_m^{\boldsymbol{x}_{\cdot j}}}{\widetilde{K}_{m+1}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) - \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) \\ = & \frac{\epsilon}{\widetilde{K}_{m+1}^{\boldsymbol{x}_{\cdot j}} - \widetilde{K}_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right) \\ = & \frac{\epsilon \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right)}{K_m^{\boldsymbol{x}_{\cdot j}} + \epsilon - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \quad . \end{aligned} \tag{B.70}$$

For the third case, where $m-l \leq k \leq m-1$, the result is

$$\begin{aligned}
& \frac{K_m^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) + \frac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - K_m^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - \widetilde{K}_{k+1}^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) \\
- & \frac{I_m^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) - \frac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - I_m^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - \widetilde{K}_{k+1}^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) \\
= & \frac{K_m^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j} - I_m^{\boldsymbol{x}\cdot j} + \widetilde{K}_k^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) \\
& + \frac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - K_m^{\boldsymbol{x}\cdot j} - \widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} + I_m^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - \widetilde{K}_{k+1}^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) \\
= & \frac{K_m^{\boldsymbol{x}\cdot j} - I_m^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) \\
& + \frac{I_m^{\boldsymbol{x}\cdot j} - K_m^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - \widetilde{K}_{k+1}^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) \\
= & \frac{\epsilon \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right)}{K_{k+l+1}^{\boldsymbol{x}\cdot j} - K_k^{\boldsymbol{x}\cdot j} - \epsilon} - \frac{\epsilon \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right)}{K_{k+l}^{\boldsymbol{x}\cdot j} - K_k^{\boldsymbol{x}\cdot j}} \quad .
\end{aligned} \tag{B.71}$$

For $k = m$, expression B.69 can be rewritten as

$$\begin{aligned}
\mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) - & \frac{I_m^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) \\
& - \frac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - I_m^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - \widetilde{K}_{k+1}^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) \\
= \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) - & \frac{I_m^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+1}^{\boldsymbol{x}\cdot j} - \widetilde{K}_k^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) \\
& - \frac{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - \widetilde{K}_{m+1}^{\boldsymbol{x}\cdot j}}{\widetilde{K}_{k+l+2}^{\boldsymbol{x}\cdot j} - \widetilde{K}_{m+1}^{\boldsymbol{x}\cdot j}} \cdot \mathbf{B}_{k+1}^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) \\
= - & \frac{\epsilon \cdot \mathbf{B}_k^l\left(x_{ij}, \widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right)}{K_{k+l}^{\boldsymbol{x}\cdot j} - K_m^{\boldsymbol{x}\cdot j}} \quad .
\end{aligned} \tag{B.72}$$

Due to definitions B.66 and B.67, the knot vector $\widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}$ tends to $\boldsymbol{K}^{\boldsymbol{x}\cdot j}$ with duplicated knot $K_m^{\boldsymbol{x}\cdot j}$ for $\epsilon \to 0$:

$$\lim_{\epsilon \to 0}\left(\widetilde{\boldsymbol{K}}^{\boldsymbol{x}\cdot j}\right) = \overline{\boldsymbol{K}}^{\boldsymbol{x}\cdot j} := \left[K_1^{\boldsymbol{x}\cdot j} \quad \cdots \quad K_m^{\boldsymbol{x}\cdot j} \quad K_m^{\boldsymbol{x}\cdot j} \quad \cdots \quad K_u^{\boldsymbol{x}\cdot j}\right] \quad . \tag{B.73}$$

Combining these results, the derivative defined in equation B.65 is

$$\frac{\partial\left(\mathbf{B}_k^l\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right)}{\partial\left(K_m^{\boldsymbol{x}_{\cdot j}}\right)} = \begin{cases} 0 & \text{if} \quad 0 < k < m-l-1 \\ \dfrac{\mathbf{B}_{k+1}^l\left(x_{ij}, \overline{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right)}{K_m^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} & \text{if} \quad k = m-l-1 \\ \dfrac{\mathbf{B}_{k+1}^l\left(x_{ij}, \overline{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \dfrac{\mathbf{B}_k^l\left(x_{ij}, \overline{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+l}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} & \text{if} \quad m-l \leq k \leq m-1 \\ -\dfrac{\mathbf{B}_k^l\left(x_{ij}, \overline{\boldsymbol{K}}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+l}^{\boldsymbol{x}_{\cdot j}} - K_m^{\boldsymbol{x}_{\cdot j}}} & \text{if} \quad k = m \\ 0 & \text{if} \quad m < k < u \quad . \end{cases} \tag{B.74}$$

which is equality 5.70 (cf. Piegl and Tiller, 1998, p. 931).

B.4.4.2 Derivatives of B-splines With Respect to the Input Variable

In the following, derivatives for B-spline functions up to degree $l = 2$ with respect to the input variables $\boldsymbol{X}$ are provided. Derivatives of higher order splines follow recursively.

For B-spline base functions defined in equations 5.58, the derivative for a base function of degree $l = 0$ w.r.t. $\boldsymbol{X}$ is given by

$$\frac{\partial\left(\mathbf{B}_k^0\left(x_{ij}, \boldsymbol{K}^{x_{ij}}\right)\right)}{\partial\left(x_{ij}\right)} = 0 \tag{B.75}$$

Using this result for base functions of degree $l = 1$, one obtains

$$\begin{aligned} \frac{\partial\left(\mathbf{B}_k^1\left(x_{ij}, \boldsymbol{K}^{x_{ij}}\right)\right)}{\partial\left(x_{ij}\right)} &= \frac{\partial\left(\dfrac{x_{ij} - K_k^{\boldsymbol{x}_{\cdot j}}}{K_{k+1}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}}\right)}{\partial\left(x_{ij}\right)} \cdot \mathbf{B}_k^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \\ &\quad - \frac{x_{ij} - K_k^{\boldsymbol{x}_{\cdot j}}}{K_{k+1}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \cdot \frac{\partial\left(\mathbf{B}_k^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right)}{\partial\left(x_{ij}\right)} \\ &\quad + \frac{\partial\left(\dfrac{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - x_{ij}}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}}\right)}{\partial\left(x_{ij}\right)} \cdot \mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) \\ &\quad - \frac{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - x_{ij}}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \frac{\partial\left(\mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right)}{\partial\left(x_{ij}\right)} \\ &= \frac{\mathbf{B}_k^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+1}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \quad . \end{aligned} \tag{B.76}$$

For base functions of degree $l = 2$, the result is

$$
\begin{aligned}
&\frac{\partial\left(\mathbf{B}_k^2\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right)}{\partial\left(x_{ij}\right)} \\
=&\frac{\partial\left(\dfrac{x_{ij} - K_k^{\boldsymbol{x}_{\cdot j}}}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}}\right)}{\partial\left(x_{ij}\right)} \cdot \mathbf{B}_k^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) + \frac{x_{ij} - K_k^{\boldsymbol{x}_{\cdot j}}}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \cdot \frac{\partial\left(\mathbf{B}_k^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right)}{\partial\left(x_{ij}\right)} \\
&+ \frac{\partial\left(\dfrac{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - x_{ij}}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}}\right)}{\partial\left(x_{ij}\right)} \cdot \mathbf{B}_{k+1}^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) + \frac{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - x_{ij}}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \frac{\partial\left(\mathbf{B}_{k+1}^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right)}{\partial\left(x_{ij}\right)} \\
=&\frac{\mathbf{B}_k^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+1}^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \\
&+ \frac{x_{ij} - K_k^{\boldsymbol{x}_{\cdot j}}}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \cdot \left(\frac{\mathbf{B}_k^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+1}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}}\right) \\
&+ \frac{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - x_{ij}}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \left(\frac{\mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+2}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+2}^{\boldsymbol{x}_{\cdot j}}}\right) \\
=&\frac{\mathbf{B}_k^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+1}^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \\
&+ \frac{1}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \cdot \left(\frac{x_{ij} - K_k^{\boldsymbol{x}_{\cdot j}}}{K_{k+1}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) + \frac{K_k^{\boldsymbol{x}_{\cdot j}} - x_{ij}}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right) \\
&+ \frac{1}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \cdot \left(\frac{x_{ij} - K_k^{\boldsymbol{x}_{\cdot j}}}{K_{k+1}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) + \frac{K_k^{\boldsymbol{x}_{\cdot j}} - x_{ij}}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right) \\
&- \frac{1}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \left(\frac{x_{ij} - K_{k+3}^{\boldsymbol{x}_{\cdot j}}}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right) + \frac{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - x_{ij}}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+2}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+2}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right) \\
=&\frac{\mathbf{B}_k^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+1}^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \\
&+ \frac{1}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \cdot \left(\frac{x_{ij} - K_k^{\boldsymbol{x}_{\cdot j}}}{K_{k+1}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_k^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right. \\
&\qquad \left. + \frac{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - x_{ij} - K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right) \\
&- \frac{1}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \left(\frac{x_{ij} - K_{k+1}^{\boldsymbol{x}_{\cdot j}} + K_{k+1}^{\boldsymbol{x}_{\cdot j}} - K_{k+3}^{\boldsymbol{x}_{\cdot j}}}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right. \\
&\qquad \left. + \frac{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - x_{ij}}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+2}^{\boldsymbol{x}_{\cdot j}}} \cdot \mathbf{B}_{k+2}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right) \\
=&\frac{\mathbf{B}_k^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+1}^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} + \frac{\mathbf{B}_k^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} \\
&- \frac{\mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+1}^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} + \frac{\mathbf{B}_{k+1}^0\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}} \\
=&2 \cdot \left(\frac{\mathbf{B}_k^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+2}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+1}^1\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+3}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}}\right) \qquad .
\end{aligned}
\tag{B.77}
$$

The same pattern recursively occurs for derivatives of higher order B-splines ($l > 2$) , such that

$$\frac{\partial\left(\mathbf{B}_k^l\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)\right)}{\partial\left(x_{ij}\right)} = l \cdot \left(\frac{\mathbf{B}_k^{l-1}\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+l}^{\boldsymbol{x}_{\cdot j}} - K_k^{\boldsymbol{x}_{\cdot j}}} - \frac{\mathbf{B}_{k+1}^{l-1}\left(x_{ij}, \boldsymbol{K}^{\boldsymbol{x}_{\cdot j}}\right)}{K_{k+l+1}^{\boldsymbol{x}_{\cdot j}} - K_{k+1}^{\boldsymbol{x}_{\cdot j}}}\right) \quad , \tag{B.78}$$

which constitutes equality 5.67 and can be proven in different ways (cf. de Boor, 1972; 1978, p. 138; Procházková and Procházka, 2007, p. 6)

B.4.5 Derivation of Support Vector Machines

B.4.5.1 Derivation of the Support Vector Classifier

The support vector classifier for a variable $\boldsymbol{y}_{\cdot l}^{\mathsf{s}}$ with values $y_{il} \in \{-1; 1\}$ is defined by

$$\widehat{\boldsymbol{y}}_{\cdot l}^{\mathsf{s}} = \operatorname{sign}\left(\widetilde{\boldsymbol{X}}^{\mathsf{s}}\boldsymbol{\beta}\right) \quad , \tag{B.79}$$

where $\widetilde{\boldsymbol{X}}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}} \times h}$ with $\widetilde{\boldsymbol{x}}_{\cdot 1}^{\mathsf{s}} \coloneqq \mathbf{1}_{n^{\mathsf{s}} \times 1}$ as intercept column. The coefficients $\boldsymbol{\beta}$ are determined by

$$\begin{aligned} \boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} & \left(\frac{1}{2} \cdot \sum_{j=2}^{h} \beta_j^2 + \sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i\right) \\ \text{s.t.} \quad & \xi_i \geq 0 \quad \text{for all} \quad i = 1, \ldots, n^{\mathsf{s}} \\ & y_{il} \cdot \left(\widetilde{\boldsymbol{x}}_{i\cdot}^{\mathsf{s}}\boldsymbol{\beta}\right) \geq 1 - \xi_i \quad \text{for all} \quad i = 1, \ldots, n^{\mathsf{s}} \quad , \end{aligned} \tag{B.80}$$

where $\boldsymbol{\xi} \in \mathbb{R}_{\geq 0}^{n^{\mathsf{s}}}$ is a slack-variable and $\boldsymbol{c} \in \mathbb{R}_{\geq 0}^{n^{\mathsf{s}}}$ a corresponding penalty-parameter. The primal Lagrange function is

$$\mathrm{L}_P\left(\boldsymbol{\beta}\right) = \frac{1}{2} \cdot \sum_{j=2}^{h} \beta_j^2 + \sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \left(y_{il}\left(\widetilde{\boldsymbol{x}}_{i\cdot}^{\mathsf{s}}\boldsymbol{\beta}\right) - \left(1 - \xi_i\right)\right) - \sum_{i=1}^{n^{\mathsf{s}}} \lambda_i \cdot \xi_i \quad , \tag{B.81}$$

where $\boldsymbol{\alpha}, \boldsymbol{\lambda} \in \mathbb{R}^{n^{\mathsf{s}}}$ are vectors of Lagrange multipliers for the constraints in problem B.80. The saddle point at the minimum with respect to $\left[\boldsymbol{\beta}^{\mathsf{T}} \quad \boldsymbol{\xi}^{\mathsf{T}}\right]^{\mathsf{T}}$ and the maximum w.r.t. $\left[\boldsymbol{\alpha}^{\mathsf{T}} \quad \boldsymbol{\lambda}^{\mathsf{T}}\right]^{\mathsf{T}}$ is an optimum of problem B.81 (cf. Geiger and Kanzow, 2002, p. 316). Calculating the corresponding derivatives yields

$$\frac{\partial(\mathrm{L}_P\left(\boldsymbol{\beta}\right))}{\partial\left(\beta_1\right)} = \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \cdot \frac{\partial(\beta_1)}{\partial\left(\beta_1\right)} = \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \tag{B.82a}$$

$$\begin{aligned} \frac{\partial(\mathrm{L}_P\left(\boldsymbol{\beta}\right))}{\partial\left(\beta_j\right)} &= \frac{\partial\left(\frac{1}{2} \cdot \sum_{j=2}^{h} \beta_j^2\right)}{\partial\left(\beta_j\right)} - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \cdot \frac{\partial(\widetilde{\boldsymbol{x}}_{i\cdot}^{\mathsf{s}}\boldsymbol{\beta})}{\partial\left(\beta_j\right)} \\ &= \beta_j - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \cdot \widetilde{x}_{ij}^{\mathsf{s}} \quad \text{for all} \quad j = 2, \ldots, h \end{aligned} \tag{B.82b}$$

$$\begin{aligned} \frac{\partial(\mathrm{L}_P\left(\boldsymbol{\beta}\right))}{\partial\left(\xi_i\right)} &= \frac{\partial\left(\sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i\right)}{\partial\left(\xi_i\right)} + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \frac{\partial(1 - \xi_i)}{\partial\left(\xi_i\right)} - \sum_{i=1}^{n^{\mathsf{s}}} \lambda_i \cdot \frac{\partial(\xi_i)}{\partial\left(\xi_i\right)} \\ &= c_i - \alpha_i - \lambda_i \quad . \end{aligned} \tag{B.82c}$$

Setting these derivatives to zero results in

$$0 = \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \quad , \tag{B.83a}$$

$$\beta_j = \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \cdot \widetilde{x}_{ij}^{\mathsf{s}} \quad \text{for all} \quad j > 1 \tag{B.83b}$$

and

$$0 = c_i - \alpha_i - \lambda_i \quad . \tag{B.83c}$$

Plugging equalities B.83 into the components of equation B.81, the result are

$$\begin{aligned} \frac{1}{2} \cdot \sum_{j=2}^{h} \beta_j^2 &= \frac{1}{2} \cdot \sum_{j=2}^{h} \left(\sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \cdot \widetilde{x}_{ij}^{\mathsf{s}} \right)^2 \\ &= \frac{1}{2} \cdot \sum_{k=2}^{h} \left(\left(\sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \cdot \widetilde{x}_{ik}^{\mathsf{s}} \right) \cdot \left(\sum_{j=1}^{n^{\mathsf{s}}} \alpha_j \cdot y_{jl} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \right) \right) \\ &= \frac{1}{2} \cdot \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} \sum_{k=2}^{h} \alpha_i \cdot \alpha_j \cdot y_{il} \cdot y_{jl} \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \end{aligned} \tag{B.84a}$$

and

$$\begin{aligned} &\sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \left(y_{il} \left(\widetilde{\boldsymbol{x}}_{i.}^{\mathsf{s}} \boldsymbol{\beta} \right) - (1 - \xi_i) \right) \\ &= \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \left(y_{il} \left(\beta_1 + \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \left(\sum_{k=1}^{n^{\mathsf{s}}} \alpha_k \cdot y_{kl} \cdot \widetilde{x}_{kj}^{\mathsf{s}} \right) \right) \right) - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot (1 - \xi_i) \\ &= \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \left(\sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \left(\sum_{k=1}^{n^{\mathsf{s}}} \alpha_k \cdot y_{kl} \cdot \widetilde{x}_{kj}^{\mathsf{s}} \right) \right) + \beta_1 \cdot \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot (1 - \xi_i) \\ &= \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} \sum_{k=2}^{h} \alpha_i \cdot \alpha_j \cdot y_{il} \cdot y_{jl} \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \xi_i \quad . \end{aligned} \tag{B.84b}$$

Using equalities B.84, the resulting dual Lagrange function is

$$\begin{aligned} \mathrm{L}_D(\boldsymbol{\alpha}) &= \frac{1}{2} \cdot \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} \sum_{k=2}^{h} \alpha_i \cdot \alpha_j \cdot y_{il} \cdot y_{jl} \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \\ &\quad - \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} \sum_{k=2}^{h} \alpha_i \cdot \alpha_j \cdot y_{il} \cdot y_{jl} \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \\ &\quad + \sum_{i=1}^{n^{\mathsf{s}}} (c_i - \alpha_i - \lambda_i) \cdot \xi_i \\ &= \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i - \frac{1}{2} \cdot \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} \sum_{k=2}^{h} \alpha_i \cdot \alpha_j \cdot y_{il} \cdot y_{jl} \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \quad . \end{aligned} \tag{B.85}$$

Thus, the saddle-point of problem B.81 can be found by

$$\begin{aligned} \boldsymbol{\alpha}^* \;&=\; \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \left(\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\mathrm{L}_D \left(\boldsymbol{\alpha} \right) \right) \right) \;=\; \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left(-\mathrm{L}_D \left(\boldsymbol{\alpha} \right) \right) \\ &\text{s.t.} \qquad 0 \;\leq\; \alpha_i \;\leq\; c_i \qquad \text{for all} \quad i \;= 1, \ldots, n^{\mathsf{s}} \\ &\qquad\quad \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \;\overset{!}{=}\; 0 \end{aligned} \tag{B.86}$$

since $\mathrm{L}_D \left(\boldsymbol{\alpha} \right)$ is independent of $\boldsymbol{\beta}$.

Computation of parameters

From equations B.83, β_j can be directly computed as

$$\beta_j \;=\; \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il} \cdot \widetilde{x}_{ij}^{\mathsf{s}} \quad \text{for all} \quad j > 1 \qquad . \tag{B.87}$$

The Karush-Kuhn-Tucker optimality criteria (cf. Karush, 1939; Kuhn and Tucker, 1951) in combination with equations B.83 furthermore require that

$$\xi_i \;\geq 0 \tag{B.88a}$$

$$\alpha_i \cdot \left(y_{il} \left(\beta_1 + \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \beta_j \right) - \left(1 - \xi_i \right) \right) \;= 0 \tag{B.88b}$$

$$\lambda_i \cdot \xi_i \;=\; \left(c_i - \alpha_i \right) \cdot \xi_i \;= 0 \qquad . \tag{B.88c}$$

For $0 < \alpha_i$, condition B.88b can be reformulated as

$$\beta_1 \;= y_{il} - \sum_{k=2}^{h} \widetilde{x}_{ik}^{\mathsf{s}} \cdot \beta_k - \xi_i \quad \text{for all } i \in \{k : y_{kl} = \;\; 1\} \tag{B.89a}$$

$$\beta_1 \;= y_{il} - \sum_{k=2}^{h} \widetilde{x}_{ik}^{\mathsf{s}} \cdot \beta_k + \xi_i \quad \text{for all } i \in \{k : y_{kl} = -1\} \qquad . \tag{B.89b}$$

It is evident that β_1 can be computed directly from equations B.89a and B.89b if any $\xi_i = 0$. As follows from equations B.88c, this is the case if any $0 < \alpha_i < c_i$ holds. For numerical stability, the average of all values fulfilling the above conditions is used:

$$\beta_1 \;=\; \mathrm{E} \left(y_{il} - \sum_{j=2}^{h} x_{ij} \cdot \beta_j \middle| 0 < \alpha_i < c_i \right) \qquad . \tag{B.90}$$

If the condition is not met by any α_i, so $\{i : 0 < \alpha_i < c_i\} = \emptyset$, the mid-point of the interval determined by inequality B.88a is used as approximation:

$$\begin{aligned} \beta_1 \;\approx\; \frac{1}{2} \cdot \Bigg(&\operatorname{Max} \left(y_{il} - \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \beta_j \middle| \alpha_i = c_i, y_{il} = -1 \right) \\ &+ \operatorname{Min} \left(y_{il} - \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \beta_j \middle| \alpha_i = c_i, y_{il} = 1 \right) \Bigg) \end{aligned} \tag{B.91}$$

(cf. Chang and Lin, 2011, p. 10; Smola and Schölkopf, 2004, p. 201).

B.4.5.2 Derivation of the Support Vector Regression

The predictions in a support vector regression are defined by

$$\widehat{\boldsymbol{y}}^{\mathsf{s}}_{.l} = \widetilde{\boldsymbol{X}}^{\mathsf{s}}\boldsymbol{\beta} \quad , \tag{B.92}$$

where, as before, $\widetilde{\boldsymbol{X}}^{\mathsf{s}} \in \mathbb{R}^{n^{\mathsf{s}} \times h}$ with $\widetilde{\boldsymbol{x}}^{\mathsf{s}}_{.1} \coloneqq \mathbf{1}_{n^{\mathsf{s}} \times 1}$ as intercept column. The parameters are determined by

$$\begin{aligned} \boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} & \left(\frac{1}{2} \cdot \sum_{j=2}^{h} \beta_j^2 + \sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i + \sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i^* \right) \\ \text{s. t.} \quad & \xi_i \,,\ \xi_i^* \geq 0 \qquad \text{for all} \quad i = 1, \ldots, n^{\mathsf{s}} \\ & y^{\mathsf{s}}_{il} - \widetilde{\boldsymbol{x}}^{\mathsf{s}}_{i.}\boldsymbol{\beta} \leq e + \xi_i \quad \text{for all} \quad i = 1, \ldots, n^{\mathsf{s}} \\ & -y^{\mathsf{s}}_{il} + \widetilde{\boldsymbol{x}}^{\mathsf{s}}_{i.}\boldsymbol{\beta} \leq e + \xi_i^* \quad \text{for all} \quad i = 1, \ldots, n^{\mathsf{s}} \quad , \end{aligned} \tag{B.93}$$

where $\boldsymbol{\xi}, \boldsymbol{\xi}^* \in \mathbb{R}^{n^{\mathsf{s}}}_{\geq 0}$ are slack-variables and $\boldsymbol{c} \in \mathbb{R}^{n^{\mathsf{s}}}_{\geq 0}$ a penalty-parameter attributed to them. The primal Lagrange function is

$$\begin{aligned} \mathrm{L}_P(\boldsymbol{\beta}) = & \frac{1}{2} \cdot \sum_{j=2}^{h} \beta_j^2 + \sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i + \sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i^* - \sum_{i=1}^{n^{\mathsf{s}}} \lambda_i \cdot \xi_i - \sum_{i=1}^{n^{\mathsf{s}}} \lambda_i^* \cdot \xi_i^* \\ & - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot (e + \xi_i - y^{\mathsf{s}}_{il} + \widetilde{\boldsymbol{x}}^{\mathsf{s}}_{i.}\boldsymbol{\beta}) \\ & - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot (e + \xi_i^* + y^{\mathsf{s}}_{il} - \widetilde{\boldsymbol{x}}^{\mathsf{s}}_{i.}\boldsymbol{\beta}) \quad , \end{aligned} \tag{B.94}$$

with Lagrange multipliers $\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}, \boldsymbol{\lambda}^* \in \mathbb{R}^{n^{\mathsf{s}}}$ for the constraints in problem B.93. As before, the saddle point at the minimum with respect to $\left[\boldsymbol{\beta}^{\mathsf{T}} \quad \boldsymbol{\xi}^{\mathsf{T}} \quad (\boldsymbol{\xi}^*)^{\mathsf{T}}\right]^{\mathsf{T}}$ and the maximum w.r.t. $\left[\boldsymbol{\alpha}^{\mathsf{T}} \quad (\boldsymbol{\alpha}^*)^{\mathsf{T}} \quad \boldsymbol{\lambda}^{\mathsf{T}} \quad (\boldsymbol{\lambda}^*)^{\mathsf{T}}\right]^{\mathsf{T}}$ is an optimum of problem B.94 (cf. Geiger and Kanzow, 2002, p. 316). Calculating the corresponding derivatives yields

$$\frac{\partial(\mathrm{L}_P(\boldsymbol{\beta}))}{\partial(\beta_1)} = -\sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \frac{\partial(\beta_1)}{\partial(\beta_1)} - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot \frac{\partial(-\beta_1)}{\partial(\beta_1)} = \sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i^* - \alpha_i) \tag{B.95a}$$

$$\begin{aligned} \frac{\partial(\mathrm{L}_P(\boldsymbol{\beta}))}{\partial(\beta_j)} = & \frac{\partial\left(\frac{1}{2} \cdot \sum_{j=2}^{h} \beta_j^2\right)}{\partial(\beta_j)} - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \frac{\partial\left(\sum_{j=2}^{h} \widetilde{x}^{\mathsf{s}}_{ij} \cdot \beta_j\right)}{\partial(\beta_j)} - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot \frac{\partial\left(-\sum_{j=2}^{h} \widetilde{x}^{\mathsf{s}}_{ij} \cdot \beta_j\right)}{\partial(\beta_j)} \\ = & \beta_j - \sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \widetilde{x}^{\mathsf{s}}_{ij} \end{aligned} \tag{B.95b}$$

$$\frac{\partial(\mathrm{L}_P(\boldsymbol{\beta}))}{\partial(\xi_i)} = \frac{\partial\left(\sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i\right)}{\partial(\xi_i)} - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \frac{\partial(\xi_i)}{\partial(\xi_i)} - \sum_{i=1}^{n^{\mathsf{s}}} \lambda_i \cdot \frac{\partial(\xi_i)}{\partial(\xi_i)} = c_i - \alpha_i - \lambda_i \tag{B.95c}$$

$$\frac{\partial(\mathrm{L}_P(\boldsymbol{\beta}))}{\partial(\xi_i^*)} = \frac{\partial\left(\sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i^*\right)}{\partial(\xi_i^*)} - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot \frac{\partial(\xi_i^*)}{\partial(\xi_i^*)} - \sum_{i=1}^{n^{\mathsf{s}}} \lambda_i^* \cdot \frac{\partial(\xi_i^*)}{\partial(\xi_i^*)} = c_i - \alpha_i^* - \lambda_i^* \ . \tag{B.95d}$$

Setting these derivatives to zero results in

$$0 \;=\; \sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i^* - \alpha_i) \tag{B.96a}$$

$$\beta_j \;=\; \sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \widetilde{x}_{ij}^{\mathsf{s}} \tag{B.96b}$$

$$0 \;= c_i - \alpha_i - \lambda_i \;= c_i - \alpha_i^* - \lambda_i^* \qquad . \tag{B.96c}$$

By plugging these results into the components of equation B.94, one obtains

$$\begin{aligned} \frac{1}{2} \cdot \sum_{j=2}^{h} \beta_j^2 \;&= \frac{1}{2} \cdot \sum_{j=2}^{h} \left(\sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \widetilde{x}_{ij}^{\mathsf{s}} \right)^2 \\ &= \frac{1}{2} \cdot \sum_{k=2}^{h} \left(\left(\sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \right) \cdot \left(\sum_{j=1}^{n^{\mathsf{s}}} \left(\alpha_j - \alpha_j^* \right) \cdot \widetilde{x}_{jk}^{\mathsf{s}} \right) \right) \\ &= \frac{1}{2} \cdot \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \left(\alpha_j - \alpha_j^* \right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \qquad , \end{aligned} \tag{B.97a}$$

$$\begin{aligned} \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot (e + \xi_i - y_{il}^{\mathsf{s}} + \widetilde{\boldsymbol{x}}_{i.}^{\mathsf{s}} \boldsymbol{\beta}) \;&= \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \beta_1 + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot e + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \xi_i - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il}^{\mathsf{s}} \\ &\quad + \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} \alpha_i \cdot \left(\alpha_j - \alpha_j^* \right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \end{aligned} \tag{B.97b}$$

and

$$\begin{aligned} \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot (e + \xi_i + y_{il}^{\mathsf{s}} - \widetilde{\boldsymbol{x}}_{i.}^{\mathsf{s}} \boldsymbol{\beta}) \;&= -\sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot \beta_1 + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot e + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot \xi_i^* + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot y_{il}^{\mathsf{s}} \\ &\quad - \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot \left(\alpha_j - \alpha_j^* \right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \qquad . \end{aligned} \tag{B.97c}$$

The resulting dual Lagrange function is hence

$$\begin{aligned} \mathrm{L}_D(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) \;&= \frac{1}{2} \cdot \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \left(\alpha_j - \alpha_j^* \right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \\ &\quad + \sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i + \sum_{i=1}^{n^{\mathsf{s}}} c_i \cdot \xi_i^* - \sum_{i=1}^{n^{\mathsf{s}}} \lambda_i \cdot \xi_i - \sum_{i=1}^{n^{\mathsf{s}}} \lambda_i^* \cdot \xi_i^* \\ &\quad + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il}^{\mathsf{s}} - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \beta_1 + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot e + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot \xi_i \\ &\quad - \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} \alpha_i \cdot \left(\alpha_j - \alpha_j^* \right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \\ &\quad - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot y_{il}^{\mathsf{s}} + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot \beta_1 + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot e + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot \xi_i^* \\ &\quad + \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot \left(\alpha_j - \alpha_j^* \right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2} \cdot \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \left(\alpha_j - \alpha_j^*\right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \\
&\quad + \sum_{i=1}^{n^{\mathsf{s}}} (c_i - \alpha_i - \lambda_i) \cdot \xi_i + \sum_{i=1}^{n^{\mathsf{s}}} (c_i - \alpha_i^* - \lambda_i^*) \cdot \xi_i^* \\
&\quad + \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} (\alpha_i^* - \alpha_i) \cdot \left(\alpha_j - \alpha_j^*\right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \\
&\quad + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il}^{\mathsf{s}} + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot e \\
&\quad - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot y_{il}^{\mathsf{s}} + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot e + \beta_1 \cdot \sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i^* - \alpha_i) \\
&= \frac{1}{2} \cdot \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \left(\alpha_j - \alpha_j^*\right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \\
&\quad - \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \left(\alpha_j - \alpha_j^*\right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \\
&\quad - \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot y_{il}^{\mathsf{s}} + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i^* \cdot e + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot y_{il}^{\mathsf{s}} + \sum_{i=1}^{n^{\mathsf{s}}} \alpha_i \cdot e \\
&= -\frac{1}{2} \cdot \sum_{k=2}^{h} \sum_{i=1}^{n^{\mathsf{s}}} \sum_{j=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \left(\alpha_j - \alpha_j^*\right) \cdot \widetilde{x}_{ik}^{\mathsf{s}} \cdot \widetilde{x}_{jk}^{\mathsf{s}} \\
&\quad + \sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot y_{il}^{\mathsf{s}} - e \cdot \sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i + \alpha_i^*) \qquad .
\end{aligned}
\tag{B.98}
$$

Thus, the saddle-point of problem B.94 can be found by

$$
\begin{aligned}
(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)^* \;=\; & \underset{(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)}{\operatorname{argmax}} \left(\underset{\beta}{\operatorname{argmin}} \left(\mathrm{L}_D \left(\boldsymbol{\alpha}, \boldsymbol{\alpha}^* \right) \right) \right) \;=\; \underset{(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)^{\mathsf{T}}}{\operatorname{argmin}} \left(-\mathrm{L}_D \left(\boldsymbol{\alpha}, \boldsymbol{\alpha}^* \right) \right) \\
\text{s. t.} \quad & 0 \;\leq\; \alpha_i, \;\leq\; c_i \qquad \text{for all} \quad i \;= 1, \ldots, n^{\mathsf{s}} \\
& 0 \;\leq\; \alpha_i^* \;\leq\; c_i \qquad \text{for all} \quad i \;= 1, \ldots, n^{\mathsf{s}} \\
& \sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \;=\; 0
\end{aligned}
\tag{B.99}
$$

since $\mathrm{L}_D \left(\boldsymbol{\alpha}, \boldsymbol{\alpha}^* \right)$ does not depend on $\boldsymbol{\beta}$.

Computation of parameters

From equation B.96b, β_j can be directly computed as

$$\beta_j \;=\; \sum_{i=1}^{n^{\mathsf{s}}} (\alpha_i - \alpha_i^*) \cdot \widetilde{x}_{ij}^{\mathsf{s}} \qquad . \tag{B.100}$$

The KKT-conditions introduced in section 4.2 (cf. Karush, 1939; Kuhn and Tucker, 1951) and equations B.96c furthermore state that

$$\xi_i\,,\; \xi_i^* \;\geq 0 \tag{B.101a}$$

$$\begin{aligned} \alpha_i \cdot \left(e + \xi_i \; - y_{il}^{\mathsf{s}} + \beta_1 + \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \beta_j \right) \;&= 0 \\ \alpha_i^* \cdot \left(e + \xi_i^* + y_{il}^{\mathsf{s}} - \beta_1 - \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \beta_j \right) \;&= 0 \end{aligned} \tag{B.101b}$$

$$\begin{aligned} \lambda_i \cdot \xi_i \;&=\; (c_i - \alpha_i) \cdot \xi_i \;= 0 \\ \lambda_i^* \cdot \xi_i^* \;&=\; (c_i - \alpha_i^*) \cdot \xi_i^* \;= 0 \qquad . \end{aligned} \tag{B.101c}$$

Transformation of conditions B.101b yields

$$\begin{aligned} \beta_1 \;&= y_{il}^{\mathsf{s}} + e + \xi_i^* - \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \beta_j \quad \text{for all } i \in \{k : 0 < \alpha_k^*\} \\ \beta_1 \;&= y_{kl}^{\mathsf{s}} - e - \xi_k - \sum_{j=2}^{h} \widetilde{x}_{kj}^{\mathsf{s}} \cdot \beta_j \quad \text{for all } i \in \{k : 0 < \alpha_k\} \qquad . \end{aligned} \tag{B.102}$$

It is evident that β_1 can be computed directly from equations B.102 if any $\xi_i = 0$ or $\xi_i^* = 0$. As follows from equations B.101c, this is the case if any $0 < \alpha_i < c_i$ or $0 < \alpha_i^* < c_i$ holds. For numerical stability, the average of all values fulfilling the above conditions is used:

$$\begin{aligned} \beta_1 \;= \frac{1}{2} \cdot \Bigg(& \mathrm{E}\left(y_{il}^{\mathsf{s}} + e - \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \beta_j \middle| 0 < \alpha_i^* < c_i \right) \\ & + \mathrm{E}\left(y_{il}^{\mathsf{s}} - e - \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \beta_j \middle| 0 < \alpha_i \; < c_i \right) \Bigg) \qquad . \end{aligned} \tag{B.103}$$

If the conditions are not met by any α_i or α_i^*, so $\{k : 0 < \alpha_k < c_k\} = \{k : 0 < \alpha_k^* < c_k\} = \emptyset$, the mid-point of the interval determined by inequalities B.101a is chosen as an approximation:

$$\begin{aligned} \beta_1 \;\approx \frac{1}{2} \cdot \Bigg(& \mathrm{Max}\left(y_{il}^{\mathsf{s}} + e - \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \beta_j \middle| \alpha_i^* = c_i \right) \\ & + \mathrm{Min}\left(y_{il}^{\mathsf{s}} - e - \sum_{j=2}^{h} \widetilde{x}_{ij}^{\mathsf{s}} \cdot \beta_j \middle| \alpha_i \; = c_i \right) \Bigg) \end{aligned} \tag{B.104}$$

(cf. Chang and Lin, 2011, p. 10; Smola and Schölkopf, 2004, p. 201).

B.4.6 Derivation of Shrinkage Methods

B.4.6.1 Unconstrained Shrinkage Methods

As discussed in section 5.1.11, shrinkage methods induce an (additional) constraint to the optimization problem, which has the form

$$\boldsymbol{p}\left(\boldsymbol{\Theta}\right) \leq b \tag{B.105}$$

for some prespecified constant $b \geq 0$. In case of unconstrained penalization, where

$$\boldsymbol{\Theta}^{*} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left(\delta\left(\boldsymbol{\Theta}\right) + \lambda \cdot \boldsymbol{p}\left(\boldsymbol{\Theta}\right)\right) \quad , \tag{B.106}$$

it holds that b is a monotonically decreasing function of λ.

Consider the opposite, such that $\lambda_2 > \lambda_1$ and $\boldsymbol{p}\left(\boldsymbol{\Theta}_2^{*}\right) > \boldsymbol{p}\left(\boldsymbol{\Theta}_1^{*}\right)$. It then follows that

$$\begin{aligned} & \left(\lambda_2 - \lambda_1\right)\left(\boldsymbol{p}\left(\boldsymbol{\Theta}_2^{*}\right) - \boldsymbol{p}\left(\boldsymbol{\Theta}_1^{*}\right)\right) > 0 \\ \Rightarrow \quad & \lambda_1 \boldsymbol{p}\left(\boldsymbol{\Theta}_1^{*}\right) + \lambda_2 \boldsymbol{p}\left(\boldsymbol{\Theta}_2^{*}\right) > \lambda_2 \boldsymbol{p}\left(\boldsymbol{\Theta}_1^{*}\right) + \lambda_1 \boldsymbol{p}\left(\boldsymbol{\Theta}_2^{*}\right) \\ \Rightarrow \quad & \delta\left(\boldsymbol{\Theta}_1^{*}\right) + \lambda_1 \boldsymbol{p}\left(\boldsymbol{\Theta}_1^{*}\right) + \delta\left(\boldsymbol{\Theta}_2^{*}\right) + \lambda_2 \boldsymbol{p}\left(\boldsymbol{\Theta}_2^{*}\right) > \delta\left(\boldsymbol{\Theta}_1^{*}\right) + \lambda_2 \boldsymbol{p}\left(\boldsymbol{\Theta}_1^{*}\right) + \delta\left(\boldsymbol{\Theta}_2^{*}\right) + \lambda_1 \boldsymbol{p}\left(\boldsymbol{\Theta}_2^{*}\right) , \end{aligned} \tag{B.107}$$

which contradicts optimality B.106 and hence cannot be true (cf. Hastie, Tibshirani and Friedman, 2008, pp. 61 ff; Hoerl and Kennard, 1970; Tibshirani, 1994; 1996).

B.4.6.2 Shrinkage Parameter Selection

As described in section 5.1.11, leave-one-out cross-validation results from setting $b = 1$ in algorithm 14. Consequently, a model $\mathbf{m}_{\mathrm{i}}\left(\boldsymbol{x}_{i\cdot}^{\mathsf{s}}, \boldsymbol{\Theta}^{(\mathrm{i})}\left(\lambda\right)\right)$ excluding only a single observation i is fit in step 6 of algorithm 14 for all $i = 1, \ldots, n^{\mathsf{s}}$ and each candidate value of λ. In analogy to equality 5.127, let $\boldsymbol{B} \in \mathbb{R}^{n^{\mathsf{s}} \times n^{\mathsf{s}}}$ be the smoother matrix resulting from this model that predicts all observations in $\mathcal{S}^{\mathsf{s}}$, such that

$$\mathbf{m}_{\mathrm{i}}\left(\boldsymbol{x}_{i\cdot}^{\mathsf{s}}, \boldsymbol{\Theta}^{(\mathrm{i})}\left(\lambda_k\right)\right) = \boldsymbol{b}_{i\cdot}\boldsymbol{Y}^{\mathsf{s}} \tag{B.108}$$

is the prediction for element i. When predictions are linear in $\boldsymbol{Y}$ (cf. equality 5.127 and the related discussion), equality B.108 may equivalently be written as

$$\mathbf{m}_{\mathrm{i}}\left(\boldsymbol{x}_{i\cdot}^{\mathsf{s}}, \boldsymbol{\Theta}^{(\mathrm{i})}\left(\lambda_k\right)\right) = \boldsymbol{a}_{i\cdot}\widetilde{\boldsymbol{Y}}^{\mathsf{s}} \tag{B.109a}$$

for the adjusted dependent variable

$$\begin{aligned} \widetilde{\boldsymbol{Y}}^{\mathsf{s}} &\coloneqq \boldsymbol{Y}^{\mathsf{s}} + \boldsymbol{V}\left(\mathbf{m}_{\mathrm{i}}\left(\boldsymbol{X}^{\mathsf{s}}, \boldsymbol{\Theta}^{(\mathrm{i})}\left(\lambda_k\right)\right) - \boldsymbol{Y}^{\mathsf{s}}\right) \\ &= \left[\left(\boldsymbol{y}_{1\cdot}^{\mathsf{s}}\right)^{\mathsf{T}} \ \ldots \ \left(\boldsymbol{y}_{i-1\cdot}^{\mathsf{s}}\right)^{\mathsf{T}} \ \left(\mathbf{m}_{\mathrm{i}}\left(\boldsymbol{x}_{i\cdot}^{\mathsf{s}}, \boldsymbol{\Theta}^{(\mathrm{i})}\left(\lambda_k\right)\right)\right)^{\mathsf{T}} \ \left(\boldsymbol{y}_{i+1\cdot}^{\mathsf{s}}\right)^{\mathsf{T}} \ \ldots \ \left(\boldsymbol{y}_{n^{\mathsf{s}}\cdot}^{\mathsf{s}}\right)^{\mathsf{T}}\right]^{\mathsf{T}} , \end{aligned} \tag{B.109b}$$

using an indicator matrix $\boldsymbol{V} \in \{0;1\}^{n^{\mathsf{s}} \times n^{\mathsf{s}}}$ with elements defined by

$$v_{jk} \coloneqq \mathbb{I}\left(j = k = i\right) \quad . \tag{B.109c}$$

Together, equalities 5.127, B.108 and B.109 result in

$$\begin{aligned} \boldsymbol{b}_{i\cdot}\boldsymbol{Y}^{\mathsf{s}} &= \boldsymbol{a}_{i\cdot}\widetilde{\boldsymbol{Y}}^{\mathsf{s}} = \boldsymbol{a}_{i\cdot}\boldsymbol{Y}^{\mathsf{s}} + \boldsymbol{a}_{i\cdot}\boldsymbol{V}\left(\mathbf{m}_{\mathrm{i}}\left(\boldsymbol{X}^{\mathsf{s}}, \boldsymbol{\Theta}^{(\mathrm{i})}\left(\lambda_k\right)\right) - \boldsymbol{Y}^{\mathsf{s}}\right) \\ &= \widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}} + a_{ii} \cdot \boldsymbol{b}_{i\cdot}\boldsymbol{Y}^{\mathsf{s}} - a_{ii} \cdot \boldsymbol{y}_{i\cdot}^{\mathsf{s}} \\ &= \left(1 - a_{ii}\right)^{-1} \cdot \left(\widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}} - a_{ii} \cdot \boldsymbol{y}_{i\cdot}^{\mathsf{s}}\right) \quad . \end{aligned} \tag{B.110}$$

Denoting the matrix containing only the diagonal elements of $\boldsymbol{A}$ by $\boldsymbol{C} \coloneqq \mathbf{diag}\left(\mathbf{diag}\left(\boldsymbol{A}\right)\right)$,

$$\boldsymbol{B} \coloneqq \left(\boldsymbol{A} - \boldsymbol{C}\right) \oslash \left(\left(\mathbf{I}_{n^{\mathsf{s}}} - \boldsymbol{C}\right) \mathbf{1}_{n^{\mathsf{s}} \times n^{\mathsf{s}}}\right) \tag{B.111}$$

is the matrix that can be used to calculate $\mathbf{m}_{\mathrm{i}}\left(\boldsymbol{x}_{i\cdot}^{\mathsf{s}}, \boldsymbol{\Theta}^{(\mathrm{i})}\left(\lambda_k\right)\right)$ for all elements $i = 1, \ldots, n^{\mathsf{s}}$ at once using equality B.108 (cf. Craven and Wahba, 1979; Hastie and Tibshirani, 1990, pp. 46 ff; Golub, Heath and Wahba, 1979; Wood, 2017, pp. 169 ff).

From equalities B.110, it furthermore follows directly that

$$\begin{aligned} \boldsymbol{b}_{i\cdot}\boldsymbol{Y}^{\mathsf{s}} - \boldsymbol{y}_{i\cdot}^{\mathsf{s}} &= \left(1 - a_{ii}\right)^{-1} \cdot \left(\widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}} - a_{ii} \cdot \boldsymbol{y}_{i\cdot}^{\mathsf{s}} - \left(1 - a_{ii}\right) \cdot \boldsymbol{y}_{i\cdot}^{\mathsf{s}}\right) \\ &= \left(1 - a_{ii}\right)^{-1} \cdot \left(\widehat{\boldsymbol{y}}_{i\cdot}^{\mathsf{s}} - \boldsymbol{y}_{i\cdot}^{\mathsf{s}}\right) \quad . \end{aligned} \tag{B.112}$$

For generalized cross-validation, a_{ii} in the denominator of the right-hand side of equality B.112 is replaced by its expectation $\mathrm{E}\left(a_{ii}\right) = \mathrm{tr}\left(\boldsymbol{A}\right)/n^{\mathsf{s}}$ to simplify computation for the residual sum of squares (cf. Craven and Wahba, 1979; Hastie and Tibshirani, 1990, pp. 46 ff; Golub, Heath and Wahba, 1979; Wood, 2017, pp. 169 ff).

B.5 Mathematical Background of Calibrated Artificial Neural Networks

B.5.1 Gradient Information for Optimization

B.5.1.1 Derivatives of the Distance Function

The distance function for estimating calibrated ANNs is defined in equation 5.160 as

$$\delta\left(\boldsymbol{\Theta}\right) \coloneqq \mathrm{v}_1 \cdot \delta_{\mathrm{m}}\left(\boldsymbol{\omega}\right) + \frac{1}{2} \cdot \mathbf{v}_{\mathcal{J}}^{\mathsf{T}} \left(\boldsymbol{\Theta} - \mathbf{C}_{\boldsymbol{\Theta}}\right)^{\circ 2} \quad , \tag{B.113}$$

where $\delta_{\mathrm{m}}\left(\boldsymbol{\omega}\right)$ is the distance function of the artificial neural network defined in equation 5.82, and $\mathcal{J} = \{2, \ldots, u+1\}$ is used to subset all elements of $\mathbf{v}$ but the first. As the derivative of a general weighted squared function is given by

$$\begin{aligned} \frac{\partial}{\partial\left(s\right)}\left(h \cdot \frac{\left(s-u\right)^2}{2}\right) &= \frac{h}{2} \cdot \frac{\partial\left(\left(s-u\right)^2\right)}{\partial\left(s-u\right)} \cdot \frac{\partial\left(s-u\right)}{\partial\left(s\right)} \\ &= h \cdot \left(s-u\right) \quad , \end{aligned} \tag{B.114}$$

the derivative of the quadratic parts of the distance function are given by

$$\frac{\partial}{\partial\left(\Theta_j\right)}\left(\mathrm{v}_{(j+1)} \cdot \frac{\left(\Theta_j - \mathrm{C}_{\Theta_j}\right)^2}{2}\right) = \mathrm{v}_{(j+1)} \cdot \left(\Theta_j - \mathrm{C}_{\Theta_j}\right) \quad . \tag{B.115}$$

Hence, the resulting Jacobian for the distance function $\delta\left(\boldsymbol{\Theta}\right)$ is

$$\mathbf{J}_{\delta}\left(\boldsymbol{\Theta}\right) = \mathrm{v}_1 \cdot \begin{bmatrix} \mathbf{J}_{\delta_{\mathrm{m}}}\left(\boldsymbol{\omega}\right) & \mathbf{0}_{1\times p} & \mathbf{0}_{1\times r} \end{bmatrix} + \left(\mathbf{v}_{\mathcal{J}} \circ \left(\boldsymbol{\Theta} - \mathbf{C}_{\boldsymbol{\Theta}}\right)\right)^{\mathsf{T}} \quad , \tag{B.116}$$

where $\mathbf{J}_{\delta_{\mathrm{m}}}\left(\boldsymbol{\omega}\right)$ is the Jacobi matrix of the artificial neural network distance function defined in equation 5.83.

B.5.1.2 Derivatives of the Constraints

Recall that for the weighting model, the estimated participation propensities are given by the predictions of an artificial neural network $\widehat{\boldsymbol{p}}^{\text{nps}}(\boldsymbol{\omega})$ for target variables $\boldsymbol{r}^{\text{nps}}$ representing the binary inclusion indicator for the non-probability sample (cf. equations 2.2 and 5.152). The correction weights are determined by the inverse of these propensities, such that

$$\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}(\boldsymbol{\omega}) = \boldsymbol{w}^{\text{nps}} \oslash \widehat{\boldsymbol{p}}^{\text{nps}}(\boldsymbol{\omega}) \tag{B.117}$$

are the calibration weights with derivatives defined by

$$\frac{\partial(\widetilde{\boldsymbol{w}})}{\partial(\widehat{\boldsymbol{p}}^{\text{nps}})} = -\operatorname{\mathbf{diag}}\left(\boldsymbol{w}^{\text{nps}} \oslash (\widehat{\boldsymbol{p}}^{\text{nps}}(\boldsymbol{\omega}))^{\circ 2}\right) \quad . \tag{B.118}$$

Therefore, the Jacobian matrix of $\widetilde{\boldsymbol{w}}(\boldsymbol{\omega})$ is obtained as

$$\begin{aligned} \mathbf{J}_{\widetilde{w}}(\boldsymbol{\omega}) &= -\operatorname{\mathbf{diag}}\left(\boldsymbol{w}^{\text{nps}} \oslash (\widehat{\boldsymbol{p}}^{\text{nps}}(\boldsymbol{\omega}))^{\circ 2}\right) \mathbf{J}_{\widetilde{w}}(\widehat{\boldsymbol{p}}^{\text{nps}}(\boldsymbol{\omega})) \\ &= -\left(\mathbf{1}_{1\times\dim(\boldsymbol{\omega})} \otimes \left(\boldsymbol{w}^{\text{nps}} \circ (\widehat{\boldsymbol{p}}^{\text{nps}}(\boldsymbol{\omega}))^{\circ(-2)}\right)\right) \circ \mathbf{J}_{\widehat{\boldsymbol{p}}^{\text{nps}}}(\boldsymbol{\omega}) \quad , \end{aligned} \tag{B.119}$$

where $\dim(\boldsymbol{\omega})$ is again the number of weighting parameters, and $\mathbf{J}_{\widehat{\boldsymbol{p}}^{\text{nps}}}(\boldsymbol{\omega})$ is the Jacobian matrix of ANN predictions defined in equation 5.84. To achieve a compact and coherent presentation, multiplication of a vector with the inverse of a scalar is denoted as a fraction in the following derivations, i.e. $\frac{\boldsymbol{a}}{h} \coloneqq \boldsymbol{a} \cdot h^{-1}$. Furthermore, a variable which is centered around its mean is again denoted by $\mathbf{e}(\boldsymbol{x}_{\cdot k})$ (cf. equations 3.6).

The following equations provide the components for calculating the Jacobi matrix of the constraint functions used in problem 5.158, as defined in 5.164 and 5.163. The derivatives for the sums of the weights and squared weights respectively are

$$\frac{\partial\left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)}{\partial(\widetilde{\boldsymbol{w}})} = \mathbf{1}_{n^{\text{nps}}\times 1} \tag{B.120a}$$

and

$$\frac{\partial\left(\widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}^{\circ 2}\right)\right)}{\partial(\widetilde{\boldsymbol{w}})} = 2 \cdot \widetilde{\boldsymbol{w}} \quad . \tag{B.120b}$$

Based on equalities B.120, differentiation of the bias correction factor for covariances yields

$$\begin{aligned} &\frac{\partial(\nu(\widetilde{\boldsymbol{w}}))}{\partial(\widetilde{\boldsymbol{w}})} \\ &= \left(\frac{\partial\left(\widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}^{\circ 2}\right)\right)}{\partial(\widetilde{\boldsymbol{w}})} \cdot \left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)^2 - \frac{\partial\left(\left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)^2\right)}{\partial(\widetilde{\boldsymbol{w}})} \cdot \widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}^{\circ 2}\right)\right) \cdot \left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)^{-4} \\ &= \left(2 \cdot \widetilde{\boldsymbol{w}} \cdot \left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)^2 - 2 \cdot \mathbf{1}_{n^{\text{nps}}\times 1} \cdot \widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}}) \cdot \widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}^{\circ 2}\right)\right) \cdot \left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)^{-4} \\ &= 2 \cdot \left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)^{-3} \cdot \left(\widetilde{\boldsymbol{w}} \cdot \widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}}) - \mathbf{1}_{n^{\text{nps}}\times 1} \cdot \widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}^{\circ 2}\right)\right) \quad . \end{aligned} \tag{B.121}$$

The gradients for totals and means can be calculated as

$$\frac{\partial(\widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}(\widetilde{\boldsymbol{w}}))}{\partial(\widetilde{\boldsymbol{w}})} = \boldsymbol{x}_{\cdot k}^{\mathsf{nps}} \tag{B.122}$$

and

$$\begin{aligned}\frac{\partial\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}(\widetilde{\boldsymbol{w}})\right)}{\partial(\widetilde{\boldsymbol{w}})} &= \left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}}) \cdot \frac{\partial(\widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}(\widetilde{\boldsymbol{w}}))}{\partial(\widetilde{\boldsymbol{w}})} - \widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}(\widetilde{\boldsymbol{w}}) \cdot \frac{\partial\left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)}{\partial(\widetilde{\boldsymbol{w}})}\right) \cdot \widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})^{-2} \\ &= \left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}}) \cdot \boldsymbol{x}_{\cdot k}^{\mathsf{nps}} - \widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}(\widetilde{\boldsymbol{w}})\right) \cdot \left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)^{-2} \\ &= \left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)^{-1} \cdot \mathbf{e}(\boldsymbol{x}_{\cdot k}) \quad .\end{aligned} \tag{B.123}$$

Using equality B.123, the derivative of the ML covariance estimate for two variables $\boldsymbol{x}_{\cdot k}$ and $\boldsymbol{x}_{\cdot l}$ is

$$\begin{aligned}&\frac{\partial\left(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}(\widetilde{\boldsymbol{w}})\right]_{kl}\right)}{\partial(\widetilde{\boldsymbol{w}})} \\ &= \frac{\partial(\widehat{\boldsymbol{\mu}}(\boldsymbol{x}_{\cdot k}^{\mathsf{nps}} \circ \boldsymbol{x}_{\cdot l}^{\mathsf{nps}}, \widetilde{\boldsymbol{w}}))}{\partial(\widetilde{\boldsymbol{w}})} \\ &\quad - \frac{\partial\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}(\widetilde{\boldsymbol{w}})\right)}{\partial(\widetilde{\boldsymbol{w}})} \cdot \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}_{\cdot\mathrm{l}}}(\widetilde{\boldsymbol{w}}) - \frac{\partial\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{x}_{\cdot\mathrm{l}}}(\widetilde{\boldsymbol{w}})\right)}{\partial(\widetilde{\boldsymbol{w}})} \cdot \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}(\widetilde{\boldsymbol{w}}) \\ &= \left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)^{-1} \cdot \left(\mathbf{e}(\boldsymbol{x}_{\cdot k} \circ \boldsymbol{x}_{\cdot l}) - \boldsymbol{x}_{\cdot k}^{\mathsf{nps}} \cdot \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}_{\cdot\mathrm{l}}}(\widetilde{\boldsymbol{w}}) - \boldsymbol{x}_{\cdot l}^{\mathsf{nps}} \cdot \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}(\widetilde{\boldsymbol{w}})\right) \quad .\end{aligned} \tag{B.124}$$

Equalities B.124 and B.121 can be used for differentiation of the corresponding unbiased estimate for the covariance of $\boldsymbol{x}_{\cdot k}$ and $\boldsymbol{x}_{\cdot l}$, which yields

$$\begin{aligned}\frac{\partial\left(\left[\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}(\widetilde{\boldsymbol{w}})\right]_{kl}\right)}{\partial(\widetilde{\boldsymbol{w}})} &= \left(\frac{\partial\left(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}(\widetilde{\boldsymbol{w}})\right]_{kl}\right)}{\partial(\widetilde{w}_i)} \cdot (1 - \nu(\widetilde{\boldsymbol{w}})) \right. \\ &\quad \left. - \frac{\partial(1 - \nu(\widetilde{\boldsymbol{w}}))}{\partial(\widetilde{\boldsymbol{w}})} \cdot \left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}(\widetilde{\boldsymbol{w}})\right]_{kl}\right) \cdot (1 - \nu(\widetilde{\boldsymbol{w}}))^{-2} \\ &= (1 - \nu(\widetilde{\boldsymbol{w}}))^{-1} \cdot \left(\frac{\partial\left(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}(\widetilde{\boldsymbol{w}})\right]_{kl}\right)}{\partial(\widetilde{\boldsymbol{w}})} \right. \\ &\quad \left. + 2 \cdot \frac{\left[\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}(\widetilde{\boldsymbol{w}})\right]_{kl}}{\left(\widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}})\right)^3} \cdot \left(\widetilde{\boldsymbol{w}} \cdot \widehat{\mathrm{N}}(\widetilde{\boldsymbol{w}}) - \mathbf{1}_{n^{\mathsf{nps}} \times 1} \cdot \widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}^{\circ 2}\right)\right)\right) \quad .\end{aligned} \tag{B.125}$$

The derivative of the product of the standard deviations of $\boldsymbol{x}_{\cdot k}$ and $\boldsymbol{x}_{\cdot l}$ is

$$\begin{aligned}
&\frac{\partial\Big(\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}}\cdot\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} \\
&\quad= \frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\cdot\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}^{0.5}\Big)}{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}\Big)}\cdot\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}} \\
&\qquad+\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\cdot\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}^{0.5}\Big)}{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}\Big)}\cdot\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}} \\
&\quad= \frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\cdot\frac{\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}}}{2\cdot\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}}}+\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\cdot\frac{\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}}}{2\cdot\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}}}\quad.
\end{aligned}\tag{B.126}$$

Using equalities B.125 and B.126, differentiation of a correlation with respect to the new weights results in

$$\begin{aligned}
\frac{\partial([\widehat{\boldsymbol{\rho}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)]_{kl})}{\partial\left(\widetilde{\boldsymbol{w}}\right)} &= \frac{\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}}\cdot\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}}}{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}\cdot\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}}\cdot\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} \\
&\quad-\frac{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl}}{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}\cdot\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}}\cdot\frac{\partial\Big(\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}}\cdot\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)} \\
&= \left(\left(\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}}\cdot\sqrt{\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}}\right)^{-1}\cdot\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\right) \\
&\quad-\frac{[\widehat{\boldsymbol{\rho}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)]_{kl}}{2}\cdot\Bigg(\quad\left(\left(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}\right)^{-1}\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kk}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\right) \\
&\qquad\qquad+\quad\left(\left(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}\right)^{-1}\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{ll}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\right)\Bigg)\quad.
\end{aligned}\tag{B.127}$$

The derivatives of weighted totals, covariances and correlations are composed of these components. From equality B.122, it follows that

$$\frac{\partial(\widehat{\boldsymbol{\tau}}\left(\boldsymbol{X}^{\mathsf{nps}},\widetilde{\boldsymbol{w}}\right))}{\partial\left(\widetilde{\boldsymbol{w}}\right)} = \left[\frac{\partial(\widehat{\boldsymbol{\tau}}\left(\boldsymbol{x}_{\cdot 1}^{\mathsf{nps}},\widetilde{\boldsymbol{w}}\right))}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\quad\cdots\quad\frac{\partial\Big(\widehat{\boldsymbol{\tau}}\left(\boldsymbol{x}_{\cdot p}^{\mathsf{nps}},\widetilde{\boldsymbol{w}}\right)\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\right]=\boldsymbol{X}^{\mathsf{nps}}\quad,\tag{B.128}$$

and equation B.124 determines the elements of

$$\frac{\partial\Big(\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right)\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}=\left[\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{11}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\quad\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{12}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\quad\cdots\quad\frac{\partial\Big(\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{pp}\Big)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\right].\tag{B.129}$$

The resulting Jacobian for the equality constraint function

$$\bar{\mathbf{g}}\left(\boldsymbol{\Theta}\right) \coloneqq \begin{bmatrix} \left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right)^{\mathsf{T}} \\ \left(\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right)\right)^{\mathsf{T}} \end{bmatrix} - \begin{bmatrix} \left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right)^{\mathsf{T}} \\ \left(\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right)\right)^{\mathsf{T}} \end{bmatrix} \circ \begin{bmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\varepsilon} \end{bmatrix} \tag{B.130}$$

defined in equation 5.164 is thus

$$\mathbf{J}_{\bar{\mathbf{g}}}\left(\boldsymbol{\Theta}\right) = \begin{bmatrix} \left(\boldsymbol{X}^{\mathsf{nps}}\right)^{\mathsf{T}} \mathbf{J}_{\widetilde{\boldsymbol{w}}}\left(\boldsymbol{\omega}\right) & -\mathbf{diag}\left(\widehat{\boldsymbol{\tau}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right) & \mathbf{0}_{p\times r} \\ \left(\frac{\partial\left(\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right)\right)}{\partial\left(\widetilde{\boldsymbol{w}}\right)}\right)^{\mathsf{T}} \mathbf{J}_{\widetilde{\boldsymbol{w}}}\left(\boldsymbol{\omega}\right) & \mathbf{0}_{r\times p} & -\mathbf{diag}\left(\mathbf{vec}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right)\right) \end{bmatrix}. \tag{B.131}$$

Equivalent expressions for calibrating unbiased covariances or correlations follow from respectively replacing $\widetilde{\boldsymbol{\Sigma}}$ by $\widehat{\boldsymbol{\Sigma}}$ or $\widehat{\boldsymbol{\rho}}$ in equations B.129 and B.131. The required derivatives are then determined by B.125 and B.127, respectively.

Because the inequality constraints are defined as

$$\widetilde{\mathbf{g}}\left(\boldsymbol{\Theta}\right) \coloneqq \begin{bmatrix} \boldsymbol{L}_{\boldsymbol{\Theta}} \\ \boldsymbol{\Theta} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Theta} \\ \boldsymbol{U}_{\boldsymbol{\Theta}} \end{bmatrix} \tag{B.132}$$

in equation 5.163, the corresponding Jacobian matrix is simply

$$\mathbf{J}_{\widetilde{\mathbf{g}}}\left(\boldsymbol{\Theta}\right) = \begin{bmatrix} -\mathbf{I}_u \\ \mathbf{I}_u \end{bmatrix} \quad . \tag{B.133}$$

B.5.2 The Link Between Covariance Calibration and Post-stratification

The maximum likelihood covariance estimator for two variables $\boldsymbol{x}_{\cdot k}$ and $\boldsymbol{x}_{\cdot l}$ using weights $\widetilde{\boldsymbol{w}}$ for the non-probability sample nps defined in equation 2.18c can be written as

$$\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl} = \frac{\widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{k}}\circ\boldsymbol{x}_{\cdot\mathrm{l}}}\left(\widetilde{\boldsymbol{w}}\right)}{\widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}\right)} - \frac{\widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}\left(\widetilde{\boldsymbol{w}}\right)}{\widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}\right)} \cdot \frac{\widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{l}}}\left(\widetilde{\boldsymbol{w}}\right)}{\widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}\right)} \quad . \tag{B.134}$$

Assuming that matrix $\boldsymbol{X}$ contains an intercept $\boldsymbol{x}_{\cdot 1} \coloneqq \mathbf{1}_{\mathrm{N}\times 1}$ as well as an interaction term $\boldsymbol{x}_{\cdot k} \circ \boldsymbol{x}_{\cdot l}$ in addition to the main effects $\boldsymbol{x}_{\cdot k}$ and $\boldsymbol{x}_{\cdot l}$, constraints 5.142 imply that

$$\begin{aligned} \widehat{\mathrm{N}}\left(\widetilde{\boldsymbol{w}}\right) &\overset{!}{=} \widehat{\mathrm{N}}\left(\boldsymbol{w}^{\mathsf{cal}}\right) \\ \widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}\left(\widetilde{\boldsymbol{w}}\right) &\overset{!}{=} \widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{k}}}\left(\boldsymbol{w}^{\mathsf{cal}}\right) \\ \widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{l}}}\left(\widetilde{\boldsymbol{w}}\right) &\overset{!}{=} \widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{l}}}\left(\boldsymbol{w}^{\mathsf{cal}}\right) \\ \widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{k}}\circ\boldsymbol{x}_{\cdot\mathrm{l}}}\left(\widetilde{\boldsymbol{w}}\right) &\overset{!}{=} \widehat{\boldsymbol{\tau}}_{\boldsymbol{x}_{\cdot\mathrm{k}}\circ\boldsymbol{x}_{\cdot\mathrm{l}}}\left(\boldsymbol{w}^{\mathsf{cal}}\right) \quad . \end{aligned} \tag{B.135}$$

From constraints B.135, it follows directly that the ML covariance estimator defined in equation B.134 is calibrated as well because all components required for its calculation coincide:

$$\left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\widetilde{\boldsymbol{w}}\right)\right]_{kl} = \left[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\left(\boldsymbol{w}^{\mathsf{cal}}\right)\right]_{kl} \quad . \tag{B.136}$$

Therefore, covariance-based measures of association can be adjusted by using terms that represent main and interaction effects. This, for example, is the case when it comes to post-stratification or raking to known cell counts in cross-tables (cf. Deville and Särndal, 1992, p. 380; Deville, Särndal and Sautory, 1993, p. 1015). In this manner, even correlations may be aligned by calibrating all component (co-)variances (cf. equation 2.18d). Due to equality 2.18f, however, this is not the case for unbiased covariance estimators.

B.6 Rationale of MSE-intervals

A vector of arbitrary weights $\widetilde{\boldsymbol{w}} \in \mathbb{R}^{n^{\mathsf{nps}}}$ is used for the non-probability sample. The coefficient of variation for these weights is

$$\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right) \coloneqq \sqrt{\mathrm{V}\left(\widetilde{\boldsymbol{w}}\right) \Big/ \left(\mathrm{E}\left(\widetilde{\boldsymbol{w}}\right)\right)^2} \quad . \tag{B.137a}$$

The weighted version $\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \in \mathbb{R}^{\mathrm{N}}$ of the sample inclusion indicator $\boldsymbol{r}^{\mathsf{nps}}$ is defined by

$$\widetilde{r}_i^{\,\mathsf{nps}} \coloneqq \begin{cases} \widetilde{w}_i & \text{if} \quad i \in \mathcal{S}^{\mathsf{nps}} \\ 0 & \text{else} \end{cases} \quad , \tag{B.137b}$$

and the sampling fraction is

$$fr^{\mathsf{nps}} \coloneqq \mathrm{E}\left(\boldsymbol{r}^{\mathsf{nps}}\right) = n^{\mathsf{nps}}/\mathrm{N} \quad . \tag{B.137c}$$

Based on definitions B.137, it holds that

$$\begin{aligned} \mathrm{E}\left(\mathrm{V}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \,|\, \boldsymbol{r}^{\mathsf{nps}}\right)\right) &= fr^{\mathsf{nps}} \cdot \mathrm{V}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \,|\, r_i^{\mathsf{nps}} = 1\right) \\ &= fr^{\mathsf{nps}} \cdot \mathrm{V}\left(\widetilde{\boldsymbol{w}}\right) \end{aligned} \tag{B.138a}$$

and

$$\begin{aligned} \mathrm{V}\left(\mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \,|\, \boldsymbol{r}^{\mathsf{nps}}\right)\right) &= fr^{\mathsf{nps}} \cdot \Big(\mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \,|\, r_i^{\mathsf{nps}} = 1\right) \\ &\qquad - fr^{\mathsf{nps}} \cdot \mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \,|\, r_i^{\mathsf{nps}} = 1\right) \Big)^2 \\ &\quad + \left(1 - fr^{\mathsf{nps}}\right) \cdot \left(fr^{\mathsf{nps}} \cdot \mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \,|\, r_i^{\mathsf{nps}} = 1\right)\right)^2 \\ &= fr^{\mathsf{nps}} \cdot \left(1 - fr^{\mathsf{nps}}\right) \cdot \left(\mathrm{E}\left(\widetilde{\boldsymbol{w}}\right)\right)^2 \cdot \left(1 - fr^{\mathsf{nps}}\right) \\ &\quad + fr^{\mathsf{nps}} \cdot \left(1 - fr^{\mathsf{nps}}\right) \cdot \left(\mathrm{E}\left(\widetilde{\boldsymbol{w}}\right)\right)^2 \cdot fr^{\mathsf{nps}} \\ &= fr^{\mathsf{nps}} \cdot \left(1 - fr^{\mathsf{nps}}\right) \cdot \left(\mathrm{E}\left(\widetilde{\boldsymbol{w}}\right)\right)^2 \quad . \end{aligned} \tag{B.138b}$$

It then follows that

$$\begin{aligned} \mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right) &= \mathrm{E}\left(\mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \,|\, \boldsymbol{r}^{\mathsf{nps}}\right)\right) \\ &= fr^{\mathsf{nps}} \cdot \mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \,|\, r_i^{\mathsf{nps}} = 1\right) \\ &= fr^{\mathsf{nps}} \cdot \mathrm{E}\left(\widetilde{\boldsymbol{w}}\right) \quad , \end{aligned} \tag{B.138c}$$

and, by the law of total variance (cf. e.g. Blitzstein and Hwang, 2013, p. 401),

$$\begin{aligned} \mathrm{V}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right) &= \mathrm{E}\left(\mathrm{V}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \,|\, \boldsymbol{r}^{\mathsf{nps}}\right)\right) + \mathrm{V}\left(\mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}} \,|\, \boldsymbol{r}^{\mathsf{nps}}\right)\right) \\ &= fr^{\mathsf{nps}} \cdot \left(\mathrm{V}\left(\widetilde{\boldsymbol{w}}\right) + \left(1 - fr^{\mathsf{nps}}\right) \cdot \left(\mathrm{E}\left(\widetilde{\boldsymbol{w}}\right)\right)^2\right) \quad . \end{aligned} \tag{B.138d}$$

Based on equalities B.137 and B.138, $\mathrm{CV}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right)$ can hence be written as

$$\begin{aligned} \mathrm{CV}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right) &= \sqrt{\mathrm{V}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right)} \Big/ \mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right) \\ &= \sqrt{fr^{\mathsf{nps}} \cdot \left(\mathrm{V}\left(\widetilde{\boldsymbol{w}}\right) + \left(1 - fr^{\mathsf{nps}}\right) \cdot \left(\mathrm{E}\left(\widetilde{\boldsymbol{w}}\right)\right)^2\right)} \Big/ \left(fr^{\mathsf{nps}} \cdot \mathrm{E}\left(\widetilde{\boldsymbol{w}}\right)\right) \\ &= \sqrt{\left(1 - fr^{\mathsf{nps}} + \mathrm{V}\left(\widetilde{\boldsymbol{w}}\right) \Big/ \left(\mathrm{E}\left(\widetilde{\boldsymbol{w}}\right)\right)^2\right) \Big/ fr^{\mathsf{nps}}} \\ &= \sqrt{\left(1 - fr^{\mathsf{nps}} + \left(\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)\right)^2\right) \Big/ fr^{\mathsf{nps}}} \quad . \end{aligned} \tag{B.139}$$

Furthermore, it holds that

$$\begin{aligned} \frac{\sqrt{\left(1 - fr^{\mathsf{nps}} + \left(\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)\right)^2\right) \Big/ fr^{\mathsf{nps}}}}{\sqrt{\left(1 - fr^{\mathsf{nps}}\right) / fr^{\mathsf{nps}}}} &= \frac{\sqrt{1 - fr^{\mathsf{nps}} + \left(\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)\right)^2}}{\sqrt{1 - fr^{\mathsf{nps}}}} \\ &= \sqrt{\frac{1 - fr^{\mathsf{nps}} + \left(\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)\right)^2}{1 - fr^{\mathsf{nps}}}} \\ &= \sqrt{1 + \frac{\left(\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)\right)^2}{1 - fr^{\mathsf{nps}}}} \quad . \end{aligned} \tag{B.140}$$

For design linear estimators, deviations between estimated and true statistics $\widehat{\vartheta}_k$ and ϑ_k for all $k = 1, \ldots, h$ elements of $\widehat{\boldsymbol{\vartheta}} \in \mathbb{R}^h$ can then be written using equalities B.139 and B.140:

$$\begin{aligned} \widehat{\vartheta}_k - \vartheta_k &= \mathrm{E}\left(\boldsymbol{t}_k\left(\boldsymbol{Y}\right) \circ \widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right) \cdot \left(\mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right)\right)^{-1} - \mathrm{E}\left(\boldsymbol{t}_k\left(\boldsymbol{Y}\right)\right) \\ &= \left(\mathrm{E}\left(\boldsymbol{t}_k\left(\boldsymbol{Y}\right) \circ \widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right) - \mathrm{E}\left(\boldsymbol{t}_k\left(\boldsymbol{Y}\right)\right) \cdot \mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right)\right) \cdot \left(\mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right)\right)^{-1} \\ &= \boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}} \cdot \sqrt{\mathrm{V}\left(\boldsymbol{t}_k\left(\boldsymbol{Y}\right)\right)} \cdot \sqrt{\mathrm{V}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right)} \Big/ \mathrm{E}\left(\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}\right) \\ &= \boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}} \cdot \sqrt{\mathrm{V}\left(\boldsymbol{t}_k\left(\boldsymbol{Y}\right)\right)} \cdot \sqrt{\left(1 - fr^{\mathsf{nps}} + \left(\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)\right)^2\right) \Big/ fr^{\mathsf{nps}}} \\ &= \boldsymbol{\rho}_{\boldsymbol{t}_k(\boldsymbol{Y})\widetilde{\boldsymbol{r}}^{\,\mathsf{nps}}} \cdot \sqrt{\frac{1 - fr^{\mathsf{nps}}}{fr^{\mathsf{nps}}}} \cdot \sqrt{\mathrm{V}\left(\boldsymbol{t}_k\left(\boldsymbol{Y}\right)\right)} \cdot \sqrt{1 + \frac{\left(\mathrm{CV}\left(\widetilde{\boldsymbol{w}}\right)\right)^2}{1 - fr^{\mathsf{nps}}}} \quad . \end{aligned} \tag{B.141}$$

(cf. Meng, 2018, pp. 690, 702).

Appendix C Documentation of R-packages

As summarized in section 6.1.2, three R-packages are developed in the context of this thesis. In the current appendix C, an overview of the functionality of these packages is provided. To that end, excerpts from the R-package manuals are subsequently presented, which contain the descriptions of functions most relevant for this thesis. Mathematical and computational details of the implementations chapters 4 and 5.

Note that while an alpha version for package `sqp` is already available via the Comprehensive R Archive Network (CRAN; `https://cran.r-project.org/`), packages `ann` and `calmod` are still under development. Therefore, package documentations and functionalities may be subject to updates before stable releases are available.

C.1 Documentation for Package `sqp`

Package 'sqp'

15.04.2021

Type Package

Title (Sequential) Quadratic Programming

Version 0.5

Date 2020-03-25

Author Simon Lenau

Maintainer Simon Lenau <lenau@uni-trier.de>

Description Solving procedures for quadratic programming with optional equality and inequality constraints, which can be used for by sequential quadratic programming (SQP). Similar to Newton-Raphson methods in the unconstrained case, sequential quadratic programming solves non-linear constrained optimization problems by iteratively solving linear approximations of the optimality conditions of such a problem. The Hessian matrix in this strategy is commonly approximated by the BFGS method in its damped modification (cf. Powell, 1978; Nocedal and Wright, 1999). All methods in this package are implemented in C++ as header-only library, such that they are easily usable in other packages.

License GPL-3

Imports Matrix, Rdpack

LinkingTo Rcpp, RcppArmadillo, RcppEigen

SystemRequirements C++11, GNU Make

NeedsCompilation yes

RdMacros Rdpack

Encoding UTF-8

RoxygenNote 7.1.0

`bfgs_init` *(Initial) Hessian approximation based on finite differences*

Description

The `BFGS-update` for approximating the Hessian matrix $\mathbf{H}_{\mathbf{f}}(\boldsymbol{\Theta}^{(0)})$ of a function $\mathbf{f}$ requires an initial 'guess' for this matrix. This is usually either a scalar multiple of the identity matrix or an approximation based on finite differences. This function generates a finite difference approximation for the Hessian matrix, based on a vector of initial parameters $\boldsymbol{\Theta}^{(0)}$ and a function $\nabla_{\mathbf{f}}(\boldsymbol{\Theta}^{(0)})$ generating the gradient. The finite difference approximation is defined by

$$\widetilde{\mathbf{H}}_{ij}^{(0)}(\boldsymbol{\Theta}^{(0)}) = \frac{\nabla_{\mathbf{f}}(\boldsymbol{\Theta}^{(0)} + \epsilon) - \nabla_{\mathbf{f}}(\boldsymbol{\Theta}^{(0)})}{\epsilon_i} \ ,$$

where ϵ is a vector of the same dimension as $\boldsymbol{\Theta}^{(0)}$ that contains exactly one non-zero value, which is its i-th element $\epsilon_i > 0$.

Usage

```
bfgs_init(
parm,
gradient_function,
eps = 1e-4,
force_symmetric = TRUE
)
```

Arguments

`parm`
: **Numeric vector** of size `N`:
Initial / current parameters $\boldsymbol{\Theta}^{(0)}$ at which the finite differences approximation should take place.

`gradient_function`
: **Function**:
The gradient function $\nabla_{\mathbf{f}}$, which must accept argument `parm` as input, and must return a scalar or vector.

`eps`
: **Numeric value**:
The difference ($\epsilon_i > 0$) for the finite differences approximation.

`force_symmetric`
: **Boolean value**:
Whether to force the result to be symmetric. This is done by taking the mean of the approximation and its transpose.

Value

A **dense matrix** of size $N \times N$:
The Hessian approximation $\widetilde{\mathbf{H}}_{ij}^{(0)}(\boldsymbol{\Theta}^{(0)})$ based on finite differences

Examples

```
library(sqp)
set.seed(3)
N <- 5

start.parm <- cbind(runif(N))

## Analytical distance function, gradient and Hessian matrix
distance.function <- function(parm)
  1/5*sum(parm^5) + 1/3*sum(parm^3)

gradient.function <- function(parm)
  cbind((parm^4) + (parm^2))

hessian.function <- function(parm)
  diag(c(4*parm^3+2*parm))

## Finite difference approximation
H <- bfgs_init(parm=start.parm,
               gradient_function=gradient.function,
               eps = 1e-12,
               force_symmetric = TRUE)

## Compare analytical Hessian matrix with finite difference approximation
hessian.function(start.parm)
print(H)
```

`bfgs_update` *(Damped) BFGS Hessian approximation*

Description

`BFGS` update for approximation of the Hessian matrix $\mathbf{H_f}(\mathbf{\Theta}$ of a function $\mathbf{f}$ (cf. Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970) in its damped version proposed by Powell (1978). The approximation is based on first-order information (differences in parameters $\mathbf{\Theta}$ and gradients $\nabla_\mathbf{f}(\mathbf{\Theta}$ between two iterations). The update in iteration k is defined by

$$\widetilde{\mathbf{H}}_{ij}^{(k)}(\mathbf{\Theta}^{(k)}) = \widetilde{\mathbf{H}}_{ij}^{(k-1)}(\mathbf{\Theta}^{(k-1)}) + \frac{\mathbf{y}\mathbf{y}^T}{\mathbf{y}^T\mathbf{s}} - \frac{\widetilde{\mathbf{H}}_{ij}^{(k-1)}(\mathbf{\Theta}^{(k-1)})\mathbf{s}\mathbf{s}^T\widetilde{\mathbf{H}}_{ij}^{(k-1)}(\mathbf{\Theta}^{(k-1)})}{\mathbf{s}^T\widetilde{\mathbf{H}}_{ij}^{(k-1)}(\mathbf{\Theta}^{(k-1)})\mathbf{s}} \ ,$$

where

$$\mathbf{s} = \mathbf{\Theta}^{(k)} - \mathbf{\Theta}^{(k-1)}$$

is the difference in parameters between two subsequent iterations, and

$$\mathbf{y} = \nabla_\mathbf{f}(\mathbf{\Theta}^{(k)}) - \nabla_\mathbf{f}(\mathbf{\Theta}^{(k-1)})$$

is the corresponding difference in gradients.

Usage

```
bfgs_update(
hessian,
old_y,
new_y,
old_gradient,
new_gradient,
constraint_adjustment = TRUE
)
```

Arguments

`hessian`

Dense matrix of size $N \times N$:
Current approximation $\widetilde{H}_{ij}^{(k-1)}$ of the Hessian matrix, which is updated by reference. Needs to be symmetric positive definite. A common starting point for the `BFGS` algorithm is the identity matrix or an approximation by finite differences.

`old_y, new_y, old_gradient, new_gradient`

Numeric vectors of size `N`:
Parameters **old_y**,**new_y** ($\boldsymbol{\Theta}^{(k-1)}$ and $\boldsymbol{\Theta}^{(k)}$), and corresponding gradients **old_gradient**, **new_gradient** ($\nabla_{\mathbf{f}}(\boldsymbol{\Theta}^{(k-1)}$ and $\nabla_{\mathbf{f}}(\boldsymbol{\Theta}^{(k)})$ from previous and current iteration.

`constraint_adjustment`

Boolean:
Whether to enforce positive definiteness, mainly for constrained optimization.

Value

`NULL`. Argument 'hessian' is updated by reference.

References

Broyden CG (1970). "The convergence of a class of double-rank minimization algorithms: 2. The new algorithm." *IMA journal of applied mathematics*, **6**(3), pp. 222–231. doi: 10.1093/imamat/6.3.222.

Fletcher R (1970). "A new approach to variable metric algorithms." *The computer journal*, **13**(3), pp. 317–322. doi: 10.1093/comjnl/13.3.317.

Goldfarb D (1970). "A family of variable-metric methods derived by variational means." *Mathematics of computation*, **24**(109), pp. 23–26. doi: 10.1090/S00255718-197002582496.

Powell MJ (1978). "A fast algorithm for nonlinearly constrained optimization calculations." In *Numerical analysis*, pp. 144–157. Springer. doi: 10.1007/BFb0067703.

Shanno DF (1970). "Conditioning of quasi-Newton methods for function minimization." *Mathematics of computation*, **24**(111), pp. 647–656. doi: 10.1090/S00255718-19700274029X.

Examples

```
library(sqp)
set.seed(3)
N <- 5

start.parm <- cbind(runif(N))

## Analytical distance function, gradient and Hessian matrix
distance.function <- function(parm)
  1/5*sum(parm^5) + 1/3*sum(parm^3)

gradient.function <- function(parm)
  cbind((parm^4) + (parm^2))

hessian.function <- function(parm)
  diag(c(4*parm^3+2*parm))

## Finite difference approximation
H <- bfgs_init(parm=start.parm,
               gradient_function=gradient.function,
               eps = 1e-12,
               force_symmetric = TRUE)

## Compare analytical Hessian matrix with finite difference approximation
hessian.function(start.parm)
print(H)

## Make first update
parm.1 <- start.parm - solve(H)%*%gradient.function(start.parm)

## Check decrease of distance function
distance.function(start.parm)
distance.function(parm.1)

## BFGS update for Hessian approximation
bfgs_update(hessian=H,
            old_y = start.parm,
            new_y = parm.1,
            old_gradient=gradient.function(start.parm),
            new_gradient=gradient.function(parm.1),
            constraint_adjustment = FALSE)

## Make second update
parm.2 <- parm.1 - solve(H)%*%gradient.function(parm.1)

## Check decrease of distance function
```

```
distance.function(parm.1)
distance.function(parm.2)
```

`qp_solver` *Quadratic optimization solver*

Description

Dense & Sparse solvers for linearly constrained quadratic optimization problems (cf. Fletcher 1971; Nocedal and Wright 1999; Powell 1978; Wilson 1963). This function iteratively solves quadratic optimization problems under N_{eq} linear equality and N_{ineq} linear inequality constraints. Slack variables $\boldsymbol{\xi} = [(\boldsymbol{\xi}_{eq}^{+})^T, (\boldsymbol{\xi}_{eq}^{-})^T, (\boldsymbol{\xi}_{ineq})^T]^T$ can be used for infeasible or temporarily violated constraints. In summary, such problems have the form

$$argmin_{\boldsymbol{\Theta},\boldsymbol{\xi}}(\boldsymbol{\Theta}^T\mathbf{Q}\boldsymbol{\Theta} + \boldsymbol{\Theta}^T\mathbf{l} + \varsigma \cdot ||\boldsymbol{\xi}||_1)$$

$$s.t.$$

$$\mathbf{C}_{eq}\boldsymbol{\Theta} + \boldsymbol{\xi}_{eq}^{+} - \boldsymbol{\xi}_{eq}^{-} = \mathbf{t}_{eq}$$

$$\mathbf{C}_{ineq}\boldsymbol{\Theta} - \boldsymbol{\xi}_{ineq} \leq \mathbf{t}_{ineq}$$

$$\boldsymbol{\xi} \geq \mathbf{0}$$

for a vector of N unknown parameters $\boldsymbol{\Theta} = [\Theta_1, \ldots, \Theta_N]^T$.

Usage

```
qp_solver(
  Q,
  C_eq = NULL,
  C_ineq = NULL,
  l = NULL,
  t_eq = NULL,
  t_ineq = NULL,
  x = NULL,
  penalty = 1e+10,
  tol = 1e-07,
  max_iter = 500,
  fast = FALSE,
  all_slack = FALSE,
  debug = FALSE,
  solver = 0
)
```

Arguments

`Q, C_eq, C_ineq`

Dense or sparse numeric matrices:

Q $N \times N$-**matrix**:
Quadratic distance multiplier **Q** for the optimization problem. Typically a (damped) BFGS approximation of the Hessian matrix.

C_eq $N_{eq} \times N$-**matrix**:
Equality constraint multiplier $\mathbf{C}_{eq}$ for the N_{eq} equality constraints.

C_ineq $N_{ineq} \times N$-**matrix**:
Inequality constraint multiplier $\mathbf{C}_{ineq}$ for the N_{ineq} inequality constraints.

`l, t_eq, t_ineq`
Numeric vectors:

l **Vector** of size N:
Linear distance multiplier **l** for the optimization problem .

t_eq **Vector** of size N_{eq}:
Targets $\mathbf{t}_{eq}$ for equality constraints.

t_ineq **Vector** of size N_{ineq}:
upper bounds $\mathbf{t}_{ineq}$ for inequality constraints.

`x` **Numeric vector** of size `N`:
Initial values for optimization parameters $\boldsymbol{\Theta}$. Slack variables are only used for constraints violated by this `x` unless `all_slack` is `TRUE`.

`penalty`
Numeric value:
Penalty multiplier ς for slack variables in distance function.

`tol` **Numeric value**:
Tolerance for assessing convergence criteria & constraints.

`max_iter`
Integer value:
Maximum number of iterations for the active set strategy in presence of inequality constraints.

`fast`
Boolean:
Whether to use faster (but lower quality) solver (cf. Armadillo documentation): "fast mode: disable determining solution quality via rcond, disable iterative refinement, disable equilibration".

`all_slack`
Boolean:
Whether to use slack variables for all constraints instead of
only for the ones violated by the initial values

`debug`
Boolean:
Whether to print debugging status messages.

`solver`
Solver identification used for optimization in the **dense** matrix case. Not yet used.

Details

Sequential quadratic programming relies on iteratively solving linear approximations of the optimality conditions to update parameters $\boldsymbol{\Theta}$ (cf. Kjeldsen 2000; Kuhn and Tucker 1951). This is equivalent to minimizing a quadratic approximation of the distance function under linearized constraint functions. Therefore, `qp_solver` can be used to solve such quadratic sub-problem in sequential quadratic programming.

Solving a quadratic problem under linear equality constraints is equivalent to solving a system of linear equations. The inequality constraints are handled by an active set strategy, where the binding ones are treated as equalities, and the active set is found iteratively (cf. Fletcher 1971; Nocedal and Wright 1999; Powell 1978; Wilson 1963).

Value

A **named list** with values

x Final values for optimization parameters

lagrange_eq, lagrange_ineq Lagrange multipliers for equality and inequality constraints

slack_eq_positive, slack_eq_negative Positive and negative slack variables for equality constraints

slack_ineq Slack variables for inequalities constraints

lagrange_slack_eq_positive, lagrange_slack_eq_negative, lagrange_slack_ineq
Lagrange multipliers for positivity of slack variables

Note

Although there is already an implementation for using the SuperLU sparse solver within this package, the solver itself is currently not included due to licensing considerations. Therefore, sparse matrices are converted to dense ones in the solving procedure. Hopefully, this can be updated in the near future.

References

Fletcher R (1971). "A general quadratic programming algorithm." *IMA Journal of Applied Mathematics*, **7**(1), pp. 76–91. doi: 10.1093/imamat/7.1.76.

Kjeldsen TH (2000). "A contextualized historical analysis of the Kuhn-Tucker theorem in nonlinear programming: the impact of World War II." *Historia mathematica*, **27**(4), pp. 331-361. doi: 10.1006/hmat.2000.2289.

Kuhn HW and Tucker AW (1951). "Nonlinear programming." In Neyman J (ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. `http://web.math.ku.dk/~moller/undervisning/MASO2010/kuhntucker1950.pdf`.

Nocedal J and Wright SJ (1999). *Numerical optimization*. Springer, New York. ISBN 387987932.

Powell MJ (1978). "A fast algorithm for nonlinearly constrained optimization calculations." In *Numerical analysis*, pp. 144–157. Springer. doi: 10.1007/BFb0067703.

Wilson RB (1963). *A simplicial algorithm for concave programming*. PhD thesis, Harvard University.

Examples

```
set.seed(1)
n <- 5

x_init <- cbind(runif(n))

w <- runif(n)

## minimize sum(3*x^2 + 3*x)
Q <- 3*diag(n)
l <- cbind(rep(3,n))

## Equality constraints: sum(x) == 1 and sum(w*x) == 5
C_eq <- rbind(1,w)
t_eq <- rbind(1,5)
## Inequality constraints: all(x >= -4) & all(x <= 4)
C_ineq <- rbind(diag(n),-diag(n))
t_ineq <- cbind(rep(c(4,4),each=n))

output <- qp_solver(Q = Q,
               C_eq = C_eq,
               C_ineq = C_ineq,
               l=l,
               t_eq = t_eq,
               t_ineq = t_ineq,
               x = x_init,
               tol = 1e-15)


## Check equality constraints I:  sum(x) == 1
sum(output$x)
## Check equality constraints II: sum(w*x) == 5

## Check inequality constraints I:  all(x >= -4)
all(output$x >= -4)
## Check inequality constraints II: all(x <= 4)
```

C.2 Documentation for Package ann

Package 'ann'

15.04.2021

Type Package

Title Artificial Neural Networks

Version 1.0

Date 2020-03-25

Author Simon Lenau

Maintainer Simon Lenau <lenau@uni-trier.de>

Description Object-oriented implementation of artificial neural networks. This package implements ANNs in C++ and provides pointers to R. Neural networks can be fit by gradient descent or the BFGS algorithm. Both options can be combined with batch learning. Semi-parametric components in form of B-Spline layers may be used, which allow optimizing knots as coefficients of the ANN.

License GPL-3

LinkingTo Rcpp, RcppArmadillo, RcppEigen, sqp

SystemRequirements C++11, GNU Make

RdMacros Rdpack

Encoding UTF-8

RoxygenNote 7.1.0

---

ann *Semi-parametric artificial neural network object*

Description

An artificial neural network (ANN) is composed of one or multiple layers $i = l, \ldots, L$, specified in the constructor. It maps input variables $\mathbf{X}$ to output variables $\mathbf{Y}$, using layer-specific activation functions $\mathbf{t}^{(l)}$ and coefficients $\mathbf{\Theta}^{(l)}$. Predictions of the ANN are obtained by

$$\widehat{\mathbf{Y}} = \mathbf{t}^{(L)}(\sum_{j=0}^{L} \widetilde{\mathbf{X}}^{(j)} \mathbf{\Theta}^{(j)})$$

$$\widetilde{\mathbf{X}}^{(l)} = \mathbf{t}^{(l)}(\sum_{j=0}^{L} \widetilde{\mathbf{X}}^{(j)} \mathbf{\Theta}^{(j)})$$

$$\widetilde{\mathbf{X}}^{(0)} = \mathbf{X} \ .$$

(cf. Bishop 1995; Hagan et al. 1996). This package implements ANNs in an object-oriented programming context, where the R-object is solely a pointer to a C++ object.

The artificial neural network itself is completely implemented in C++. For ease of fitting and use of ANNs in R, various utility wrapper functions and methods are available (see examples and details below).

Fitting of the neural network can be done by gradient descent or by using the BFGS algorithm implemented in package sqp. Both options can be used for batch learning as well (see examples and details below).

As an important feature of this package, semi-parametric components in form of B-Spline layers may be used, which allow optimizing knots as coefficients of the ANN. Further details on theory and implementation beyond the examples and details provided below can be found in Lenau (2021).

Usage

```
ann <- ann(x = x,
           y = y,
           layer_spec = layer_spec,
           weight = weight,
           distance_function = distance_function)
```

Arguments

`x` **Numeric matrices**:
The independent / input variables for the `ann`

`y` **Numeric matrices**:
The dependent / target / output variables for the `ann`

`layer_spec`
List:
The declaration of layers for the `ann`. Each list element of `layer_spec` is itself a list, with named elements

add_bias Boolean value: Whether a bias / intercept column should be added to the layers output.

input_layers Vector of indices for the input layers, where 0 corresponds to the input layer. A single numeric value unless `link_function='concatenation'`.

link_function String value: Link function of the layer that is applied to an input row-vector η_i for each observation i. One of

"linear" The identity activation function
$\mathbf{t}^{(l)}(\eta_i) = \eta_i$.

"softmax" The softmax activation function
$\mathbf{t}^{(l)}(\eta_i) = \mathbf{exp}(\eta_i)/||\mathbf{exp}(\eta_i)||_1$.

"raking" The raking activation function
$\mathbf{t}^{(l)}(\eta_i) = \mathbf{exp}(-\eta_i)$.

"bspline" The B-spline base functions of degree d
$\mathbf{t}^{(l)}(\eta_i) = \mathbf{B}^d(\eta_i, \mathbf{\Theta}^{(l)})$,
where $\mathbf{\Theta}^{(l)}$ are the spline's *knots*.

"concatenation" Activation function that can be used to concatenate (combine) the outputs of two or more layers, such that e.g.
$\mathbf{t}^{(l)}(\eta_i^k, \eta_i^l) = [\eta_i^k, \eta_l^k]$,
where η_i^k is the output of layer k .

"split" Activation function that can be used to split (select certain columns of) the output of a layer, such that e.g.
$\mathbf{t}^{(l)}(\eta_i) = [[\eta_i]_1, \ldots, [\eta_i]_j]$
for `j<length(`η_i`)`. In this context, $[\eta_i]_j$ denotes the j-th element of η_i.

coefficients *(semi-optional)* Initial values for the layer's coefficients, in matrix form.

dim *(semi-optional)* Output dimension (number of output columns) of the layer.

Note: At least one of `coefficients` *or* `dim` *must be specified to define the layer's output dimension.*

input_columns *(optional)* Indices of the input columns, where 0 corresponds to the first column. Ignored unless `link_function='split'`.

n_knots *(optional)* Number of knots in a B-spline layer.
Ignored unless `link_function='bspline'`.

degree *(optional)* Degree of the B-spline base function.
Ignored unless `link_function='bspline'`.

optimize_knots *(optional)* Boolean value: Whether to optimize the knot locations (TRUE) or keeping them fixed (FALSE).
Ignored unless `link_function='bspline'`.

`weights`

Numeric vector:
The case-weights applied for the distance function in the fitting process. If `NULL` (the default), all observations are given a weight of one.

`distance_function`

String value:
The distance function to be used for fitting. One of:

"squared" Squared loss function, i.e. residual sum of squares over all dependent variables $\mathbf{Y}$

"entropy" Deviance / cross-entropy, which is the negative binomial log-likelihood in case of a single binary dependent variable $\mathbf{Y}$)

Details

For fitting the neural network, a distance function $\delta(\mathbf{\Theta}^{(1)}, \ldots, \mathbf{\Theta}^{(L)})$ is minimized w.r.t. the optimization parameters $\mathbf{\Theta}^{(l)}$ for all $l = 1, \ldots, L$. These parameters are updated by

$$\mathbf{\Theta}^{(l)}_{new} = \mathbf{\Theta}^{(l)}_{old} - r \cdot \Delta_{\mathbf{\Theta}^{(l)}} \ ,$$

where

$$\Delta_{\mathbf{\Theta}^{(l)}} = \mathbf{Q}^{-1} \frac{\partial \ \delta(\mathbf{\Theta}^{(1)}, \ldots, \mathbf{\Theta}^{(L)})}{\partial \ \mathbf{\Theta}^{(l)}}$$

is the *step direction* and r is the *step size*. Furthermore $\mathbf{Q}$ is a symmetric matrix whose inverse is multiplied with the distance function's derivative w.r.t. $\mathbf{\Theta}^{(l)}$. In **backpropagation**, r is the *learning rate*, and $\mathbf{Q}$ is an identity matrix of appropriate dimension. In the **BFGS-update**, r is typically determined by a line search, and $\mathbf{Q}$ is an approximation for the Hessian matrix of δ (cf. Armijo 1966; Jarre and Stoer 2004; Nocedal and Wright 1999; Nesterov 2004).

After each update of all parameters, the predictions are updated by

$$\widehat{\mathbf{Y}} = \mathbf{t}^{(L)}(\sum_{j=0}^{L} \widetilde{\mathbf{X}}^{(j)} \mathbf{\Theta}^{(j)})$$

$$\widetilde{\mathbf{X}}^{(l)} = \mathbf{t}^{(l)}(\sum_{j=0}^{L} \widetilde{\mathbf{X}}^{(j)} \mathbf{\Theta}^{(j)})$$

$$\widetilde{\mathbf{X}}^{(0)} = \mathbf{X} \ .$$

In **Batch learning**, parameter updates are based on a subset of all observations. It can be achieved by providing one of the optional arguments `batch_size` or `batch_index` to functions `bfgs` or `backward_propagation` (see examples).
`batch_index` is a vector of observation indices that identifies observations used for the current parameter update.
`batch_size` is an integer value that determines the size of a batch that is selected from all observations by simple random sampling without replacement.

Value

`ann` is the constructor for an artificial neural network object.

`layer_declaration` returns a list that describes the layers of the `ann` object.

`input` and `target` respectively return numeric matrices of input and target variables.

`input_centering`, `input_scaling` `target_centering` and `target_scaling` return boolean values that respectively indicate whether input / target variables are currently centered / scaled.

`input_center`, `input_scale`, `target_center`, `target_scale` return numeric vectors of current centering / scaling values for input / target variables, respectively.

`weights` returns a numeric vector of current observation weights.

`learning_rate` returns a numeric value that constitutes the learning rate (fixed step width).

`coef` returns a list, where each list element contains coefficients $\boldsymbol{\Theta}^{(l)}$ for all layers $l = 0, \ldots, l$.

`coef_vec` returns a vector that contains all coefficients $\boldsymbol{\Theta}^{(l)}$ for all layers $l = 0, \ldots, l$ in a vectorized form.

`distance_type` returns a string value that describes the current loss .

`copy.ann` and `as.ann` return an `ann` object.

`parameter_randomization`, `forward_evaluation`, `backward_propagation`, `bfgs_init` and `bfgs` return NULL because they are used to update an `ann` object by reference.

`converged` returns a boolean value that indicates whether the change in individual loss contributions for all observations and target variables is smaller than argument `tol`.

`distance_value` returns a numeric matrix of individual loss contributions for all observations and target variables.

`fitted` and `residuals` return numeric matrices of respectively predictions / prediction errors for all observations and target variables.

`intercept` returns an integer vector containing the indices of intercept (bias) parameters in the vectorized coefficients.

`as.list` returns a list that contains all relevant information of the `ann` object.

References

Armijo L (1966). "Minimization of Functions Having Lipschitz Continuous First Partial Derivatives." *Pacific Journal of Mathematics*, **16**(1), pp. 1–3. doi: 10.2140/pjm.1966.16.1.

Bishop CM (1995). *Neural Networks for Pattern Recognition*. Calrendon Press, Oxford. ISBN 978-0-19-853864-6.

Hagan MT, Demuth HB, Beale MH and De Jesus O (1996). *Neural Network Design*, 2 edition. Pws Pub., Boston. ISBN 0-9717321-1-6.

Jarre F and Stoer J (2004). *Optimierung*. Springer, Berlin. ISBN 978-3-642-18785-8.

Lenau S (2021). *Statistical and Machine Learning Methods for Handling Selectivity in Non-Probability Samples*. PhD thesis, University of Trier.

Nesterov YE (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, New York. ISBN 978-1-4613-4691-3, doi: 10.1007/9781441988539.

Nocedal J and Wright SJ (1999). *Numerical Optimization*. Springer, New York. ISBN 978-0-387-22742-9.

Examples

```
library(ann)
set.seed(3)

#####################################
## Generate data for basic examples
#####################################

n <- 400
## Independent variable, including intercept
x <- cbind(1,rnorm(n,30,0.5))
## Dependent variable
y <- cbind(50 +
             5*x[,2] +
             4*(x[,2]-mean(x[,2]))^2 +
             1*(x[,2]-mean(x[,2]))^3 +
             rnorm(n,0,2))
## Observation weights
w <- 1/runif(n)

## Plot independent and dependent variable
plot(x[,2],y)

#####################################
## Construct ANN
#####################################

## Using a single layer with linear activation function,
## this ANN represents a simple linear regression structure
ann <- ann(x = x,
           y = y,
           layer_spec =
             list(list(dim = 1,
                       link_function = "linear",
                       input_layers = 0,
                       add_bias = FALSE)),
           weight = w,
           distance_function = "squared")

## Inspect layer specification of the ANN
layer_declaration(ann)

## Get current input (independent) variables
input(ann)
## Set new input (independent) variables (same as old one in this case)
input(ann) <- x

## Get current output (dependent) variables
target(ann)
## Set new output (dependent) variables (same as old one in this case)
```

```
target(ann) <- y

## Check whether input variables are currently centered and scaled
## using weighted means and standard deviations
input_centering(ann)
input_scaling(ann)
## Check current centering and scaling values for input variables
## (constant columns are kept unchanged)
input_center(ann)
input_scale(ann)
## Set whether input variables are centered and scaled
input_centering(ann) <- TRUE
input_scaling(ann) <- TRUE

## Check whether output variables are currently centered and scaled
## using weighted means and standard deviations
target_centering(ann)
target_scaling(ann)
## Check current centering and scaling values for target variables
## (constant columns are kept unchanged)
target_center(ann)
target_scale(ann)
## Set whether output variables are centered and scaled
target_centering(ann) <- TRUE
target_scaling(ann) <- TRUE

## Get current observation weights
weights(ann)
## Set new observation weights (same as old one in this case)
weights(ann) <- w

## Get current learning rate (multiplier for gradient descent)
learning_rate(ann)
## Set new learning rate (usually <= 1)
learning_rate(ann) <- 0.1

## Get current coefficients
coef(ann) # coefficients in layer-structure
coef_vec(ann) # coefficients in combined vector
## Randomize initial coefficients
## (gradient descent may fail if these are zero)
parameter_randomization(ann)
coef(ann)

## Get current distance (loss) function
distance_type(ann)
## Set new distance function (same as old one in this case)
distance_type(ann) <- "squared"

#######################################
## Fit ANN
```

```
#######################################

## Make copy of ANN
## to compare fitting via backpropagation (gradient descent) and BFGS
ann2 <- copy.ann(ann)

## Fit using backpropagation (gradient descent)
for(i in 1:1000)
{
  ## Forward-pass (update predictions from coefficients)
  forward_evaluation(ann)
  ## Backward-pass (update coefficients from gradients)
  backward_propagation(ann)
  ## Check whether change in distance function is larger than tolerance
  if(i>1 & converged(ann,tolerance=1e-12))
  {
    forward_evaluation(ann)
    cat("Backpropagation converged after ",i," iterations\n")
    break()
  }
}

## Fit using BFGS updates
## (alternative to backpropagation, using approximate Hessian matrix)

## Initialize BFGS information
bfgs_init(ann2)
for(i in 1:1000)
{
  ## Forward-pass (update predictions from coefficients)
  forward_evaluation(ann2)
  ## BFGS-Update
  bfgs(ann2)
  ## Check whether change in distance function is larger than tolerance
  if(i>1 & converged(ann2,tolerance=1e-12))
  {
    forward_evaluation(ann2)
    cat("BFGS converged after ",i," iterations\n")
    break()
  }
}

## Get current distance function values as sums of individual components
sum(distance_value(ann))
sum(distance_value(ann2))

## Get current fitted values of the ANNs
fitted(ann)
## Add fitted values to plot
points(x=x[,2],y=fitted(ann),col="red",pch=4)
```

```
## Check whether backpropagation and BFGS lead to the same results
all.equal(fitted(ann),fitted(ann2))
all.equal(coef_vec(ann),coef_vec(ann2))

## Get current residuals of the ANN
residuals(ann)

## Make out-of-sample predictions for different input data
x_newdata <- cbind(1,rnorm(n,30,0.5))
predict(ann,x_newdata)
## Add predictions to plot
points(x=x_newdata[,2],y=predict(ann,x_newdata),pch=1,col="blue")

#####################################
## Miscellaneous functions for ANNs
#####################################

## Determine intercept parameters
intercept(ann) # Intercept's position
coef_vec(ann)[intercept(ann)] # Intercept's value

## Convert ann to list, e.g.\ for saving to .RData file
ann_list <- as.list(ann)
## re-create ann from list
ann3 <- as.ann(ann_list)
```

C.3 Documentation for Package `calmod`

Package 'calmod'

15.04.2021

Type Package

Title Calibrated semi-parametric neural networks as integrated response and calibration models

Version 1.0

Date 2020-03-25

Author Simon Lenau

Maintainer Simon Lenau <lenau@uni-trier.de>

Description Implementation of calibrated semi-parametric neural networks which allow for incorporation of total, covariance and correlation calibration in response models. This package provides an integration and extensions of response (propensity) and calibration weighting, which are widespread techniques for different purposes, such as non-response adjustments and estimation from non-probability samples.

License GPL-3

Imports Rdpack

LinkingTo Rcpp, RcppArmadillo, RcppEigen, RcppParallel, data.table (>= 1.11.4), sqp, ann

SystemRequirements C++11, GNU Make

NeedsCompilation yes

RdMacros Rdpack

Encoding UTF-8

RoxygenNote 7.1.0

`cov_calib` *General response and/or calibration models*

Description

Response propensity and calibration weighting are widespread techniques for different purposes, such as non-response adjustments and estimation from non-probability samples. `cov_calib` provides an integration and extension of these approaches.

The general aim is to obtain a vector of response or calibration weights $\widetilde{\mathbf{w}}$ of size `n` for a sample of interest. These weights are determined as a function

$$\widetilde{\mathbf{w}} = \mathbf{f}(\mathbf{Z}, \omega) \circ \mathbf{w}$$

of a matrix of auxiliary variables $\mathbf{Z}$, where ω is a vector of parameters to be estimated. Furthermore, $\mathbf{w}$ denotes a vector of initial weights and $\circ$ is element-wise multiplication. The function $\mathbf{f}$ is referred to as the *response model*. To allow a flexible specification, this response model can be an arbitrary semi-parametric artificial neural network (ANN). which incorporates highly relevant special cases, such as generalized linear and additive regression models. Semi-parametric artificial neural network are implemented in package ann and allow for highly flexible specification of response models $\mathbf{f}(\mathbf{Z}, \omega) = \widehat{\mathbf{p}}^{\circ(-1)}$. In this context, $\widehat{\mathbf{p}}$ is the estimated response propensity obtained from the ANN, and $^\circ$ denotes element-wise power of vector elements.

At the same time, calibration constraints may be used to enforce similarity or coincidence between weighted estimates and some benchmark information. For example, soft calibration of totals induces constraints of the form

$$\widehat{\tau}_{\mathbf{X}}(\widetilde{\mathbf{w}}) = \tau_{\mathbf{X}} \circ \epsilon,$$

where $\mathbf{X}$ is a matrix of calibration variables for the sample. Further, $\tau_{\mathbf{X}}$ is a vector of known population totals (benchmarks) for these variables, and $\widehat{\tau}_{\mathbf{X}}(\widetilde{\mathbf{w}})$ are estimates of these totals that depend on $\widetilde{\mathbf{w}}$. The deviation of estimates from benchmarks is expressed by a vector ϵ of corresponding dimension. (cf. Chang and Kott 2008; Estevao and Saerndal 2000; Estevao and Saerndal 2006; Folsom and Singh 2000; Kott 2006). Analogous calibration of weighted covariances or correlations is possible as well, i.e.

$$\widehat{\mathbf{\Sigma}}_{\mathbf{X}}(\widetilde{\mathbf{w}}) = \mathbf{\Sigma}_{\mathbf{X}} \circ \mathbf{E}$$

or

$$\widehat{\rho}_{\mathbf{X}}(\widetilde{\mathbf{w}}) = \rho_{\mathbf{X}} \circ \mathbf{E},$$

where $\widehat{\mathbf{\Sigma}}_{\mathbf{X}}(\widetilde{\mathbf{w}})$, $\mathbf{\Sigma}_{\mathbf{X}}$, $\widehat{\rho}_{\mathbf{X}}(\widetilde{\mathbf{w}})$, and $\rho_{\mathbf{X}}$ denote estimated and true covariances or correlations, respectively. In analogy to ϵ, $\mathbf{E}$ represents a matrix of multipliers to express calibration error. Since this matrix is symmetric by definition of covariances and correlations, the unique elements of $\mathbf{E}$ are denoted by ε.

In general, the model is fit by finding optimization parameters

$$\mathbf{\Theta} = [\omega^T, \epsilon^T, \varepsilon^T]^T$$

through solving the optimization problem

$$\delta(\mathbf{\Theta}) = u \cdot \delta_{ANN}(\omega) + \frac{1}{2} \cdot \mathbf{v} \circ (\mathbf{\Theta} - \mathbf{C}_{\mathbf{\Theta}})^{\circ 2} \quad ,$$

$$s.t.$$

$$\widehat{\tau}_{\mathbf{X}}(\widetilde{\mathbf{w}}) = \tau_{\mathbf{X}} \circ \epsilon$$

$$\widehat{\mathbf{\Sigma}}_{\mathbf{X}}(\widetilde{\mathbf{w}}) = \mathbf{\Sigma}_{\mathbf{X}} \circ \mathbf{E}$$

$$\mathbf{L}_{\mathbf{\Theta}} \leq \mathbf{\Theta} \leq \mathbf{U}_{\mathbf{\Theta}} \ ,$$

where correlations $\widehat{\rho}_{\mathbf{X}}$ may be calibrated instead of covariances $\widehat{\mathbf{\Sigma}}_{\mathbf{X}}$, as summarized above. In this context, u and $\mathbf{v}$ constitute importance weights, $\mathbf{C}_{\mathbf{\Theta}}$ is a vector of centering constants for parameters $\mathbf{\Theta}$ of corresponding dimension, and $\mathbf{L}_{\mathbf{\Theta}}$ as well as $\mathbf{U}_{\mathbf{\Theta}}$ are vectors of lower and upper bounds for $\mathbf{\Theta}$ of the same dimension. Inequalities are applied element-wise. Furthermore, $\delta_{ANN}(\omega)$ is the scalar-valued loss function of the ANN response model. This distance component is used only when a reference sample is available, in which case it typically corresponds to the negative binomial log-likelihood (deviance) of the model. If no reference sample is available, $\delta_{ANN}(\omega)$ is ignored by setting $u = 0$. The implementation for fitting these models with respect to calibration constraints for totals, covariances and correlations is based on sequential quadratic programming (`SQP`), which is implemented in package sqp.

The resulting weights are flexible and can be adapted to various degrees of auxiliary information for weighting adjustments in the context of non-probability samples or non-response. They allow representing many important weighting methods as special cases, such as the generalized regression estimator (cf. Deville and Saerndal 1992) or response propensity weighting (cf. Rosenbaum and Rubin 1983).

Further details on theory and implementation can be found in Lenau (2021).

Usage

```
cov_calib(
  calibration_variables = NULL,
  weight = NULL,
  model_variables = NULL,
  parm_lb = NULL,
  parm_ub = NULL,
  parm_center = NULL,
  calibration_variables_reference = NULL,
  weight_reference = NULL,
  model_variables_reference = NULL,
  tau_lb = NULL,
  tau_ub = NULL,
  cov_lb = NULL,
  cov_ub = NULL,
  tau_target = NULL,
  cov_target = NULL,
  loss_weight = NULL,
  parm_weight = NULL,
  tau_weight = NULL,
  cov_weight = NULL,
  cutoff_min_tau = NULL,
  cutoff_min_cov = NULL,
  ann_specification = NULL,
  GRID = NULL,
  penalty_start = 0,
  penalty_constant = 1e-05,
  step_limitation = 0.2,
  step_basis = 0.9,
  maxeval = 500L,
  maxeval_local = 500L,
  maxeval_stepsize = 500L,
  maxeval_initial_values = 50L,
  weight_rescale_type = 0L,
  solver_no = 0L,
  batch_size = NULL,
  type = "direct",
  sparse = TRUE,
  adjust_range = TRUE,
  batch_constraints = FALSE,
  cor = FALSE,
  ML = FALSE,
  calibrate_reference_union = FALSE,
  scale_intercept = TRUE,
  bfgs_init_type = 1L,
  simplify_bfgs_update = FALSE,
  bfgs_initial_values = FALSE,
```

```
    step_information = FALSE,
    debug = FALSE,
    greg_init = 0L,
    raking_p_constraints = TRUE,
    tolerance = 1e-08
  )
```

Arguments

`calibration_variables`

Numeric matrix:
Matrix of calibration variables ($\mathbf{X}$) observed in the sample of interest. Estimates for these variables are adjusted towards calibration benchmarks provided as arguments `tau_target` and `cov_target`.
Default: NULL, such that no calibration constraints are used.

`weight`

Numeric vector or value:
Vector of initial weights ($\mathbf{w}$) for the sample of interest, e.g. design weights.
Default: NULL, corresponding to constant weights for all observations.

`model_variables`

Numeric matrix:
Matrix of independent variables in the response model ($\mathbf{Z}$) observed for the sample of interest. These are the input variables for the artificial neural network constructing the weights, such that resulting weights are a function of these variables.
Default: NULL, in which case an identity matrix is used. This works only when `type='direct'`.

`parm_lb, parm_ub`

Numeric vectors or values:
Lower and upper bounds $\mathbf{L}_\omega$ and $\mathbf{U}_\omega$ for the neural network parameters ω, defining the feasible range of these parameters as entries of $\mathbf{L}_{\mathbf{\Theta}}$ and $\mathbf{U}_{\mathbf{\Theta}}$.
Default: NULL, indicating no boundaries.

`parm_center`

Numeric vector or value:
Centering constants $\mathbf{C}_\omega$ for neural network parameter s ω, which are entries of $\mathbf{C}_{\mathbf{\Theta}}$. As described above, ridge penalization is applied for parameters deviating from these constants.
Default: NULL, indicating centering constants to be zero (except for the intercept(s) if `scale_intercept=TRUE`).

`calibration_variables_reference`

Numeric matrix:
Matrix of calibration variables ($\mathbf{X}$) observed in the reference sample. See argument `calibration_variables`. Used in conjunction with `calibration_variables` only if `calibrate_reference_union=TRUE`.
Default: NULL, corresponding to no calibration of the reference sample.

`weight_reference`

Numeric vector:
Vector of observation weights ($\mathbf{w}$) for the reference sample. See argument `weight`. Used for weighting the response model's loss function $\delta_{ANN}(\omega)$, and for the resulting calibration weights if `calibrate_reference_union=TRUE`.
Default: NULL, corresponding to constant weights for all observations.

`model_variables_reference`

Numeric matrix:
Matrix of response model variables ($\mathbf{Z}$) observed in the reference sample. See argument `model_variables`. Used in conjunction with `model_variables`, to calculate the ANN's loss function $\delta_{ANN}(\omega)$.
Default: NULL, in which case the reference sample is not used for estimating parameters $\boldsymbol{\Theta}$.

`tau_lb, cov_lb, tau_ub, cov_ub`

Numeric vectors or values:
Lower and upper bounds $\mathbf{L}_\epsilon$, $\mathbf{L}_\varepsilon$, $\mathbf{U}_\epsilon$, and $\mathbf{U}_\varepsilon$, for ϵ and ε. As components of $\mathbf{L}_{\boldsymbol{\Theta}}$, and $\mathbf{U}_{\boldsymbol{\Theta}}$, these boundaries can be used to limit the permissible deviations from calibration benchmarks.
Default: NULL, indicating no boundaries for the deviation from benchmarks.

`tau_target`

Numeric vector:
Vector of total calibration benchmarks $\tau_{\mathbf{X}}$.
Default: NULL, corresponding to no total calibration.

`cov_target`

Numeric matrix:
Matrix of covariance or correlation calibration benchmarks $\boldsymbol{\Sigma}_{\mathbf{X}}$ or $\rho_{\mathbf{X}}$.
Default: NULL, corresponding to no covariance of correlation calibration.

`loss_weight`

Numeric value:
Importance weight u for the ANN's loss-function as component of the overall distance measure. This value determines the importance of the ANN's loss function $\delta_{ANN}(\omega)$ in optimization when estimating parameters $\boldsymbol{\Theta}$.
Default: NULL.

`parm_weight`

Numeric vector or value:
Importance weights (entries of $\mathbf{v}$) for the squared distance $(\omega - \mathbf{C}_\omega)^{\circ \mathbf{2}}$ between response model parameters and `parm_center` as component of the overall distance measure. These values determine the importance of the ridge penalty terms in optimization when estimating parameters $\boldsymbol{\Theta}$.
Default: NULL.

`tau_weight`

Numeric vector or value:
Importance weights (entries of $\mathbf{v}$) for the squared distance between estimated and benchmark totals $(\epsilon - \mathbf{1})^{\circ 2}$ as component of the overall distance measure. This value determines the importance of the deviation from total benchmarks in optimization when estimating parameters $\boldsymbol{\Theta}$.
Default: NULL.

`cov_weight`

Numeric vector or value:
Importance weights (entries of $\mathbf{v}$) for the squared distance between estimated and benchmark covariances or correlations $(\varepsilon - \mathbf{1})^{\circ 2}$ as component of the overall distance measure. This value determines the importance of the deviation from covariance or correlation benchmarks in optimization when estimating parameters $\boldsymbol{\Theta}$.
Default: NULL.

`cutoff_min_tau, cutoff_min_cov`

Numeric value:
Minimal values for benchmarks to be considered for total or covariance calibration. Benchmarks below these cutoff values are ignored because relative deviations ϵ or ε for values close to zero can cause numerical instabilities.
Default: NULL, indicating no cutoff value.

`ann_specification`

List:
The layer specification for the artificial neural network, as described for argument `layer_spec` in package `ann`. Not used unless `type='ann'`. Default: NULL, which does only work if `type!='ann'`.

`GRID`

List:
A named list for performing grid-searches of hyper-parameters. Currently not used.
Default: NULL.

`penalty_start`

Numeric value:
Initial penalty multiplier for slack variables in sequential quadratic programming. The penalty multiplier is described as argument `penalty` for function `qp_solver` in package `sqp`
Default: 0

`penalty_constant`

Numeric value:
Minimal increase of `penalty_start` in each `SQP` iteration.
Default: 0.00001

`step_limitation`

Numeric value:
Lower bound for the decrease of the loss-function in the Armijo-type line search.
Default: 0.2

`step_basis`

Numeric value:
Multiplier for the step-size in the Armijo-type line search.
Default: 0.9

`maxeval`

Integer value:
Maximum number of iterations for the `SQP`-algorithm.
Default: 500

`maxeval_local`

Integer value:
Maximum number of iterations for the `QP`-algorithm, which is used within the `SQP`-algorithm.
Default: 500

`maxeval_stepsize`

Integer value:
Maximum number of iterations for the Armijo-type line search.
Default: 500

`weight_rescale_type`

Integer value:
Type of weight rescaling that is applied to **w** before optimization. One of:

0 No rescaling

1 Weights are scaled to sum to the number of observations in target and reference sample

2 Weights are scaled to sum to the number of observations in the target sample

3 Weights are scaled to sum to one

4 Weights are scaled to all be smaller than one

Default: 0

`solver_no`

Integer value:
Solver identification used for optimization if `sparse=FALSE`, i.e. in the **dense** matrix case. Not yet used.
Default: 0

`batch_size`

Integer value:
Size of batches (sub-samples) for stochastic gradient descent.
Default: NULL, corresponding to using all observations.

`type`

String value:
The type of weighting model to be used. One of:

"direct" GREG-type weighting model, using one parameter for each observation in the sample

"logit" Logistic regression model, corresponding to a generalized linear model with logit link.

"raking" A generalized raking model, which is similar to the logit model with a different link function, (cf. Kott 2006)

"ann" A general semi-parametric ANN, which requires layers to be specified in argument `ann_specification`.

Default: direct

`sparse`

Boolean:
Whether to use a sparse solver for optimization in the `SQP`-algorithm.
Default: TRUE

`adjust_range`

Boolean:
Whether the range of distance metric components should be scaled to have the same maximum. This is achieved by rescaling importance weights u and $\mathbf{v}$. If `adjust_range=TRUE`, the ANN's loss function $\delta_{ANN}(\omega)$, ridge penalty for ANN coefficients $(\omega - \mathbf{C}_\omega)^{\circ 2}$, and squared relative deviations from calibration targets $(\epsilon - \mathbf{1})^{\circ 2}$ and $(\varepsilon - \mathbf{1})^{\circ 2}$ are all ≤ 1.
Default: TRUE

`batch_constraints`

Boolean:
If batch learning (stochastic gradient descent) is used (`batch_size!=NULL`): Whether updates of calibration constraints in `SQP` should also be based on the batches rather than all observations.
Default: FALSE

`cor` **Boolean**:
Whether argument `cov_target` contains correlations $\widehat{\rho}_{\mathbf{X}}$ (`TRUE`) instead of covariances $\widehat{\mathbf{\Sigma}}_{\mathbf{X}}$ (`FALSE`).
Default: FALSE

`ML` **Boolean**:
Whether covariance calibration is to be done for maximum likelihood (`TRUE`) or unbiased (`FALSE`) covariance estimates $\widehat{\mathbf{\Sigma}}_{\mathbf{X}}$.
Default: FALSE

`calibrate_reference_union`

Boolean:
Whether calibration should be done for the union of target and reference sample.
Default: FALSE, in which case only the target sample is calibrated.

`scale_intercept`

Boolean:
Whether intercept parameter(s) (if any) should be scaled before optimization to account for the population size.
Default: TRUE

`bfgs_init_type`

Integer value:
Type of 'initial guess' used for initializing the BFGS Hessian approximation. One of:

0 Finite differences approximation for the Hessian of the *Lagrange function*

1 Finite differences approximation for the Hessian of the *loss function*

2 Identity matrix

Higher values result in faster but less accurate initialization of the algorithm.
Default: 1

`simplify_bfgs_update`
: **Boolean**:
Whether to use the BFGS update only for the weighting neural network parameters ω, and keep the Hessian fixed for the calibration error parameters ϵ and ε.
Default: FALSE

`step_information`
: **Boolean**:
Whether information about each optimization step should be included in the output.
Default: FALSE

`debug`
: **Boolean**:
Whether debug information should be printed while fitting the weighting model.
Default: FALSE

`greg_init`
: **Integer value**:
Type of parameter initialization. One of:

0 Parameters are chosen such that all initial weights are equal.

1 Parameters are chosen such that initial weights are (close to) GREG weights.

2 Parameters are chosen such that initial weights are (close to) GREG weights, but larger than one.

Default: 1

`raking_p_constraints`
: **Boolean**:
Whether probabilities $\widehat{\mathbf{p}}$ in the raking model should be explicitly constrained to be numerically greater than zero. Only used if `type='raking'`.
Default: TRUE

`tolerance`
: **Numeric value**:
Tolerance for numerical optimization.
Default: 0.00000001

Value

A list, containing the vector of weights $\widetilde{\mathbf{w}}$ and information about the weighting model from which it is obtained.

References

Chang T and Kott PS (2008). "Using calibration weighting to adjust for nonresponse under a plausible model." *Biometrika*, **95**(3), pp. 555–571. doi: 10.1093/biomet/asn022.

Deville J and Saerndal C (1992). "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association*, **87**(418), pp. 376–382. doi: 10.1080/01621459.1992.10475217.

Estevao VM and Saerndal C (2000). "A functional form approach to calibration." *Journal of Official Statistics*, **16**(4), pp. 379–399. ISSN 0282423X.

Estevao VM and Saerndal C (2006). "Survey estimates by calibration on complex auxiliary information." *International Statistical Review*, **74**(2), pp. 127–147. doi: 10.1111/j.17515823.2006.tb00165.x.

Folsom RE and Singh AC (2000). "The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification." In *Proceedings of the Survey Research Methods Section, American Statistical Association (2000)*, pp. 598–603.

Kott PS (2006). "Using calibration weighting to adjust for nonresponse and coverage errors." *Survey Methodology*, **32**(2), pp. 133–142.

Lenau S (2021). *Statistical and Machine Learning Methods for Handling Selectivity in Non-Probability Samples*. PhD thesis, University of Trier.

Rosenbaum PR and Rubin DB (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, **70**(1), pp. 41–55. doi: 10.1093/biomet/70.1.41.

Examples

```
library(calmod)

set.seed(3)
N <- 50000
n <- 500

## Generate calibration variables
x0 <- cbind(rep(1,N))
x1 <- cbind(rlnorm(N))
x2 <- 0.5*x1+rlnorm(N)
X <- cbind(x0,x1,x2)

## Generate response model variables
z0 <- x0
z1 <- 0.5*x1+0.5*rlnorm(N)
z2 <- 0.75*x2+0.25*rlnorm(N)
Z <- cbind(z0,z1,z2)

## Generate response probabilities
```

```
pi <- 1/(1+exp(-rowSums(Z)))
pi <- pi/sum(pi)*n

## draw sample (Poisson sampling)
smp <- which(runif(N) <= pi)

## draw reference sample (simple random sampling)
smp_ref <- sample.int(N,n)

## Fit calibrated propensity (logit) model
weights <- cov_calib(
  calibration_variables = X[smp,],
  model_variables = Z[smp,],
  model_variables_reference = Z[smp_ref,],
  tau_target = colSums(X),
  cov_target = cov(X),
  tau_lb = 0.9,
  tau_ub = 1.1,
  cov_lb = 0.9,
  cov_ub = 1.1,
  maxeval = 1000,
  type = "logit",
  cor=FALSE,
  ML=TRUE
)
## Print summary:
## weights, weighting parameters and calibration constraints
summary(weights)
```

Appendix D Additional Results for the German WageIndicator Web Survey

Considering the same data and variables, the following tables provide results that are complementary to those discussed in chapter 7. Outcomes for combinations of model- and pseudo-design-based approaches for point estimation are summarized in tables D.1 to D.4, considering the mean absolute error over all income classes. The individual estimated frequencies for each class are provided in tables D.5 and D.6. Corresponding estimated standard deviations of the point estimates are presented in tables D.7 and D.8. Since it is not clear which inferential approach is the most adequate for non-probability samples, these results are based on the most conservative approach, which are Monte Carlo bootstrap estimates of the total variance (cf. section 5.4). Similar as in chapter 7, each of the tables is organized in blocks that encompass methods using the same kind of auxiliary information.

Table D.1: Mean absolute errors for income class frequencies (in percentage points) estimated by weighted aggregation of predictions in the WI (weighting methods without response propensity model)

	No model (weighted WI)	GLM (OLS)	GLM (Ridge)	GLM (LASSO)	GLM (Elastic net)	GAM (fix)	Regression tree	MARS	ANN (opt.)	SVM	GLMM	GAMM (fix)	Heckman model	Plain pseudo-design based estimates
Calibration benchmarks: None														
Unweighted	2.8	3.8	3.9	3.8	3.8	3.8	7.0	5.5	3.9	4.2	3.8	3.8	3.3	2.8
Calibration benchmarks: Totals														
Sub-sampling	3.3	4.5	4.5	4.5	4.5	4.5	6.6	5.1	4.5	4.5	4.5	4.5	5.4	3.3
Post-stratification	2.8	3.9	4.0	3.9	3.9	3.9	7.0	5.4	4.0	4.5	3.9	3.9	5.0	2.8
GREG	3.0	4.9	4.9	4.9	4.9	4.9	6.9	7.2	5.4	5.4	4.9	4.9	5.8	3.0
cal. ANN (1 par./obs.)	3.0	4.3	4.4	4.3	4.3	4.3	7.1	6.0	4.5	4.7	4.3	4.3	3.6	3.0
cal. ANN (par.)	2.9	4.5	4.5	4.5	4.5	4.5	6.9	6.2	4.8	4.8	4.5	4.5	4.7	2.9
cal. ANN (fix)	2.9	4.5	4.5	4.5	4.5	4.5	6.9	6.3	4.8	4.8	4.5	4.5	4.8	2.9
cal. ANN (opt.)	3.1	4.7	4.8	4.7	4.7	4.7	7.1	6.7	5.0	5.2	4.7	4.7	3.9	3.1
Calibration benchmarks: Covariances														
GREG	3.0	4.4	4.4	4.4	4.4	4.4	7.1	6.4	4.7	4.9	4.4	4.4	3.6	3.0
cal. ANN (1 par./obs.)	2.9	4.5	4.5	4.5	4.5	4.5	7.1	5.7	4.5	4.6	4.5	4.5	3.8	2.9
cal. ANN (par.)	3.5	5.2	5.2	5.2	5.2	5.2	7.1	6.1	5.2	5.3	5.2	5.2	3.8	3.5
cal. ANN (fix)	2.4	6.6	6.6	6.6	6.6	6.6	7.1	7.4	6.6	6.3	6.6	6.6	7.2	2.4
cal. ANN (opt.)	3.2	4.5	4.5	4.5	4.5	4.5	7.1	6.8	4.8	5.1	4.5	4.5	4.5	3.2
Calibration benchmarks: Totals and covariances														
GREG	3.5	5.5	5.5	5.4	5.4	5.5	7.0	7.2	5.8	6.0	5.5	5.5	5.2	3.5
cal. ANN (1 par./obs.)	3.8	6.5	6.4	6.4	6.4	6.5	7.5	5.6	5.5	5.5	6.5	6.5	5.1	3.8
cal. ANN (par.)	2.5	3.8	3.9	3.8	3.8	3.8	6.9	5.8	3.7	5.0	3.8	3.8	6.1	2.5
cal. ANN (fix)	2.9	6.9	6.9	6.9	6.9	6.9	7.1	7.6	6.9	7.6	6.9	6.9	7.6	2.9
cal. ANN (opt.)	3.2	4.5	4.5	4.5	4.5	4.5	7.1	6.8	4.9	5.1	4.5	4.5	4.6	3.2
Plain model-based estimates	2.8	3.8	3.9	3.8	3.8	3.8	7.0	5.5	3.9	4.2	3.8	3.8	3.3	

Highest mean absolute error

Lowest mean absolute error

cal. ANN: calibrated ANN
(1 par./obs.): one parameter per observation (as for the GREG)
(par.): parametric
(fix): non-parametric, fixed knots
(opt.): non-parametric, optimized knots

Plain model- / design-based estimates: Values in the last row and column are the same as in the first, but the coloring respectively concerns exclusively model- and pseudo-design-based estimates. All other colors in the table concern the comparison of all possible cross-combinations of weighting and prediction.

Table D.2: Mean absolute errors for income class frequencies (in percentage points) estimated by weighted aggregation of predictions in the WI (weighting methods with response propensity model)

	No model (weighted WI)	GLM (OLS)	GLM (Ridge)	GLM (LASSO)	GLM (Elastic net)	GAM (fix)	Regression tree	MARS	ANN (opt.)	SVM	GLMM	GAMM (fix)	Heckman model	Plain pseudo-design based estimates
Calibration benchmarks: None														
Unweighted	2.8	3.8	3.9	3.8	3.8	3.8	7.0	5.5	3.9	4.2	3.8	3.8	3.3	2.8
Logit (par.)	2.8	4.2	4.2	4.2	4.2	4.2	6.8	5.6	4.5	4.7	4.2	4.2	5.3	2.8
Pseudo-Weights (par.)	2.7	3.9	4.0	3.9	3.9	3.9	7.0	5.8	4.0	4.5	3.9	3.9	4.1	2.7
cal. ANN (par.)	3.0	4.4	4.4	4.4	4.4	4.4	7.1	6.0	4.6	4.7	4.4	4.4	3.6	3.0
Logit (fix)	2.8	4.0	4.0	4.0	4.0	4.0	6.6	4.7	4.1	4.3	4.0	4.0	5.3	2.8
Pseudo-Weights (fix)	2.7	3.6	3.6	3.6	3.6	3.6	6.9	5.3	3.7	4.5	3.6	3.6	4.3	2.7
cal. ANN (fix)	3.0	4.4	4.4	4.4	4.4	4.4	7.1	6.0	4.6	4.7	4.4	4.4	3.6	3.0
cal. ANN (opt.)	3.0	4.4	4.4	4.4	4.4	4.4	7.1	6.0	4.6	4.7	4.4	4.4	3.6	3.0
Calibration benchmarks: Totals														
Logit (par.) and GREG	3.0	4.9	4.9	4.9	4.9	4.9	7.0	7.1	5.3	5.2	4.9	4.9	5.4	3.0
cal. ANN (par.)	4.5	7.8	7.8	7.8	7.8	7.8	7.4	5.8	7.8	5.8	7.8	7.8	5.5	4.5
Logit (fix) and GREG	2.8	4.2	4.2	4.2	4.2	4.2	6.7	5.8	4.5	4.7	4.2	4.2	5.1	2.8
cal. ANN (fix)	4.4	7.8	7.8	7.8	7.8	7.8	7.4	5.6	7.8	5.8	7.8	7.8	5.6	4.4
cal. ANN (opt.)	3.0	4.4	4.5	4.4	4.4	4.4	7.1	6.0	4.6	4.7	4.4	4.4	3.6	3.0
Calibration benchmarks: Covariances														
Logit (par.) and GREG	2.9	4.1	4.2	4.2	4.2	4.1	7.1	6.2	4.4	5.0	4.1	4.1	3.4	2.9
cal. ANN (par.)	3.2	4.8	4.8	4.8	4.8	4.8	7.1	5.5	4.7	4.5	4.8	4.8	3.8	3.2
Logit (fix) and GREG	2.8	4.2	4.2	4.2	4.2	4.2	7.1	5.6	4.3	4.3	4.2	4.2	3.5	2.8
cal. ANN (fix)	2.7	4.1	4.2	4.2	4.2	4.1	7.0	6.7	4.2	4.7	4.1	4.1	3.9	2.7
cal. ANN (opt.)	3.1	4.4	4.4	4.4	4.4	4.4	7.1	6.1	4.6	4.8	4.4	4.4	3.8	3.1
Calibration benchmarks: Totals and covariances														
Logit (par.) and GREG	3.1	5.3	5.4	5.4	5.4	5.3	6.9	7.2	5.5	6.0	5.3	5.3	6.0	3.1
cal. ANN (par.)	4.5	7.8	7.8	7.8	7.8	7.8	7.4	5.8	7.8	5.8	7.8	7.8	5.5	4.5
Logit (fix) and GREG	2.9	4.3	4.3	4.2	4.2	4.3	6.9	5.9	4.7	5.0	4.3	4.3	4.7	2.9
cal. ANN (fix)	4.4	7.8	7.8	7.8	7.8	7.8	7.4	5.6	7.8	5.8	7.8	7.8	5.6	4.4
cal. ANN (opt.)	3.1	4.4	4.4	4.4	4.4	4.4	7.1	6.2	4.7	4.9	4.4	4.4	3.9	3.1
Plain model-based estimates	2.8	3.8	3.9	3.8	3.8	3.8	7.0	5.5	3.9	4.2	3.8	3.8	3.3	

Logit: Weights from GLM with logit link
cal. ANN: calibrated ANN
Logit and GREG: Weights from GLM with logit link, calibrated using the GREG

(par.): parametric
(fix): non-parametric, fixed knots
(opt.): non-parametric, optimized knots

Plain model- / design-based estimates: Values in the last row and column are the same as in the first, but the coloring respectively concerns exclusively model- and pseudo-design-based estimates. All other colors in the table concern the comparison of all possible cross-combinations of weighting and prediction.

Table D.3: Mean absolute errors for income class frequencies (in percentage points) estimated from the imputed Microcensus, using a weighted loss function for prediction models (weighting methods without response propensity model)

	No model (weighted WI)	Matching	GLM (OLS)	GLM (Ridge)	GLM (LASSO)	GLM (Elastic net)	GAM (fix)	Regression tree	MARS	ANN (opt.)	SVM	GLMM	GAMM (fix)	Heckman model	Plain pseudo-design based estimates
Calibration benchmarks: None															
Unweighted	2.8	3.7	4.6	4.7	4.6	4.6	4.6	6.8	5.7	4.0	4.4	4.6	6.5	5.4	2.8
Calibration benchmarks: Totals															
Sub-sampling	3.3		5.0	5.0	5.0	5.0	5.0	7.1	7.7	8.0	4.0	5.2	5.5	8.0	3.3
Post-stratification	2.8		4.6	4.8	4.6	4.6	4.6	6.7	4.6	4.4	4.5	4.6	6.3	3.1	2.8
GREG	3.0		4.5	8.0	8.0	8.0	4.5	8.0	8.0	8.0	8.0	4.5	6.5	8.0	3.0
cal. ANN (1 par./obs.)	3.0		4.6	4.7	4.6	4.6	4.6	6.8	4.3	3.8	4.8	4.6	6.3	4.5	3.0
cal. ANN (par.)	2.9		4.5	4.5	4.5	4.5	4.4	6.8	5.4	4.0	5.3	4.5	6.5	4.5	2.9
cal. ANN (fix)	2.9		4.4	4.4	4.4	4.4	4.4	6.7	4.4	4.3	5.4	4.5	7.7	5.5	2.9
cal. ANN (opt.)	3.1		4.5	4.5	4.5	4.5	4.4	7.3	4.0	4.0	5.0	4.5	6.3	5.0	3.1
Calibration benchmarks: Covariances															
GREG	3.0		4.7	8.0	8.0	8.0	4.7	8.0	8.0	8.0	8.0	4.7	6.3	8.0	3.0
cal. ANN (1 par./obs.)	2.9		4.5	4.5	4.5	4.5	4.5	6.4	5.3	4.1	4.6	4.5	6.5	4.9	2.9
cal. ANN (par.)	3.5		4.4	4.5	4.4	4.4	4.4	6.8	4.2	3.9	5.2	4.4	6.5	8.0	3.5
cal. ANN (fix)	2.4		4.6	5.5	4.7	4.8	5.1	6.6	5.4	8.0	3.6	4.6	6.3	8.0	2.4
cal. ANN (opt.)	3.2		4.8	4.8	4.8	4.8	4.8	7.3	4.5	3.9	4.5	4.8	6.3	8.0	3.2
Calibration benchmarks: Totals and covariances															
GREG	3.5		4.5	8.0	8.0	8.0	4.5	8.0	8.0	8.0	8.0	4.5	6.3	8.0	3.5
cal. ANN (1 par./obs.)	3.8		4.5	8.0	8.0	8.0	4.5	8.0	8.0	8.0	8.0	4.5	6.5	8.0	3.8
cal. ANN (par.)	2.5		4.3	5.4	4.3	4.3	4.3	6.9	6.5	3.8	5.7	4.3	7.7	5.2	2.5
cal. ANN (fix)	2.9		6.4	7.4	7.4	7.4	4.6	7.4	8.0	8.0	3.5	8.0	8.0	8.0	2.9
cal. ANN (opt.)	3.2		4.7	4.7	4.7	4.7	4.7	7.3	6.5	4.5	6.3	4.7	6.3	4.9	3.2
Plain model-based estimates	2.8	3.7	4.6	4.7	4.6	4.6	4.6	6.8	5.7	4.0	4.4	4.6	6.5	5.4	

Highest mean absolute error

Lowest mean absolute error

cal. ANN: calibrated ANN
(1 par./obs.): one parameter per observation (as for the GREG)
(par.): parametric
(fix): non-parametric, fixed knots
(opt.): non-parametric, optimized knots

Plain model- / design-based estimates: Values in the last row and column are the same as in the first, but the coloring respectively concerns exclusively model- and pseudo-design-based estimates. All other colors in the table concern the comparison of all possible cross-combinations of weighting and prediction.

Table D.4: Mean absolute errors for income class frequencies (in percentage points) estimated from the imputed Microcensus, using a weighted loss function for prediction models (weighting methods with response propensity model)

	No model (weighted WI)	Matching	GLM (OLS)	GLM (Ridge)	GLM (LASSO)	GLM (Elastic net)	GAM (fix)	Regression tree	MARS	ANN (opt.)	SVM	GLMM	GAMM (fix)	Heckman model	Plain pseudo-design based estimates
Calibration benchmarks: None															
Unweighted	2.8	3.7	4.6	4.7	4.6	4.6	4.6	6.8	5.7	4.0	4.4	4.6	6.5	5.4	2.8
Logit (par.)	2.8		4.5	4.5	4.5	4.5	4.5	6.8	5.0	4.6	7.5	4.5	6.5	7.6	2.8
Pseudo-Weights (par.)	2.7		4.6	4.7	4.6	4.6	4.7	6.8	8.0	4.5	7.5	4.7	6.5	5.5	2.7
cal. ANN (par.)	3.0		4.6	4.6	4.6	4.6	4.6	6.8	4.3	4.1	4.9	4.6	6.3	5.3	3.0
Logit (fix)	2.8		4.5	4.6	4.5	4.5	4.1	6.4	3.9	4.1	7.5	4.6	6.5	5.7	2.8
Pseudo-Weights (fix)	2.7		4.7	4.7	4.7	4.7	4.7	6.4	8.0	4.9	7.5	4.7	6.3	5.6	2.7
cal. ANN (fix)	3.0		4.6	4.6	4.6	4.6	4.6	6.8	4.3	4.0	4.9	4.6	6.3	5.3	3.0
cal. ANN (opt.)	3.0		4.6	4.6	4.6	4.6	4.6	6.8	4.3	4.0	4.9	4.6	6.3	5.3	3.0
Calibration benchmarks: Totals															
Logit (par.) and GREG	3.0		4.4	8.0	8.0	8.0	4.4	8.0	8.0	8.0	8.0	4.4	7.7	8.0	3.0
cal. ANN (par.)	4.5		6.7	5.4	6.3	6.2	6.6	7.3	8.0	5.5	3.9	8.0	8.0	8.0	4.5
Logit (fix) and GREG	2.8		4.5	8.0	8.0	8.0	4.3	8.0	8.0	8.0	8.0	4.6	6.3	8.0	2.8
cal. ANN (fix)	4.4		6.6	5.3	6.2	6.2	6.6	7.3	8.0	5.7	3.6	8.0	8.0	8.0	4.4
cal. ANN (opt.)	3.0		4.6	4.6	4.6	4.6	4.6	6.8	4.3	4.7	5.0	4.6	6.3	5.0	3.0
Calibration benchmarks: Covariances															
Logit (par.) and GREG	2.9		4.7	8.0	8.0	8.0	4.7	8.0	8.0	8.0	8.0	4.7	6.5	8.0	2.9
cal. ANN (par.)	3.2		4.4	4.4	4.4	4.4	4.7	7.3	3.9	4.3	5.3	4.4	6.5	8.0	3.2
Logit (fix) and GREG	2.8		4.4	8.0	8.0	8.0	4.5	8.0	8.0	8.0	8.0	4.4	6.3	8.0	2.8
cal. ANN (fix)	2.7		4.5	4.9	4.5	4.5	4.3	6.5	6.5	3.8	5.1	4.5	7.7	8.0	2.7
cal. ANN (opt.)	3.1		4.6	4.7	4.6	4.6	4.6	7.3	4.4	3.5	5.0	4.6	6.3	8.0	3.1
Calibration benchmarks: Totals and covariances															
Logit (par.) and GREG	3.1		4.5	8.0	8.0	8.0	4.5	8.0	8.0	8.0	8.0	4.5	6.5	8.0	3.1
cal. ANN (par.)	4.5		6.7	5.5	6.3	6.2	6.6	7.3	8.0	5.6	3.9	8.0	8.0	8.0	4.5
Logit (fix) and GREG	2.9		4.6	8.0	8.0	8.0	4.4	8.0	8.0	8.0	8.0	4.7	6.2	8.0	2.9
cal. ANN (fix)	4.4		6.6	5.3	6.3	6.2	6.6	7.3	8.0	8.0	3.6	8.0	8.0	8.0	4.4
cal. ANN (opt.)	3.1		4.7	4.7	4.7	4.7	4.7	7.3	4.5	4.5	5.0	4.7	6.3	5.1	3.1
Plain model-based estimates	2.8	3.7	4.6	4.7	4.6	4.6	4.6	6.8	5.7	4.0	4.4	4.6	6.5	5.4	

Highest mean absolute error

Lowest mean absolute error

Logit: Weights from GLM with logit link
cal. ANN: calibrated ANN
Logit and GREG: Weights from GLM with logit link, calibrated using the GREG

(par.): parametric
(fix): non-parametric, fixed knots
(opt.): non-parametric, optimized knots

Plain model- / design-based estimates: Values in the last row and column are the same as in the first, but the coloring respectively concerns exclusively model- and pseudo-design-based estimates. All other colors in the table concern the comparison of all possible cross-combinations of weighting and prediction.

Table D.5: Income class frequencies (in percentage points) estimated by weighted aggregation of predictions in the WI

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: unweighted, using no auxiliary information																									
None (weighted WI)	0.0	1.0	0.6	2.6	2.4	4.8	8.2	10.1	11.7	11.2	13.6	10.9	6.6	3.9	3.4	2.9	1.9	1.6	0.6	0.5	0.2	0.2	0.1	0.1	0.8
GLM (OLS)	0.1	0.3	0.4	1.1	1.9	2.3	3.2	5.8	8.4	8.7	11.7	16.1	15.1	9.3	6.3	5.0	1.5	0.8	0.4	0.6	0.6	0.3	0.0	0.0	0.0
GLM (Ridge)	0.1	0.1	0.4	1.0	2.0	1.9	3.2	5.9	8.6	8.3	12.5	16.6	14.9	8.9	6.2	5.3	1.1	0.8	0.4	0.7	0.7	0.2	0.0	0.0	0.0
GLM (LASSO)	0.1	0.2	0.4	1.0	2.0	2.3	3.0	5.9	8.6	8.3	12.5	15.7	15.1	9.3	6.3	5.3	1.2	0.8	0.4	0.6	0.7	0.3	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.2	0.4	1.0	2.0	2.3	3.0	5.9	8.6	8.3	12.5	15.7	15.1	9.3	6.3	5.3	1.2	0.8	0.4	0.6	0.7	0.3	0.0	0.0	0.0
GAM (fix knots)	0.1	0.3	0.4	1.1	1.9	2.3	3.2	5.8	8.4	8.7	11.7	16.1	15.1	9.3	6.3	5.0	1.5	0.8	0.4	0.6	0.6	0.3	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	35.4	0.0	1.1	0.0	61.8	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.7	0.0
MARS	0.2	0.1	0.2	0.5	1.1	2.5	2.4	2.8	3.2	27.8	3.4	6.1	49.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.2	0.3	0.1
ANN (opt. knots)	0.1	0.4	0.4	1.2	2.1	1.8	3.7	5.0	11.5	5.9	13.8	15.8	13.4	8.2	9.5	2.8	1.4	1.2	0.5	0.5	0.9	0.2	0.0	0.0	0.0
SVM	0.0	0.0	0.2	0.2	0.5	1.4	4.7	15.8	9.6	9.3	32.5	7.6	7.0	10.2	0.4	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.3	0.4	1.1	1.9	2.3	3.2	5.8	8.4	8.7	11.7	16.1	15.1	9.3	6.3	5.0	1.5	0.8	0.4	0.6	0.6	0.3	0.0	0.0	0.0
GAMM (fix knots)	0.1	0.3	0.4	1.1	1.9	2.3	3.2	5.8	8.4	8.7	11.7	16.1	15.1	9.3	6.3	5.0	1.5	0.8	0.4	0.6	0.6	0.3	0.0	0.0	0.0
Heckman model	31.8	3.3	4.9	7.8	8.5	8.1	3.8	2.7	5.5	8.1	9.5	5.6	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (parametric), using propensity weights																									
None (weighted WI)	0.0	1.0	0.2	2.4	2.7	5.5	8.8	9.4	12.0	11.7	13.0	11.5	6.9	4.2	3.0	2.9	1.5	1.4	0.3	0.2	0.1	0.3	0.1	0.1	0.7
GLM (OLS)	0.0	0.0	0.0	0.4	1.4	1.4	2.4	6.9	9.6	7.5	12.0	18.3	15.7	13.3	5.8	4.2	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.4	1.4	1.1	2.6	6.9	9.6	7.3	12.5	19.1	15.2	12.9	5.8	4.5	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.4	1.4	1.4	2.4	6.9	9.6	7.4	12.5	18.0	15.7	13.3	5.8	4.3	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.4	1.4	1.4	2.4	6.9	9.6	7.4	12.5	18.0	15.7	13.3	5.8	4.3	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.4	1.4	1.4	2.4	6.9	9.6	7.5	12.0	18.3	15.7	13.3	5.8	4.2	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.5	0.0	4.0	0.0	54.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0
MARS	0.0	0.0	0.0	0.1	0.7	3.0	2.8	2.4	2.7	35.2	2.8	6.3	43.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.7	1.5	1.3	3.0	3.4	14.6	4.9	16.3	13.0	14.8	17.1	5.0	3.4	0.7	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.2	1.9	19.9	16.6	7.2	27.2	11.3	11.7	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.4	1.4	1.4	2.4	6.9	9.6	7.5	12.0	18.3	15.7	13.3	5.8	4.2	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.4	1.4	1.4	2.4	6.9	9.6	7.5	12.0	18.3	15.7	13.3	5.8	4.2	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	65.3	2.9	2.1	3.9	6.0	8.8	1.8	0.2	1.3	2.2	3.3	1.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Pseudo-Weights (parametric), using propensity weights																									
None (weighted WI)	0.0	1.0	0.4	2.5	2.7	5.5	9.1	10.3	12.3	11.1	13.1	10.6	6.6	3.8	3.1	2.6	1.7	1.5	0.5	0.3	0.2	0.2	0.1	0.1	0.7
GLM (OLS)	0.0	0.1	0.2	0.9	1.8	2.3	3.3	6.4	9.2	8.7	12.5	15.9	14.8	10.3	6.2	5.1	1.5	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.1	0.2	0.9	1.9	1.9	3.4	6.5	9.3	8.6	13.2	16.5	14.5	9.8	6.1	5.4	1.2	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.3	0.8	1.9	2.3	3.2	6.5	9.2	8.5	13.2	15.4	14.8	10.3	6.3	5.3	1.3	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.3	0.8	1.9	2.3	3.2	6.5	9.2	8.5	13.2	15.4	14.8	10.3	6.3	5.3	1.3	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.2	0.9	1.8	2.3	3.3	6.4	9.2	8.7	12.5	15.9	14.8	10.3	6.2	5.1	1.5	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	42.1	0.0	1.9	0.0	55.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0
MARS	0.1	0.1	0.2	0.4	0.9	2.3	2.1	2.2	2.6	36.3	2.4	4.4	46.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.2	0.3	1.1	2.1	1.9	3.5	4.5	14.0	6.0	14.4	14.2	13.8	10.5	7.1	3.8	1.5	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.1	0.1	0.3	0.9	3.7	19.6	13.6	7.7	30.7	8.2	8.5	6.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.2	0.9	1.8	2.3	3.3	6.4	9.2	8.7	12.5	15.9	14.8	10.3	6.2	5.1	1.5	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.1	0.2	0.9	1.8	2.3	3.3	6.4	9.2	8.7	12.5	15.9	14.8	10.3	6.2	5.1	1.5	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	48.2	3.4	3.6	5.8	7.9	9.4	4.5	1.2	2.8	4.3	4.9	3.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using propensity weights																									
None (weighted WI)	0.0	0.9	0.4	1.9	1.8	4.3	7.3	9.6	11.7	11.1	14.3	11.9	7.3	4.0	3.8	3.1	2.2	1.8	0.6	0.4	0.2	0.2	0.1	0.1	0.8
GLM (OLS)	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.5	1.5	1.1	1.7	4.6	9.0	8.3	10.9	22.2	18.5	9.1	7.6	4.2	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.5	1.5	1.3	1.5	4.6	9.0	8.4	11.3	20.5	19.1	9.7	7.6	4.1	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.5	1.5	1.3	1.5	4.6	9.0	8.4	11.3	20.5	19.1	9.7	7.6	4.1	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.7	0.0	0.2	0.0	71.8	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.5	1.5	1.5	1.6	2.4	22.2	2.7	6.9	60.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.1	0.8	1.5	0.9	1.7	2.7	13.5	6.0	12.5	17.4	20.9	7.7	11.9	1.2	0.7	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	2.3	15.8	6.8	7.7	41.3	6.9	5.8	12.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Heckman model	23.8	1.2	3.4	7.8	10.6	14.3	2.9	1.3	5.7	9.3	13.2	6.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots), using propensity weights																									
None (weighted WI)	0.0	1.4	0.4	2.7	2.6	4.6	8.4	8.8	12.0	11.6	13.2	11.8	7.6	4.3	2.8	2.9	1.5	1.4	0.3	0.2	0.2	0.3	0.0	0.1	0.7
GLM (OLS)	0.1	0.1	0.4	0.9	1.7	1.7	3.7	6.9	7.0	6.8	11.4	16.9	16.6	11.7	5.7	6.6	1.7	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.1	0.0	0.3	1.0	1.8	1.5	3.8	6.9	7.1	6.5	11.8	17.2	16.4	11.5	5.7	7.0	1.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.0	0.4	0.9	1.8	1.7	3.6	6.9	7.1	6.5	11.8	16.7	16.6	11.7	5.7	6.7	1.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.0	0.4	0.9	1.8	1.7	3.6	6.9	7.1	6.5	11.8	16.7	16.6	11.7	5.7	6.7	1.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.1	0.4	0.9	1.7	1.7	3.7	6.9	7.0	6.8	11.4	16.9	16.6	11.7	5.7	6.6	1.7	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.7	0.0	6.8	0.0	55.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0
MARS	0.6	0.2	0.3	0.6	2.1	5.5	4.9	4.1	3.9	26.5	4.5	8.5	38.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.2	0.2	0.2	1.3	1.6	2.6	4.3	4.3	9.8	4.7	15.5	13.0	14.1	15.1	5.4	5.8	1.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.2	0.1	0.3	0.4	0.7	1.6	3.8	15.0	15.2	10.1	24.2	11.9	11.6	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	0.4	0.9	1.7	1.7	3.7	6.9	7.0	6.8	11.4	16.9	16.6	11.7	5.7	6.6	1.7	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.1	0.1	0.4	0.9	1.7	1.7	3.7	6.9	7.0	6.8	11.4	16.9	16.6	11.7	5.7	6.6	1.7	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	64.8	5.0	4.1	4.4	4.2	4.6	1.5	0.3	1.4	3.0	4.2	2.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.5: Income class frequencies (in percentage points) estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Pseudo-Weights (fix knots), using propensity weights																									
None (weighted WI)	0.0	1.1	0.4	2.6	2.7	5.5	9.8	11.4	13.5	11.2	13.2	10.1	6.1	3.4	2.6	2.1	1.4	1.1	0.4	0.2	0.1	0.2	0.1	0.1	0.6
GLM (OLS)	0.0	0.2	0.2	1.1	3.1	3.6	4.1	7.3	9.7	9.9	12.4	14.8	13.6	8.9	4.7	4.4	1.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.2	1.1	3.2	3.0	4.5	7.3	9.9	9.7	13.2	15.2	13.3	8.6	4.6	4.7	1.2	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.2	1.1	3.2	3.6	4.0	7.3	9.8	9.6	13.2	14.3	13.6	8.9	4.7	4.6	1.3	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.2	1.1	3.2	3.6	4.0	7.3	9.8	9.6	13.2	14.3	13.6	8.9	4.7	4.6	1.3	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.2	0.2	1.1	3.1	3.6	4.1	7.3	9.7	9.9	12.4	14.8	13.6	8.9	4.7	4.4	1.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	43.4	0.0	2.4	0.0	53.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0
MARS	0.0	0.0	0.1	0.3	1.2	3.4	3.2	3.2	3.6	34.1	3.7	6.6	40.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.2	0.3	1.6	3.6	2.7	4.2	5.4	14.6	7.1	13.5	13.3	13.1	8.8	5.2	3.9	1.4	0.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.7	4.1	22.8	14.1	10.2	27.9	8.5	6.9	4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.2	0.2	1.1	3.1	3.6	4.1	7.3	9.7	9.9	12.4	14.8	13.6	8.9	4.7	4.4	1.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.2	0.2	1.1	3.1	3.6	4.1	7.3	9.7	9.9	12.4	14.8	13.6	8.9	4.7	4.4	1.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	51.2	3.4	3.8	7.9	7.8	7.6	4.2	1.5	2.0	3.7	4.1	2.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using propensity weights																									
None (weighted WI)	0.0	0.9	0.4	1.9	1.8	4.3	7.3	9.6	11.7	11.1	14.3	11.9	7.3	4.0	3.8	3.1	2.2	1.8	0.6	0.4	0.2	0.2	0.1	0.1	0.8
GLM (OLS)	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.5	1.5	1.1	1.7	4.6	9.0	8.3	10.9	22.2	18.5	9.1	7.6	4.2	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.5	1.5	1.3	1.5	4.6	9.0	8.4	11.3	20.5	19.1	9.7	7.6	4.1	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.5	1.5	1.3	1.5	4.6	9.0	8.4	11.3	20.5	19.1	9.7	7.6	4.1	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.7	0.0	0.2	0.0	71.8	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.5	1.5	1.5	1.6	2.4	22.2	2.7	6.9	60.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.1	0.8	1.5	0.9	1.7	2.7	13.5	6.0	12.5	17.4	20.9	7.7	11.9	1.2	0.7	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	2.3	15.8	6.8	7.7	41.3	6.9	5.8	12.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Heckman model	23.8	1.2	3.4	7.8	10.6	14.3	2.9	1.3	5.7	9.3	13.2	6.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using propensity weights																									
None (weighted WI)	0.0	0.9	0.4	1.9	1.8	4.3	7.3	9.6	11.7	11.1	14.3	11.9	7.3	4.0	3.8	3.1	2.2	1.8	0.6	0.4	0.2	0.2	0.1	0.1	0.8
GLM (OLS)	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.5	1.5	1.1	1.7	4.6	9.0	8.3	10.9	22.2	18.5	9.1	7.6	4.2	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.5	1.5	1.3	1.5	4.6	9.0	8.4	11.3	20.5	19.1	9.7	7.6	4.1	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.5	1.5	1.3	1.5	4.6	9.0	8.4	11.3	20.5	19.1	9.7	7.6	4.1	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.7	0.0	0.2	0.0	71.8	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.5	1.5	1.5	1.6	2.4	22.2	2.7	6.9	60.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.1	0.8	1.5	0.9	1.7	2.7	13.5	6.0	12.5	17.4	20.9	7.7	11.9	1.2	0.7	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	2.3	15.8	6.8	7.7	41.3	6.9	5.8	12.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.1	0.1	0.5	1.5	1.3	1.6	4.6	8.9	8.7	10.4	21.1	19.1	9.7	7.6	3.9	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Heckman model	23.8	1.2	3.4	7.8	10.6	14.3	2.9	1.3	5.7	9.3	13.2	6.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Subsampling, using total calibration																									
None (weighted WI)	0.0	1.2	0.0	1.4	1.7	4.8	5.7	7.2	11.5	11.7	13.4	14.3	9.5	4.5	3.1	3.6	1.9	1.9	0.5	0.0	0.2	0.2	0.0	0.2	1.4
GLM (OLS)	0.0	0.2	0.1	0.5	0.9	1.4	3.2	5.2	4.3	5.2	10.5	14.9	15.8	12.7	7.9	11.6	3.5	1.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.1	0.1	0.5	0.9	1.2	3.1	5.4	4.1	5.2	11.0	14.9	15.5	12.7	8.3	12.1	2.7	1.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.1	0.5	0.9	1.4	2.9	5.4	4.3	5.1	11.0	14.6	15.8	12.7	8.1	11.5	3.5	1.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.1	0.5	0.9	1.4	2.9	5.4	4.3	5.1	11.0	14.6	15.8	12.7	8.1	11.5	3.5	1.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.2	0.1	0.5	0.9	1.4	3.2	5.2	4.3	5.2	10.5	14.9	15.8	12.7	7.9	11.6	3.5	1.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.3	0.0	8.7	0.0	61.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9	0.0
MARS	0.1	0.0	0.2	0.7	1.5	4.0	3.3	3.4	4.0	25.6	4.8	8.6	43.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.1	0.1	0.5	1.2	2.0	3.3	3.8	5.4	4.1	14.5	11.4	12.8	17.9	5.6	10.8	3.2	2.7	0.2	0.1	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.2	0.1	0.2	1.4	3.2	9.7	14.4	9.0	25.3	15.8	17.5	3.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.2	0.1	0.5	0.9	1.4	3.2	5.2	4.3	5.2	10.5	14.9	15.8	12.7	7.9	11.6	3.5	1.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.2	0.1	0.5	0.9	1.4	3.2	5.2	4.3	5.2	10.5	14.9	15.8	12.7	7.9	11.6	3.5	1.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	67.4	8.3	3.8	4.3	4.4	4.3	1.7	0.5	0.9	1.5	1.5	1.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Post-stratification, using total calibration																									
None (weighted WI)	0.0	1.1	0.4	2.5	2.6	4.8	8.9	9.7	12.7	11.3	13.0	10.9	7.1	4.3	3.0	2.9	1.6	1.5	0.3	0.2	0.1	0.3	0.0	0.1	0.8
GLM (OLS)	0.0	0.0	0.2	0.9	1.9	2.2	3.3	6.3	8.0	8.1	12.7	17.1	14.4	12.1	6.6	4.6	1.1	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.2	0.9	1.9	1.9	3.6	6.4	8.0	7.9	13.4	17.6	13.9	11.6	6.5	4.8	0.8	0.4	0.1	0.1	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.2	0.9	1.9	2.2	3.3	6.3	8.0	7.9	13.3	16.6	14.4	12.1	6.6	4.6	1.0	0.4	0.1	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.2	0.9	1.9	2.2	3.3	6.3	8.0	7.9	13.3	16.6	14.4	12.1	6.6	4.6	1.0	0.4	0.1	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.2	0.9	1.9	2.2	3.3	6.3	8.0	8.1	12.7	17.1	14.4	12.1	6.6	4.6	1.1	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.7	0.0	1.6	0.0	60.3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.4	0.0
MARS	0.1	0.1	0.2	0.5	1.3	2.8	2.8	2.9	3.3	29.8	4.1	6.9	45.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.2	1.2	2.2	2.2	3.4	4.4	11.6	5.9	15.1	13.8	13.0	15.6	5.4	3.6	1.3	0.7	0.1	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.1	0.1	0.4	1.3	3.2	17.3	15.7	8.6	24.9	12.0	11.8	4.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.2	0.9	1.9	2.2	3.3	6.3	8.0	8.1	12.7	17.1	14.4	12.1	6.6	4.6	1.1	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.2	0.9	1.9	2.2	3.3	6.3	8.0	8.1	12.7	17.1	14.4	12.1	6.6	4.6	1.1	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0
Heckman model	61.7	2.9	3.0	6.2	7.3	8.1	3.2	0.4	0.5	1.7	3.1	2.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.5: Income class frequencies (in percentage points) estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using total calibration																									
None (weighted WI)	0.0	0.9	0.0	1.6	2.4	5.3	8.4	7.1	9.4	11.1	11.6	12.0	8.0	5.7	3.9	4.1	2.8	2.9	0.7	0.5	0.2	0.4	0.2	0.1	0.7
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.1	1.4	5.5	7.4	2.0	7.3	20.9	16.8	16.2	10.0	10.0	2.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	1.4	5.7	7.4	2.0	7.2	22.3	16.6	15.0	10.0	10.8	1.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.1	1.2	5.8	7.2	2.0	7.4	20.8	16.8	16.2	10.0	10.6	1.6	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.1	1.2	5.8	7.2	2.0	7.4	20.8	16.8	16.2	10.0	10.6	1.6	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.1	1.4	5.5	7.4	2.0	7.3	20.9	16.8	16.2	10.0	10.0	2.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	43.1	0.0	2.8	0.0	53.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	40.6	0.0	0.0	59.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.2	1.8	0.4	14.3	0.7	11.4	15.7	15.3	19.9	11.6	5.8	2.3	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	19.3	8.6	0.0	37.5	9.9	12.5	12.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.1	1.4	5.5	7.4	2.0	7.3	20.9	16.8	16.2	10.0	10.0	2.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.1	1.4	5.5	7.4	2.0	7.3	20.9	16.8	16.2	10.0	10.0	2.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	61.4	3.8	0.0	0.0	5.5	15.6	4.7	0.0	0.0	0.0	0.0	8.9	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using total calibration																									
None (weighted WI)	0.0	0.9	0.3	1.9	1.9	4.5	7.8	10.3	12.5	11.6	14.4	11.6	7.0	3.7	3.4	2.7	1.9	1.5	0.5	0.3	0.2	0.2	0.1	0.1	0.7
GLM (OLS)	0.0	0.1	0.1	0.6	1.7	1.5	1.8	5.1	9.6	8.9	11.2	22.6	18.6	8.7	5.7	3.0	0.5	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.5	1.7	1.2	2.0	5.1	9.7	8.6	11.6	23.9	18.0	8.1	5.7	3.1	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.5	1.7	1.5	1.7	5.2	9.6	8.7	12.1	22.0	18.6	8.7	5.7	3.1	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.5	1.7	1.5	1.7	5.2	9.6	8.7	12.1	22.0	18.6	8.7	5.7	3.1	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	0.6	1.7	1.5	1.8	5.1	9.6	8.9	11.2	22.6	18.6	8.7	5.7	3.0	0.5	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.8	0.0	0.3	0.0	71.7	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.6	1.6	1.4	1.5	2.1	22.8	2.7	6.5	60.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.1	0.9	1.8	1.0	2.0	3.0	14.4	6.3	13.6	17.3	21.6	7.3	8.7	1.0	0.5	0.2	0.1	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	2.5	17.6	6.7	7.7	44.6	6.2	5.1	9.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	0.6	1.7	1.5	1.8	5.1	9.6	8.9	11.2	22.6	18.6	8.7	5.7	3.0	0.5	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.1	0.1	0.6	1.7	1.5	1.8	5.1	9.6	8.9	11.2	22.6	18.6	8.7	5.7	3.0	0.5	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Heckman model	27.2	1.3	2.9	7.6	10.8	16.1	3.3	1.4	6.4	8.7	9.6	4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using total calibration																									
None (weighted WI)	0.0	0.9	0.2	2.0	2.2	4.9	8.0	8.7	11.2	11.9	13.4	12.1	7.7	4.8	3.4	3.2	1.8	1.7	0.4	0.3	0.1	0.4	0.1	0.1	0.8
GLM (OLS)	0.0	0.0	0.0	0.2	0.7	0.9	1.7	5.1	8.0	6.3	11.8	20.1	16.9	14.8	7.3	4.9	0.8	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.2	0.7	0.7	1.7	5.1	8.1	6.2	11.8	21.0	16.7	14.3	7.3	5.2	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.2	0.7	0.9	1.5	5.2	8.0	6.2	12.2	19.8	16.9	14.8	7.4	5.0	0.7	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.2	0.7	0.9	1.5	5.2	8.0	6.2	12.2	19.8	16.9	14.8	7.4	5.0	0.7	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.2	0.7	0.9	1.7	5.1	8.0	6.3	11.8	20.1	16.9	14.8	7.3	4.9	0.8	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.6	0.0	2.6	0.0	63.3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.4	0.0
MARS	0.0	0.0	0.0	0.0	0.3	1.5	1.5	1.4	1.5	31.0	1.9	4.6	56.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.3	0.8	0.7	2.3	2.6	12.0	4.0	17.8	13.5	15.6	19.0	6.1	3.8	1.0	0.4	0.1	0.0	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	1.6	15.5	13.5	5.4	33.7	11.1	14.2	4.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.2	0.7	0.9	1.7	5.1	8.0	6.3	11.8	20.1	16.9	14.8	7.3	4.9	0.8	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.2	0.7	0.9	1.7	5.1	8.0	6.3	11.8	20.1	16.9	14.8	7.3	4.9	0.8	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	56.7	3.6	1.9	2.7	5.6	10.4	2.8	0.6	3.9	5.6	3.6	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using total calibration																									
None (weighted WI)	0.0	0.9	0.1	2.1	2.3	5.0	8.1	8.6	11.0	11.6	13.2	12.0	7.6	4.8	3.5	3.2	1.8	1.8	0.4	0.3	0.1	0.4	0.1	0.1	0.8
GLM (OLS)	0.0	0.0	0.0	0.2	0.6	0.8	1.5	4.8	8.2	6.4	11.7	18.8	16.9	14.8	8.0	5.7	1.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.2	0.7	0.6	1.5	4.9	8.3	6.3	11.7	19.7	16.7	14.3	7.9	6.0	0.7	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.2	0.7	0.8	1.4	4.9	8.2	6.3	12.0	18.6	16.9	14.8	8.0	5.8	0.9	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.2	0.7	0.8	1.4	4.9	8.2	6.3	12.0	18.6	16.9	14.8	8.0	5.8	0.9	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.2	0.6	0.8	1.5	4.8	8.2	6.4	11.7	18.8	16.9	14.8	8.0	5.7	1.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	35.9	0.0	2.7	0.0	60.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0
MARS	0.0	0.0	0.0	0.0	0.3	1.5	1.4	1.3	1.4	33.7	1.6	3.7	54.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.3	0.7	0.6	2.1	2.3	12.7	4.0	17.3	12.6	15.6	18.6	6.8	4.4	1.3	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	1.4	15.9	14.6	5.1	31.9	10.4	14.8	5.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.2	0.6	0.8	1.5	4.8	8.2	6.4	11.7	18.8	16.9	14.8	8.0	5.7	1.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.2	0.6	0.8	1.5	4.8	8.2	6.4	11.7	18.8	16.9	14.8	8.0	5.7	1.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	57.9	3.9	1.9	2.4	5.2	10.4	3.4	0.4	3.0	5.1	3.5	2.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using total calibration																									
None (weighted WI)	0.0	0.8	0.2	1.6	1.7	4.1	7.2	9.0	11.1	11.1	13.8	11.8	7.5	4.3	4.2	3.3	2.7	2.5	0.8	0.5	0.3	0.2	0.2	0.2	0.8
GLM (OLS)	0.0	0.0	0.0	0.2	0.6	0.5	0.9	3.6	8.3	6.0	8.5	23.3	20.9	9.2	10.0	6.6	1.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.2	0.6	0.4	0.8	3.7	8.5	5.8	8.2	25.3	20.6	8.2	9.9	6.9	0.6	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.2	0.6	0.5	0.8	3.7	8.4	5.9	8.9	23.0	20.9	9.2	10.0	6.9	0.6	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.2	0.6	0.5	0.8	3.7	8.4	5.9	8.9	23.0	20.9	9.2	10.0	6.9	0.6	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.2	0.6	0.5	0.9	3.6	8.3	6.0	8.5	23.3	20.9	9.2	10.0	6.6	1.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.9	0.0	0.1	0.0	73.8	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.3	0.6	0.6	0.6	0.8	24.0	1.0	2.1	69.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.3	0.6	0.3	1.4	1.7	12.8	3.6	13.3	17.9	22.4	6.8	15.5	1.8	1.1	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.2	1.6	14.6	5.2	2.7	49.3	4.2	4.9	17.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.2	0.6	0.5	0.9	3.6	8.3	6.0	8.5	23.3	20.9	9.2	10.0	6.6	1.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.2	0.6	0.5	0.9	3.6	8.3	6.0	8.5	23.3	20.9	9.2	10.0	6.6	1.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
Heckman model	23.0	1.4	1.9	2.9	8.4	19.8	4.8	0.5	6.8	10.3	9.4	10.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.5: Income class frequencies (in percentage points) estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using covariance calibration																									
None (weighted WI)	0.0	0.9	0.3	1.8	2.0	4.6	7.5	9.9	11.4	11.6	14.2	11.5	7.1	4.0	3.6	2.9	2.3	1.9	0.6	0.4	0.2	0.2	0.2	0.2	0.7
GLM (OLS)	0.0	0.0	0.0	0.5	1.4	1.3	1.6	4.9	9.3	7.7	11.1	22.5	17.4	8.5	7.2	5.1	1.1	0.1	0.0	0.1	0.2	0.1	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.4	1.4	1.0	1.7	4.8	9.4	7.3	11.0	23.6	17.5	7.9	7.1	5.6	0.7	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.4	1.4	1.3	1.5	4.9	9.4	7.4	11.8	22.1	17.4	8.5	7.2	5.5	0.8	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.4	1.4	1.3	1.5	4.9	9.4	7.4	11.8	22.1	17.4	8.5	7.2	5.5	0.8	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.5	1.4	1.3	1.6	4.9	9.3	7.7	11.1	22.5	17.4	8.5	7.2	5.1	1.1	0.1	0.0	0.1	0.2	0.1	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.2	0.0	0.0	0.0	68.6	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.4	1.0	1.1	1.1	1.7	25.9	1.2	3.3	63.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0
ANN (opt. knots)	0.0	0.0	0.1	0.7	1.5	0.8	1.7	2.3	14.6	4.8	16.4	17.6	17.3	7.1	11.3	2.0	1.2	0.3	0.1	0.1	0.2	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.2	2.1	17.8	5.6	4.5	46.4	5.3	5.3	12.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.5	1.4	1.3	1.6	4.9	9.3	7.7	11.1	22.5	17.4	8.5	7.2	5.1	1.1	0.1	0.0	0.1	0.2	0.1	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.5	1.4	1.3	1.6	4.9	9.3	7.7	11.1	22.5	17.4	8.5	7.2	5.1	1.1	0.1	0.0	0.1	0.2	0.1	0.0	0.0	0.0
Heckman model	25.2	1.8	3.1	4.9	8.4	13.3	3.6	1.4	8.9	12.1	9.7	7.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using covariance calibration																									
None (weighted WI)	0.0	0.9	0.4	2.3	2.1	4.5	7.6	9.6	12.1	11.4	14.3	12.0	7.0	4.1	3.4	2.9	1.9	1.5	0.5	0.3	0.2	0.2	0.1	0.1	0.7
GLM (OLS)	0.0	0.1	0.1	0.4	1.1	0.9	1.5	4.7	10.5	10.0	10.3	19.1	20.0	10.5	5.7	3.9	0.8	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.4	1.1	0.7	1.4	4.7	10.7	9.7	10.8	19.9	19.7	10.0	5.7	4.2	0.5	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.4	1.1	0.9	1.4	4.7	10.6	9.7	11.0	18.7	20.0	10.5	5.7	4.1	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.4	1.1	0.9	1.4	4.7	10.6	9.7	11.0	18.7	20.0	10.5	5.7	4.1	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	0.4	1.1	0.9	1.5	4.7	10.5	10.0	10.3	19.1	20.0	10.5	5.7	3.9	0.8	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	29.5	0.0	0.9	0.0	69.4	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.5	1.6	1.6	1.9	2.7	26.0	4.7	9.9	50.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.1	0.6	1.1	0.7	2.0	3.3	15.2	6.8	12.1	17.5	18.8	10.1	8.1	2.2	0.8	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.1	0.0	0.2	0.5	2.6	16.5	10.1	9.2	35.2	9.5	7.6	8.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	0.4	1.1	0.9	1.5	4.7	10.5	10.0	10.3	19.1	20.0	10.5	5.7	3.9	0.8	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.1	0.1	0.4	1.1	0.9	1.5	4.7	10.5	10.0	10.3	19.1	20.0	10.5	5.7	3.9	0.8	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Heckman model	31.9	2.3	4.1	8.7	8.8	10.5	3.0	1.2	3.9	10.3	10.9	4.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using covariance calibration																									
None (weighted WI)	0.0	0.9	0.2	1.4	0.8	2.2	6.2	7.8	11.5	13.1	15.0	13.9	9.1	5.6	3.3	3.6	1.9	1.8	0.4	0.2	0.1	0.3	0.1	0.1	0.8
GLM (OLS)	0.0	0.0	0.0	0.1	0.4	0.4	0.6	1.8	4.2	5.6	10.9	27.3	18.6	16.0	6.0	6.7	1.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.4	0.3	0.6	1.7	4.2	5.4	11.0	28.3	18.2	15.5	6.0	7.2	0.6	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.4	0.4	0.6	1.7	4.3	5.5	11.5	26.7	18.6	16.0	6.0	6.8	1.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.4	0.4	0.6	1.7	4.3	5.5	11.5	26.7	18.6	16.0	6.0	6.8	1.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.4	0.4	0.6	1.8	4.2	5.6	10.9	27.3	18.6	16.0	6.0	6.7	1.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.9	0.0	0.1	0.0	91.9	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.2	0.4	0.5	0.8	1.2	7.4	6.0	12.1	71.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.2	0.5	0.3	0.7	1.6	5.5	4.1	15.5	20.0	17.7	21.4	5.1	5.7	1.1	0.4	0.0	0.0	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.2	0.9	5.4	3.6	8.2	46.2	16.6	14.6	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.4	0.4	0.6	1.8	4.2	5.6	10.9	27.3	18.6	16.0	6.0	6.7	1.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.4	0.4	0.6	1.8	4.2	5.6	10.9	27.3	18.6	16.0	6.0	6.7	1.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	40.5	5.1	2.4	5.9	8.4	13.0	2.7	1.6	6.9	7.2	4.4	1.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using covariance calibration																									
None (weighted WI)	0.0	15.2	1.7	12.4	9.7	2.4	3.6	5.1	16.9	4.1	19.4	3.5	1.9	1.0	0.9	0.7	0.5	0.4	0.1	0.1	0.0	0.0	0.0	0.1	0.2
GLM (OLS)	0.1	0.0	0.0	0.0	61.2	36.8	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.1	0.0	0.0	0.0	61.2	36.8	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.0	0.0	0.0	61.2	36.8	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.0	0.0	0.0	61.2	36.8	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.0	0.0	0.0	61.2	36.8	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.1	0.0	0.0	0.0	62.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	37.1	0.0	0.0	0.0	62.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.0	0.0	0.0	0.0	61.2	0.0	36.8	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	38.5	61.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.0	0.0	0.0	61.2	36.8	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.1	0.0	0.0	0.0	61.2	36.8	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	37.1	0.0	0.0	0.0	0.0	61.2	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using covariance calibration																									
None (weighted WI)	0.0	0.8	0.2	1.2	1.7	4.1	6.5	9.0	10.4	11.6	14.6	12.4	8.4	4.7	4.3	4.4	2.2	1.3	0.4	0.4	0.2	0.3	0.1	0.1	0.7
GLM (OLS)	0.0	0.0	0.0	0.2	0.5	1.2	3.8	5.9	5.3	6.8	18.8	18.4	13.7	18.4	5.0	1.2	0.3	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.5	0.8	4.2	5.9	5.1	5.8	21.5	17.8	13.1	18.2	5.0	1.3	0.2	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.5	1.2	3.8	5.9	5.3	5.9	21.3	16.8	13.7	18.4	5.0	1.2	0.2	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.5	1.2	3.8	5.9	5.3	5.9	21.3	16.8	13.7	18.4	5.0	1.2	0.2	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.2	0.5	1.2	3.8	5.9	5.3	6.8	18.8	18.4	13.7	18.4	5.0	1.2	0.3	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.3	0.0	0.1	0.0	72.3	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.3	0.4	0.4	1.0	25.6	0.6	2.5	68.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.2	0.5	0.9	2.4	3.4	9.5	4.6	16.4	18.3	17.2	16.1	9.5	0.4	0.3	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	1.1	14.8	3.2	10.8	36.5	10.0	13.7	9.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.2	0.5	1.2	3.8	5.9	5.3	6.8	18.8	18.4	13.7	18.4	5.0	1.2	0.3	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.2	0.5	1.2	3.8	5.9	5.3	6.8	18.8	18.4	13.7	18.4	5.0	1.2	0.3	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
Heckman model	22.9	0.5	1.1	8.6	19.8	8.3	0.7	2.9	4.8	2.8	25.3	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.5: Income class frequencies (in percentage points) estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model and total calibration																									
None (weighted WI)	0.0	0.7	0.0	1.8	2.2	5.5	7.6	7.6	9.8	11.1	12.7	12.5	8.2	5.3	4.2	3.8	2.0	2.2	0.4	0.4	0.2	0.5	0.1	0.1	0.9
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	3.1	8.6	6.3	12.0	17.5	16.5	18.2	10.2	5.9	1.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	3.1	8.7	6.3	12.2	18.8	16.0	17.4	10.2	6.3	0.7	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	8.6	6.3	12.3	17.3	16.5	18.2	10.2	6.1	0.9	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	8.6	6.3	12.3	17.3	16.5	18.2	10.2	6.1	0.9	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	3.1	8.6	6.3	12.0	17.5	16.5	18.2	10.2	5.9	1.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	38.5	0.0	1.9	0.0	58.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	41.2	0.0	0.9	57.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	14.7	3.7	17.1	11.1	15.8	22.6	8.2	4.4	1.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.9	17.3	2.7	28.7	11.1	18.5	6.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.1	3.1	8.6	6.3	12.0	17.5	16.5	18.2	10.2	5.9	1.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	3.1	8.6	6.3	12.0	17.5	16.5	18.2	10.2	5.9	1.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	62.4	3.0	0.6	0.8	6.0	13.1	3.8	0.1	1.2	2.1	3.1	3.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model and total calibration																									
None (weighted WI)	0.0	0.0	1.3	1.7	0.0	4.1	4.1	1.2	0.7	13.4	14.7	14.8	5.8	3.6	6.4	8.7	2.6	2.7	3.2	3.3	1.4	1.5	0.0	0.0	4.7
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	25.7	8.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	32.4	14.5	20.9	28.2	4.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	27.4	6.3	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	27.4	6.3	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	25.7	8.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	49.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.7	0.0	0.0	0.0	0.0	13.2	0.0
MARS	0.1	0.0	0.0	0.0	0.0	0.0	1.0	4.8	6.2	27.5	0.0	2.1	25.6	0.0	0.0	4.0	0.0	1.7	0.0	3.9	0.0	5.1	8.5	8.2	1.4
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	31.0	17.5	15.7	33.0	2.8	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	9.1	23.1	17.5	10.4	7.4	14.4	17.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	25.7	8.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	25.7	8.0	0.0	0.0	0.0
Heckman model	41.1	6.8	1.2	0.1	0.0	0.0	0.0	17.2	24.1	8.2	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model and total calibration																									
None (weighted WI)	0.0	1.1	0.2	2.3	2.6	5.3	8.5	8.9	11.3	11.0	12.6	11.5	8.6	4.3	3.3	3.0	1.7	1.7	0.3	0.3	0.2	0.4	0.1	0.1	0.7
GLM (OLS)	0.0	0.0	0.2	0.7	1.4	1.3	2.2	6.0	8.3	6.8	11.4	15.6	15.8	13.1	7.3	6.8	2.5	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.2	0.8	1.4	1.0	2.4	6.0	8.3	6.7	11.8	16.2	15.5	12.7	7.3	7.3	1.9	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.2	0.7	1.4	1.3	2.2	6.0	8.3	6.7	11.8	15.4	15.8	13.1	7.3	7.1	2.3	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.2	0.7	1.4	1.3	2.2	6.0	8.3	6.7	11.8	15.4	15.8	13.1	7.3	7.1	2.3	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.2	0.7	1.4	1.3	2.2	6.0	8.3	6.8	11.4	15.6	15.8	13.1	7.3	6.8	2.5	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	42.3	0.0	5.3	0.0	51.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0
MARS	0.1	0.1	0.1	0.2	0.9	3.3	2.7	2.4	2.6	36.5	1.8	4.2	45.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.1	1.2	1.5	1.3	2.6	3.0	12.6	4.5	15.7	11.2	14.5	16.1	6.6	6.2	2.1	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.5	2.0	17.2	17.3	7.5	24.8	10.3	14.0	6.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.2	0.7	1.4	1.3	2.2	6.0	8.3	6.8	11.4	15.6	15.8	13.1	7.3	6.8	2.5	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.2	0.7	1.4	1.3	2.2	6.0	8.3	6.8	11.4	15.6	15.8	13.1	7.3	6.8	2.5	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	63.8	4.5	3.2	3.6	5.4	7.3	2.7	0.3	0.7	2.0	3.2	3.1	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model and total calibration																									
None (weighted WI)	0.0	0.0	1.1	1.6	0.0	4.5	4.4	1.3	1.4	14.6	14.9	15.3	5.7	3.7	6.4	7.8	2.3	2.3	2.9	3.1	1.2	1.2	0.0	0.0	4.3
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	23.4	6.4	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	36.0	13.5	21.6	25.2	3.2	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	24.8	5.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	24.8	5.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	23.4	6.4	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	51.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.3	0.0	0.0	0.0	0.0	10.9	0.0
MARS	0.6	0.0	0.0	0.0	0.0	0.0	1.3	5.6	7.0	28.9	0.0	2.4	25.9	0.0	0.0	3.7	0.0	1.6	0.0	3.5	0.0	4.5	7.3	6.7	1.1
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	34.5	16.3	16.7	29.8	2.2	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.4	0.2	0.0	0.0	0.0	0.3	10.6	25.2	16.9	10.7	7.4	13.3	15.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	23.4	6.4	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	23.4	6.4	0.0	0.0	0.0
Heckman model	44.8	6.0	1.0	0.1	0.0	0.0	0.0	18.4	22.1	6.7	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model and total calibration																									
None (weighted WI)	0.0	0.9	0.3	1.8	1.8	4.3	7.3	9.6	11.7	11.1	14.3	11.9	7.3	4.0	3.9	3.1	2.2	1.8	0.6	0.4	0.2	0.2	0.1	0.1	0.8
GLM (OLS)	0.0	0.1	0.1	0.5	1.4	1.3	1.6	4.6	8.9	8.6	10.4	21.1	19.2	9.7	7.6	4.0	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.5	1.5	1.1	1.7	4.6	9.0	8.3	10.9	22.3	18.5	9.1	7.6	4.2	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.4	1.5	1.3	1.5	4.6	8.9	8.3	11.3	20.6	19.2	9.7	7.6	4.2	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.4	1.5	1.3	1.5	4.6	8.9	8.3	11.3	20.6	19.2	9.7	7.6	4.2	0.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	0.5	1.4	1.3	1.6	4.6	8.9	8.6	10.4	21.1	19.2	9.7	7.6	4.0	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.7	0.0	0.2	0.0	71.9	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.5	1.5	1.5	1.6	2.3	22.3	2.6	6.7	60.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.1	0.7	1.5	0.9	1.7	2.7	13.4	5.9	12.5	17.4	21.0	7.7	12.0	1.2	0.7	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	2.3	15.7	6.7	7.6	41.6	6.8	5.8	12.9	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	0.5	1.4	1.3	1.6	4.6	8.9	8.6	10.4	21.1	19.2	9.7	7.6	4.0	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.1	0.1	0.5	1.4	1.3	1.6	4.6	8.9	8.6	10.4	21.1	19.2	9.7	7.6	4.0	0.6	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Heckman model	23.8	1.2	3.4	7.7	10.5	14.5	2.9	1.3	5.7	9.3	13.2	6.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.5: Income class frequencies (in percentage points) estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model and covariance calibration																									
None (weighted WI)	0.0	1.0	0.3	2.1	2.1	4.7	8.0	10.1	11.8	11.1	13.8	11.3	6.8	3.9	3.6	2.8	2.3	1.9	0.6	0.5	0.2	0.2	0.2	0.1	0.7
GLM (OLS)	0.0	0.1	0.1	0.7	1.8	1.6	2.7	6.2	9.3	7.5	9.0	20.1	18.0	8.0	7.9	5.7	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.6	1.8	1.3	2.6	6.2	9.5	7.2	9.0	21.3	18.2	7.3	7.8	6.2	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.6	1.8	1.6	2.5	6.3	9.4	7.2	9.6	19.8	18.0	8.0	7.9	6.1	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.6	1.8	1.6	2.5	6.3	9.4	7.2	9.6	19.8	18.0	8.0	7.9	6.1	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	0.7	1.8	1.6	2.7	6.2	9.3	7.5	9.0	20.1	18.0	8.0	7.9	5.7	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.0	0.0	0.1	0.0	66.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.1	0.2	0.9	2.3	2.0	1.5	1.9	25.1	1.4	3.0	61.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.1	1.0	1.8	1.3	3.6	3.4	14.2	4.9	12.6	17.2	18.0	6.1	11.8	2.4	1.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	3.6	19.6	3.8	4.9	44.6	4.0	6.2	12.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	0.7	1.8	1.6	2.7	6.2	9.3	7.5	9.0	20.1	18.0	8.0	7.9	5.7	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.1	0.1	0.7	1.8	1.6	2.7	6.2	9.3	7.5	9.0	20.1	18.0	8.0	7.9	5.7	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	24.4	3.1	3.6	5.0	8.5	13.4	4.1	1.1	7.8	11.6	9.7	7.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model and covariance calibration																									
None (weighted WI)	0.0	0.8	0.5	2.3	1.5	3.1	5.9	7.7	10.7	10.3	15.8	13.9	8.2	4.8	4.4	3.4	2.2	1.7	0.6	0.4	0.3	0.2	0.1	0.1	0.9
GLM (OLS)	0.0	0.0	0.0	0.3	0.8	0.6	0.9	2.6	7.1	8.1	7.7	17.3	28.7	12.4	7.9	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.3	0.8	0.5	0.9	2.6	7.2	7.9	8.2	17.8	28.2	12.2	7.9	5.2	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.3	0.8	0.6	0.9	2.5	7.2	7.9	8.4	16.9	28.7	12.4	7.9	5.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.3	0.8	0.6	0.9	2.5	7.2	7.9	8.4	16.9	28.7	12.4	7.9	5.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.3	0.8	0.6	0.9	2.6	7.1	8.1	7.7	17.3	28.7	12.4	7.9	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.6	0.0	0.4	0.0	87.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.1	0.4	1.3	1.6	2.4	3.5	10.7	8.6	17.7	53.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.0	0.5	0.8	0.6	1.1	2.9	8.8	6.0	9.2	18.9	25.3	9.3	12.9	2.9	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.3	0.0	0.4	1.0	2.8	8.1	8.0	11.4	31.1	14.3	9.1	13.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.3	0.8	0.6	0.9	2.6	7.1	8.1	7.7	17.3	28.7	12.4	7.9	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.3	0.8	0.6	0.9	2.6	7.1	8.1	7.7	17.3	28.7	12.4	7.9	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	10.1	2.6	4.3	9.9	7.1	7.2	1.6	1.4	4.5	21.3	25.0	4.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model and covariance calibration																									
None (weighted WI)	0.0	1.1	0.5	2.5	2.5	4.8	8.1	10.0	11.6	11.2	13.5	11.3	6.8	4.0	3.2	2.8	2.0	1.6	0.6	0.4	0.2	0.2	0.1	0.2	0.6
GLM (OLS)	0.2	0.2	0.3	0.9	1.7	1.6	2.1	5.7	9.9	9.3	10.8	18.9	16.1	8.6	5.3	5.8	2.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.2	0.0	0.4	0.9	1.8	1.4	2.1	5.8	10.0	9.0	11.4	19.6	15.9	8.2	5.2	6.5	1.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.2	0.1	0.4	0.8	1.8	1.7	1.9	5.8	10.0	9.0	11.6	18.6	16.1	8.6	5.4	6.4	1.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.2	0.1	0.4	0.8	1.8	1.7	1.9	5.8	10.0	9.0	11.6	18.6	16.1	8.6	5.4	6.4	1.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.2	0.2	0.3	0.9	1.7	1.6	2.1	5.7	9.9	9.3	10.8	18.9	16.1	8.6	5.3	5.8	2.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.9	0.0	0.0	0.0	66.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.5	0.1	0.2	0.3	1.1	1.6	2.3	2.2	2.7	27.7	4.2	7.8	49.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.3	0.3	0.2	1.4	1.6	1.7	2.4	3.5	15.4	7.2	11.2	17.5	15.5	6.7	9.3	3.5	2.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.2	0.2	0.6	0.3	0.8	1.5	3.8	18.6	7.3	7.6	35.3	8.0	5.1	10.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.2	0.2	0.3	0.9	1.7	1.6	2.1	5.7	9.9	9.3	10.8	18.9	16.1	8.6	5.3	5.8	2.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.2	0.2	0.3	0.9	1.7	1.6	2.1	5.7	9.9	9.3	10.8	18.9	16.1	8.6	5.3	5.8	2.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	30.6	3.6	5.3	7.4	8.7	10.5	3.8	0.9	3.8	7.9	10.2	6.8	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model and covariance calibration																									
None (weighted WI)	0.0	1.1	0.4	2.2	2.7	6.4	8.9	10.5	11.6	12.8	13.9	10.8	6.5	3.3	2.6	1.9	1.6	1.1	0.4	0.2	0.1	0.1	0.2	0.1	0.6
GLM (OLS)	0.1	0.1	0.1	0.3	0.8	1.3	3.1	8.3	11.2	7.6	13.1	25.0	15.7	6.1	4.1	2.5	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.1	0.0	0.1	0.2	0.9	1.0	2.2	8.5	11.9	7.5	12.2	25.8	16.5	5.8	4.1	2.6	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.0	0.1	0.2	0.9	1.3	2.4	9.0	11.2	7.4	13.5	24.7	15.7	6.1	4.1	2.6	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.0	0.1	0.2	0.9	1.3	2.4	9.0	11.2	7.4	13.5	24.7	15.7	6.1	4.1	2.6	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.1	0.1	0.3	0.8	1.3	3.1	8.3	11.2	7.6	13.1	25.0	15.7	6.1	4.1	2.5	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.6	0.0	1.0	0.0	60.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0
MARS	0.2	0.0	0.0	0.1	0.4	0.8	0.7	0.7	1.0	36.2	0.9	2.4	56.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.1	0.1	0.4	0.7	0.7	6.2	5.5	14.5	4.6	23.5	15.6	14.2	6.1	5.5	1.5	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.2	0.1	0.1	0.1	0.2	5.7	21.3	7.7	3.7	47.7	3.7	4.4	5.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	0.1	0.3	0.8	1.3	3.1	8.3	11.2	7.6	13.1	25.0	15.7	6.1	4.1	2.5	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.1	0.1	0.1	0.3	0.8	1.3	3.1	8.3	11.2	7.6	13.1	25.0	15.7	6.1	4.1	2.5	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	29.6	6.4	5.4	3.2	5.4	9.4	2.1	1.0	11.8	16.9	5.9	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model and covariance calibration																									
None (weighted WI)	0.0	0.8	0.3	1.6	1.7	4.2	7.0	9.3	11.4	11.2	14.6	12.6	7.7	4.2	4.0	3.3	2.1	1.6	0.5	0.4	0.2	0.2	0.1	0.1	0.8
GLM (OLS)	0.0	0.0	0.0	0.3	1.0	1.8	2.0	4.7	7.7	9.6	12.5	19.9	19.3	10.9	6.7	2.8	0.4	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.3	1.1	1.3	2.4	4.6	7.8	8.9	13.6	20.7	18.5	10.5	6.7	2.9	0.2	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.3	1.1	1.8	2.0	4.6	7.8	9.0	13.8	19.2	19.3	10.9	6.7	2.9	0.3	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.3	1.1	1.8	2.0	4.6	7.8	9.0	13.8	19.2	19.3	10.9	6.7	2.9	0.3	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.3	1.0	1.8	2.0	4.7	7.7	9.6	12.5	19.9	19.3	10.9	6.7	2.8	0.4	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.6	0.0	0.2	0.0	71.9	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.3	0.9	1.1	1.4	2.6	22.5	1.9	7.7	61.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.5	1.2	1.3	1.8	2.7	12.1	7.2	13.1	16.8	21.9	9.0	10.6	0.9	0.4	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.2	1.6	15.1	6.1	10.4	38.5	9.8	7.2	11.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.3	1.0	1.8	2.0	4.7	7.7	9.6	12.5	19.9	19.3	10.9	6.7	2.8	0.4	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.3	1.0	1.8	2.0	4.7	7.7	9.6	12.5	19.9	19.3	10.9	6.7	2.8	0.4	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Heckman model	23.3	1.0	2.7	9.4	12.7	12.3	1.8	1.9	5.2	7.6	17.4	4.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.5: Income class frequencies (in percentage points) estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using total and covariance calibration																									
None (weighted WI)	0.0	1.1	0.0	0.0	0.9	3.7	6.6	6.0	7.7	11.3	12.7	13.6	10.3	5.7	4.8	5.0	3.9	3.4	1.0	0.5	0.3	0.1	0.3	0.0	0.8
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	2.0	2.8	6.2	0.0	0.0	5.6	37.0	15.1	10.8	13.1	7.2	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	1.5	3.2	6.7	0.0	0.0	6.0	38.6	13.3	10.3	13.1	7.2	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	2.0	2.4	7.1	0.0	0.0	6.9	35.9	14.9	10.7	13.0	7.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	2.0	2.4	7.1	0.0	0.0	6.9	35.9	14.9	10.7	13.0	7.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	2.0	2.8	6.2	0.0	0.0	5.6	37.0	15.1	10.8	13.1	7.2	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.8	0.0	1.4	0.0	69.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	17.5	0.0	0.0	81.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.3	0.9	3.0	0.0	2.7	0.0	11.0	21.2	25.4	10.5	22.2	2.5	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.0	0.0	0.0	57.1	5.8	2.5	20.5	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	2.0	2.8	6.2	0.0	0.0	5.6	37.0	15.1	10.8	13.1	7.2	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	2.0	2.8	6.2	0.0	0.0	5.6	37.0	15.1	10.8	13.1	7.2	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.0	0.0
Heckman model	31.6	4.1	0.0	0.0	13.8	23.9	0.0	0.0	3.6	0.0	13.6	9.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using total and covariance calibration																									
None (weighted WI)	0.0	0.9	0.0	- 1.1	0.2	3.0	4.9	7.3	9.2	12.3	15.7	14.2	11.0	4.6	5.0	4.6	3.2	2.2	0.7	0.2	0.3	0.2	0.3	0.0	1.0
GLM (OLS)	0.1	1.2	1.0	9.5	0.5	0.0	0.0	0.0	0.0	68.7	0.0	0.0	0.0	0.0	0.0	6.9	10.0	1.4	0.0	0.9	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.1	0.3	1.6	9.6	1.7	0.0	0.0	0.0	0.0	66.3	0.0	0.0	0.0	0.0	0.0	11.3	6.9	1.4	0.0	0.9	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.6	1.5	9.4	1.5	0.0	0.0	0.0	0.0	66.6	0.0	0.0	0.0	0.0	0.0	11.6	6.3	1.4	0.0	0.9	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.6	1.5	9.4	1.5	0.0	0.0	0.0	0.0	66.6	0.0	0.0	0.0	0.0	0.0	11.6	6.3	1.4	0.0	0.9	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	1.2	1.0	9.5	0.5	0.0	0.0	0.0	0.0	68.7	0.0	0.0	0.0	0.0	0.0	6.9	10.0	1.4	0.0	0.9	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.2	0.2	0.6	1.8	7.7	12.5	0.0	0.0	0.0	0.0	47.2	29.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
ANN (opt. knots)	0.2	1.5	1.3	14.6	0.0	3.1	0.0	11.9	30.3	0.0	0.0	0.0	0.0	0.0	0.0	22.4	11.4	2.5	0.0	0.8	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.6	0.3	1.9	7.1	13.3	0.0	45.5	31.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	1.2	1.0	9.5	0.5	0.0	0.0	0.0	0.0	68.7	0.0	0.0	0.0	0.0	0.0	6.9	10.0	1.4	0.0	0.9	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.1	1.2	1.0	9.5	0.5	0.0	0.0	0.0	0.0	68.7	0.0	0.0	0.0	0.0	0.0	6.9	10.0	1.4	0.0	0.9	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	11.8	12.2	0.0	0.0	43.1	14.4	0.0	15.2	0.0	2.4	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using total and covariance calibration																									
None (weighted WI)	0.0	1.4	0.6	4.3	5.9	11.7	13.6	13.3	12.8	11.4	10.5	7.4	2.9	1.1	1.1	0.6	0.6	0.1	0.1	0.1	0.1	0.0	0.0	0.1	0.2
GLM (OLS)	0.0	0.3	0.2	1.0	2.3	3.1	6.3	15.3	18.1	13.0	17.5	12.9	7.5	1.4	0.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.1	0.3	0.7	2.5	2.5	5.7	15.7	18.7	12.6	18.3	13.3	6.9	1.3	0.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.3	0.8	2.6	3.1	5.5	16.1	18.1	12.4	18.1	12.9	7.5	1.4	0.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.3	0.8	2.6	3.1	5.5	16.1	18.1	12.4	18.1	12.9	7.5	1.4	0.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.3	0.2	1.0	2.3	3.1	6.3	15.3	18.1	13.0	17.5	12.9	7.5	1.4	0.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	92.6	0.0	5.7	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0
MARS	0.0	0.0	0.1	0.2	1.3	4.8	4.9	4.2	4.8	78.6	0.0	0.1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.3	0.2	1.3	2.4	2.1	10.0	9.3	25.2	7.4	23.9	10.1	5.4	1.0	0.6	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.6	8.3	37.9	31.1	8.0	13.6	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.3	0.2	1.0	2.3	3.1	6.3	15.3	18.1	13.0	17.5	12.9	7.5	1.4	0.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.3	0.2	1.0	2.3	3.1	6.3	15.3	18.1	13.0	17.5	12.9	7.5	1.4	0.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	72.1	6.1	7.5	5.8	6.5	1.3	0.0	0.0	0.1	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using total and covariance calibration																									
None (weighted WI)	0.0	0.8	0.5	2.5	2.4	4.9	7.3	7.3	8.6	10.6	12.1	12.3	8.4	6.3	5.1	3.4	2.1	2.5	0.4	0.4	0.1	0.7	0.1	0.1	0.9
GLM (OLS)	0.0	0.0	0.0	0.0	5.4	94.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	5.4	94.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	5.4	94.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	5.4	94.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	5.4	94.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	94.6	0.0	0.0	0.0	5.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	94.6	0.0	0.0	0.0	5.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	5.4	0.0	94.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	94.6	5.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	5.4	94.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	5.4	94.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	94.6	0.0	0.0	0.0	0.0	5.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using total and covariance calibration																									
None (weighted WI)	0.0	0.9	0.2	1.2	1.7	4.1	6.6	8.8	10.7	11.3	14.6	12.6	8.5	4.5	4.5	4.2	2.2	1.4	0.4	0.5	0.2	0.3	0.0	0.1	0.7
GLM (OLS)	0.0	0.0	0.0	0.1	0.2	1.4	3.1	6.8	5.4	7.0	17.6	21.4	13.1	16.8	5.8	0.6	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.3	0.9	3.5	6.9	5.3	6.0	20.2	20.8	12.2	16.7	5.8	0.7	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.3	1.4	3.1	6.8	5.4	6.3	19.9	19.8	13.1	16.8	5.8	0.6	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.3	1.4	3.1	6.8	5.4	6.3	19.9	19.8	13.1	16.8	5.8	0.6	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.2	1.4	3.1	6.8	5.4	7.0	17.6	21.4	13.1	16.8	5.8	0.6	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.6	0.0	0.1	0.0	71.9	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	1.4	25.8	0.3	3.6	68.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.1	0.3	1.0	1.9	3.0	10.4	5.3	14.7	18.8	18.8	14.1	10.7	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.9	15.2	3.1	10.0	38.3	10.4	11.1	10.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.2	1.4	3.1	6.8	5.4	7.0	17.6	21.4	13.1	16.8	5.8	0.6	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.2	1.4	3.1	6.8	5.4	7.0	17.6	21.4	13.1	16.8	5.8	0.6	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0
Heckman model	23.8	0.5	0.8	8.5	20.5	9.2	0.2	2.7	5.6	2.3	24.5	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.5: Income class frequencies (in percentage points) estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model, total and covariance calibration																									
None (weighted WI)	0.0	0.2	0.0	1.1	2.8	6.4	9.0	5.3	8.5	9.8	10.7	12.0	7.9	7.7	4.4	5.3	3.4	3.2	0.7	0.3	0.2	0.2	0.1	0.1	0.8
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.2	2.8	13.4	8.4	0.0	0.0	15.2	23.3	23.9	6.4	4.3	1.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	2.6	13.8	8.7	0.0	0.0	15.7	22.8	23.5	6.4	5.3	1.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.2	2.3	14.5	7.9	0.0	0.0	14.9	23.4	24.0	6.4	4.7	1.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.2	2.3	14.5	7.9	0.0	0.0	14.9	23.4	24.0	6.4	4.7	1.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.2	2.8	13.4	8.4	0.0	0.0	15.2	23.3	23.9	6.4	4.3	1.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	45.0	0.0	3.4	0.0	51.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.4	0.0	0.0	62.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	5.5	3.5	12.6	0.0	0.0	15.9	18.7	32.8	8.2	0.9	1.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.8	0.0	0.0	35.0	17.7	12.7	8.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.2	2.8	13.4	8.4	0.0	0.0	15.2	23.3	23.9	6.4	4.3	1.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.2	2.8	13.4	8.4	0.0	0.0	15.2	23.3	23.9	6.4	4.3	1.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	57.3	3.9	0.0	0.0	10.9	19.2	0.3	0.0	0.0	0.8	0.0	7.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model, total and covariance calibration																									
None (weighted WI)	0.0	0.0	1.3	1.7	0.0	4.1	4.1	1.2	0.7	13.4	14.7	14.8	5.8	3.6	6.4	8.7	2.6	2.7	3.2	3.3	1.4	1.5	0.0	0.0	4.7
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	25.7	8.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	32.4	14.5	20.9	28.2	4.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	27.4	6.3	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	27.4	6.3	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	25.7	8.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	49.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.7	0.0	0.0	0.0	0.0	13.2	0.0
MARS	0.1	0.0	0.0	0.0	0.0	0.0	1.0	4.8	6.2	27.5	0.0	2.1	25.6	0.0	0.0	4.0	0.0	1.7	0.0	3.9	0.0	5.1	8.5	8.2	1.4
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	31.0	17.5	15.7	33.0	2.8	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	9.1	23.1	17.5	10.4	7.4	14.4	17.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	25.7	8.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.9	14.4	19.9	25.7	8.0	0.0	0.0	0.0
Heckman model	41.1	6.8	1.2	0.1	0.0	0.0	0.0	17.2	24.1	8.2	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model, total and covariance calibration																									
None (weighted WI)	0.0	0.9	0.2	2.2	2.2	4.8	8.5	8.2	11.0	11.4	12.9	12.6	6.5	5.6	3.1	3.6	2.2	2.0	0.5	0.3	0.1	0.3	0.1	0.1	0.8
GLM (OLS)	0.1	0.2	0.0	0.1	0.5	1.5	3.5	7.0	8.7	6.8	5.4	16.5	17.5	16.7	6.9	7.4	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.1	0.0	0.0	0.0	0.6	1.1	3.6	7.1	8.8	6.4	6.1	16.9	17.4	16.4	6.8	8.1	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.1	0.0	0.0	0.6	1.5	3.3	7.4	8.6	6.3	6.1	16.3	17.5	16.7	6.9	7.7	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.1	0.0	0.0	0.6	1.5	3.3	7.4	8.6	6.3	6.1	16.3	17.5	16.7	6.9	7.7	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.2	0.0	0.1	0.5	1.5	3.5	7.0	8.7	6.8	5.4	16.5	17.5	16.7	6.9	7.4	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.6	0.0	2.5	0.0	59.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.2	0.3	2.9	1.6	1.8	2.4	26.1	2.8	7.2	54.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.1	0.9	0.6	4.6	3.3	14.2	3.7	5.0	17.3	13.8	21.7	7.6	4.9	1.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	2.7	21.1	4.6	4.9	30.0	16.7	13.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.2	0.0	0.1	0.5	1.5	3.5	7.0	8.7	6.8	5.4	16.5	17.5	16.7	6.9	7.4	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.1	0.2	0.0	0.1	0.5	1.5	3.5	7.0	8.7	6.8	5.4	16.5	17.5	16.7	6.9	7.4	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	58.3	6.6	2.9	3.1	6.4	7.8	2.8	0.0	0.0	2.2	5.0	4.6	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model, total and covariance calibration																									
None (weighted WI)	0.0	0.0	1.1	1.6	0.0	4.5	4.4	1.3	1.4	14.6	14.9	15.3	5.7	3.7	6.4	7.8	2.3	2.3	2.9	3.1	1.2	1.2	0.0	0.0	4.3
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	23.4	6.4	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	36.0	13.5	21.6	25.2	3.2	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	24.8	5.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	24.8	5.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	23.4	6.4	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	51.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.3	0.0	0.0	0.0	0.0	10.9	0.0
MARS	0.6	0.0	0.0	0.0	0.0	0.0	1.3	5.6	7.0	28.9	0.0	2.4	25.9	0.0	0.0	3.7	0.0	1.6	0.0	3.5	0.0	4.5	7.3	6.7	1.1
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	34.5	16.3	16.7	29.8	2.2	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.4	0.2	0.0	0.0	0.0	0.3	10.6	25.2	16.9	10.7	7.4	13.3	15.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	23.4	6.4	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	35.4	13.5	20.6	23.4	6.4	0.0	0.0	0.0
Heckman model	44.8	6.0	1.0	0.1	0.0	0.0	0.0	18.4	22.1	6.7	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model, total and covariance calibration																									
None (weighted WI)	0.0	0.8	0.3	1.5	1.7	4.2	6.9	9.1	11.3	11.2	14.6	12.7	7.8	4.3	4.1	3.4	2.1	1.5	0.5	0.4	0.2	0.2	0.1	0.1	0.8
GLM (OLS)	0.0	0.0	0.0	0.3	0.9	1.7	2.2	4.7	7.4	9.3	13.2	19.8	19.0	11.4	6.6	2.5	0.4	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.2	0.9	1.2	2.6	4.7	7.4	8.6	14.5	20.6	18.2	11.1	6.6	2.6	0.2	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.2	0.9	1.7	2.1	4.7	7.5	8.7	14.6	19.0	19.0	11.4	6.6	2.6	0.3	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.2	0.9	1.7	2.1	4.7	7.5	8.7	14.6	19.0	19.0	11.4	6.6	2.6	0.3	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.3	0.9	1.7	2.2	4.7	7.4	9.3	13.2	19.8	19.0	11.4	6.6	2.5	0.4	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.6	0.0	0.2	0.0	72.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.2	0.7	0.9	1.3	2.5	22.9	1.6	7.3	62.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.4	1.0	1.3	1.9	2.6	11.8	7.1	13.5	16.7	21.9	9.6	10.5	0.8	0.4	0.2	0.1	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	1.4	14.9	5.7	10.6	38.2	10.2	7.7	11.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.3	0.9	1.7	2.2	4.7	7.4	9.3	13.2	19.8	19.0	11.4	6.6	2.5	0.4	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.3	0.9	1.7	2.2	4.7	7.4	9.3	13.2	19.8	19.0	11.4	6.6	2.5	0.4	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Heckman model	23.3	0.9	2.4	9.4	13.5	12.0	1.6	2.0	5.3	6.9	18.5	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.6: Income class frequencies (in percentage points) estimated from the imputed Micro-census, using a weighted loss function for prediction models

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: unweighted, using no auxiliary information																									
Matching	0.0	0.0	0.0	0.0	1.6	4.1	4.6	7.2	9.9	11.8	17.2	12.3	5.8	13.4	5.0	2.1	0.4	0.2	0.1	0.0	0.2	0.8	0.4	1.7	1.1
GLM (OLS)	0.1	0.1	0.7	1.4	0.5	1.3	2.8	4.4	3.5	3.7	9.7	10.0	11.7	11.7	9.2	9.8	8.3	6.0	3.1	1.8	0.3	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.1	0.3	1.8	0.6	1.2	2.8	4.5	3.5	3.6	9.9	10.4	11.6	11.8	9.3	10.1	8.1	5.7	3.2	1.4	0.2	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.1	0.7	1.4	0.6	1.3	2.7	4.5	3.5	3.5	9.8	9.9	11.7	11.7	9.3	10.0	8.1	6.0	3.1	1.8	0.2	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.1	0.7	1.4	0.6	1.3	2.7	4.5	3.5	3.5	9.8	9.9	11.7	11.7	9.3	10.0	8.1	6.0	3.1	1.8	0.2	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.1	0.7	1.4	0.5	1.3	2.8	4.4	3.5	3.7	9.7	10.0	11.7	11.7	9.2	9.8	8.3	6.0	3.1	1.8	0.3	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.4	0.0	4.0	0.0	48.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.1	0.0
MARS	0.1	0.2	0.9	1.2	1.4	2.2	2.0	3.3	2.9	42.8	2.2	2.9	38.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.4	0.7	0.9	3.8	0.7	4.6	3.3	4.0	3.5	4.8	13.2	7.6	7.2	12.3	8.3	8.2	6.1	7.0	1.6	2.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.4	3.4	3.9	7.0	14.8	11.7	21.3	15.6	16.5	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	0.7	1.4	0.5	1.3	2.8	4.4	3.5	3.7	9.7	10.0	11.7	11.7	9.2	9.8	8.3	6.0	3.1	1.8	0.3	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	62.2	6.0	8.4	7.2	4.8	3.6	2.2	1.2	0.9	1.1	1.3	0.7	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (parametric), using propensity weights																									
GLM (OLS)	0.1	0.0	0.7	1.5	0.5	2.6	2.6	3.9	3.3	5.3	9.9	10.1	12.1	11.3	9.3	10.0	7.8	4.9	2.8	1.2	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.3	1.8	0.5	2.1	2.4	4.4	3.4	5.1	10.5	10.2	12.7	12.0	9.1	10.0	7.5	4.5	2.7	0.7	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.1	0.7	1.5	0.5	2.6	2.5	4.0	3.4	5.3	9.9	10.1	12.4	11.7	8.5	10.4	7.4	4.9	3.0	1.0	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.1	0.7	1.5	0.5	2.6	2.5	4.0	3.4	5.3	9.9	10.1	12.4	11.7	8.5	10.4	7.4	4.9	3.0	1.0	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.1	0.7	1.6	0.4	3.2	2.1	3.8	3.6	4.2	13.1	6.6	7.7	11.5	9.9	8.6	7.0	9.4	3.0	2.0	0.9	0.5	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.4	0.0	4.0	0.0	48.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.1	0.0
MARS	0.0	0.0	0.0	0.0	1.8	3.5	3.3	2.4	3.4	29.2	5.9	6.0	35.9	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.4	0.6	1.4	3.6	1.1
ANN (opt. knots)	0.1	0.3	1.0	1.1	0.6	2.3	3.8	2.8	4.1	3.3	17.2	9.3	8.5	12.6	10.0	7.6	5.0	7.7	1.4	1.4	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	1.5	0.8	0.8	2.5	3.3	3.1	2.8	5.7	10.4	8.8	11.6	12.3	9.9	8.7	8.4	5.1	2.6	1.3	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	94.9	0.3	0.4	0.5	0.5	0.6	0.7	1.4	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Pseudo-Weights (parametric), using propensity weights																									
GLM (OLS)	0.0	0.1	0.1	1.8	0.8	0.7	2.7	4.3	4.1	3.9	9.4	11.4	10.8	11.4	8.5	8.2	7.9	6.1	4.5	2.8	0.7	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	1.5	0.9	0.6	2.2	4.2	4.4	4.0	10.5	11.8	11.5	11.3	8.6	8.6	7.4	6.0	4.0	2.3	0.2	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.1	1.8	0.8	0.7	2.7	4.0	4.5	3.9	9.5	11.3	10.8	11.5	8.8	7.8	7.9	6.4	4.1	2.8	0.7	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.1	1.8	0.8	0.7	2.7	4.0	4.5	3.9	9.5	11.3	10.8	11.5	8.8	7.8	7.9	6.4	4.1	2.8	0.7	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.1	1.2	1.1	0.4	1.1	3.2	4.3	3.7	2.3	13.8	12.5	6.2	15.7	5.6	9.2	5.9	5.8	3.9	3.9	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.4	0.0	0.0	11.7	0.0	30.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.4	0.0	6.1	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.1	0.4	1.3	1.1	3.9	3.0	7.5	3.2	15.8	10.2	6.0	17.0	4.1	5.3	7.9	4.8	5.7	2.5	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	0.3	1.8	0.5	0.9	3.1	4.3	3.2	3.2	11.7	10.5	8.8	13.1	7.6	9.2	7.2	7.0	3.9	2.7	0.7	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	34.9	0.0	4.1	24.1	23.1	0.0	1.5	1.8	1.1	6.2	0.4	0.7	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using propensity weights																									
GLM (OLS)	0.0	0.1	0.1	1.5	0.9	0.7	3.1	3.6	4.5	4.5	12.7	13.3	13.6	12.9	9.5	10.1	5.4	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	1.2	1.2	0.7	3.0	3.1	5.1	4.6	12.7	14.1	13.9	12.7	9.5	10.1	4.9	2.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.1	1.5	0.9	0.7	3.0	3.7	4.4	4.6	12.7	13.4	14.2	12.5	9.7	10.0	5.3	2.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.1	1.5	0.9	0.7	3.0	3.7	4.4	4.6	12.7	13.4	14.2	12.5	9.7	10.0	5.3	2.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	1.7	0.8	1.2	2.8	4.2	3.7	4.9	13.0	15.1	13.8	12.3	9.3	10.1	4.4	2.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	51.5	0.0	0.0	21.1	0.0	27.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	2.4	3.1	3.7	3.0	6.0	35.3	4.4	5.8	7.5	26.7	0.5	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.1	0.7	1.4	0.5	2.0	5.9	4.2	6.5	7.0	14.6	13.3	9.6	15.3	8.6	5.1	5.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	3.2	10.1	13.7	22.3	19.6	29.6	0.9	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	1.5	0.9	0.7	3.1	3.6	4.5	4.5	12.7	13.3	13.6	12.9	9.5	10.1	5.4	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.1	1.8	0.8	0.7	1.0	3.4	4.0	5.6	7.9	12.4	11.1	11.7	13.1	10.5	9.2	4.1	2.3	0.4	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots), using propensity weights																									
GLM (OLS)	0.1	0.1	0.3	1.7	0.7	2.2	3.0	3.6	3.2	5.5	9.3	11.1	10.8	12.8	10.8	10.6	7.5	4.1	2.1	0.5	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.1	0.3	1.7	0.6	2.1	2.6	4.0	3.2	5.1	10.3	11.0	11.5	13.3	10.8	10.3	7.1	4.0	1.8	0.3	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.1	0.3	1.7	0.7	2.1	3.0	3.7	3.2	5.5	9.8	10.6	11.1	12.6	10.8	10.6	7.5	4.1	2.2	0.4	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.1	0.3	1.7	0.7	2.1	3.0	3.7	3.2	5.5	9.8	10.6	11.1	12.6	10.8	10.6	7.5	4.1	2.2	0.4	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	3.5	0.7	1.4	1.2	1.4	3.1	4.1	3.6	3.8	6.2	13.2	8.1	7.9	14.4	8.8	5.5	7.9	2.9	1.8	0.5	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.4	10.6	4.0	0.0	0.0	37.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.8	0.0
MARS	5.2	2.0	1.7	3.4	3.3	4.1	4.3	4.6	5.6	5.9	12.4	2.2	2.3	25.0	0.2	0.4	0.4	0.5	0.7	0.5	0.5	2.2	3.4	6.9	2.1
ANN (opt. knots)	0.3	0.6	0.5	1.0	1.2	4.6	4.7	4.3	3.4	5.5	12.9	12.4	7.1	15.7	8.0	7.3	6.5	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	0.7	1.4	0.7	2.5	3.2	3.4	2.0	5.6	11.0	9.1	11.4	14.9	9.0	10.9	7.5	4.1	1.9	0.4	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	58.4	11.8	12.0	10.7	4.4	0.0	0.0	0.2	1.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.6: Income class frequencies (in percentage points) estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Pseudo-Weights (fix knots), using propensity weights																									
GLM (OLS)	0.0	0.1	0.1	0.8	1.5	0.7	2.4	3.5	4.8	3.9	10.3	11.2	12.6	12.7	9.1	10.3	7.3	5.0	2.9	0.8	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.3	1.8	0.7	1.8	3.7	5.1	4.1	11.0	11.9	13.0	13.1	8.8	10.2	6.7	4.7	2.6	0.4	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.8	1.5	0.6	2.4	3.5	4.8	3.9	10.4	11.2	12.6	13.2	8.6	10.5	7.1	5.0	2.9	0.8	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.8	1.5	0.6	2.4	3.5	4.8	3.9	10.4	11.2	12.6	13.2	8.8	10.2	7.1	5.0	2.9	0.8	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	1.2	1.2	0.7	2.6	4.0	4.0	3.5	10.8	12.6	12.4	12.8	8.7	10.1	6.7	5.3	2.9	0.5	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.4	0.0	4.0	36.5	11.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.8	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.1	0.1	0.7	0.7	2.0	4.4	4.6	3.3	15.8	13.3	7.7	15.7	6.4	11.5	4.6	5.6	3.6	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	1.2	1.1	0.6	2.6	4.2	3.8	3.5	11.9	10.2	11.3	14.3	9.1	9.5	7.7	5.0	2.9	0.8	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	42.5	57.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	34.9	0.0	0.0	1.8	11.4	38.1	0.0	3.4	1.5	6.2	0.0	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using propensity weights																									
GLM (OLS)	0.0	0.1	0.1	1.5	0.9	0.7	3.1	3.6	4.5	4.5	12.7	13.3	13.6	12.9	9.5	10.1	5.4	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	1.2	1.2	0.7	3.0	3.1	5.1	4.6	12.7	14.1	13.9	12.7	9.5	10.1	4.9	2.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.1	1.5	0.9	0.7	3.0	3.7	4.4	4.6	12.7	13.4	14.2	12.5	9.7	10.0	5.3	2.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.1	1.5	0.9	0.7	3.0	3.7	4.4	4.6	12.7	13.4	14.2	12.5	9.7	10.0	5.3	2.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	1.7	0.8	1.2	2.8	4.2	3.7	4.9	13.0	15.1	13.8	12.3	9.3	10.1	4.4	2.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	51.5	0.0	0.0	21.1	0.0	27.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	2.4	3.1	3.7	3.0	6.0	35.3	4.4	5.8	7.5	26.7	0.5	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.1	0.4	5.0	1.1	6.2	2.7	5.8	5.6	21.3	7.5	9.3	18.8	3.9	7.8	3.8	0.6	0.2	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	3.2	10.1	13.7	22.3	19.6	29.6	0.9	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	1.5	0.9	0.7	3.1	3.6	4.5	4.5	12.7	13.3	13.6	12.9	9.5	10.1	5.4	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.1	1.8	0.8	0.7	1.0	3.4	4.0	5.6	7.9	12.4	11.1	11.7	13.1	10.5	9.2	4.1	2.3	0.4	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using propensity weights																									
GLM (OLS)	0.0	0.1	0.1	1.5	0.9	0.7	3.1	3.6	4.5	4.5	12.7	13.3	13.6	12.9	9.5	10.1	5.4	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	1.2	1.2	0.7	3.0	3.1	5.1	4.6	12.7	14.1	13.9	12.7	9.5	10.1	4.9	2.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.1	1.5	0.9	0.7	3.0	3.7	4.4	4.6	12.7	13.4	14.2	12.5	9.7	10.0	5.3	2.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.1	1.5	0.9	0.7	3.0	3.7	4.4	4.6	12.7	13.4	14.2	12.5	9.7	10.0	5.3	2.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	1.7	0.8	1.2	2.8	4.2	3.7	4.9	13.0	15.1	13.8	12.3	9.3	10.1	4.4	2.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	51.5	0.0	0.0	21.1	0.0	27.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	2.4	3.1	3.7	3.0	6.0	35.3	4.4	5.8	7.5	26.7	0.5	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.4	1.2	4.5	0.5	6.3	2.5	6.4	6.1	17.5	12.9	7.5	15.2	8.7	5.1	5.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	3.2	10.1	13.7	22.3	19.6	29.6	0.9	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	1.5	0.9	0.7	3.1	3.6	4.5	4.5	12.7	13.3	13.6	12.9	9.5	10.1	5.4	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.1	1.8	0.8	0.7	1.0	3.4	4.0	5.6	7.9	12.4	11.1	11.7	13.1	10.5	9.2	4.1	2.3	0.4	0.0	0.0	0.0	0.0
Weighting model: Subsampling, using total calibration																									
GLM (OLS)	3.0	1.5	1.5	1.6	1.9	2.1	2.2	2.2	2.1	2.2	3.9	4.0	4.7	5.1	5.5	7.2	7.6	8.1	9.1	7.5	6.4	9.4	1.5	0.0	0.0
GLM (Ridge)	0.0	0.0	0.6	1.0	0.7	2.2	2.6	2.6	3.3	3.1	4.5	5.7	6.7	7.4	8.5	10.5	9.6	11.7	9.6	5.5	2.7	1.5	0.0	0.0	0.0
GLM (LASSO)	2.4	0.7	2.1	1.8	1.7	2.1	2.3	2.2	2.5	2.5	3.4	4.0	4.7	5.5	6.0	7.8	7.1	8.5	9.3	7.4	6.6	8.4	1.3	0.0	0.0
GLM (Elastic net)	2.4	1.2	1.6	1.8	1.8	2.1	2.2	2.4	2.2	2.1	4.0	4.0	4.8	5.1	6.1	7.8	7.0	8.4	9.2	8.0	6.0	8.8	1.3	0.0	0.0
GAM (fix knots)	3.0	1.5	1.5	1.6	1.9	2.1	2.2	2.2	2.1	2.2	3.9	4.0	4.7	5.1	5.5	7.2	7.6	8.1	9.1	7.5	6.4	9.4	1.5	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.0	0.0	0.0	29.7	0.0	0.0	3.6	0.0	0.0	0.0	0.0	0.0	33.4	0.0	8.3	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	69.6	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	1.5	2.7	24.6	
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.8	3.9	3.5	6.3	12.4	6.5	10.4	13.9	16.2	12.3	6.1	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	4.2	1.2	1.7	1.2	2.1	2.1	2.1	1.8	1.1	1.9	3.6	4.4	5.9	7.3	5.0	5.9	4.1	7.5	7.0	6.2	6.0	13.8	3.9	0.0	0.0
GAMM (fix knots)	18.9	23.6	38.2	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100
Weighting model: Post-stratification, using total calibration																									
GLM (OLS)	0.1	0.1	1.5	0.8	0.5	2.2	3.0	3.9	2.8	4.0	8.8	9.4	11.1	10.5	9.8	10.3	8.4	6.6	3.3	2.2	0.6	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	1.6	0.7	0.7	2.6	4.4	3.3	3.8	10.4	10.3	12.2	12.0	9.9	10.7	7.7	5.7	2.8	1.2	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.1	1.4	0.8	0.4	2.1	3.1	3.9	2.8	4.0	9.3	8.9	11.1	10.5	9.9	10.3	8.4	6.6	3.3	2.3	0.6	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.1	1.4	0.8	0.4	2.1	3.1	3.9	2.8	4.0	9.3	8.9	11.1	10.5	9.9	10.3	8.4	6.6	3.3	2.3	0.6	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.1	1.5	0.8	0.5	2.2	3.0	3.9	2.8	4.0	8.8	9.4	11.1	10.5	9.8	10.3	8.4	6.6	3.3	2.2	0.6	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.4	0.0	4.0	0.0	36.5	11.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.8	0.0
MARS	0.0	0.0	0.0	0.3	3.4	2.8	2.4	2.4	3.1	7.4	10.7	22.9	6.6	7.9	19.9	0.0	1.8	2.8	5.4	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.7	0.5	1.0	0.6	1.7	3.9	2.9	4.4	5.0	13.6	9.2	7.3	11.8	7.7	10.7	4.1	10.0	2.0	1.6	1.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.4	5.4	8.8	11.7	11.9	22.3	18.9	19.5	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.3	1.5	0.6	0.5	2.4	3.1	3.6	2.5	4.0	10.5	7.4	9.4	12.1	9.3	11.5	7.0	8.1	3.1	2.3	0.7	0.0	0.0	0.0	0.0
GAMM (fix knots)	42.5	57.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.1	0.2	0.2	2.3	4.4	6.7	8.5	10.1	9.1	10.4	12.8	13.4	12.4	8.1	1.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.6: Income class frequencies (in percentage points) estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using total calibration																									
GLM (OLS)	0.1	0.1	1.5	0.9	0.5	2.2	2.9	4.7	3.5	4.4	11.3	12.1	11.4	12.0	9.7	9.4	7.3	4.1	1.8	0.3	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.4	1.2	0.6	0.4	1.1	2.8	3.5	3.2	3.0	5.4	12.4	14.0	10.9	11.1	10.2	9.2	6.1	3.9	0.8	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.3	1.5	0.6	0.6	2.5	2.6	4.7	3.1	4.1	12.3	11.3	11.2	12.9	9.1	9.6	7.3	4.1	1.8	0.3	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using total calibration																									
GLM (OLS)	0.0	0.0	0.1	0.8	1.6	0.6	2.8	3.4	4.6	5.1	12.5	14.6	14.2	12.5	9.6	9.9	4.7	2.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.4	1.9	0.7	2.7	2.9	5.2	5.1	12.5	15.2	14.3	12.6	9.5	9.9	4.4	2.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.8	1.6	0.6	2.8	3.2	4.8	5.1	12.4	14.6	14.2	12.5	9.6	9.9	4.9	2.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.8	1.6	0.6	2.8	3.2	4.8	5.1	12.4	14.6	14.2	12.5	9.6	9.9	4.9	2.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	0.8	1.6	0.6	2.8	3.4	4.6	5.1	12.5	14.8	14.5	12.2	9.2	10.4	4.7	2.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	51.5	0.0	0.0	7.7	0.0	40.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.7	3.6	3.7	4.3	4.4	35.4	5.1	6.6	8.5	25.6	0.6	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	3.0	1.2	0.3	4.2	1.4	2.7	4.0	6.4	4.7	6.8	18.5	10.8	8.4	14.2	5.2	5.8	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	3.2	10.3	12.7	22.6	19.5	28.5	2.5	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	0.8	1.6	0.6	2.8	3.4	4.6	5.1	12.5	14.6	14.2	12.5	9.6	9.9	4.7	2.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.1	0.4	1.9	0.8	3.4	3.1	5.2	5.9	14.6	14.6	15.4	11.5	9.2	9.0	3.7	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using total calibration																									
GLM (OLS)	0.1	0.7	1.1	0.8	0.6	2.5	3.4	3.4	3.2	4.9	9.3	10.0	10.8	11.6	8.4	9.9	7.8	5.9	3.2	1.9	0.4	0.0	0.0	0.0	0.0
GLM (Ridge)	0.1	0.1	1.5	0.8	0.5	2.2	2.7	4.4	3.5	4.3	10.1	10.1	12.0	11.6	8.8	9.7	7.8	5.3	2.9	1.5	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.7	1.1	0.7	0.6	2.6	2.9	3.9	3.2	4.8	9.4	9.9	11.1	11.4	8.4	9.9	7.9	6.0	2.9	1.9	0.3	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.7	1.1	0.7	0.6	2.6	2.9	3.9	3.2	4.8	9.4	9.9	11.1	11.4	8.4	9.9	7.9	6.0	2.9	1.9	0.3	0.0	0.0	0.0	0.0
GAM (fix knots)	0.2	1.7	0.6	0.4	1.9	2.2	3.8	2.5	3.1	4.9	11.2	6.5	7.1	11.3	7.0	11.0	6.0	10.1	4.0	2.0	1.4	1.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.4	0.0	4.0	0.0	48.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.1	0.0
MARS	0.0	0.1	0.6	1.7	2.1	2.2	1.4	2.4	33.7	2.3	3.7	3.1	38.2	0.0	0.2	0.2	0.2	0.2	0.4	0.2	0.2	1.1	1.7	3.4	0.6
ANN (opt. knots)	2.1	0.2	0.6	2.4	2.1	5.1	2.2	4.8	4.9	4.0	11.4	9.0	8.0	14.2	3.8	9.6	5.6	6.5	1.6	1.9	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.4	17.6	33.5	21.0	19.8	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.4	1.5	0.4	0.4	1.6	2.5	3.9	2.2	3.0	4.3	10.6	8.4	10.4	11.7	8.8	10.1	6.8	7.4	2.9	2.1	0.6	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	44.4	16.8	0.3	0.8	2.6	0.0	4.6	20.1	0.0	3.2	4.5	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using total calibration																									
GLM (OLS)	0.2	1.1	1.0	0.5	0.9	2.7	3.4	3.7	3.0	4.5	9.5	8.8	10.3	11.4	8.2	9.2	8.1	6.4	3.8	2.3	0.8	0.1	0.0	0.0	0.0
GLM (Ridge)	0.1	0.3	1.5	0.7	0.5	2.6	2.9	3.9	3.4	5.0	9.1	9.9	11.3	11.7	8.0	9.5	7.5	6.3	3.3	2.2	0.5	0.0	0.0	0.0	0.0
GLM (LASSO)	0.2	1.1	0.9	0.5	0.9	2.7	3.4	3.6	3.1	4.5	9.5	8.8	10.3	11.4	8.4	9.3	7.8	6.7	3.5	2.3	0.8	0.1	0.0	0.0	0.0
GLM (Elastic net)	0.2	1.1	0.9	0.5	0.9	2.7	3.4	3.6	3.1	4.5	9.5	8.8	10.3	11.4	8.4	9.3	7.8	6.7	3.5	2.3	0.8	0.1	0.0	0.0	0.0
GAM (fix knots)	0.3	1.9	0.5	0.4	2.3	2.1	4.2	2.0	3.1	4.7	11.0	5.8	6.3	11.2	5.3	11.0	6.7	10.3	4.7	2.7	1.6	1.9	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.4	0.0	4.0	0.0	36.5	11.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.8	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.3	6.9	5.3	7.7	7.2	1.5	12.1	15.6	9.6	1.4	2.1	2.6	2.8	2.9	3.1	2.8	6.9	5.6	2.7	0.9
ANN (opt. knots)	0.1	0.1	0.4	0.6	0.6	2.9	5.8	3.6	5.7	2.8	14.5	6.8	8.8	11.3	4.6	10.4	4.5	8.6	3.9	1.6	1.1	1.3	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.8	16.9	34.8	23.3	18.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	1.3	0.9	0.3	0.5	2.3	2.4	3.7	2.0	2.9	4.9	9.8	7.7	9.3	11.8	7.9	10.7	6.4	8.1	3.3	2.4	1.2	0.2	0.0	0.0	0.0
GAMM (fix knots)	80.7	0.0	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	61.7	1.5	1.6	0.0	7.7	14.5	2.5	0.1	2.2	3.0	4.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using total calibration																									
GLM (OLS)	0.0	0.1	0.1	1.8	0.9	0.8	2.4	3.7	5.0	5.6	10.0	13.0	12.3	11.2	9.1	9.4	6.5	5.0	2.7	0.3	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	1.6	1.0	0.8	2.3	3.6	4.7	6.2	10.0	13.7	12.3	11.6	8.8	9.5	6.6	4.6	2.6	0.1	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.1	1.8	0.9	0.7	2.5	3.7	5.1	5.6	9.9	13.6	12.1	11.2	8.8	9.7	6.6	4.8	2.6	0.2	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.1	1.8	0.9	0.7	2.5	3.7	5.1	5.6	9.9	13.6	12.1	11.2	8.8	9.7	6.6	4.8	2.6	0.2	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.1	1.5	0.8	0.5	1.3	3.3	3.7	5.0	5.1	13.9	13.0	9.4	9.7	10.2	10.0	3.7	6.6	2.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	59.2	0.0	0.0	0.0	40.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.1	1.7	3.1	3.7	3.1	3.2	7.6	32.1	3.7	5.8	7.5	26.5	0.5	0.5	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.3	0.2	0.4	3.2	0.7	3.1	3.5	3.5	7.2	6.4	14.9	11.4	6.9	10.3	11.5	5.4	2.3	8.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	2.5	10.2	13.7	23.8	19.6	28.8	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	1.8	0.9	0.8	2.4	3.7	5.0	5.6	10.0	13.0	12.3	11.2	9.1	9.4	6.5	5.0	2.7	0.3	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.1	0.4	2.0	0.8	0.7	1.0	4.2	4.9	7.2	10.4	12.3	9.2	9.6	10.7	8.2	6.5	6.1	4.1	1.6	0.0	0.0	0.0	0.0

Table D.6: Income class frequencies (in percentage points) estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using covariance calibration																									
GLM (OLS)	0.0	0.0	0.1	0.2	2.2	0.8	1.1	3.8	4.7	5.9	12.5	16.5	14.2	12.4	10.6	9.8	4.2	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	0.4	2.0	0.7	1.2	4.0	4.9	5.5	13.0	17.1	13.5	12.3	10.5	10.3	3.9	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	0.2	2.2	0.8	1.1	3.8	4.7	5.9	12.5	16.5	14.2	12.4	10.6	9.8	4.2	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using covariance calibration																									
GLM (OLS)	0.1	0.1	0.3	1.8	0.6	1.7	3.0	3.9	4.1	4.8	11.8	12.8	12.9	12.8	9.5	9.9	5.8	3.3	0.9	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.1	0.1	1.9	0.7	1.2	3.1	4.2	4.2	4.8	12.4	13.0	12.7	13.3	9.5	9.6	5.7	2.8	0.7	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.3	1.7	0.7	1.8	3.0	3.8	4.1	4.8	11.8	12.9	13.1	12.6	9.6	9.8	5.9	3.3	0.8	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.3	1.7	0.7	1.8	3.0	3.8	4.1	4.8	11.8	12.9	13.2	12.6	9.6	9.8	5.9	3.3	0.8	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.3	1.3	0.8	0.7	2.2	3.5	3.3	3.4	6.3	12.8	14.3	12.9	12.3	9.2	9.7	4.6	2.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	51.5	0.0	10.6	4.5	0.0	0.0	33.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	1.8	0.6	0.6	1.5	2.2	2.5	2.1	2.2	33.0	1.5	1.9	28.5	0.2	2.5	0.4	0.4	0.4	0.4	0.7	0.7	0.5	2.2	3.3	7.4	2.3
ANN (opt. knots)	0.0	0.1	0.1	1.1	1.1	6.3	2.1	5.2	3.9	7.3	14.5	13.4	10.1	14.5	7.0	8.1	3.8	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	4.8	9.2	12.4	20.5	19.6	23.3	9.3	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	0.3	1.8	0.6	1.7	3.0	3.9	4.1	4.8	11.8	12.8	12.9	12.8	9.5	9.9	5.8	3.3	0.9	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.1	0.1	1.9	0.7	0.9	3.1	4.5	3.9	7.4	12.8	12.9	12.8	12.4	11.1	8.8	4.6	1.9	0.2	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using covariance calibration																									
GLM (OLS)	0.1	0.3	1.5	0.6	0.6	2.1	3.8	3.3	3.6	5.8	10.0	10.8	12.3	12.2	11.3	9.9	6.6	4.4	0.9	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.1	0.3	1.3	0.8	0.5	2.1	3.9	2.8	4.0	5.4	10.4	11.2	12.4	12.5	11.0	10.1	6.6	3.8	0.7	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.3	1.5	0.6	0.7	2.1	3.8	3.3	3.6	5.8	10.0	10.8	12.6	12.1	11.1	9.7	6.8	4.2	0.9	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.3	1.5	0.6	0.7	2.1	3.8	3.3	3.6	5.8	10.0	10.8	12.6	12.1	11.1	9.7	6.8	4.2	0.9	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.7	1.2	0.5	0.8	2.5	3.4	3.2	3.7	5.7	10.2	11.0	12.9	12.7	11.0	10.0	6.4	3.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.8	0.0	14.8	0.0	44.9	18.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2
MARS	0.0	0.3	0.9	2.5	1.9	1.9	3.1	3.8	20.8	3.7	6.1	22.4	4.5	5.5	6.1	5.3	4.2	3.9	2.4	0.6	0.1	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.0	2.0	0.4	0.8	3.2	6.6	2.8	5.5	7.0	9.8	14.7	10.8	10.1	9.8	9.0	3.2	2.4	1.3	0.2	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.3	18.3	30.6	19.1	22.6	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.3	1.5	0.6	0.6	2.1	3.8	3.3	3.6	5.8	10.0	10.8	12.3	12.2	11.3	9.9	6.6	4.4	0.9	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.1	0.2	3.0	5.0	4.0	3.4	3.9	6.7	10.0	10.3	10.4	10.2	12.1	10.9	7.2	2.3	0.4	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	1.1	10.8	21.5	35.6	28.9	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	3.0	4.4	9.8	11.2	17.5	20.0	14.2	14.2	5.2	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	3.0	4.5	9.8	11.2	19.4	19.9	15.4	12.8	3.8	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	1.8	5.6	1.7	1.8	1.4	2.2	7.7	22.5	17.2	21.0	11.5	4.8	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	12.5	0.0	0.0	1.1	0.0	58.9	0.0	0.0	27.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	1.1	3.3	1.4	3.2	2.8	2.6	1.7	28.4	51.1	0.3	1.4	1.4	0.8	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	6.2	0.2	0.6	1.1	3.2	3.6	5.1	8.4	12.6	7.7	10.1	14.3	11.2	14.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.2	3.0	5.0	4.0	3.4	3.9	6.7	10.0	10.3	10.4	10.2	12.1	10.9	7.2	2.3	0.4	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	57.1	42.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using covariance calibration																									
GLM (OLS)	0.0	0.0	0.1	0.4	1.9	0.6	1.4	3.9	4.4	4.1	10.2	11.7	13.2	12.6	9.7	10.4	7.6	4.7	2.7	0.5	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.2	2.0	0.7	1.0	3.9	4.7	4.4	10.0	12.4	13.2	13.1	9.5	10.6	7.0	4.7	2.4	0.3	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.3	1.8	0.7	1.4	3.9	4.5	4.1	10.1	12.0	13.0	13.3	9.1	10.7	7.2	4.8	2.6	0.4	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.3	1.8	0.7	1.4	3.9	4.5	4.1	10.1	12.0	13.0	13.2	9.1	10.7	7.2	4.8	2.6	0.4	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	0.4	1.9	0.6	1.4	3.9	4.4	4.1	10.2	11.7	13.2	12.6	9.7	10.4	7.6	4.7	2.7	0.5	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.9	0.0	0.0	79.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.1	3.0	4.3	3.7	4.0	9.6	32.5	6.5	32.6	1.8	0.6	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	1.2	0.4	0.3	1.1	2.1	3.4	3.9	3.4	8.3	7.2	13.6	11.9	8.3	15.5	4.8	6.7	5.7	2.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.4	5.4	8.6	12.4	10.6	22.7	18.3	20.4	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	0.4	1.9	0.6	1.4	3.9	4.4	4.1	10.2	11.7	13.2	12.6	9.7	10.4	7.6	4.7	2.7	0.5	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	57.5	42.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100

Table D.6: Income class frequencies (in percentage points) estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model and total calibration																									
GLM (OLS)	0.2	1.5	0.7	0.5	1.2	2.7	3.3	3.7	3.4	4.5	9.1	9.8	9.0	9.4	7.4	7.7	6.4	6.7	4.8	4.0	2.5	1.4	0.0	0.0	0.0
GLM (Ridge)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.7	0.4	0.3	2.3	2.8	2.8	2.7	2.0	4.7	3.7	12.5	4.0	3.6	9.3	4.9	4.7	5.7	9.8	5.8	9.0	2.2	4.2	1.7	0.0	0.0
Regression tree	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	2.3	0.3	0.3	1.9	2.6	3.1	2.1	2.0	3.9	4.6	10.1	6.0	7.6	11.3	7.8	6.2	7.0	6.5	5.2	4.2	2.6	2.3	0.0	0.0	0.0
GAMM (fix knots)	80.7	0.0	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model and total calibration																									
GLM (OLS)	2.4	0.7	0.0	0.7	0.7	0.7	0.0	0.7	0.7	0.7	0.8	0.9	1.7	2.0	1.3	1.4	2.8	2.9	2.9	3.2	3.4	10.0	17.5	39.2	2.4
GLM (Ridge)	7.0	0.9	1.0	1.5	1.4	1.5	2.3	1.4	2.6	1.5	2.8	3.2	4.0	3.3	3.3	4.9	5.2	5.6	6.9	5.2	6.7	14.8	12.0	1.0	0.0
GLM (LASSO)	8.1	0.0	0.9	1.2	0.6	0.7	1.3	1.3	1.4	0.7	2.1	1.4	1.4	1.4	2.3	2.5	3.4	3.3	3.4	3.3	3.3	10.6	15.7	29.3	0.3
GLM (Elastic net)	6.0	0.7	0.7	0.8	0.2	1.2	1.3	1.3	0.7	1.3	2.1	2.1	1.4	2.1	2.2	2.3	3.3	3.4	4.4	3.3	4.1	11.7	17.5	25.9	0.0
GAM (fix knots)	57.9	0.0	0.0	1.0	0.7	0.0	1.6	1.4	1.5	2.8	0.9	0.0	2.9	4.5	0.0	0.7	0.0	0.8	0.7	2.4	3.5	1.9	3.7	1.7	9.4
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.0	0.0	0.0	0.0	0.0	19.0	0.0	0.0	0.0	0.0	0.0	0.0	6.9	15.7	33.4
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	15.9	1.0	0.8	1.0	1.1	1.1	1.7	2.0	3.0	0.8	1.5	4.5	3.5	1.9	4.8	4.8	2.4	3.1	4.2	1.5	1.8	5.7	5.0	9.4	17.3
SVM	0.0	0.0	0.1	0.5	2.0	4.1	5.8	10.5	14.2	8.3	12.8	16.9	16.2	8.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model and total calibration																									
GLM (OLS)	0.1	0.1	0.3	1.8	0.7	1.8	2.9	3.5	4.4	4.5	10.6	10.2	12.8	11.5	8.3	10.0	7.2	5.3	2.9	1.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	4.2	0.3	0.5	1.1	3.0	0.9	5.6	2.8	4.7	4.3	15.5	6.8	8.2	11.6	6.8	8.0	4.6	7.5	1.5	2.1	0.0	0.0	0.0	0.0	0.0
Regression tree	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	1.2	1.1	0.5	2.7	2.5	4.0	2.8	4.5	12.0	9.7	10.8	13.1	8.8	9.0	7.4	5.4	3.0	1.2	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	42.5	57.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model and total calibration																									
GLM (OLS)	2.4	0.0	0.7	0.7	0.7	0.7	0.0	0.7	0.7	0.7	0.8	1.7	1.1	1.9	2.1	2.8	2.2	2.8	4.0	3.4	3.4	11.5	19.3	35.2	0.6
GLM (Ridge)	8.0	1.3	1.0	1.2	1.6	2.0	1.9	1.9	1.8	1.9	3.4	3.0	3.4	3.2	3.2	4.9	4.8	5.7	6.5	5.3	6.6	13.7	12.4	1.2	0.0
GLM (LASSO)	5.2	0.7	0.0	0.7	0.7	0.8	0.9	1.2	1.3	1.3	2.1	2.1	2.1	2.2	2.2	2.3	3.3	4.4	4.1	4.2	4.0	12.6	19.7	21.9	0.0
GLM (Elastic net)	6.0	0.7	0.0	0.7	0.8	0.9	0.7	1.1	1.3	1.3	2.1	2.1	1.4	2.2	2.2	3.2	3.4	3.4	4.1	4.2	3.2	12.4	19.2	23.5	0.0
GAM (fix knots)	57.9	0.0	0.0	1.0	0.7	0.7	2.4	0.7	3.0	0.0	1.5	1.5	0.7	4.5	0.7	0.7	0.0	0.8	0.7	3.4	2.5	1.9	3.7	1.7	9.4
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.0	0.0	0.0	0.0	0.0	19.0	0.0	0.0	0.0	0.0	0.0	0.0	6.9	15.7	33.4
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	16.5	1.0	1.5	0.5	0.8	1.1	1.7	0.7	2.0	1.9	2.4	3.0	5.1	3.3	5.4	5.3	1.6	1.5	3.1	2.0	2.2	6.1	5.0	11.3	15.1
SVM	2.5	0.5	0.6	1.2	2.2	3.9	5.5	9.2	14.2	7.2	10.5	14.6	14.7	10.7	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model and total calibration																									
GLM (OLS)	0.0	0.1	0.1	1.6	0.9	0.7	3.1	3.7	4.5	4.5	12.7	13.1	13.9	12.5	9.8	10.1	5.4	2.8	0.6	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	1.2	1.2	0.7	3.1	3.0	5.1	4.7	12.6	14.1	13.9	12.7	9.5	10.1	4.9	2.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.1	1.5	0.9	0.7	3.1	3.6	4.4	4.7	12.6	13.4	14.2	12.5	9.7	10.0	5.2	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.1	1.5	0.9	0.7	3.1	3.6	4.4	4.7	12.6	13.4	14.2	12.5	9.7	10.0	5.2	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	1.7	0.7	1.2	2.8	4.2	3.6	4.9	13.0	15.3	13.5	12.0	9.6	9.8	4.7	2.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	51.5	0.0	0.0	21.1	0.0	27.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	2.4	3.1	3.7	3.0	6.0	35.3	4.4	5.8	7.5	26.7	0.5	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.1	0.1	1.9	0.7	1.2	2.9	4.2	4.4	3.7	18.6	13.1	9.4	15.9	8.4	9.3	4.1	1.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	2.8	10.4	13.3	23.2	19.5	29.4	0.9	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	1.6	0.9	0.7	3.1	3.7	4.5	4.5	12.7	13.1	13.9	12.5	9.8	10.1	5.4	2.8	0.6	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.1	0.3	1.8	0.6	0.8	2.1	4.9	3.8	6.4	11.5	12.9	12.1	11.5	12.1	8.9	6.1	3.2	0.8	0.0	0.0	0.0	0.0	0.0

Table D.6: Income class frequencies (in percentage points) estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model and covariance calibration																									
GLM (OLS)	0.0	0.0	0.1	0.8	1.6	0.7	0.9	3.6	5.1	4.5	8.7	14.0	11.2	9.9	9.2	8.9	6.4	6.3	4.8	2.8	0.3	0.0	0.0	0.0	0.0
GLM (Ridge)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	1.5	0.9	0.6	1.5	3.5	5.1	4.1	9.4	14.0	11.3	9.8	8.9	9.2	6.0	5.8	5.4	2.7	0.1	0.0	0.0	0.0	0.0
Regression tree	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	0.8	1.6	0.7	0.9	3.6	5.1	4.5	8.7	14.0	11.2	9.9	9.2	8.9	6.4	6.3	4.8	2.8	0.3	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model and covariance calibration																									
GLM (OLS)	0.1	0.1	1.1	1.0	1.2	2.5	3.3	2.5	4.1	5.4	9.4	9.4	11.0	10.8	10.7	11.5	8.3	5.6	1.8	0.3	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.1	0.1	1.1	0.9	0.8	2.6	3.3	2.9	4.1	5.4	9.6	9.7	10.6	11.5	11.6	10.9	8.3	4.9	1.6	0.2	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.1	1.1	1.0	1.2	2.5	3.3	2.5	4.1	5.4	9.4	9.7	10.8	10.7	10.7	12.0	8.1	5.3	1.8	0.3	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.1	1.1	1.0	1.2	2.5	3.3	2.5	4.1	5.4	9.4	9.7	10.8	10.7	10.7	12.0	8.1	5.3	1.8	0.3	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	1.2	0.6	0.4	0.9	2.8	1.7	2.3	2.9	4.1	3.9	9.9	16.9	16.5	13.6	11.0	8.7	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	35.6	0.0	0.0	0.0	64.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	12.8	1.5	3.0	3.3	4.8	5.1	4.0	5.4	4.2	5.1	6.3	41.0	1.8	0.4	0.4	0.5	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.7	1.3	1.1	2.0	3.3	2.6	5.0	4.0	5.8	16.6	16.9	13.2	14.9	7.1	5.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.4	17.6	33.4	20.5	20.2	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	1.1	1.0	1.2	2.5	3.3	2.5	4.1	5.4	9.4	9.4	11.0	10.8	10.7	11.5	8.3	5.6	1.8	0.3	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100
Weighting model: Logit model (fix knots) and GREG, using a propensity model and covariance calibration																									
GLM (OLS)	0.0	0.0	0.1	0.4	2.0	1.7	3.4	2.9	5.5	7.2	14.4	14.6	16.0	11.2	8.9	7.7	3.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.0	0.7	1.4	1.1	2.2	2.6	3.9	4.6	9.0	21.8	10.5	13.6	16.1	3.4	9.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	0.4	2.0	1.7	3.4	2.9	5.5	7.2	14.4	14.6	16.0	11.2	8.9	7.7	3.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model and covariance calibration																									
GLM (OLS)	0.0	0.0	0.1	0.8	1.6	0.7	2.8	3.4	5.1	5.3	11.5	13.0	11.9	8.5	5.5	5.0	1.1	1.7	3.8	5.6	6.1	6.3	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	1.2	1.7	1.8	5.6	8.5	16.0	17.6	12.3	7.9	3.6	0.3	0.7	6.5	7.4	7.4	1.4	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.8	1.6	0.7	2.8	3.3	5.3	5.3	12.2	12.3	12.4	8.1	5.9	4.8	0.9	1.7	4.1	5.4	6.3	6.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.4	2.0	0.8	2.3	3.7	5.3	5.4	12.4	12.7	12.2	8.5	5.5	4.7	0.7	1.9	4.1	5.6	6.3	5.7	0.0	0.0	0.0
GAM (fix knots)	0.8	1.0	0.4	0.5	1.8	3.1	2.6	2.6	5.1	8.0	16.3	10.5	5.4	12.8	4.5	0.9	0.1	2.0	2.2	7.9	3.9	7.4	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	38.0	0.0	0.0	0.0	32.5	22.4	6.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1
MARS	0.6	0.0	0.2	0.1	0.0	0.0	0.4	38.0	0.0	0.2	0.2	35.8	0.4	0.0	3.2	0.0	0.5	0.2	0.2	0.4	0.5	19.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.2	0.4	1.2	1.3	4.8	3.0	8.5	6.2	9.5	13.1	8.6	7.5	10.1	1.5	0.7	0.0	2.1	3.5	7.5	5.7	4.7	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	2.0	7.8	16.8	29.9	18.8	23.6	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	1.2	1.2	0.4	2.4	2.8	4.1	2.7	5.1	14.3	10.0	10.9	10.2	5.6	4.4	1.1	2.4	2.5	5.9	6.4	6.4	0.0	0.0	0.0
GAMM (fix knots)	80.7	0.0	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model and covariance calibration																									
GLM (OLS)	0.0	0.0	0.1	1.2	1.2	0.7	3.1	3.3	4.3	5.0	11.8	13.8	13.8	12.5	10.4	9.8	5.5	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.7	1.5	0.7	2.7	3.4	4.7	4.8	12.1	14.5	14.5	12.5	9.9	10.2	4.8	2.6	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	1.2	1.2	0.6	3.1	3.4	4.3	5.0	11.9	14.3	13.3	13.6	9.4	10.5	5.1	2.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	1.2	1.2	0.6	3.1	3.4	4.3	5.0	11.9	14.3	13.3	13.6	9.4	10.5	5.1	2.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	1.2	1.2	0.7	3.1	3.3	4.3	5.0	11.8	13.8	13.8	12.5	10.4	9.8	5.5	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	70.5	0.0	0.0	29.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	1.2	3.7	3.7	3.7	4.6	36.6	3.8	6.5	7.5	26.8	0.6	0.5	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.0	1.8	1.1	1.6	6.5	5.1	7.0	5.1	8.7	18.3	8.6	10.8	14.1	3.9	5.9	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	2.3	9.3	15.3	26.0	19.2	26.6	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	1.2	1.2	0.7	3.1	3.3	4.3	5.0	11.8	13.8	13.8	12.5	10.4	9.8	5.5	2.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	89.3	9.3

Table D.6: Income class frequencies (in percentage points) estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using total and covariance calibration																									
GLM (OLS)	0.1	0.1	0.7	1.4	0.6	2.5	2.9	3.7	3.2	5.2	10.1	10.5	11.3	13.0	10.5	10.4	7.3	4.4	1.8	0.3	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.7	1.1	0.5	1.6	2.4	3.0	2.6	3.8	5.2	10.9	14.1	13.1	12.5	10.0	9.7	5.4	3.2	0.3	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	0.7	1.4	0.6	2.5	2.9	3.7	3.2	5.2	10.1	10.5	11.3	13.0	10.5	10.4	7.6	4.1	1.8	0.3	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.1	1.2	1.2	1.0	3.5	2.9	5.5	4.9	13.3	13.0	14.9	11.7	9.2	10.0	4.7	2.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	1.7	0.7	1.7	2.9	3.9	4.1	5.4	14.0	15.0	13.4	12.2	8.8	9.3	4.1	2.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	1.2	1.2	1.0	3.5	2.9	5.5	4.9	13.3	13.0	14.9	11.7	9.2	10.0	4.7	2.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.0	0.2	2.3	4.1	4.3	6.0	7.5	17.0	12.6	9.7	7.8	4.6	3.8	4.7	4.6	7.3	3.5	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9	27.1	11.9	32.7	2.9	10.1	7.8	5.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.2	2.2	4.1	4.3	5.9	7.9	16.7	12.6	10.1	7.8	4.2	3.8	4.7	4.8	7.3	3.3	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.2	2.2	4.1	4.3	5.9	7.9	16.7	12.6	10.1	7.8	4.2	3.8	4.7	4.8	7.3	3.3	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	2.1	2.9	2.4	2.1	8.5	10.8	12.2	9.9	14.5	5.4	4.7	3.2	5.7	3.7	8.1	3.2	0.4	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	76.5	0.0	8.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.2	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	72.0	0.0	6.5	0.5	0.5	0.0	0.5	0.5	0.5	0.5	0.9	0.4	0.4	2.0	3.4	8.3	3.5
ANN (opt. knots)	0.0	0.0	0.0	0.3	1.8	3.0	3.1	8.6	9.3	10.4	12.0	7.9	12.0	6.9	1.2	2.6	5.2	4.6	9.3	1.8	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	16.3	39.3	33.7	7.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	1.8	2.4	3.6	1.6	6.5	9.9	13.8	11.1	11.8	6.8	5.6	3.7	4.7	4.7	7.3	4.0	0.6	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	0.0	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	63.2	6.7	4.9	6.1	2.8	3.8	5.9	3.6	1.7	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	3.1	4.1	4.4	4.2	4.4	4.2	5.4	8.4	10.0	9.4	8.2	19.1	14.3	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	52.2	8.8	6.9	7.9	7.8	8.0	6.6	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	7.0	0.5	0.2	1.0	2.0	4.1	5.1	8.2	13.3	7.2	9.1	13.0	11.0	10.7	3.0	0.9	1.5	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.1	1.5	0.9	0.6	3.0	3.9	3.7	3.9	11.0	9.8	12.1	12.8	8.8	10.1	7.5	5.5	3.4	1.3	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	1.2	1.1	0.6	2.7	3.9	3.9	4.1	11.1	10.7	12.7	12.8	8.2	10.2	7.1	5.4	3.2	1.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	1.5	0.8	0.7	2.7	4.3	3.7	4.0	10.9	9.8	12.6	12.6	9.0	10.1	7.1	5.7	3.1	1.3	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	1.5	0.8	0.7	2.7	4.3	3.7	4.0	10.9	9.8	12.5	12.7	8.9	10.2	7.1	5.7	3.1	1.3	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	1.5	0.9	0.6	3.0	3.9	3.7	3.9	11.0	9.8	12.1	12.8	8.8	10.1	7.5	5.5	3.4	1.3	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.9	0.0	0.0	79.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	48.1	0.0	0.0	1.7	46.1	0.3	1.7	0.5	0.5	0.3	0.3	0.2	0.1	0.1	0.1	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.5	1.8	4.8	7.8	7.0	16.7	11.7	11.1	15.2	4.1	5.6	7.3	5.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.3	39.2	48.7	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	1.5	0.9	0.6	3.0	3.9	3.7	3.9	11.0	9.8	12.1	12.8	8.8	10.1	7.5	5.5	3.4	1.3	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	57.5	42.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.3	1.8	0.6	0.6	3.9	4.0	3.7	8.0	11.2	11.8	11.5	10.2	9.0	8.8	7.0	4.4	2.6	0.5	0.0	0.0	0.0	0.0

Table D.6: Income class frequencies (in percentage points) estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model, total and covariance calibration																									
GLM (OLS)	0.0	0.1	0.1	1.8	0.8	0.7	3.4	3.0	5.3	5.2	11.8	14.1	12.9	12.1	9.6	10.1	5.8	2.8	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	1.8	0.8	0.7	3.4	3.0	5.3	5.2	11.8	14.1	12.9	12.1	9.6	10.1	5.8	2.8	0.5	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	1.8	0.8	0.7	3.4	3.0	5.3	5.2	11.8	14.1	12.9	12.1	9.6	10.1	5.8	2.8	0.5	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	80.7	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model, total and covariance calibration																									
GLM (OLS)	2.4	0.7	0.0	0.7	0.7	0.7	0.0	0.7	0.7	0.7	0.8	0.9	1.7	2.0	1.3	1.4	2.8	2.9	2.9	3.2	3.4	10.0	17.5	39.2	2.4
GLM (Ridge)	9.9	0.9	1.0	2.0	1.6	1.3	1.6	2.5	1.5	1.6	2.9	2.7	3.1	3.2	3.2	3.8	4.8	5.8	5.3	5.9	5.5	13.8	13.7	2.6	0.0
GLM (LASSO)	8.1	0.0	0.9	1.2	0.6	0.7	1.3	1.3	1.4	0.7	2.1	1.4	1.4	1.4	2.3	2.5	3.4	3.3	3.4	3.3	3.3	10.6	15.7	29.3	0.3
GLM (Elastic net)	7.3	0.1	0.7	0.9	1.2	0.6	1.3	0.7	1.3	1.4	1.4	2.2	2.1	1.4	1.5	3.2	2.4	3.4	4.4	3.3	3.3	10.6	17.8	27.3	0.2
GAM (fix knots)	57.9	0.0	0.0	1.0	0.7	0.0	1.6	1.4	1.5	2.8	0.9	0.0	2.9	4.5	0.0	0.7	0.0	0.8	0.7	2.4	3.5	1.9	3.7	1.7	9.4
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.0	0.0	0.0	0.0	0.0	19.0	0.0	0.0	0.0	0.0	0.0	0.0	6.9	15.7	33.4
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	21.4	0.1	1.3	1.1	0.3	0.6	2.6	2.4	2.2	2.9	2.7	2.0	4.4	1.9	1.1	6.8	2.4	1.6	2.3	3.4	0.7	6.1	4.2	8.5	16.8
SVM	0.0	0.0	0.1	0.5	2.0	4.1	5.8	10.2	14.5	8.3	12.8	16.7	16.4	8.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model, total and covariance calibration																									
GLM (OLS)	0.0	0.1	0.1	1.2	1.2	0.8	3.1	3.8	4.6	4.5	12.2	14.1	13.7	13.2	9.6	9.8	4.9	2.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	1.1	0.8	0.4	1.1	2.5	3.1	3.5	4.3	10.2	13.1	15.1	7.7	23.6	2.2	4.4	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	2.0	0.6	1.5	2.9	4.6	2.6	4.2	16.5	12.4	11.4	17.6	6.8	9.8	4.6	2.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	18.9	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model, total and covariance calibration																									
GLM (OLS)	2.4	0.0	0.7	0.7	0.7	0.7	0.0	0.7	0.7	0.7	0.8	1.7	1.1	1.9	2.1	2.8	2.2	2.8	4.0	3.4	3.4	11.5	19.3	35.2	0.6
GLM (Ridge)	8.0	1.3	1.0	1.2	1.6	2.0	1.9	1.9	1.8	1.9	3.4	3.0	3.4	3.2	3.2	4.9	4.8	5.7	6.5	5.3	6.6	13.7	12.4	1.2	0.0
GLM (LASSO)	6.7	0.7	0.0	0.7	0.8	0.9	1.1	1.3	1.3	0.6	2.1	1.4	2.2	2.1	1.5	3.2	3.4	3.4	3.4	4.0	4.2	11.6	17.5	25.9	0.0
GLM (Elastic net)	1.8	0.0	0.6	1.4	0.7	0.7	0.7	0.7	1.4	0.9	2.4	1.3	2.8	2.8	2.9	2.9	4.8	5.1	5.1	4.9	5.2	15.9	21.1	13.7	0.0
GAM (fix knots)	57.9	0.0	0.0	1.0	0.7	0.7	2.4	0.7	3.0	0.0	1.5	1.5	0.7	4.5	0.7	0.7	0.0	0.8	0.7	3.4	2.5	1.9	3.7	1.7	9.4
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.0	0.0	0.0	0.0	0.0	19.0	0.0	0.0	0.0	0.0	0.0	0.0	6.9	15.7	33.4
MARS	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	2.5	0.5	0.6	1.2	2.2	4.0	5.4	9.8	13.6	7.2	10.5	14.6	14.7	10.7	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model, total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.1	1.2	1.2	0.6	2.8	3.7	4.3	4.4	12.2	13.3	14.3	12.4	10.5	9.8	5.6	3.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.8	1.5	0.7	2.7	3.1	5.0	4.4	12.3	14.1	14.1	13.1	9.9	10.2	4.9	2.7	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	1.2	1.2	0.6	2.8	3.7	4.3	4.4	12.2	13.3	14.3	12.4	10.5	10.4	5.0	3.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	1.2	1.2	0.6	2.8	3.7	4.3	4.4	12.2	13.3	14.3	12.4	10.5	10.4	5.0	3.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	1.2	1.2	0.6	2.8	3.7	4.3	4.4	12.2	13.3	14.3	12.4	10.5	9.8	5.6	3.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	70.5	0.0	0.0	29.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	1.2	3.7	3.7	3.1	5.2	35.9	4.4	6.6	32.4	1.9	0.6	0.5	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.1	0.8	1.1	4.2	2.5	8.2	5.8	22.4	8.6	11.7	19.0	4.6	7.7	2.9	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	2.3	8.2	16.9	27.8	19.0	24.7	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	1.2	1.2	0.6	2.8	3.7	4.3	4.4	12.2	13.3	14.3	12.4	10.5	9.8	5.6	3.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	61.8	38.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.1	1.9	0.7	0.7	1.7	5.2	3.6	5.9	11.6	13.2	13.0	11.3	11.8	9.3	6.2	3.0	0.8	0.0	0.0	0.0	0.0	0.0

Table D.7: Estimated standard deviations (in percentage points) for income class frequencies estimated by weighted aggregation of predictions in the WI

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: unweighted, using no auxiliary information																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.3	0.4	0.3	0.4	0.4	0.8	0.3	0.3	0.3	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.1	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.4	0.2	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.1	0.1	0.3	0.3	0.3	0.4	0.4	0.6	0.6	0.6	1.5	1.3	0.4	0.3	0.3	0.2	0.2	0.1	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.3	0.3	0.3	0.4	0.2	0.2	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (parametric), using propensity weights																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.5	0.6	0.6	0.5	0.3	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.2	0.2	0.3	0.4	0.5	0.4	0.7	0.7	1.2	0.8	0.4	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.5	0.6	0.6	0.5	0.3	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.5	0.6	0.6	0.5	0.3	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.5	0.6	0.6	0.5	0.3	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.4	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.3	0.3	0.2	0.2	0.7	0.2	0.4	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.3	0.3	0.3	0.5	0.4	0.8	0.5	0.8	1.9	1.7	0.6	0.4	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.6	0.6	0.4	0.6	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.5	0.6	0.6	0.5	0.3	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.5	0.6	0.6	0.5	0.3	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Pseudo-Weights (parametric), using propensity weights																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.3	0.4	0.3	0.5	0.4	0.8	0.4	0.3	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.1	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.6	0.1	0.2	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.0	0.1	0.3	0.3	0.3	0.4	0.4	0.7	0.6	0.6	1.5	1.4	0.4	0.3	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.3	0.2	0.5	0.3	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using propensity weights																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.3	0.4	0.5	0.5	0.7	1.1	0.5	0.4	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.3	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.3	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.5	0.2	0.3	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.3	0.3	0.2	0.2	0.4	0.8	0.7	0.7	2.7	2.6	0.6	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.5	0.3	0.3	0.6	0.3	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots), using propensity weights																									
None (weighted WI)	0.4	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.3	0.3	0.5	0.6	0.7	0.5	0.3	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.2	0.3	0.3	0.4	0.5	0.5	0.4	0.6	0.6	1.1	0.7	0.4	0.6	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.3	0.3	0.5	0.6	0.7	0.5	0.3	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.3	0.3	0.5	0.6	0.7	0.5	0.3	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.3	0.3	0.5	0.6	0.7	0.5	0.3	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.6	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.2	0.1	0.1	0.1	0.3	0.4	0.4	0.4	0.3	0.6	0.3	0.5	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.2	0.1	0.1	0.3	0.4	0.6	1.0	0.6	0.6	0.6	0.9	1.8	1.5	0.6	0.6	0.5	0.2	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0
SVM	0.2	0.1	0.1	0.2	0.2	0.2	0.3	0.6	0.5	0.6	0.7	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.3	0.3	0.5	0.6	0.7	0.5	0.3	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.3	0.3	0.5	0.6	0.7	0.5	0.3	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.7: Estimated standard deviations (in percentage points) for income class frequencies estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Pseudo-Weights (fix knots), using propensity weights																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.2	0.3	0.3	0.3	0.3	0.4	0.4	0.6	0.4	0.7	0.4	0.3	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.2	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.2	0.2	0.5	0.2	0.2	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.1	0.1	0.5	0.5	0.4	0.5	0.5	0.8	0.7	0.6	1.4	1.2	0.4	0.3	0.3	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.5	0.3	0.3	0.5	0.3	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using propensity weights																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.3	0.4	0.5	0.5	0.7	1.1	0.5	0.4	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.3	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.3	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.5	0.2	0.3	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.3	0.3	0.2	0.2	0.4	0.8	0.7	0.7	2.7	2.6	0.6	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.5	0.3	0.3	0.6	0.3	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using propensity weights																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.3	0.4	0.5	0.5	0.7	1.1	0.5	0.4	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.3	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.3	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.5	0.2	0.3	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.3	0.3	0.2	0.2	0.4	0.8	0.7	0.7	2.7	2.6	0.6	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.5	0.3	0.3	0.6	0.3	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Subsampling, using total calibration																									
None (weighted WI)	1.9	0.2	0.1	0.3	0.3	0.4	0.6	0.6	0.7	0.7	0.8	0.8	0.6	0.5	0.4	0.4	0.3	0.3	0.2	0.1	0.1	0.1	0.0	0.0	0.2
GLM (OLS)	0.0	0.1	0.1	0.3	0.4	0.5	0.7	0.9	0.9	1.0	1.3	1.5	1.5	1.5	1.0	1.2	0.6	0.5	0.3	0.1	0.1	0.1	0.0	0.0	0.0
GLM (Ridge)	0.0	0.1	0.1	0.3	0.4	0.5	0.7	0.9	1.0	0.9	1.4	1.5	1.7	1.4	1.1	1.3	0.6	0.5	0.2	0.1	0.2	0.1	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.1	0.3	0.4	0.5	0.7	0.9	0.9	1.0	1.3	1.5	1.5	1.5	1.1	1.2	0.6	0.5	0.3	0.1	0.2	0.1	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.1	0.3	0.4	0.5	0.7	0.9	0.9	1.0	1.3	1.5	1.5	1.5	1.1	1.2	0.6	0.5	0.3	0.1	0.2	0.1	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.1	0.3	0.4	0.5	0.7	0.9	0.9	1.0	1.3	1.5	1.5	1.5	1.0	1.2	0.6	0.5	0.3	0.1	0.1	0.1	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9	0.0	1.0	0.0	2.1	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.5	0.0
MARS	0.1	0.1	0.2	0.3	0.4	0.7	0.7	0.8	0.7	1.8	0.9	1.2	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.1	0.1
ANN (opt. knots)	0.1	0.1	0.1	0.4	0.4	0.6	1.0	0.9	1.1	0.8	1.6	1.8	1.7	1.6	1.4	1.4	0.6	0.6	0.2	0.2	0.2	0.1	0.0	0.0	0.0
SVM	0.0	0.0	0.1	0.1	0.2	0.4	0.7	1.3	1.5	1.2	1.8	1.5	1.5	0.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.1	0.3	0.4	0.5	0.7	0.9	0.9	1.0	1.3	1.5	1.5	1.5	1.0	1.2	0.6	0.5	0.3	0.1	0.1	0.1	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.1	0.1	0.3	0.4	0.5	0.7	0.9	0.9	1.0	1.3	1.5	1.5	1.5	1.0	1.2	0.6	0.5	0.3	0.1	0.1	0.1	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Post-stratification, using total calibration																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.3	0.4	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.4	0.4	0.3	0.6	0.5	0.8	0.5	0.3	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.3	0.4	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.3	0.4	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.2	0.2	0.7	0.2	0.2	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.1	0.1	0.1	0.4	0.4	0.4	0.5	0.5	0.7	0.7	0.8	1.4	1.2	0.4	0.5	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.4	0.3	0.2	0.4	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.3	0.4	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.3	0.4	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.7: Estimated standard deviations (in percentage points) for income class frequencies estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using total calibration																									
None (weighted WI)	0.5	0.1	0.0	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.6	0.5	0.5	0.8	0.8	0.7	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.4	0.6	0.6	0.6	0.9	1.6	0.9	0.7	0.8	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.6	0.5	0.5	0.8	0.9	0.7	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.6	0.5	0.5	0.7	0.8	0.7	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.6	0.5	0.5	0.8	0.8	0.7	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.3	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.9	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.3	1.3	0.8	0.9	2.3	2.5	1.0	0.7	0.6	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.5	0.0	0.9	0.6	0.6	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.6	0.5	0.5	0.8	0.8	0.7	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.6	0.5	0.5	0.8	0.8	0.7	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using total calibration																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.2	0.1	0.2	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.3	0.4	0.5	0.5	0.8	1.1	0.5	0.3	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.3	0.4	0.6	0.5	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.3	0.4	0.6	0.5	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.1	0.2	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.1	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.5	0.2	0.3	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.3	0.3	0.2	0.2	0.4	0.8	0.7	0.8	3.0	2.8	0.5	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.5	0.3	0.3	0.6	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.2	0.1	0.2	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.1	0.2	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using total calibration																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.3	0.4	0.3	0.5	0.6	0.6	0.6	0.4	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.3	0.4	0.3	0.6	0.7	1.4	0.8	0.5	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.3	0.4	0.3	0.5	0.6	0.6	0.6	0.4	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.3	0.4	0.3	0.5	0.6	0.6	0.6	0.4	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.3	0.4	0.3	0.5	0.6	0.6	0.6	0.4	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.2	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.1	0.6	0.2	0.3	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.3	0.3	0.7	0.4	0.9	2.1	1.8	0.7	0.5	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.5	0.3	0.7	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.3	0.4	0.3	0.5	0.6	0.6	0.6	0.4	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.3	0.4	0.3	0.5	0.6	0.6	0.6	0.4	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using total calibration																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.5	0.7	0.6	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.2	0.1	0.3	0.4	0.5	0.4	0.7	0.8	1.3	0.9	0.6	0.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.6	0.7	0.6	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.6	0.7	0.6	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.5	0.7	0.6	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.4	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.3	0.3	0.2	0.2	0.8	0.2	0.4	1.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.1	0.2	0.2	0.4	0.3	0.8	0.5	0.9	2.0	1.7	1.0	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.8	0.6	0.5	0.8	0.5	0.9	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.5	0.7	0.6	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.4	0.4	0.4	0.5	0.7	0.6	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using total calibration																									
None (weighted WI)	0.5	0.1	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.1	0.3	0.6	0.3	0.3	0.6	0.9	1.1	0.8	2.2	1.8	1.1	1.3	1.5	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.2	0.7	0.3	0.3	0.6	0.9	1.0	1.1	2.8	2.8	1.1	1.4	1.4	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.3	0.6	0.3	0.3	0.5	0.9	1.0	0.9	2.2	1.7	1.1	1.3	1.5	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.3	0.6	0.3	0.3	0.5	0.9	1.0	0.9	2.2	1.8	1.1	1.3	1.5	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	0.3	0.6	0.3	0.3	0.6	0.9	1.1	0.8	2.2	1.8	1.1	1.3	1.5	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.1	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.1	0.1	0.3	0.7	0.6	0.6	0.8	1.6	1.0	1.9	4.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.3	0.6	0.6	0.4	0.7	0.7	1.3	1.4	1.4	3.1	3.3	0.6	1.8	0.9	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.1	0.0	0.1	0.3	0.8	1.1	1.1	2.9	4.5	1.2	0.7	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	0.3	0.6	0.3	0.3	0.6	0.9	1.1	0.8	2.2	1.8	1.1	1.3	1.5	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.1	0.3	0.6	0.3	0.3	0.6	0.9	1.1	0.8	2.2	1.8	1.1	1.3	1.5	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.7: Estimated standard deviations (in percentage points) for income class frequencies estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using covariance calibration																									
None (weighted WI)	0.7	0.0	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.0	0.1	0.0	0.1
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.2	1.6	0.9	0.8	1.2	0.7	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	2.0	3.0	1.4	1.7	1.8	1.3	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.3	1.6	0.9	0.8	1.2	0.6	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.2	1.6	0.9	0.8	1.2	0.6	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.2	1.6	0.9	0.8	1.2	0.7	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9	0.0	3.1	4.7	6.9	1.7	1.8	2.1	1.1	0.3	0.0	0.0	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.2	0.0	0.4	1.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.2	1.6	0.9	0.8	1.2	0.7	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.2	1.6	0.9	0.8	1.2	0.7	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using covariance calibration																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.3	0.4	0.4	0.4	0.5	0.6	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.3	0.5	0.5	0.5	0.6	1.1	0.6	0.3	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.3	0.4	0.4	0.4	0.5	0.6	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.3	0.4	0.4	0.4	0.5	0.6	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.3	0.4	0.4	0.4	0.5	0.6	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.5	0.3	0.4	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.2	0.2	0.2	0.2	0.5	0.9	0.8	0.6	2.1	2.1	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.5	0.4	0.4	0.6	0.4	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.3	0.4	0.4	0.4	0.5	0.6	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.3	0.4	0.4	0.4	0.5	0.6	0.4	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using covariance calibration																									
None (weighted WI)	1.6	0.1	0.1	0.3	0.3	0.2	0.2	0.2	0.2	0.5	0.4	0.6	0.5	0.4	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
GLM (OLS)	0.1	0.1	0.3	0.7	0.6	0.5	0.8	3.1	3.2	0.7	1.0	4.4	1.8	4.1	1.0	1.0	1.5	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.1	0.0	0.2	0.6	0.7	0.5	0.7	2.9	3.7	0.9	1.0	4.6	2.6	3.9	0.9	1.2	1.0	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.1	0.1	0.3	0.7	0.6	0.5	0.8	3.1	3.2	0.7	1.0	4.3	1.8	4.1	0.9	1.0	1.3	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.1	0.3	0.7	0.6	0.5	0.8	3.1	3.2	0.7	1.0	4.3	1.8	4.1	0.9	1.0	1.3	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.1	0.1	0.3	0.7	0.6	0.5	0.8	3.1	3.2	0.7	1.0	4.4	1.8	4.1	1.0	1.0	1.5	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.8	0.0	0.9	0.0	7.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
MARS	0.2	0.1	0.2	0.3	0.8	2.0	2.9	2.7	2.4	5.6	1.7	2.5	14.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.2	0.2	0.2	1.0	0.7	0.7	1.1	2.8	3.8	1.3	2.4	3.2	2.8	6.7	2.0	1.1	1.0	0.3	0.3	0.2	0.1	0.0	0.0	0.0	0.0
SVM	0.1	0.1	1.5	0.3	1.0	2.5	3.1	4.2	1.9	0.7	7.5	4.3	4.3	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	0.3	0.7	0.6	0.5	0.8	3.1	3.2	0.7	1.0	4.4	1.8	4.1	1.0	1.0	1.5	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.1	0.1	0.3	0.7	0.6	0.5	0.8	3.1	3.2	0.7	1.0	4.4	1.8	4.1	1.0	1.0	1.5	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using covariance calibration																									
None (weighted WI)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using covariance calibration																									
None (weighted WI)	2.1	0.1	0.1	0.4	0.3	0.2	0.3	0.4	0.4	0.4	0.5	0.6	0.5	0.3	0.3	0.3	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.1
GLM (OLS)	0.2	0.2	0.7	2.0	2.1	0.9	0.8	2.9	3.4	3.7	3.6	3.9	3.8	4.6	2.5	1.7	0.2	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.1	0.1	0.4	1.8	2.6	1.1	0.9	2.6	4.2	3.8	4.6	4.3	4.1	4.4	2.5	1.5	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.2	0.1	0.8	2.0	2.1	0.9	0.8	2.9	3.4	3.7	4.3	3.7	3.7	4.6	2.5	1.7	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.2	0.1	0.8	2.0	2.1	0.9	0.8	2.9	3.4	3.7	4.3	3.7	3.8	4.6	2.5	1.7	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.2	0.2	0.7	2.0	2.1	0.9	0.8	2.9	3.4	3.7	3.6	3.9	3.8	4.6	2.5	1.7	0.2	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.2	0.0	2.6	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.6	0.5	0.8	1.0	2.1	2.8	4.0	3.2	2.8	4.8	3.3	5.9	16.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.8	0.5	0.7	2.6	1.9	1.2	1.6	3.2	3.1	2.8	2.7	3.4	4.3	3.8	4.0	0.5	0.2	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
SVM	0.1	0.1	1.8	0.5	1.7	4.0	4.7	2.5	3.7	4.4	9.1	3.4	3.3	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.2	0.2	0.7	2.0	2.1	0.9	0.8	2.9	3.4	3.7	3.6	3.9	3.8	4.6	2.5	1.7	0.2	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.2	0.2	0.7	2.0	2.1	0.9	0.8	2.9	3.4	3.7	3.6	3.9	3.8	4.6	2.5	1.7	0.2	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.7: Estimated standard deviations (in percentage points) for income class frequencies estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model and total calibration																									
None (weighted WI)	0.5	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.5	0.4	0.6	0.7	0.6	0.7	0.6	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.5	0.4	0.8	0.9	1.4	0.9	0.6	0.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.5	0.4	0.6	0.7	0.6	0.7	0.6	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.5	0.4	0.6	0.7	0.6	0.7	0.6	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.5	0.4	0.6	0.7	0.6	0.7	0.6	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.3	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.8	0.1	0.2	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	1.0	0.7	1.0	2.2	2.0	0.9	0.7	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.7	0.4	0.7	0.6	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.5	0.4	0.6	0.7	0.6	0.7	0.6	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.5	0.4	0.6	0.7	0.6	0.7	0.6	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model and total calibration																									
None (weighted WI)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model and total calibration																									
None (weighted WI)	0.4	0.1	0.0	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.1	0.2	0.4	0.3	0.4	0.6	0.8	0.5	0.6	1.0	0.8	1.2	0.9	0.8	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.3	0.4	0.3	0.4	0.7	0.9	0.5	0.7	1.1	1.3	1.3	1.0	0.9	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.2	0.4	0.3	0.4	0.7	0.8	0.5	0.6	1.0	0.9	1.2	0.9	0.8	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.2	0.4	0.3	0.4	0.7	0.8	0.5	0.6	1.0	0.8	1.2	0.9	0.8	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	0.2	0.4	0.3	0.4	0.6	0.8	0.5	0.6	1.0	0.8	1.2	0.9	0.8	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.8	0.0	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.1	0.0	0.1	0.1	0.3	0.6	0.4	0.4	0.3	1.7	0.3	0.6	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.2	0.1	0.1	0.4	0.4	0.4	0.7	0.5	1.7	0.5	1.3	1.8	1.6	1.7	1.0	0.8	0.4	0.4	0.3	0.2	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.1	0.1	0.2	0.5	1.9	0.8	0.7	1.9	1.2	1.2	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.1	0.2	0.4	0.3	0.4	0.6	0.8	0.5	0.6	1.0	0.8	1.2	0.9	0.8	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.1	0.2	0.4	0.3	0.4	0.6	0.8	0.5	0.6	1.0	0.8	1.2	0.9	0.8	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model and total calibration																									
None (weighted WI)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model and total calibration																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.3	0.4	0.5	0.5	0.7	1.1	0.5	0.4	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.3	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.3	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.5	0.2	0.3	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.3	0.3	0.2	0.2	0.4	0.8	0.7	0.7	2.7	2.6	0.6	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.5	0.3	0.3	0.6	0.3	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.3	0.4	0.4	0.4	0.6	0.5	0.4	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.7: Estimated standard deviations (in percentage points) for income class frequencies estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model and covariance calibration																									
None (weighted WI)	0.6	0.0	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.1
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.6	0.7	1.5	1.3	1.1	0.8	0.8	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.6	0.7	1.9	2.2	1.7	1.1	1.3	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.6	0.7	1.5	1.3	1.1	0.8	0.9	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.6	0.7	1.5	1.3	1.1	0.8	0.9	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.6	0.7	1.5	1.3	1.1	0.8	0.8	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	1.1	1.7	6.8	7.3	0.6	1.2	1.2	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.0	1.1	0.0	0.9	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.6	0.7	1.5	1.3	1.1	0.8	0.8	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.6	0.7	1.5	1.3	1.1	0.8	0.8	0.3	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model and covariance calibration																									
None (weighted WI)	1.2	0.0	0.1	0.2	0.4	0.8	0.6	0.5	0.2	0.3	0.7	0.8	0.6	0.4	0.3	0.3	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1
GLM (OLS)	0.0	0.0	0.0	0.1	0.4	0.6	1.1	3.8	6.6	3.0	2.7	4.9	7.6	3.8	2.0	1.9	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.4	0.5	1.0	3.8	6.9	3.4	2.6	5.4	7.9	3.6	1.9	2.0	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.4	0.6	0.9	4.0	6.6	3.0	2.6	4.8	7.6	3.8	2.0	1.9	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.4	0.6	0.9	4.0	6.6	3.0	2.6	4.8	7.6	3.8	2.0	1.9	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.4	0.6	1.1	3.8	6.6	3.0	2.7	4.9	7.6	3.8	2.0	1.9	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	26.9	0.0	0.5	0.0	27.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.3	0.4	0.6	1.0	25.4	3.3	6.6	17.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.3	0.4	0.3	1.6	1.3	12.1	1.4	3.9	5.0	6.1	5.0	4.3	1.6	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.2	0.0	0.2	0.4	1.0	15.2	6.1	2.8	10.3	5.2	3.6	4.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.4	0.6	1.1	3.8	6.6	3.0	2.7	4.9	7.6	3.8	2.0	1.9	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.4	0.6	1.1	3.8	6.6	3.0	2.7	4.9	7.6	3.8	2.0	1.9	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model and covariance calibration																									
None (weighted WI)	1.2	0.1	0.0	0.1	0.3	0.7	0.4	0.6	0.3	0.4	0.4	0.5	0.5	0.4	0.4	0.3	0.2	0.2	0.1	0.1	0.0	0.1	0.0	0.0	0.1
GLM (OLS)	0.8	0.1	0.1	0.2	0.5	0.5	0.7	1.6	4.9	3.5	2.7	3.8	4.6	4.0	2.6	3.1	2.7	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.5	0.1	0.3	0.2	0.5	0.4	0.6	1.8	4.9	3.3	3.0	3.8	5.1	3.9	3.0	3.9	2.2	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.8	0.1	0.1	0.2	0.5	0.5	0.6	1.7	4.9	3.3	2.9	3.7	4.7	4.0	2.6	3.3	2.3	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.8	0.1	0.1	0.2	0.5	0.5	0.6	1.7	4.9	3.3	2.9	3.7	4.6	4.0	2.6	3.3	2.3	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.8	0.1	0.1	0.2	0.5	0.5	0.7	1.6	4.9	3.5	2.7	3.8	4.6	4.0	2.6	3.1	2.7	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.7	0.0	0.0	0.0	17.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
MARS	0.5	0.0	0.0	0.1	0.2	0.0	0.0	0.0	0.0	12.7	0.0	0.0	12.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.6	0.2	0.1	0.4	0.4	0.6	2.0	1.0	7.3	1.7	5.2	4.2	6.2	3.1	3.5	2.8	1.6	1.0	0.5	0.3	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.1	0.3	0.0	0.3	1.3	7.0	1.9	0.0	4.8	1.6	3.4	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.8	0.1	0.1	0.2	0.5	0.5	0.7	1.6	4.9	3.5	2.7	3.8	4.6	4.0	2.6	3.1	2.7	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.8	0.1	0.1	0.2	0.5	0.5	0.7	1.6	4.9	3.5	2.7	3.8	4.6	4.0	2.6	3.1	2.7	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model and covariance calibration																									
None (weighted WI)	0.5	0.1	0.0	0.2	0.2	0.4	0.4	0.3	0.4	0.3	0.3	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.1
GLM (OLS)	0.1	0.1	0.3	0.4	0.5	0.4	0.7	1.8	3.2	2.2	1.6	3.5	1.6	1.1	3.4	3.0	0.4	0.3	0.3	0.2	0.5	0.2	0.0	0.0	0.0
GLM (Ridge)	0.1	0.1	0.2	0.4	0.6	0.4	0.5	1.7	3.4	2.4	1.3	3.6	2.0	1.2	3.7	2.9	0.2	0.4	0.4	0.2	0.5	0.1	0.0	0.0	0.0
GLM (LASSO)	0.1	0.1	0.3	0.4	0.5	0.4	0.5	2.0	3.2	2.3	1.6	3.6	1.6	1.1	3.4	3.1	0.2	0.3	0.3	0.2	0.5	0.1	0.0	0.0	0.0
GLM (Elastic net)	0.1	0.1	0.3	0.4	0.5	0.4	0.5	2.0	3.2	2.3	1.6	3.6	1.6	1.1	3.4	3.1	0.2	0.3	0.3	0.2	0.5	0.1	0.0	0.0	0.0
GAM (fix knots)	0.1	0.1	0.3	0.4	0.5	0.4	0.7	1.8	3.2	2.2	1.6	3.5	1.6	1.1	3.4	3.0	0.4	0.3	0.3	0.2	0.5	0.2	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.6	0.0	1.0	0.0	12.9	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.5	0.0
MARS	0.3	0.1	0.1	0.2	0.5	1.2	2.0	1.9	1.9	12.4	3.1	3.4	7.9	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.2	0.2	0.2	0.0
ANN (opt. knots)	0.2	0.1	0.2	0.6	0.6	0.7	2.1	1.8	5.5	1.5	5.6	4.9	2.5	1.3	6.0	1.4	0.4	0.3	0.4	0.1	0.6	0.1	0.0	0.0	0.0
SVM	0.1	0.1	0.9	0.3	0.8	1.6	1.6	6.7	3.1	2.6	7.7	1.8	1.1	7.3	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.1	0.1	0.3	0.4	0.5	0.4	0.7	1.8	3.2	2.2	1.6	3.5	1.6	1.1	3.4	3.0	0.4	0.3	0.3	0.2	0.5	0.2	0.0	0.0	0.0
GAMM (fix knots)	0.1	0.1	0.3	0.4	0.5	0.4	0.7	1.8	3.2	2.2	1.6	3.5	1.6	1.1	3.4	3.0	0.4	0.3	0.3	0.2	0.5	0.2	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model and covariance calibration																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.7	0.5	0.7	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.1	0.2	0.3	0.3	0.4	0.8	0.7	0.8	1.3	0.8	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.7	0.6	0.7	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.7	0.6	0.7	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.7	0.5	0.7	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.3	0.6	0.2	0.9	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.2	0.2	0.3	0.2	0.3	0.7	0.9	0.8	2.5	2.7	1.0	0.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.4	0.7	1.0	0.8	0.5	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.7	0.5	0.7	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	0.7	0.5	0.7	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.7: Estimated standard deviations (in percentage points) for income class frequencies estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using total and covariance calibration																									
None (weighted WI)	0.5	0.1	0.0	0.0	0.1	0.1	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.6	0.5	0.0	0.6	1.1	1.1	0.6	0.7	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.2	0.4	1.0	0.4	0.0	1.4	1.5	1.5	0.7	0.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.6	0.4	0.0	0.6	1.2	1.1	0.6	0.7	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.6	0.4	0.0	0.6	1.1	1.1	0.6	0.7	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.6	0.5	0.0	0.6	1.1	1.1	0.6	0.7	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.2	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.2	0.6	0.4	2.5	0.0	1.5	4.5	4.3	0.6	1.1	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	1.1	0.5	0.4	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.6	0.5	0.0	0.6	1.1	1.1	0.6	0.7	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.6	0.5	0.0	0.6	1.1	1.1	0.6	0.7	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using total and covariance calibration																									
None (weighted WI)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using total and covariance calibration																									
None (weighted WI)	6.0	0.2	0.2	0.3	0.3	1.0	1.0	1.0	0.8	0.9	0.5	0.7	0.5	0.2	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
GLM (OLS)	3.1	1.2	3.3	2.6	4.9	2.6	3.8	3.6	6.5	3.4	4.2	5.1	4.9	2.7	1.7	2.8	1.4	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	2.7	1.2	2.8	3.1	3.2	4.6	4.0	3.6	7.0	3.4	4.9	5.0	4.9	2.6	1.7	2.8	1.2	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	3.1	0.9	3.4	2.6	4.5	2.8	3.9	3.8	6.5	3.2	4.4	5.1	4.9	2.7	1.8	2.9	1.2	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	3.1	0.9	3.4	2.6	4.5	2.8	3.9	3.8	6.5	3.2	4.4	5.1	4.9	2.7	1.8	2.9	1.2	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	3.1	1.2	3.3	2.6	4.9	2.6	3.8	3.6	6.5	3.4	4.2	5.1	4.9	2.7	1.7	2.8	1.4	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.0	0.0	5.1	0.0	26.6	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.3	0.0
MARS	6.5	1.1	1.6	1.9	5.4	6.6	7.2	3.7	1.8	33.8	3.6	2.9	12.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	2.7	1.7	0.8	4.1	2.6	3.9	4.2	3.2	9.6	3.8	6.3	5.0	5.4	2.6	2.0	2.3	1.1	0.7	0.1	0.2	0.0	0.0	0.0	0.0	0.0
SVM	0.5	3.5	1.8	4.8	5.6	8.3	4.9	13.4	12.0	4.8	6.1	2.6	2.4	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	3.1	1.2	3.3	2.6	4.9	2.6	3.8	3.6	6.5	3.4	4.2	5.1	4.9	2.7	1.7	2.8	1.4	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	3.1	1.2	3.3	2.6	4.9	2.6	3.8	3.6	6.5	3.4	4.2	5.1	4.9	2.7	1.7	2.8	1.4	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using total and covariance calibration																									
None (weighted WI)	1.8	0.1	0.1	0.2	0.3	0.7	0.4	0.6	1.0	0.4	0.5	0.4	0.4	0.4	0.4	0.3	0.2	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.1
GLM (OLS)	1.0	0.1	0.4	1.1	1.3	2.4	1.4	5.9	7.3	4.6	6.2	11.8	5.5	5.2	5.4	1.3	0.2	0.4	0.2	0.1	0.4	0.2	0.0	0.0	0.0
GLM (Ridge)	0.0	1.0	0.2	0.7	2.0	2.0	1.0	5.1	7.2	4.6	5.5	10.6	6.2	5.3	4.7	1.0	0.2	0.4	0.3	0.1	0.5	0.0	0.0	0.0	0.0
GLM (LASSO)	1.0	0.0	0.4	0.7	1.9	2.4	1.0	6.0	7.3	4.5	6.1	11.8	5.5	5.2	5.4	1.3	0.2	0.4	0.2	0.1	0.5	0.1	0.0	0.0	0.0
GLM (Elastic net)	1.0	0.0	0.4	0.7	1.9	2.4	1.0	6.0	7.3	4.5	6.1	11.8	5.5	5.2	5.4	1.3	0.2	0.4	0.2	0.1	0.5	0.1	0.0	0.0	0.0
GAM (fix knots)	1.0	0.1	0.4	1.1	1.3	2.4	1.4	5.9	7.3	4.6	6.2	11.8	5.5	5.2	5.4	1.3	0.2	0.4	0.2	0.1	0.4	0.2	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.0	0.0	3.0	0.0	17.3	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.9	0.0
MARS	1.1	0.2	0.3	0.4	0.6	1.1	2.6	3.9	5.8	21.2	0.8	6.4	15.8	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.2	0.3	0.3	0.0
ANN (opt. knots)	0.3	1.0	0.4	1.1	1.8	1.6	5.8	4.3	11.6	6.7	11.1	6.3	5.4	7.1	3.2	0.7	0.2	0.4	0.3	0.0	0.6	0.1	0.0	0.0	0.0
SVM	0.0	0.3	0.4	1.1	0.9	1.4	3.1	10.3	9.3	5.9	6.5	2.9	8.3	1.8	0.1	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	1.0	0.1	0.4	1.1	1.3	2.4	1.4	5.9	7.3	4.6	6.2	11.8	5.5	5.2	5.4	1.3	0.2	0.4	0.2	0.1	0.4	0.2	0.0	0.0	0.0
GAMM (fix knots)	1.0	0.1	0.4	1.1	1.3	2.4	1.4	5.9	7.3	4.6	6.2	11.8	5.5	5.2	5.4	1.3	0.2	0.4	0.2	0.1	0.4	0.2	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using total and covariance calibration																									
None (weighted WI)	2.3	0.1	0.1	0.4	0.3	0.2	0.3	0.6	0.5	0.4	0.5	0.6	0.6	0.3	0.3	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.1
GLM (OLS)	0.3	0.2	0.9	1.8	2.7	0.9	1.1	3.2	4.0	3.6	4.0	5.4	4.5	4.7	3.5	3.1	2.1	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.2	0.2	0.6	1.5	3.2	1.1	1.2	3.0	4.8	3.5	5.2	6.3	5.3	4.7	3.7	3.4	1.4	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.3	0.1	1.0	1.8	2.7	0.9	1.1	3.2	4.0	3.6	4.7	5.2	4.5	4.7	3.5	3.3	1.9	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.3	0.1	1.0	1.8	2.7	0.9	1.1	3.2	4.0	3.6	4.7	5.2	4.5	4.7	3.5	3.3	1.9	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.3	0.2	0.9	1.8	2.7	0.9	1.1	3.2	4.0	3.6	4.0	5.4	4.5	4.7	3.5	3.1	2.1	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.4	0.0	0.3	0.0	3.6	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
MARS	1.2	0.5	0.6	1.0	2.2	3.4	4.4	3.0	2.7	5.2	4.0	5.3	19.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.6	0.4	0.3	3.1	1.7	1.8	1.3	2.9	3.8	2.5	2.9	4.0	5.9	3.5	4.8	1.3	1.7	0.5	0.1	0.0	0.2	0.0	0.0	0.0	0.0
SVM	0.2	0.6	1.9	1.0	1.8	3.8	4.7	3.1	4.2	4.6	11.6	2.9	3.1	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.3	0.2	0.9	1.8	2.7	0.9	1.1	3.2	4.0	3.6	4.0	5.4	4.5	4.7	3.5	3.1	2.1	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.3	0.2	0.9	1.8	2.7	0.9	1.1	3.2	4.0	3.6	4.0	5.4	4.5	4.7	3.5	3.1	2.1	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.7: Estimated standard deviations (in percentage points) for income class frequencies estimated by weighted aggregation of predictions in the WI (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model, total and covariance calibration																									
None (weighted WI)	0.6	0.0	0.0	0.1	0.1	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.8	1.2	0.6	1.0	1.1	1.2	1.1	0.6	0.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.9	1.3	1.0	0.8	1.2	2.5	1.9	0.9	0.6	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.8	1.2	0.6	1.0	1.2	1.2	1.1	0.6	0.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.8	1.2	0.6	1.0	1.1	1.2	1.1	0.6	0.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.8	1.2	0.6	1.0	1.1	1.2	1.1	0.6	0.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.4	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.9	2.3	2.1	0.6	3.0	3.2	1.2	0.6	0.6	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	1.3	0.0	1.3	0.8	1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.8	1.2	0.6	1.0	1.1	1.2	1.1	0.6	0.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.8	1.2	0.6	1.0	1.1	1.2	1.1	0.6	0.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model, total and covariance calibration																									
None (weighted WI)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model, total and covariance calibration																									
None (weighted WI)	0.5	0.1	0.0	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.1	0.0	0.2	0.4	0.2	0.6	0.9	1.1	0.9	1.7	2.3	1.0	1.9	1.2	1.0	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.6	0.2	0.4	0.3	0.6	0.9	1.1	0.8	1.6	2.2	1.8	2.0	1.2	1.1	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.7	0.2	0.4	0.3	0.5	0.9	1.1	0.8	1.6	2.3	1.0	1.9	1.2	1.1	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.7	0.2	0.4	0.3	0.5	0.9	1.1	0.8	1.6	2.3	1.0	1.9	1.2	1.1	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.1	0.0	0.2	0.4	0.2	0.6	0.9	1.1	0.9	1.7	2.3	1.0	1.9	1.2	1.0	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.6	0.0	0.6	0.0	4.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.1	0.1	0.2	0.5	0.7	0.6	0.5	0.5	2.8	0.9	1.1	3.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.2	0.1	0.4	0.4	0.5	1.5	1.0	2.4	1.0	3.3	1.7	1.6	2.4	1.3	0.9	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	1.3	2.6	0.9	0.9	4.0	1.3	1.9	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.1	0.0	0.2	0.4	0.2	0.6	0.9	1.1	0.9	1.7	2.3	1.0	1.9	1.2	1.0	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.1	0.0	0.2	0.4	0.2	0.6	0.9	1.1	0.9	1.7	2.3	1.0	1.9	1.2	1.0	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model, total and covariance calibration																									
None (weighted WI)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model, total and covariance calibration																									
None (weighted WI)	0.4	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (OLS)	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.5	0.7	0.6	0.9	0.9	0.8	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.1	0.2	0.2	0.3	0.4	0.5	0.9	0.9	1.0	1.5	0.9	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.5	0.7	0.7	0.9	0.9	0.8	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.5	0.7	0.7	0.9	0.9	0.8	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.5	0.7	0.6	0.9	0.9	0.8	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.4	0.7	0.3	1.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.1	0.2	0.3	0.2	0.3	0.8	0.9	0.8	2.5	2.7	1.2	0.7	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.6	0.5	1.0	1.3	1.0	0.6	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.5	0.7	0.6	0.9	0.9	0.8	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.5	0.7	0.6	0.9	0.9	0.8	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.8: Estimated standard deviations (in percentage points) for income class frequencies estimated from the imputed Microcensus, using a weighted loss function for prediction models

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: unweighted, using no auxiliary information																									
Matching	0.0	0.0	0.0	0.1	0.3	0.4	0.4	0.4	0.5	0.5	0.6	0.6	0.4	0.4	0.3	0.2	0.1	0.1	0.0	0.0	0.1	0.2	0.1	0.2	0.1
GLM (OLS)	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.4	0.2	0.4	0.3	0.5	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.1	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.2	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
MARS	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.4	0.1	0.1	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.6	0.3	0.5	0.9	0.8	1.2	1.4	1.4	1.4	1.3	1.7	1.2	1.0	1.3	1.4	0.9	0.9	0.9	0.7	0.5	0.2	0.2	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.3	0.3	0.4	0.3	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (parametric), using propensity weights																									
GLM (OLS)	0.0	0.2	0.5	0.4	0.2	0.3	0.4	0.5	0.5	0.5	1.4	1.4	1.3	0.9	0.6	0.5	0.8	1.3	1.1	1.1	0.7	0.5	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.4	0.5	0.3	0.5	0.6	0.5	0.6	0.5	1.2	1.4	1.2	0.6	0.6	0.5	0.8	1.0	0.9	0.7	0.2	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.2	0.5	0.4	0.2	0.3	0.4	0.5	0.5	0.5	1.3	1.5	1.3	0.9	0.6	0.5	0.9	1.3	1.1	1.0	0.7	0.5	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.2	0.5	0.4	0.2	0.3	0.4	0.5	0.5	0.5	1.3	1.5	1.3	0.9	0.6	0.5	0.9	1.3	1.2	1.0	0.7	0.5	0.0	0.0	0.0
GAM (fix knots)	0.6	0.2	0.6	0.7	0.4	0.6	0.6	0.7	0.6	1.0	1.5	3.2	3.1	1.2	2.3	1.8	1.6	2.5	2.3	1.6	1.3	2.6	2.0	0.5	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.1	0.0	2.1	1.6	5.5	4.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.8	2.7	3.4
MARS	8.2	0.2	0.3	0.7	0.9	1.1	1.4	1.1	7.8	12.3	7.0	4.0	13.7	11.2	5.7	3.1	0.8	0.5	0.5	0.4	0.4	1.1	2.1	4.4	3.1
ANN (opt. knots)	0.6	0.3	0.4	0.5	0.9	1.6	1.7	1.6	1.6	1.8	2.6	2.1	1.8	1.9	2.5	2.2	2.1	2.2	2.0	1.2	0.5	0.5	0.1	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	0.3	0.4	0.4	1.0	3.1	1.8	2.3	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	16.4	18.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	16.4	18.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Pseudo-Weights (parametric), using propensity weights																									
GLM (OLS)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.3	0.5	0.4	0.5	0.4	0.3	0.4	0.3	0.3	0.2	0.3	0.1	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.3	0.3	0.1	0.4	0.3	0.4	0.2	0.4	0.6	0.5	0.5	0.3	0.4	0.3	0.2	0.4	0.4	0.2	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.4	0.5	0.4	0.5	0.4	0.3	0.4	0.3	0.3	0.3	0.3	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.4	0.5	0.4	0.5	0.4	0.3	0.4	0.3	0.3	0.3	0.3	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.2	0.2	0.1	0.3	0.3	0.3	0.2	0.3	1.0	2.1	1.5	1.4	1.4	1.6	1.8	1.6	1.2	0.9	0.8	0.4	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.2	2.5	3.6	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3	1.8	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.6	0.2	0.4	0.9	1.0	0.9	1.2	1.3	1.5	1.4	1.7	1.3	1.2	1.7	1.7	1.0	0.9	1.0	1.1	0.5	0.6	0.2	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.2	0.4	0.3	0.3	1.0	2.2	3.4	3.3	2.3	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using propensity weights																									
GLM (OLS)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	1.0	0.7	0.6	0.7	0.4	0.4	0.8	0.8	0.7	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.3	0.3	0.1	0.3	0.3	0.4	0.5	1.0	0.7	0.7	0.7	0.4	0.5	0.8	0.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	0.9	0.7	0.6	0.7	0.5	0.5	0.8	0.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	0.9	0.7	0.6	0.7	0.5	0.5	0.8	0.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.2	0.2	0.2	0.4	0.3	0.3	0.4	0.5	1.2	0.7	0.9	0.8	0.5	0.6	0.8	1.2	0.5	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	23.2	12.0	2.1	11.8	28.0	4.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	1.1	0.9	1.0	1.3	5.3	5.2	0.7	1.9	12.9	12.6	0.8	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.8	0.5	0.6	1.0	1.5	1.4	1.6	1.6	1.6	1.6	2.2	2.3	2.1	2.2	2.5	1.9	2.4	1.6	0.3	0.1	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.2	0.3	0.3	0.3	0.5	1.5	3.7	2.3	2.3	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots), using propensity weights																									
GLM (OLS)	0.0	0.2	0.5	0.4	0.2	0.4	0.4	0.6	0.4	0.5	1.4	1.0	1.5	1.2	0.6	0.7	1.1	1.3	0.9	0.6	0.3	0.1	0.0	0.0	0.0
GLM (Ridge)	0.0	0.1	0.4	0.4	0.3	0.5	0.5	0.6	0.4	0.5	1.4	1.1	1.5	1.0	0.7	0.7	1.2	1.2	0.9	0.6	0.2	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.2	0.5	0.4	0.2	0.4	0.4	0.6	0.4	0.5	1.4	1.0	1.5	1.1	0.6	0.7	1.1	1.3	0.9	0.7	0.3	0.1	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.2	0.5	0.4	0.2	0.4	0.4	0.6	0.4	0.5	1.4	1.0	1.5	1.1	0.6	0.7	1.1	1.3	0.9	0.7	0.3	0.1	0.0	0.0	0.0
GAM (fix knots)	2.9	0.3	0.4	0.5	0.4	0.6	0.7	0.6	0.5	0.8	2.3	2.8	2.2	1.8	2.2	2.1	1.6	1.8	1.1	0.9	0.9	1.4	0.3	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0	8.4	13.0	4.7	2.4	14.7	9.6	3.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.8	2.5
MARS	3.9	0.5	0.7	1.1	1.5	1.9	2.1	2.3	11.6	10.7	8.4	3.3	13.7	8.6	3.0	3.1	2.2	1.6	1.0	0.7	0.7	1.2	1.4	2.2	2.0
ANN (opt. knots)	0.2	0.2	0.2	0.9	1.0	1.4	1.6	1.6	1.7	1.9	3.0	2.3	2.9	2.5	2.0	2.3	1.9	2.0	1.3	0.7	0.4	0.7	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	0.3	0.3	0.4	1.0	3.1	1.8	2.3	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	19.1	19.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	19.4	19.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.8: Estimated standard deviations (in percentage points) for income class frequencies estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Pseudo-Weights (fix knots), using propensity weights																									
GLM (OLS)	0.0	0.0	0.0	0.2	0.2	0.1	0.2	0.3	0.3	0.2	0.4	0.3	0.4	0.4	0.3	0.4	0.3	0.3	0.2	0.2	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.3	0.2	0.1	0.2	0.3	0.4	0.3	0.4	0.5	0.5	0.4	0.4	0.3	0.3	0.3	0.2	0.2	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.2	0.2	0.1	0.2	0.3	0.3	0.2	0.4	0.3	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.2	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.2	0.2	0.1	0.2	0.3	0.3	0.2	0.4	0.3	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.2	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.2	0.1	0.1	0.2	0.2	0.1	0.3	0.3	0.4	0.2	0.8	0.7	1.0	0.6	0.9	0.4	0.9	0.4	0.3	0.3	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.2	0.4	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.3	0.2	0.2	1.0	1.1	1.2	1.2	1.4	1.3	1.4	1.6	1.4	1.1	1.2	1.1	1.0	1.1	0.9	0.5	0.4	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.2	0.3	0.3	0.4	1.5	3.8	2.4	2.4	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	6.6	6.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	6.6	6.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using propensity weights																									
GLM (OLS)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	1.0	0.7	0.6	0.7	0.4	0.4	0.8	0.8	0.7	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.3	0.3	0.1	0.3	0.3	0.4	0.5	1.0	0.7	0.7	0.7	0.4	0.5	0.8	0.9	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	0.9	0.7	0.6	0.7	0.5	0.5	0.8	0.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	0.9	0.7	0.6	0.7	0.5	0.5	0.8	0.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.2	0.2	0.2	0.4	0.3	0.3	0.4	0.5	1.2	0.7	0.9	0.8	0.5	0.6	0.8	1.2	0.5	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	23.2	12.0	2.1	11.8	28.0	4.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	1.1	0.9	1.0	1.3	5.3	5.2	0.7	1.9	12.9	12.6	0.8	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.6	0.5	0.5	1.1	1.3	1.4	1.6	1.6	1.6	1.7	2.4	2.1	2.1	2.4	2.1	2.1	2.4	1.6	0.5	0.2	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.2	0.3	0.3	0.3	0.5	1.5	3.7	2.3	2.3	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using propensity weights																									
GLM (OLS)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	1.0	0.7	0.6	0.7	0.4	0.4	0.8	0.8	0.7	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.3	0.3	0.1	0.3	0.3	0.4	0.5	1.0	0.7	0.7	0.7	0.4	0.5	0.8	0.9	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	0.9	0.7	0.6	0.7	0.5	0.5	0.8	0.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	0.9	0.7	0.6	0.7	0.5	0.5	0.8	0.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.2	0.2	0.2	0.4	0.3	0.3	0.4	0.5	1.2	0.7	0.9	0.8	0.5	0.6	0.8	1.2	0.5	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	23.2	12.0	2.1	11.8	28.0	4.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	1.1	0.9	1.0	1.3	5.3	5.2	0.7	1.9	12.9	12.6	0.8	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	1.0	0.3	0.8	1.0	1.2	1.6	1.6	1.6	1.6	1.6	2.4	2.4	2.0	2.2	2.1	2.1	2.2	1.3	0.5	0.3	0.2	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.2	0.3	0.3	0.3	0.5	1.5	3.7	2.3	2.3	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Subsampling, using total calibration																									
GLM (OLS)	2.1	0.5	0.6	0.8	0.8	1.0	1.0	1.2	1.5	2.3	4.9	4.6	3.8	2.8	1.9	2.6	2.7	3.5	3.1	2.5	2.0	3.9	1.9	0.2	0.0
GLM (Ridge)	1.1	0.2	0.5	0.7	0.7	1.1	1.3	1.7	1.9	2.8	7.5	13.1	21.1	13.1	6.7	5.2	4.3	4.2	3.3	2.2	1.6	1.6	0.0	0.0	0.0
GLM (LASSO)	1.7	0.5	0.6	0.8	0.7	0.7	1.2	1.4	1.8	3.2	14.6	27.3	36.7	21.2	5.5	5.3	4.2	4.2	3.8	3.2	2.4	3.6	0.9	0.0	0.0
GLM (Elastic net)	1.8	0.4	0.6	0.7	0.8	0.7	1.1	1.6	1.8	3.3	15.3	27.5	38.5	20.1	7.0	5.0	3.7	4.4	3.8	2.8	2.6	3.6	1.0	0.0	0.0
GAM (fix knots)	3.6	0.5	0.6	0.8	0.8	1.0	1.0	1.6	2.1	2.8	5.5	6.1	5.8	5.7	4.2	4.6	3.1	3.7	3.7	3.4	2.6	4.7	2.3	4.4	14.4
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.0	10.4	19.2	25.2	27.5	15.9	11.0	16.2	6.2	12.5	6.7	6.5	12.3	1.0	4.0	12.3	9.6	0.0
MARS	4.4	0.5	0.5	0.9	0.9	1.4	4.0	4.4	15.3	16.1	30.8	36.4	27.8	17.9	13.8	3.9	2.4	0.9	0.4	0.6	1.2	1.1	3.6	3.7	6.3
ANN (opt. knots)	49.3	0.9	0.9	2.2	3.3	2.2	3.2	3.2	3.1	4.2	5.8	5.2	4.4	4.0	3.0	3.1	3.3	2.6	2.7	1.7	1.7	2.2	2.8	3.5	0.0
SVM	0.0	0.0	0.2	0.7	1.0	1.5	2.2	3.4	4.5	4.2	4.3	4.7	4.4	4.5	2.8	2.3	1.5	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	20.5	19.7	5.9	5.7	1.6	13.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	20.5	19.7	5.9	5.7	1.6	13.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Post-stratification, using total calibration																									
GLM (OLS)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.5	0.3	0.2	0.5	0.2	0.2	0.3	0.3	0.5	0.7	0.8	0.7	0.4	0.4	0.3	0.7	0.3	0.5	0.2	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.4	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.4	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.1	0.1	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.2	0.0	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
MARS	0.0	0.0	0.1	0.2	1.3	0.5	0.2	0.8	0.2	1.7	1.8	5.7	4.1	0.2	5.4	6.8	0.4	0.1	1.2	0.2	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.6	0.2	0.3	0.8	0.9	1.2	1.2	1.4	1.5	1.4	1.3	1.3	1.0	1.0	1.1	0.9	0.7	0.7	0.6	0.4	0.4	0.3	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.2	0.3	0.2	1.1	0.3	0.9	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.8: Estimated standard deviations (in percentage points) for income class frequencies estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using total calibration																									
GLM (OLS)	0.0	0.2	0.2	0.2	0.1	0.3	0.3	0.3	0.4	0.6	1.0	0.9	0.6	0.7	0.5	0.4	0.7	0.9	0.8	0.6	0.2	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.3	0.3	0.3	0.1	0.4	0.3	0.4	0.6	0.4	0.7	1.4	1.3	1.4	1.0	0.8	0.9	1.0	1.5	1.2	0.5	0.5	0.5	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using total calibration																									
GLM (OLS)	0.0	0.0	0.0	0.3	0.3	0.1	0.4	0.3	0.3	0.5	1.0	0.7	0.6	0.7	0.4	0.5	0.8	0.9	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.2	0.2	0.1	0.4	0.2	0.3	0.5	1.1	0.7	0.6	0.7	0.5	0.6	0.8	1.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.3	0.3	0.1	0.4	0.3	0.3	0.5	1.0	0.7	0.5	0.7	0.4	0.5	0.8	0.9	0.5	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.3	0.3	0.1	0.4	0.3	0.3	0.5	1.0	0.8	0.5	0.7	0.4	0.5	0.7	0.9	0.5	0.1	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.3	0.3	0.1	0.4	0.3	0.4	0.5	1.2	0.8	0.6	0.7	0.4	0.5	0.8	1.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.0	11.9	0.0	6.6	4.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.2	0.4	0.4	0.9	0.6	0.5	0.5	3.6	0.5	3.5	12.0	12.0	3.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.8	0.3	0.5	1.1	1.0	1.4	1.5	1.8	1.7	1.7	2.5	2.6	2.0	2.2	2.6	1.9	2.3	1.4	0.5	0.4	0.4	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	0.3	0.3	0.4	1.0	3.2	1.9	2.4	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using total calibration																									
GLM (OLS)	0.2	0.5	0.4	0.3	0.3	0.4	0.5	0.6	0.5	0.5	1.3	1.4	1.3	1.0	0.6	0.7	0.7	1.2	1.2	1.1	0.9	0.9	0.0	0.0	0.0
GLM (Ridge)	0.0	0.3	0.5	0.5	0.2	0.6	0.6	0.6	0.5	0.5	1.0	1.5	1.1	0.7	0.7	0.7	0.7	1.1	1.0	0.9	0.5	0.2	0.0	0.0	0.0
GLM (LASSO)	0.2	0.5	0.4	0.3	0.3	0.4	0.5	0.6	0.5	0.5	1.3	1.4	1.3	1.0	0.7	0.7	0.7	1.2	1.2	1.1	0.9	0.9	0.0	0.0	0.0
GLM (Elastic net)	0.2	0.5	0.4	0.3	0.3	0.4	0.5	0.6	0.5	0.5	1.3	1.4	1.3	1.0	0.7	0.7	0.7	1.2	1.2	1.1	0.9	0.9	0.0	0.0	0.0
GAM (fix knots)	1.4	0.6	0.6	0.3	0.9	0.6	0.7	0.8	0.7	0.9	1.9	2.8	2.8	1.2	2.3	2.1	1.7	2.2	2.6	2.0	1.3	2.4	1.6	0.3	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.6	10.9	0.3	4.0	5.3	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	6.1	3.1
MARS	5.6	0.4	0.3	0.7	1.0	1.3	2.1	2.0	13.1	12.4	5.7	4.1	12.5	10.0	2.5	2.1	1.5	1.2	0.6	0.7	0.7	2.0	2.8	3.8	2.1
ANN (opt. knots)	0.7	0.6	0.5	0.6	1.4	1.3	1.4	1.4	1.6	1.7	3.0	1.9	1.8	2.0	2.0	2.2	2.0	2.2	2.4	1.6	1.1	0.8	0.1	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	0.3	0.3	0.4	1.0	3.1	1.8	2.4	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	5.4	7.7	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	5.4	7.7	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using total calibration																									
GLM (OLS)	0.6	0.5	0.4	0.3	0.5	0.5	0.5	0.7	0.5	0.5	1.2	1.5	1.4	1.2	0.7	0.8	0.7	1.1	1.2	1.2	1.1	1.4	0.0	0.0	0.0
GLM (Ridge)	0.3	0.5	0.5	0.5	0.3	0.8	0.7	0.8	0.8	0.6	1.0	1.9	1.4	0.7	0.8	0.9	0.7	1.1	1.0	0.9	0.6	0.2	0.0	0.0	0.0
GLM (LASSO)	0.6	0.5	0.4	0.3	0.5	0.5	0.5	0.7	0.5	0.5	1.2	1.5	1.4	1.2	0.7	0.8	0.7	1.2	1.2	1.2	1.1	1.4	0.0	0.0	0.0
GLM (Elastic net)	0.6	0.5	0.4	0.3	0.5	0.5	0.5	0.7	0.5	0.5	1.2	1.5	1.4	1.2	0.7	0.8	0.7	1.1	1.2	1.2	1.1	1.4	0.0	0.0	0.0
GAM (fix knots)	3.2	0.6	0.5	0.7	0.9	0.7	0.8	1.0	1.0	1.1	2.1	2.4	2.9	1.9	2.0	2.3	1.9	2.1	2.5	2.0	1.6	2.0	1.6	1.6	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.5	6.9	0.3	3.1	7.3	7.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	6.3	2.3
MARS	5.8	0.4	0.5	0.9	1.2	1.5	2.5	3.1	12.2	11.6	5.1	5.6	13.0	11.7	2.4	2.6	2.2	1.5	1.2	1.0	1.0	2.4	2.6	3.4	2.0
ANN (opt. knots)	0.8	0.5	0.5	1.0	1.3	1.5	1.4	1.5	1.6	1.8	2.5	1.8	1.7	2.1	2.2	2.0	1.9	2.5	2.3	1.9	1.4	1.2	0.7	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.2	0.3	0.3	0.4	1.2	3.4	2.0	2.4	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	13.3	18.7	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	13.3	18.7	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using total calibration																									
GLM (OLS)	0.0	0.0	0.1	0.5	0.3	0.3	0.8	0.4	0.5	1.0	1.6	1.2	1.6	1.1	0.9	1.1	1.2	1.6	1.7	1.5	1.4	5.9	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.5	0.4	0.2	0.8	0.4	0.5	1.1	2.3	1.4	1.7	1.1	1.1	1.1	1.1	1.7	1.7	1.5	1.5	6.6	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.1	0.5	0.3	0.2	0.8	0.7	0.7	1.1	2.4	1.5	1.6	1.4	1.2	1.4	1.0	1.6	1.7	1.5	1.4	6.5	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.1	0.5	0.3	0.2	0.7	0.5	0.5	1.2	2.1	1.3	1.6	1.2	1.1	1.4	1.0	1.6	1.7	1.5	1.4	6.5	0.0	0.0	0.0
GAM (fix knots)	0.0	0.2	0.5	0.4	0.1	0.5	0.3	0.5	0.6	1.0	1.6	1.4	2.1	1.7	1.2	1.7	1.0	2.1	2.5	2.0	2.5	7.8	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	8.9	4.4	20.2	7.5	0.0	10.3	14.8	8.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.2	0.3	0.7	1.1	1.1	1.1	1.5	13.4	13.7	5.4	5.7	13.8	11.9	3.5	2.8	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	1.0	0.4	0.8	1.0	1.2	1.5	1.7	1.5	1.6	1.7	2.5	3.1	2.3	3.0	3.0	2.3	2.1	2.6	3.0	2.1	1.7	5.3	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.3	0.4	0.4	0.6	1.2	2.4	3.6	3.0	2.4	2.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	3.3	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	4.3	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.8: Estimated standard deviations (in percentage points) for income class frequencies estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.3	0.3	0.2	0.3	0.4	0.5	0.5	0.9	1.1	0.8	0.7	0.6	0.5	0.9	0.9	0.9	0.4	0.1	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.3	0.2	0.1	0.4	0.3	0.4	0.4	0.7	1.7	0.8	1.5	1.0	0.9	1.6	0.8	1.8	0.9	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using covariance calibration																									
GLM (OLS)	0.0	0.0	0.2	0.2	0.1	0.3	0.2	0.3	0.4	0.6	0.9	0.6	0.7	0.7	0.4	0.4	0.8	0.7	0.7	0.2	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.1	0.1	0.1	0.3	0.2	0.3	0.4	0.6	0.9	0.6	0.7	0.6	0.4	0.5	0.8	0.7	0.7	0.1	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.2	0.2	0.1	0.3	0.2	0.3	0.4	0.6	0.9	0.6	0.7	0.6	0.4	0.4	0.8	0.7	0.7	0.2	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.2	0.2	0.1	0.3	0.2	0.3	0.4	0.6	0.9	0.6	0.7	0.6	0.4	0.4	0.8	0.7	0.7	0.2	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.2	0.2	0.2	0.2	0.2	0.3	0.4	0.3	0.7	1.2	0.7	1.2	0.7	0.6	0.9	0.8	1.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	4.0	10.9	2.3	0.2	0.0	0.0	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3
MARS	0.4	0.3	0.3	0.3	0.4	0.4	1.4	0.7	7.1	0.9	3.2	6.2	4.2	2.5	2.3	2.1	1.5	0.8	0.2	0.2	0.2	0.5	0.6	0.7	2.4
ANN (opt. knots)	0.8	0.4	0.8	0.8	1.1	1.4	1.5	1.6	1.6	1.9	1.9	2.1	2.3	2.2	2.3	1.8	2.1	1.5	0.6	0.4	0.2	0.1	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.2	0.4	0.4	0.4	0.8	2.1	3.8	2.9	2.2	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using covariance calibration																									
GLM (OLS)	0.0	0.2	0.6	0.3	0.3	0.4	0.7	0.6	0.9	1.2	1.5	1.4	1.1	0.6	1.1	1.0	1.3	1.2	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.1	0.5	0.2	0.3	0.4	0.7	0.7	0.9	1.5	2.1	2.0	1.8	1.0	1.5	1.5	1.5	0.9	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.2	0.5	0.3	0.3	0.4	0.6	0.6	0.9	1.4	2.0	1.8	1.2	0.6	1.3	1.5	1.6	1.1	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.2	0.5	0.3	0.2	0.4	0.7	0.6	2.2	2.3	2.2	4.3	1.2	1.3	1.5	1.7	1.5	0.9	0.3	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.7	0.2	0.3	0.4	0.3	0.3	0.5	0.6	1.1	0.6	2.3	3.5	1.7	1.5	1.3	2.7	2.3	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	18.9	3.6	29.7	9.9	0.0	15.8	1.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
MARS	4.0	0.4	0.3	0.4	0.9	1.2	0.8	2.2	7.1	3.2	2.8	8.8	6.1	2.6	2.9	2.1	1.5	1.4	0.9	0.2	0.1	0.2	0.6	2.1	3.1
ANN (opt. knots)	10.2	0.7	0.8	0.6	1.1	1.2	1.5	1.8	1.8	1.8	4.1	2.5	2.1	1.9	2.8	2.4	1.4	0.9	0.9	0.5	0.3	0.3	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.6	0.5	0.4	0.4	0.6	1.7	1.1	1.4	2.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	12.2	10.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	11.5	11.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using covariance calibration																									
GLM (OLS)	0.0	0.0	0.3	0.7	0.5	0.6	0.7	0.5	0.7	1.1	1.2	1.5	1.5	1.4	1.5	1.2	1.3	1.7	1.8	1.6	1.4	2.2	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.2	0.7	0.5	0.6	0.8	0.5	0.7	1.2	1.6	2.0	1.7	1.4	1.5	1.5	1.4	1.8	1.8	1.6	1.3	2.2	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.3	0.7	0.5	0.6	0.7	0.4	0.7	1.1	1.2	1.6	1.6	1.4	1.5	1.2	1.4	1.7	1.8	1.5	1.3	2.3	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.3	0.7	0.5	0.6	0.7	0.4	0.7	1.1	1.2	1.6	1.6	1.4	1.5	1.2	1.4	1.7	1.8	1.6	1.3	2.3	0.0	0.0	0.0
GAM (fix knots)	0.0	0.3	0.4	0.5	0.4	0.7	0.6	0.7	0.6	1.0	1.8	3.4	4.5	1.7	1.4	2.5	2.5	2.2	1.7	2.0	1.4	2.3	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.3	5.9	21.1	27.5	21.6	30.6	22.2	22.0	10.8	10.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.2
MARS	5.2	0.9	0.9	1.3	1.6	1.7	2.9	10.3	12.2	9.9	9.0	14.6	11.2	4.2	2.0	1.4	1.0	0.3	0.1	0.5	1.4	4.0	3.3	3.6	3.3
ANN (opt. knots)	25.6	0.6	0.6	0.7	1.4	1.7	2.0	1.9	2.0	2.0	4.3	3.8	4.9	3.5	2.6	2.5	2.2	2.6	2.6	3.1	1.3	5.1	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.2	0.7	0.6	0.5	0.5	2.6	2.7	2.1	3.0	2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	11.6	11.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	11.6	11.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.8: Estimated standard deviations (in percentage points) for income class frequencies estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model and total calibration																									
GLM (OLS)	0.7	0.5	0.5	0.3	0.8	0.5	0.5	0.7	0.7	0.5	1.1	1.7	1.2	1.0	0.8	1.2	1.1	1.1	1.2	1.3	1.4	3.5	1.6	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	4.2	0.5	0.5	0.7	0.9	0.8	0.7	0.9	0.9	1.2	1.7	2.4	2.6	2.0	1.9	2.3	2.0	2.0	2.1	2.3	1.9	3.7	2.4	2.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	11.5	19.4	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	11.5	19.4	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model and total calibration																									
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model and total calibration																									
GLM (OLS)	0.4	0.4	0.5	0.5	0.5	0.8	0.5	0.6	0.9	0.9	1.3	2.1	1.6	0.9	0.8	0.8	1.1	1.5	1.3	1.2	0.9	0.9	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	3.5	0.3	0.4	0.7	1.0	1.3	1.4	0.9	1.4	1.3	3.1	2.1	2.5	1.9	1.9	2.2	2.0	1.8	2.0	1.3	1.1	1.1	0.2	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	20.3	16.4	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	20.2	16.7	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model and total calibration																									
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model and total calibration																									
GLM (OLS)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	1.0	0.7	0.6	0.7	0.4	0.4	0.8	0.8	0.7	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.3	0.3	0.1	0.3	0.3	0.4	0.5	1.0	0.7	0.7	0.7	0.4	0.5	0.8	0.9	0.6	0.1	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.3	0.4	0.5	0.9	0.7	0.6	0.7	0.5	0.5	0.8	0.8	0.7	0.1	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.2	0.2	0.2	0.3	0.4	0.4	0.5	0.9	0.7	0.6	0.7	0.5	0.5	0.8	0.8	0.7	0.1	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.2	0.2	0.2	0.4	0.3	0.3	0.4	0.5	1.2	0.7	0.9	0.8	0.5	0.6	0.8	1.3	0.6	0.1	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	22.0	11.8	2.1	13.1	28.5	4.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	1.1	0.9	0.9	1.4	5.2	5.2	0.7	1.9	12.7	12.4	0.8	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	1.1	0.3	0.6	1.3	1.1	1.4	1.7	1.5	1.7	1.6	2.3	2.4	2.1	2.2	2.3	2.1	2.5	1.5	0.5	0.2	0.1	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.2	0.3	0.3	0.3	0.5	1.6	3.7	2.3	2.4	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.8: Estimated standard deviations (in percentage points) for income class frequencies estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model and covariance calibration																									
GLM (OLS)	0.0	0.0	0.4	0.5	0.4	0.3	0.3	0.5	0.8	0.7	1.0	1.0	0.9	0.7	0.5	0.6	1.0	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.4	0.4	0.4	0.3	0.3	0.5	0.8	0.6	1.1	1.3	0.9	0.8	0.7	0.6	1.0	0.8	0.5	1.6	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	11.6	11.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	11.6	11.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model and covariance calibration																									
GLM (OLS)	0.0	0.1	0.5	0.5	0.6	0.6	0.4	0.5	0.8	1.4	2.1	1.6	1.3	2.4	2.7	3.8	2.9	1.8	1.9	2.4	2.6	5.4	2.9	0.0	0.0
GLM (Ridge)	0.0	0.1	0.4	0.5	0.3	0.6	0.8	0.9	1.3	3.1	4.8	2.4	1.6	3.7	4.3	4.4	2.1	2.0	3.4	2.2	0.8	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.1	0.5	0.5	0.6	0.6	0.4	0.5	0.8	1.4	2.1	1.7	1.3	2.4	2.7	3.8	2.9	1.8	1.9	2.4	2.6	5.4	2.8	0.0	0.0
GLM (Elastic net)	0.0	0.1	0.5	0.5	0.6	0.6	0.4	0.5	0.8	1.5	2.2	1.7	1.3	2.4	2.7	3.8	2.9	1.8	2.0	2.4	2.6	5.4	2.8	0.0	0.0
GAM (fix knots)	0.5	0.4	0.5	0.4	0.8	0.5	0.7	0.5	0.8	3.3	1.9	4.0	2.0	3.7	2.8	3.3	2.3	2.0	2.4	3.1	2.4	5.2	3.5	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.1	0.4	12.2	31.0	18.3	21.8	16.5	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	6.5	1.6
MARS	4.0	0.7	0.9	1.3	1.4	1.4	1.2	4.2	14.9	2.2	4.4	14.8	11.3	3.0	2.9	2.1	1.7	1.7	1.2	0.4	0.7	0.7	3.7	3.4	2.8
ANN (opt. knots)	1.0	0.4	0.8	0.8	1.4	1.5	1.4	1.7	1.8	3.6	2.5	4.3	2.3	3.6	3.5	3.0	1.8	2.4	2.4	2.9	2.5	4.8	7.5	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.2	0.7	0.6	0.5	0.5	1.4	2.9	2.1	2.3	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	6.1	6.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	6.3	6.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model and covariance calibration																									
GLM (OLS)	0.3	0.4	0.6	0.5	0.4	0.6	0.5	0.6	1.0	0.8	1.9	1.8	1.6	1.0	0.8	0.7	1.5	1.6	1.3	0.9	0.6	0.2	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	1.5	0.3	0.4	0.7	0.7	0.7	0.7	1.1	1.9	2.0	4.4	4.5	2.1	5.1	4.4	2.8	2.5	2.5	3.0	2.9	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	6.0	9.6	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	15.2	16.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model and covariance calibration																									
GLM (OLS)	0.8	0.1	0.3	0.6	0.5	0.8	0.5	0.4	0.7	1.1	1.1	1.4	1.7	2.2	2.3	2.2	2.2	1.8	2.1	2.4	2.2	4.4	5.5	0.0	0.0
GLM (Ridge)	0.5	0.1	0.4	0.7	0.5	0.8	0.8	1.2	1.1	1.6	5.1	4.0	2.5	2.8	3.4	4.1	2.4	2.6	3.5	2.8	1.5	5.7	2.0	0.0	0.0
GLM (LASSO)	0.8	0.1	0.3	0.6	0.6	0.7	0.5	0.4	0.8	1.1	9.0	8.9	1.7	2.2	2.3	2.3	2.2	1.8	2.1	2.3	2.3	4.0	5.9	0.0	0.0
GLM (Elastic net)	0.8	0.2	0.3	0.6	0.6	0.7	0.5	0.4	0.8	3.4	9.0	9.2	1.7	2.2	2.3	2.3	2.3	1.9	2.3	2.4	2.3	4.0	5.1	0.0	0.0
GAM (fix knots)	1.2	0.4	0.4	0.5	0.6	0.6	0.5	0.6	1.5	1.8	2.9	4.1	4.6	2.3	3.2	2.9	2.3	2.3	2.4	2.9	2.2	4.5	6.1	4.8	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.1	14.3	16.4	25.4	13.1	12.8	20.0	23.1	13.0	0.0	0.1	0.3	0.0	0.0	5.3	0.5	3.6	3.3
MARS	4.5	0.6	0.7	1.0	1.3	1.7	2.6	16.5	16.2	2.4	6.4	15.2	13.6	4.9	1.4	1.5	1.7	2.1	2.4	2.1	1.1	8.5	2.2	2.2	5.2
ANN (opt. knots)	0.8	0.8	0.8	1.2	1.3	1.5	1.6	1.7	1.6	2.4	3.6	3.6	3.7	2.3	3.1	2.0	2.6	2.3	2.4	2.6	2.1	4.5	5.7	0.0	0.0
SVM	0.0	0.0	0.1	0.1	0.5	0.5	0.5	0.9	1.0	1.7	4.5	2.2	2.6	2.6	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	9.6	11.1	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	9.6	11.1	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model and covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.3	0.3	0.1	0.3	0.3	0.4	0.4	1.0	0.9	0.6	0.8	0.5	0.4	0.9	0.8	0.8	0.2	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.3	0.3	0.1	0.4	0.3	0.4	0.4	1.0	0.9	0.7	0.8	0.5	0.5	0.9	0.9	0.7	0.1	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.3	0.3	0.1	0.4	0.3	0.4	0.4	1.1	0.8	0.7	0.7	0.5	0.5	0.9	0.9	0.8	0.2	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.3	0.3	0.1	0.4	0.3	0.4	0.4	1.1	0.8	0.7	0.7	0.5	0.5	0.9	0.9	0.8	0.2	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.3	0.3	0.2	0.3	0.4	0.4	0.4	1.1	0.8	0.6	0.8	0.5	0.4	0.9	0.9	0.8	0.1	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.8	0.0	0.0	22.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.8	1.3	1.3	4.0	8.1	14.4	5.3	8.2	10.7	10.5	3.9	2.1	0.4	0.0	0.4	3.5	0.9	2.4	0.0	0.0	0.0	0.0
ANN (opt. knots)	1.5	0.4	1.0	1.3	1.3	1.4	1.7	1.5	1.9	1.7	2.6	2.1	2.1	2.4	1.7	1.8	2.3	2.0	0.4	0.1	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.3	0.3	0.4	0.4	1.0	3.1	1.8	2.3	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	2.7	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	2.7	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.8: Estimated standard deviations (in percentage points) for income class frequencies estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: GREG, using total and covariance calibration																									
GLM (OLS)	0.0	0.1	0.3	0.3	0.2	0.2	0.4	0.4	0.3	0.6	0.9	0.6	0.9	0.8	0.6	0.5	0.8	0.8	0.7	0.5	0.2	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	1.6	0.3	0.2	0.3	0.4	0.6	0.5	0.4	0.6	1.1	1.0	3.1	2.5	1.4	2.1	1.9	1.7	1.6	1.8	1.8	1.2	3.6	2.5	0.5	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	13.3	13.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	13.3	13.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (one parameter per observation, as for the GREG), using total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using total and covariance calibration																									
GLM (OLS)	2.1	0.5	0.3	0.3	0.7	0.9	0.6	0.8	1.7	2.4	5.4	3.4	2.6	2.0	1.7	2.7	2.1	2.8	2.0	2.0	2.9	5.8	2.2	0.0	0.0
GLM (Ridge)	1.7	0.5	0.4	0.5	0.6	16.3	3.0	3.4	10.8	7.1	10.9	5.3	3.8	3.9	8.5	4.0	2.9	3.3	2.5	2.2	2.4	5.2	2.2	0.0	0.0
GLM (LASSO)	2.1	0.2	0.5	0.3	0.7	11.0	0.8	1.0	24.3	22.2	5.8	3.6	2.9	2.0	10.7	2.6	2.1	2.4	1.9	2.2	2.9	5.8	2.3	0.0	0.0
GLM (Elastic net)	2.2	0.2	0.4	0.4	0.7	10.8	0.8	1.0	22.5	20.4	5.9	3.6	2.9	2.0	10.6	2.6	2.1	2.2	2.0	2.3	3.0	5.7	2.2	0.0	0.0
GAM (fix knots)	2.8	0.3	0.4	0.4	0.6	0.9	0.8	0.8	2.4	3.3	2.9	7.0	5.0	4.0	2.5	2.0	2.2	1.9	3.1	1.8	2.2	6.6	2.6	4.2	0.1
Regression tree	0.0	0.0	0.0	0.6	1.8	2.3	5.4	21.8	16.5	20.1	12.7	20.4	33.9	51.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.8	1.9	2.5
MARS	4.2	1.4	1.6	2.2	3.2	3.7	8.2	17.1	30.4	14.3	7.4	11.0	6.4	6.2	3.1	2.8	1.6	4.5	0.7	0.5	0.6	1.9	3.0	2.9	2.9
ANN (opt. knots)	49.0	0.2	0.3	0.6	1.3	1.6	2.0	2.3	2.2	3.0	5.4	6.2	3.5	2.9	2.3	2.0	1.5	1.9	2.3	2.1	2.0	3.8	0.1	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.3	0.4	0.6	2.1	2.1	12.3	10.3	6.8	5.2	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	8.7	12.4	0.3	7.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	8.7	12.4	0.3	7.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using total and covariance calibration																									
GLM (OLS)	22.1	0.6	0.7	0.9	0.8	0.9	1.0	0.7	1.0	1.5	3.6	3.3	3.0	1.7	1.2	2.3	2.9	2.1	2.3	1.9	3.4	4.8	0.0	0.0	0.0
GLM (Ridge)	17.5	0.3	0.4	0.9	0.9	0.7	1.4	1.2	1.0	2.9	7.2	5.2	4.7	1.7	3.5	3.3	2.7	2.2	5.3	0.3	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	20.0	0.7	0.8	0.7	0.7	0.9	1.0	0.7	1.0	1.6	28.9	3.6	3.0	1.4	0.8	2.7	2.8	2.1	2.3	1.9	3.7	1.5	0.0	0.0	0.0
GLM (Elastic net)	20.0	0.7	0.8	0.6	0.7	0.9	1.1	0.7	1.0	1.8	29.0	3.8	2.9	1.3	1.0	2.9	2.7	2.1	2.2	2.1	4.1	0.1	0.0	0.0	0.0
GAM (fix knots)	24.8	0.4	0.6	1.3	0.8	1.3	0.9	3.1	1.1	4.3	5.8	4.7	4.6	2.8	4.0	3.3	1.9	2.8	2.3	3.1	4.0	3.0	5.6	1.2	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	26.9	5.9	0.0	18.4	10.3	9.5	13.5	8.7	0.0	0.0	0.0	6.5	0.0	0.0	0.0	0.0	0.0	0.5
MARS	18.5	0.4	0.4	0.8	1.1	0.7	1.2	7.5	13.6	12.3	5.2	10.8	13.5	18.3	2.8	10.4	4.4	0.9	3.7	0.0	5.7	0.0	0.6	6.0	0.0
ANN (opt. knots)	25.4	0.7	0.7	1.5	2.5	2.2	2.8	1.8	3.4	1.2	5.2	5.2	4.8	3.5	3.6	2.6	1.2	2.3	2.2	0.5	3.0	0.0	0.0	0.0	0.0
SVM	10.4	0.1	0.1	0.2	0.8	1.2	1.1	1.4	1.4	6.4	3.3	3.4	6.2	3.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	10.3	16.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	10.3	16.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using total and covariance calibration																									
GLM (OLS)	0.4	0.3	0.5	0.7	0.5	0.7	0.7	0.7	0.7	1.0	2.0	1.6	1.7	1.8	1.6	2.1	1.8	2.0	1.8	1.6	1.7	4.8	4.5	6.9	0.0
GLM (Ridge)	0.2	0.2	0.5	0.7	0.5	0.7	0.7	0.8	0.9	1.2	2.1	2.7	3.2	2.0	1.9	2.2	2.1	2.1	1.9	1.7	1.8	5.0	6.0	10.0	0.0
GLM (LASSO)	0.2	0.2	0.5	0.7	0.5	0.7	0.7	0.8	1.4	1.3	3.1	1.8	9.0	3.1	1.8	2.2	2.0	2.0	1.8	1.6	1.7	4.7	5.8	10.4	0.0
GLM (Elastic net)	0.2	0.2	0.5	0.6	0.5	0.7	0.7	0.8	1.1	1.0	2.3	1.8	9.0	2.8	1.6	2.2	1.9	2.0	1.7	1.6	1.7	4.7	5.4	8.4	0.0
GAM (fix knots)	0.8	0.4	0.5	0.5	0.5	0.7	0.5	1.1	1.2	1.1	2.8	4.9	5.8	2.9	2.4	2.6	2.5	2.0	1.6	1.8	1.7	5.1	4.6	6.9	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	3.4	3.0	21.9	22.6	21.2	21.8	25.5	21.7	23.3	13.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.4
MARS	4.6	0.8	0.9	1.4	1.7	1.7	2.9	17.5	14.8	13.7	11.2	17.9	12.7	5.4	5.4	1.5	1.2	0.7	1.6	0.2	2.5	4.3	6.3	6.2	4.3
ANN (opt. knots)	24.7	0.7	0.6	1.0	1.7	1.8	1.8	2.2	2.1	2.4	4.0	4.5	3.9	3.2	3.7	2.8	2.3	2.4	2.5	2.8	3.1	6.9	5.6	7.0	0.0
SVM	0.0	0.0	0.0	0.0	0.3	1.0	1.0	0.8	1.1	6.3	2.5	3.2	6.6	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	11.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	11.7	11.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table D.8: Estimated standard deviations (in percentage points) for income class frequencies estimated from the imputed Microcensus, using a weighted loss function for prediction models (continued)

Prediction model \ Income class	< 0 €	0 - 150 €	150 - 300 €	300 - 500 €	500 - 700 €	700 - 900 €	900 - 1100 €	1 100 - 1 300 €	1 300 - 1 500 €	1 500 - 1 700 €	1 700 - 2 000 €	2 000 - 2 300 €	2 300 - 2 600 €	2 600 - 2 900 €	2 900 - 3 200 €	3 200 - 3 600 €	3 600 - 4 000 €	4 000 - 4 500 €	4 500 - 5 000 €	5 000 - 5 500 €	5 500 - 6 000 €	6 000 - 7 500 €	7 500 - 10 000 €	10 000 - 18 000 €	> 18 000 €
Weighting model: Logit model (parametric) and GREG, using a propensity model, total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.5	0.4	0.1	0.3	0.3	0.3	0.9	1.2	1.1	0.8	0.6	0.6	0.5	1.1	1.3	1.1	0.7	0.7	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.4	0.4	0.1	0.3	0.3	0.3	0.9	1.5	1.1	1.1	0.7	0.6	0.7	1.1	1.4	1.1	0.7	0.8	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	19.3	19.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (parametric), using a propensity model, total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: Logit model (fix knots) and GREG, using a propensity model, total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.1	0.4	0.4	0.2	0.3	0.5	0.8	0.8	1.3	1.3	1.1	0.7	0.6	0.7	1.3	1.3	1.1	0.7	0.6	0.1	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.5	0.4	0.3	0.2	0.3	0.5	0.5	0.5	1.6	2.5	4.9	4.6	2.0	3.9	3.6	3.0	3.0	3.4	2.7	3.7	0.6	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	9.1	3.9	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	8.2	3.8	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (fix knots), using a propensity model, total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN (opt. knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weighting model: cal. ANN (opt. knots), using a propensity model, total and covariance calibration																									
GLM (OLS)	0.0	0.0	0.0	0.3	0.3	0.1	0.4	0.4	0.4	0.4	1.0	0.9	0.7	0.8	0.6	0.5	0.9	1.0	1.0	0.4	0.2	0.0	0.0	0.0	0.0
GLM (Ridge)	0.0	0.0	0.0	0.4	0.3	0.1	0.5	0.3	0.4	0.4	1.1	1.0	0.8	0.7	0.5	0.5	0.9	1.0	0.9	0.4	0.2	0.0	0.0	0.0	0.0
GLM (LASSO)	0.0	0.0	0.0	0.4	0.3	0.1	0.4	0.4	0.4	0.4	1.0	0.9	0.8	0.8	0.6	0.5	0.9	1.0	0.9	0.4	0.2	0.0	0.0	0.0	0.0
GLM (Elastic net)	0.0	0.0	0.0	0.3	0.3	0.1	0.4	0.4	0.4	0.4	1.0	0.9	0.8	0.8	0.6	0.5	0.9	1.0	0.9	0.4	0.2	0.0	0.0	0.0	0.0
GAM (fix knots)	0.0	0.0	0.0	0.3	0.3	0.2	0.4	0.4	0.4	0.4	1.0	0.9	0.7	0.8	0.6	0.4	0.9	1.0	0.9	0.4	0.2	0.0	0.0	0.0	0.0
Regression tree	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.9	0.0	0.0	24.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MARS	0.0	0.0	0.0	0.7	1.1	1.1	3.8	7.1	14.4	8.1	8.1	10.4	10.3	4.3	2.2	0.8	0.2	0.0	1.7	2.1	1.5	5.5	0.0	0.0	0.0
ANN (opt. knots)	0.9	0.4	0.7	0.9	1.2	1.8	1.6	1.8	1.6	1.7	2.7	2.1	2.0	2.5	1.4	1.6	2.4	2.6	0.6	0.2	0.0	0.0	0.0	0.0	0.0
SVM	0.0	0.0	0.0	0.0	0.1	0.4	0.4	0.4	0.4	0.8	3.0	1.8	2.3	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLMM	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAMM (fix knots)	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heckman model	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Bibliography

Abramowitz, M. and Stegun, I. A. (1970): *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 10th ed. Washington, D.C.: National Bureau of Standards of the United States.

Ahmadi, M. H., Mohseni-Gharyehsafa, B., Farzaneh-Gord, M., Jilte, R. D., Kumar, R. and Chau, K.-W. (2019): "Applicability of Connectionist Methods to Predict Dynamic Viscosity of Silver/water Nanofluid by Using ANN-MLP, MARS and MPR Algorithms." In: *Engineering Applications of Computational Fluid Mechanics 13 (1)*, pp. 220–228.

Ai, M., Yu, J., Zhang, H. and Wang, H. (2018): "Optimal Subsampling Algorithms for Big Data Regressions." In: *Arxiv Preprint Arxiv:1806.06761*.

Akaike, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle." In: *Proceedings of the 2nd International Symposium on Information Theory*. Budapest: Akadémiai Kiado, pp. 267–281.

Alfons, A., Filzmoser, P., Hulliger, B., Kolb, J.-P., Kraft, S., Münnich, R. and Templ, M. (2011): *Synthetic Data Generation of SILC Data*. Research Project Report WP6 – D6.2. FP7-SSH-2007-217322 AMELI.

Allcott, H. and Gentzkow, M. (2017): "Social Media and Fake News in the 2016 Election." In: *Journal of Economic Perspectives 31 (2)*, pp. 211–36.

Allen, D. M. (1974): "The Relationship between Variable Selection and Data Agumentation and a Method for Prediction." In: *Technometrics 16 (1)*, pp. 125–127.

Amarov, B. and Rendtel, U. (2013): "The Recruitment of the Access Panel of German Official Statistics from a Large Survey in 2006: Empirical Results and Methodological Aspects." In: *Survey Research Methods 7 (2)*, pp. 103–114.

Amemiya, T. (1981): "Qualitative Response Models: A Survey." In: *Journal of Economic Literature 19 (4)*, pp. 1483–1536.

Amemiya, T. (1985): *Advanced Econometrics*. Cambridge et al.: Harvard university press.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. and Sorensen, D. (1999): *LAPACK Users' Guide*. 3rd ed. Philadelphia, PA: SIAM.

Anderson, T. W. and Darling, D. A. (1952): "Asymptotic Theory of Certain" Goodness of Fit" Criteria Based on Stochastic Processes." In: *The Annals of Mathematical Statistics 23 (2)*, pp. 193–212.

Anderson, T. W. and Darling, D. A. (1954): "A Test of Goodness of Fit." In: *Journal of the American Statistical Association 49 (268)*, pp. 765–769.

Andridge, R. R. and Little, R. J. (2010): "A Review of Hot Deck Imputation for Survey Non-response." In: *International Statistical Review 78 (1)*, pp. 40–64.

Andridge, R. R., West, B. T., Little, R. J., Boonstra, P. S. and Alvarado-Leiton, F. (2019): "Indices of Non-ignorable Selection Bias for Proportions Estimated from Non-probability Samples." In: *Journal of the Royal Statistical Society: Series C (applied Statistics) 68 (5)*, pp. 1465–1483.

Armijo, L. (1966): "Minimization of Functions Having Lipschitz Continuous First Partial Derivatives." In: *Pacific Journal of Mathematics 16 (1)*, pp. 1–3.

Aspin, A. A. and Welch, B. (1949): "Tables for Use in Comparisons Whose Accuracy Involves Two Variances, Separately Estimated." In: *Biometrika 36 (3/4)*, pp. 290–296.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D. A., Frankel, M. R., Garland, C. P., Groves, R. M., Kennedy, C., Krosnick, J. A., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K. and Zahs, D. (2010): *AAPOR Report on Online Panels*. American Association for Public Opinion Research.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J. and Tourangeau, R. (2013a): *Report of the AAPOR Task Force on Non-probability Sampling*. American Association for Public Opinion Research.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J. and Tourangeau, R. (2013b): "Summary Report of the AAPOR Task Force on Non-probability Sampling." In: *Journal of Survey Statistics and Methodology 1 (2)*, pp. 90–143.

Bakin, S., Hegland, M. and Osborne, M. (1997): "Can Mars Be Improved with B-splines?" In: *Computational Techniques and Applications: CTAC97. Proceedings of the Eighth Biennial Conference*. Ed. by J. Noye, M. Teubner and A. Gill. Singapore: World Scientific Publishing.

Bakin, S., Hegland, M. and Osborne, M. R. (2000): "Parallel MARS Algorithm Based on B-splines." In: *Computational Statistics 15 (4)*, pp. 463–484.

Barendregt, C., Van der Poel, A. and Van de Mheen, D. (2005): "Tracing Selection Effects in Three Non-probability Samples." In: *European Addiction Research 11 (3)*, pp. 124–131.

Barratt, M. J., Ferris, J. A. and Lenton, S. (2015): "Hidden Populations, Online Purposive Sampling, and External Validity: Taking off the Blindfold." In: *Field Methods 27 (1)*, pp. 3–21.

Bates, D. (2018): *Computational Methods for Mixed Models. Vignette for Lme4*. URL: `http://btr0xq.rz.uni-bayreuth.de/math/statlib/R/CRAN/doc/vignettes/lme4/Theory.pdf` *(Retrieved 18.11.2019)*.

Bates, D. and Eddelbuettel, D. (2013): "Fast and Elegant Numerical Linear Algebra Using the RcppEigen Package." In: *Journal of Statistical Software 52 (5)*, pp. 1–24.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015): "Fitting Linear Mixed-effects Models Using Lme4." In: *Journal of Statistical Software 67 (1)*, pp. 1–48.

Beaumont, J.-F. (2000): "An Estimation Method for Nonignorable Nonresponse." In: *Survey Methodology 26 (2)*, pp. 131–136.

Beręsewicz, M. (2015): "On Representativeness of Internet Data Sources for Real Estate Market in Poland." In: *Austrian Journal of Statistics 44 (2)*, pp. 45–57.

Beręsewicz, M. (2016): "Internet Data Sources for Real Estate Market Statistics." Doctoral dissertation. Poznań University of Economics and Business.

Berger, R. L., Hsu, J. C. et al. (1996): "Bioequivalence Trials, Intersection-union Tests and Equivalence Confidence Sets." In: *Statistical Science 11 (4)*, pp. 283–319.

Berk, R. A. (2008): *Statistical Learning from a Regression Perspective*. Heidelberg: Springer.

Berkson, J. (1944): "Application of the Logistic Function to Bio-Assay." In: *Journal of the American Statistical Association 39 (227)*, pp. 357–365.

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C. and Weimer, D. L. (2003): "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples." In: *Political Analysis 11 (1)*, pp. 1–22.

Bethell, C., Fiorillo, J., Lansky, D., Hendryx, M. and Knickman, J. (2004): "Online Consumer Surveys As a Methodology for Assessing the Quality of the United States Health Care System." In: *Journal of Medical Internet Research 6 (1)*, e2.

Bethlehem, J. (1988): "Reduction of Nonresponse Bias through Regression Estimation." In: *Journal of Official Statistics 4 (3)*, p. 251.

Bethlehem, J. (2008a): *How Accurate Are Self-selection Web Surveys?* Discussion paper 08014. Statistics Netherlands.

Bethlehem, J. (2008b): "Representativity of Web Surveys–an Illusion?" In: *Access Panels and Online Research, Panacea or Pitfall? Proceedings of the DANS Symposium*. Ed. by I. Stoop and M. Wittenberg. The Hague: DANS Symposium Publications, pp. 19–44.

Bethlehem, J. (2010): "Selection Bias in Web Surveys." In: *International Statistical Review 78 (2)*, pp. 161–188.

Bethlehem, J. and Biffignandi, S. (2012): *Handbook of Web Surveys*. Hoboken: John Wiley & Sons.

Bianchi, A. and Biffignandi, S. (2013): "Web Panel Representativeness." In: *Statistical Models for Data Analysis*. Ed. by P. Giudici, S. Ingrassia and M. Vichi. Heidelberg: Springer, pp. 37–44.

Bianchi, A., Biffignandi, S. and Lynn, P. (2017): "Web-face-to-face Mixed-mode Design in a Longitudinal Survey: Effects on Participation Rates, Sample Composition, and Costs." In: *Journal of Official Statistics 33 (2)*, pp. 385–408.

Biffignandi, S. and Artaz, R. (2012): "Online Data Collection in the Agro-Food Sector." In: *Proceedings of the 4th International Conference on Establishment Surveys*. Montreal: American Statistical Association.

Biffignandi, S. and Bethlehem, J. (2012): "Web Surveys: Methodological Problems and Research Perspectives." In: *Advanced Statistical Methods for the Analysis of Large Data-sets*. Ed. by A. Di Ciaccio, M. Coli and J. M. A. Ibanez. Heidelberg: Springer, pp. 363–373.

Biffignandi, S. and Pratesi, M. (2000): "Modelling Firm Response and Contact Probabilities in Web Surveys." In: *Proceedings of the 2nd International Conference on Establishment Surveys*. Buffalo: American Statistical Association, pp. 1528–1533.

Biffignandi, S. and Pratesi, M. (2002): "Internet Surveys: The Role of Time in Italian Firms Response Behaviour." In: *Research in Official Statistics (2)*, pp. 53–66.

Biffignandi, S. and Pratesi, M. (2003): *Potentiality of Propensity Score Matching in Inference from Web-surveys: A Simulation Study*. Working paper N. 1 2003. University of Bergamo, Dipartimento matematica, statistica, informatica e applicazioni, pp. 1–23.

Biffignandi, S., Pratesi, M., Lozar Manfreda, K. and Vehovar, V. (2002): "List Assisted Web Surveys: Quality , Timeliness and Non-response in the Steps of the Participation Flow." In: *Proceedings of the International Conference on Improving Surveys 2002*. Copenhagen.

Binder, D. A. (1983): "On the Variances of Asymptotically Normal Estimators from Complex Surveys." In: *International Statistical Review 51 (3)*, pp. 279–292.

Binder, D. A. and Roberts, G. (2009): "Design- and Model-Based Inference for Model Parameters." In: *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*. Ed. by D. Pfeffermann and C. R. Rao. Amsterdam: Elsevier, pp. 11–31.

Birnbaum, Z. W. (1952): "Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size." In: *Journal of the American Statistical Association 47 (259)*, pp. 425–441.

Bishop, C. M. (1995): *Neural Networks for Pattern Recognition*. Oxford: Calrendon Press.

Blitzstein, J. K. and Hwang, J. (2013): *Introduction to Probability*. Boca Raton: CRC Press.

Blumenstock, J., Cadamuro, G. and On, R. (2015): "Predicting Poverty and Wealth from Mobile Phone Metadata." In: *Science 350 (6264)*, pp. 1073–1076.

Boggs, P. T. and Tolle, J. W. (1995): "Sequential Quadratic Programming." In: *Acta Numerica 4 (1)*, pp. 1–51.

Böhm, W. (1980): "Inserting New Knots into B-spline Curves." In: *Computer-aided Design 12 (4)*, pp. 199–201.

Böhm, W., Farin, G. and Kahmann, J. (1984): "A Survey of Curve and Surface Methods in CAGD." In: *Computer Aided Geometric Design 1 (1)*, pp. 1–60.

Boonstra, H. J. and Buelens, B. (2011): *Model-based Estimation*. Statistische Methoden 201106. Statistics Netherlands.

Boor, C. de (2001): *A Practical Guide to Splines*. Revised. New York: Springer.

Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992): "A Training Algorithm for Optimal Margin Classifiers." In: *Proceedings of the 5th Annual Workshop on Computational Learning Theory 1992, Pittsburgh*. New York: Association for Computing Machinery, pp. 144–152.

Bowley, A. L. (1925): *Measurement of the Precision Attained in Sampling*. Cambridge et al.: Cambridge University Press.

Box, G. E. (1953): "Non-normality and Tests on Variances." In: *Biometrika 40 (3/4)*, pp. 318–335.

Brath, A., Montanari, A. and Toth, E. (2002): "Neural Networks and Non-parametric Methods for Improving Real-time Flood Forecasting through Conceptual Hydrological Models." In: *Hydrology and Earth System Sciences 6 (4)*, pp. 627–639.

Braunsberger, K., Wybenga, H. and Gates, R. (2007): "A Comparison of Reliability between Telephone and Web-based Surveys." In: *Journal of Business Research 60 (7)*, pp. 758–764.

Braver, S. L. and Bay, R. C. (1992): "Assessing and compensating for self-selection bias (non-representativeness) of the family research sample." In: *Journal of Marriage and the Family 54 (4)*, pp. 925–939.

Breidenbach, J., McRoberts, R. E. and Astrup, R. (2016): "Empirical Coverage of Model-based Variance Estimators for Remote Sensing Assisted Estimation of Stand-level Timber Volume." In: *Remote Sensing of Environment 173*, pp. 274–281.

Breidt, F. J. and Opsomer, J. D. (2009): "Nonparametric and Semiparametric Estimation in Complex Surveys." In: *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*. Ed. by D. Pfeffermann and C. R. Rao. Amsterdam: Elsevier, pp. 103–119.

Breidt, F. J. and Opsomer, J. D. (2017): "Model-assisted Survey Estimation with Modern Prediction Techniques." In: *Statistical Science 32 (2)*, pp. 190–205.

Breiman, L. (2001a): "Random Forests." In: *Machine Learning 45 (1)*, pp. 5–32.

Breiman, L. (2001b): "Statistical Modeling: The Two Cultures." In: *Statistical Science 16 (3)*, pp. 199–231.

Breslow, N. (1970): "A Generalized Kruskal-Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship." In: *Biometrika 57 (3)*, pp. 579–594.

Breslow, N. E. and Clayton, D. G. (1993): "Approximate Inference in Generalized Linear Mixed Models." In: *Journal of the American Statistical Association 88 (421)*, pp. 9–25.

Brick, M. J. (2013): "Unit Nonresponse and Weighting Adjustments: A Critical Review." In: *Journal of Official Statistics 29 (3)*, pp. 329–353.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. and Stürmer, T. (2006): "Variable Selection for Propensity Score Models." In: *American Journal of Epidemiology 163 (12)*, pp. 1149–1156.

Broyden, C. G. (1970): "The Convergence of a Class of Double-rank Minimization Algorithms: 2. The New Algorithm." In: *Ima Journal of Applied Mathematics 6 (3)*, pp. 222–231.

Bruch, C., Münnich, R. and Zins, S. (2011): *Variance Estimation for Complex Surveys*. Research Project Report WP3 – D3.1. FP7-SSH-2007-217322 AMELI.

Buchanan, T. and Smith, J. L. (1999): "Using the Internet for Psychological Research: Personality Testing on the World Wide Web." In: *British Journal of Psychology 90 (1)*, pp. 125–144.

Buelens, B., Boonstra, H. J., van den Brakel, J. and Daas, P. (2012): *Shifting Paradigms in Official Statistics: From Design-based to Model-based to Algorithmic Inferences*. Discussion paper 201218. Statistics Netherlands.

Buelens, B., Burger, J. and van den Brakel, J. A. (2015): *Predictive Inference for Non-probability Samples: A Simulation Study*. Discussion paper 2015|13. Statistics Netherlands.

Buelens, B., Burger, J. and van den Brakel, J. A. (2018): "Comparing Inference Methods for Non-probability Samples." In: *International Statistical Review 86 (2)*, pp. 322–343.

Buelens, B., Daas, P., Burger, J., Puts, M. and van den Brakel, J. (2014): *Selectivity of Big Data*. Discussion paper 2014|11. Statistics Netherlands.

Buja, A., Hastie, T. J. and Tibshirani, R. J. (1989): "Linear Smoothers and Additive Models." In: *The Annals of Statistics 17 (2)*, pp. 453–510.

Burgard, J. P. (2013): "Evaluation of Small Area Techniques for Applications in Official Statistics." Doctoral dissertation. University of Trier.

Burgard, J. P., Dörr, P. and Münnich, R. (2020): *Monte-Carlo Simulation Studies in Survey Statistics–An Appraisal*. Research Papers in Economics No 4/2020. University of Trier, Department of Economics.

Burgard, J. P., Ertz, F., Merkle, H. and Münnich, R. (2017a): *AMELIA-Data Description V0. 2.2.* AMELIA-Data Description V0. 2.2. University of Trier, Economic and Social Statistics Department. URL: `http://amelia.uni-trier.de/wp-content/uploads/2017/11/AMELIA_Data_Description_v0.2.2.1.pdf` *(Retrieved 08.05.2020).*

Burgard, J. P., Kolb, J.-P., Merkle, H. and Münnich, R. (2017b): "Synthetic Data for Open and Reproducible Methodological Research in Social Sciences and Official Statistics." In: *Asta Wirtschafts-und Sozialstatistisches Archiv 11 (3-4)*, pp. 233–244.

Burgard, J. P., Münnich, R. and Rupp, M. (2019): *A Generalized Calibration Approach Ensuring Coherent Estimates with Small Area Constraints*. Research Papers in Economics No 10/2019. University of Trier, Department of Economics.

Burgard, J. P., Münnich, R. and Rupp, M. (2020): "Qualitätszielfunktionen Für Stark Variierende Gemeindegrößen Im Zensus 2021." In: *Asta Wirtschafts-und Sozialstatistisches Archiv 14*, pp. 5–65.

Businger, P. and Golub, G. H. (1965): "Linear Least Squares Solutions by Householder Transformations." In: *Numerische Mathematik 7 (3)*, pp. 269–276.

Buskirk, T. D. and Kolenikov, S. (2015): *Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification*. Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach. URL: `https://surveyinsights.org/?p=5108` *(Retrieved 09.08.2019).*

Cajori, F. (1911): "Historical Note on the Newton-Raphson Method of Approximation." In: *The American Mathematical Monthly 18 (2)*, pp. 29–32.

Callegaro, M. (2013): "Paradata in Web Surveys." In: *Improving Surveys with Paradata: Analytic Use of Process Information*. Ed. by F. Kreuter. Hoboken: Wiley & Sons, pp. 261–279.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1976): "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations." In: *Biometrika 63 (3)*, pp. 615–620.

Ceron, A., Curini, L., Iacus, S. M. and Porro, G. (2014): "Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve Our Knowledge of Citizens' Political Preferences with an Application to Italy and France." In: *New Media & Society 16 (2)*, pp. 340–358.

Chambers, R. L., Dorfman, A. H. and Wehrly, T. E. (1993): "Bias Robust Estimation in Finite Populations Using Nonparametric Calibration." In: *Journal of the American Statistical Association 88 (421)*, pp. 268–277.

Chang, C.-C. and Lin, C.-J. (2011): "LIBSVM: A Library for Support Vector Machines." In: *Acm Transactions on Intelligent Systems and Technology 2 (3)*, pp. 1–27.

Chang, M.-W., Lin, H.-T., Tsai, M.-H., Ho, C.-H. and Yu, H.-F. (n.d.): *LIBSVM: Weights for Data Instances*. URL: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#weights_for_data_instances` *(Retrieved 12.12.2019)*.

Chang, T. and Kott, P. S. (2008): "Using Calibration Weighting to Adjust for Nonresponse under a Plausible Model." In: *Biometrika 95 (3)*, pp. 555–571.

Checchi, D. and García-Peñalosa, C. (2010): "Labour Market Institutions and the Personal Distribution of Income in the OECD." In: *Economica 77 (307)*, pp. 413–450.

Chen, J. K. T., Valliant, R. L. and Elliott, M. R. (2019): "Calibrating Non-probability Surveys to Estimated Control Totals Using LASSO, with an Application to Political Polling." In: *Journal of the Royal Statistical Society: Series C (applied Statistics) 68 (3)*, pp. 657–681.

Chiang, W.-y. K., Zhang, D. and Zhou, L. (2006): "Predicting and Explaining Patronage Behavior toward Web and Traditional Stores Using Neural Networks: A Comparative Analysis with Logistic Regression." In: *Decision Support Systems 41 (2)*, pp. 514–531.

Chipperfield, J. and Preston, J. (2007): "Efficient Bootstrap for Business Surveys." In: *Survey Methodology 33 (2)*, pp. 167–172.

Cho, D.-H. (2010): "Mixed-effects LS-SVR for Longitudinal Dat." In: *Journal of the Korean Data and Information Science Society 21 (2)*, pp. 363–369.

Choi, W., Lee, J. W., Huh, M.-H. and Kang, S.-H. (2003): "An Algorithm for Computing the Exact Distribution of the Kruskal–Wallis Test." In: *Communications in Statistics: Simulation and Computation 32 (4)*, pp. 1029–1040.

Citro, C. F. (2014): "From Multiple Modes for Surveys to Multiple Data Sources for Estimates." In: *Survey Methodology 40 (2)*, pp. 137–161.

Cochran, W. and Rubin, D. (1973): "Controlling Bias in Observational Studies." In: *Sankyha 35 (4)*, pp. 417–446.

Cochran, W. G. (1968): "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." In: *Biometrics 24 (2)*, pp. 295–313.

Cochran, W. G. and Chambers, S. P. (1965): "The Planning of Observational Studies of Human Populations." In: *Journal of the Royal Statistical Society. Series a (general) 128 (2)*, pp. 234–266.

Cochran, W. G., Moses, L. E. and Mosteller, F. (1983): *Planning and Analysis of Observational Studies*. New York et al.: John Wiley & Sons.

Cochran, W. G. (1977): *Sampling Techniques*. 3. New York et al.: John Wiley & Sons.

Cole, S. T. (2005): "Comparing Mail and Web-based Survey Distribution Methods: Results of Surveys to Leisure Travel Retailers." In: *Journal of Travel Research 43 (4)*, pp. 422–430.

Cortes, C. and Vapnik, V. (1995): "Support-vector Networks." In: *Machine Learning 20 (3)*, pp. 273–297.

Coull, B. A., Ruppert, D. and Wand, M. (2001): "Simple Incorporation of Interactions into Additive Models." In: *Biometrics 57 (2)*, pp. 539–545.

Couper, M. P. (2000): "Web Surveys: A Review of Issues and Approaches." In: *Public Opinion Quarterly 64 (4)*, pp. 464–494.

Craven, P. and Wahba, G. (1979): "Smoothing Noisy Data with Spline Functions." In: *Numerische Mathematik 31 (4)*, pp. 377–403.

Cumming, R. G. (1990): "Is Probability Sampling Always Better? A Comparison of Results from a Quota and a Probability Sample Survey." In: *Community Health Studies 14 (2)*, pp. 132–137.

Curry, H. B. and Schoenberg, I. J. (1947): "On Spline Distributions and Their Limits-the Polya Distribution Functions." In: *Bulletin of the American Mathematical Society 53 (11)*, pp. 1114–1114.

Curry, H. B. and Schoenberg, I. J. (1966): "On Pólya Frequency Functions IV: The Fundamental Spline Functions and Their Limits." In: *Journal D'analyse Mathématique 17*, pp. 71–107.

Daas, P. J., Puts, M. J., Buelens, B. and Hurk, P. A. van den (2015): "Big Data As a Source for Official Statistics." In: *Journal of Official Statistics 31 (2)*, pp. 249–262.

Darling, D. A. (1957): "The Kolmogorov-Smirnov, Cramer-von Mises Tests." In: *The Annals of Mathematical Statistics 28 (4)*, pp. 823–838.

Davidon, W. C. (1959): *Variable Metric Method for Minimization*. A.E.C. Research and Development Report. ANL-5990 (TID-4500) 14. Argonne National Laboratory.

Dawid, A. P. (1979): "Conditional Independence in Statistical Theory." In: *Journal of the Royal Statistical Society: Series B (methodological) 41 (1)*, pp. 1–15.

de Boor, C. (1972): "On Calculating with B-splines." In: *Journal of Approximation Theory 6 (1)*, pp. 50–62.

de Boor, C. (1978): *A Practical Guide to Splines*. eng. New York et al.: Springer.

de Boor, C. and Rice, J. R. (1968a): *Least Squares Cubic Spline Approximation I-Fixed Knots*. Technical Reports. Paper 141. Purdue University, Department of Computer Science.

de Boor, C. and Rice, J. R. (1968b): *Least Squares Cubic Spline Approximation, II-variable Knots*. Technical Reports. Paper 149. Purdue University, Department of Computer Science.

de Heer, W. (1999): "International Response Trends: Results of an International Survey." In: *Journal of Official Statistics 15 (2)*, p. 129.

de Heer, W. and de Leeuw, E. (2002): "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In: *Survey Nonresponse*. Ed. by R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little. New York: John Wiley, pp. 41–54.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P. and Meester, L. E. (2005): *A Modern Introduction to Probability and Statistics: Understanding Why and How*. London: Springer Science & Business Media.

Deming, W. E. and Stephan, F. F. (1940): "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known." In: *The Annals of Mathematical Statistics 11 (4)*, pp. 427–444.

Deming, W. E. (1950): *Some Theory of Sampling*. New York: John Wiley & Sons.

Demmel, J. W. (1997): *Applied Numerical Linear Algebra*. Philadelphia, PA: SIAM.

Demmel, J. W., Gilbert, J. R. and Li, X. S. (1999): "An Asynchronous Parallel Supernodal Algorithm for Sparse Gaussian Elimination." In: *Siam Journal on Matrix Analysis and Applications 20 (4)*, pp. 915–952.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977): "Maximum Likelihood from Incomplete Data Via the EM Algorithm." In: *Journal of the Royal Statistical Society: Series B (methodological) 39 (1)*, pp. 1–22.

Dever, J. A., Rafferty, A. and Valliant, R. (2008): "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?" In: *Survey Research Methods 2 (2)*, pp. 47–60.

Deville, J.-C. and Särndal, C.-E. (1992): "Calibration Estimators in Survey Sampling." In: *Journal of the American Statistical Association 87 (418)*, pp. 376–382.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993): "Generalized Raking Procedures in Survey Sampling." In: *Journal of the American Statistical Association 88 (423)*, pp. 1013–1020.

Diamond, A. and Sekhon, J. S. (2013): "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." In: *Review of Economics and Statistics 95 (3)*, pp. 932–945.

Directors-General of the National Statistical Institutes (2013): *Scheveningen Memorandum: Big Data and Official Statistics*. URL: `https://ec.europa.eu/eurostat/documents/7330775/7339365/Scheveningen-memorandum-27-09-13/2e730cdc-862f-4f27-bb43-2486c30298b6` *(Retrieved 09.08.2019)*.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. and Danforth, C. M. (2011): "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter." In: *Plos One 6 (12)*, e26752.

Dong, X., Bian, Y., Tsong, Y. and Wang, T. (2017): "Exact Test-based Approach for Equivalence Test with Parameter Margin." In: *Journal of Biopharmaceutical Statistics 27 (2)*, pp. 317–330.

Dowle, M. and Srinivasan, A. (2019): *Data.table: Extension of 'data.frame'*. R package version 1.12.8. URL: `https://cran.r-project.org/package=data.table` *(Retrieved 05.05.2021)*.

Drabble, L. A., Trocki, K. F., Korcha, R. A., Klinger, J. L., Veldhuis, C. B. and Hughes, T. L. (2018): "Comparing Substance Use and Mental Health Outcomes among Sexual Minority and Heterosexual Women in Probability and Non-probability Samples." In: *Drug and Alcohol Dependence 185*, pp. 285–292.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J. and Vapnik, V. (1997): "Support Vector Regression Machines." In: *Advances in Neural Information Processing Systems 9, 1996*. Ed. by M. C. Mozer, M. I. Jordan and T. Petsche. Cambridge: MIT Press, pp. 155–161.

Durbin, J. (1953): "Some Results in Sampling Theory When the Units Are Selected with Unequal Probabilities." In: *Journal of the Royal Statistical Society: Series B (methodological) 15 (2)*, pp. 262–269.

Durrant, G. B., D'Arrigo, J. and Müller, G. (2013): "Modeling Call Record Data: Examples from Cross-sectional and Longitudinal Surveys." In: *Improving Surveys with Paradata: Analytic Uses of Process Information*. Ed. by F. Kreuter. Hoboken: Wiley & Sons, pp. 281–308.

Eck, M. and Hadenfeld, J. (1995): "Knot Removal for B-spline Curves." In: *Computer Aided Geometric Design 12 (3)*, pp. 259–282.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J. and Bates, D. (2011): "Rcpp: Seamless `R` and `C++` Integration." In: *Journal of Statistical Software 40 (8)*, pp. 1–18.

Eddelbuettel, D. and Sanderson, C. (2014): "RcppArmadillo: Accelerating `R` with High-performance `C++` Linear Algebra." In: *Computational Statistics & Data Analysis 71*, pp. 1054–1063.

Efron, B. (1979): "Bootstrap Methods: Another Look at the Jackknife." In: *Annals of Statistics 7 (1)*, pp. 1–26.

Efron, B. (1981): "Nonparametric Standard Errors and Confidence Intervals." In: *Canadian Journal of Statistics 9 (2)*, pp. 139–158.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004): "Least Angle Regression." In: *The Annals of Statistics 32 (2)*, pp. 407–499.

Efron, B. and Tibshirani, R. (1986): "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." In: *Statistical Science 1*, pp. 54–75.

Efron, B. and Tibshirani, R. (1998): *An Introduction to the Bootstrap*. Boca Raton: Chapman & Hall/CRC.

Ehrgott, M. (2005): *Multicriteria Optimization*. Berlin: Springer.

Eilers, P. H. and Marx, B. D. (1996): "Flexible Smoothing with B-splines and Penalties." In: *Statistical Science 11 (2)*, pp. 89–102.

Einstein, H. H. and Baecher, G. B. (1983): "Probabilistic and Statistical Methods in Engineering Geology." In: *Rock Mechanics and Rock Engineering 16 (1)*, pp. 39–72.

Elliott, M. R. (2009): "Combining Data from Probability and Non-probability Samples Using Pseudo-weights." In: *Survey Practice 2 (6)*, pp. 1–7.

Elliott, M. R. and Valliant, R. (2017): "Inference for Nonprobability Samples." In: *Statistical Science 32 (2)*, pp. 249–264.

Enderle, T., Münnich, R. and Bruch, C. (2013): "On the Impact of Response Patterns on Survey Estimates from Access Panels." In: *Survey Research Methods 7 (2)*, pp. 91–101.

Estevao, V. M. and Särndal, C.-E. (2000): "A Functional Form Approach to Calibration." In: *Journal of Official Statistics 16 (4)*, pp. 379–399.

Estevao, V. M. and Särndal, C.-E. (2006): "Survey Estimates by Calibration on Complex Auxiliary Information." In: *International Statistical Review 74 (2)*, pp. 127–147.

European Statistical System (2014): *The ESS Report 2013*. European Statistical System. URL: `https://ec.europa.eu/eurostat/documents/3217494/5784877/KS-FN-13-001-EN.PDF/85767d7e-e218-4041-b442-cc9526e86f92?version=1.0` *(Retrieved 04.03.2021)*.

Faas, T. and Schoen, H. (2006): "Putting a Questionnaire on the Web Is Not Enough-A Comparison of Online and Offline Surveys Conducted in the Context of the German Federal Election 2002." In: *Journal of Official Statistics 22 (2)*, p. 177.

Fan, J. and Li, R. (2006): "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery." In: *Proceedings of the International Congress of Mathematicians 2006*. Madrid: European Mathematical Society, pp. 595–622.

Fan, R.-E., Chen, P.-H. and Lin, C.-J. (2005): "Working Set Selection Using Second Order Information for Training Support Vector Machines." In: *Journal of Machine Learning Research 6 (Dec)*, pp. 1889–1918.

Faraway, J. J. (2002): *Practical Regression and ANOVA Using R*. Bath: University of Bath.

Farebrother, R. W. (1999): *Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900*. New York: Springer Science & Business Media.

Feild, L., Pruchno, R. A., Bewley, J., Lemay, E. P. and Levinsky, N. G. (2006): "Using Probability Vs. Nonprobability Sampling to Identify Hard-to-Access Participants for Health-Related Research Costs and Contrasts." In: *Journal of Aging and Health 18 (4)*, pp. 565–583.

Feller, W. et al. (1948): "On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions." In: *The Annals of Mathematical Statistics 19 (2)*, pp. 177–189.

Ferrari, S. and Cribari-Neto, F. (2004): "Beta Regression for Modelling Rates and Proportions." In: *Journal of Applied Statistics 31 (7)*, pp. 799–815.

Fisher, R. A. (1922a): "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P." In: *Journal of the Royal Statistical Society 85 (1)*, pp. 87–94.

Fisher, R. A. (1922b): "On the Mathematical Foundations of Theoretical Statistics." In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 222 (594-604)*, pp. 309–368.

Fisher, R. A. (1925): "Theory of Statistical Estimation." In: *Mathematical Proceedings of the Cambridge Philosophical Society 22 (5)*, pp. 700–725.

Fletcher, R. and Reeves, C. M. (1964): "Function Minimization by Conjugate Gradients." In: *The Computer Journal 7 (2)*, pp. 149–154.

Fletcher, R. (1970): "A New Approach to Variable Metric Algorithms." In: *The Computer Journal 13 (3)*, pp. 317–322.

Fletcher, R. (1971): "A General Quadratic Programming Algorithm." In: *Ima Journal of Applied Mathematics 7 (1)*, pp. 76–91.

Fletcher, R. (1972): "An Algorithm for Solving Linearly Constrained Optimization Problems." In: *Mathematical Programming 2 (1)*, pp. 133–165.

Fletcher, R. and Powell, M. J. (1963): "A Rapidly Convergent Descent Method for Minimization." In: *The Computer Journal 6 (2)*, pp. 163–168.

Folgheraiter, M. (2016): "A Combined B-spline-neural-network and ARX Model for Online Identification of Nonlinear Dynamic Actuation Systems." In: *Neurocomputing 175*, pp. 433–442.

Folsom, R. E. and Singh, A. C. (2000): "The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratification." In: *JSM Proceedings, Survey Research Methods Section*. Alexandria: American Statistical Association, pp. 598–603.

Fondeur, Y. and Karamé, F. (2013): "Can Google Data Help Predict French Youth Unemployment?" In: *Economic Modelling 30*, pp. 117–125.

Forster, J. J. and Smith, P. W. (1998): "Model-based Inference for Categorical Survey Data Subject to Non-ignorable Non-response." In: *Journal of the Royal Statistical Society: Series B (statistical Methodology) 60 (1)*, pp. 57–70.

Frank, O. and Snijders, T. (1994): "Estimating the Size of Hidden Populations Using Snowball Sampling." In: *Journal of Official Statistics 10 (1)*, pp. 53–53.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. et al. (2007): "Pathwise Coordinate Optimization." In: *The Annals of Applied Statistics 1 (2)*, pp. 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2010): "Regularization Paths for Generalized Linear Models Via Coordinate Descent." In: *Journal of Statistical Software 33 (1)*, p. 1.

Friedman, J. H. (1996): *Another Approach to Polychotomous Classification*. Technical Report. Stanford University, Department of Statistics.

Friedman, J. H. and Stuetzle, W. (1981): "Projection Pursuit Regression." In: *Journal of the American Statistical Association 76 (376)*, pp. 817–823.

Friedman, J. H. (1991a): *Estimating Functions of Mixed Ordinal and Categorical Variables Using Adaptive Splines*. Technical Report No. 108. Stanford University, Department of Statistics.

Friedman, J. H. (1991b): "Multivariate Adaptive Regression Splines." In: *The Annals of Statistics 19 (1)*, pp. 1–67.

Friedman, J. H. (1993): *Fast MARS*. Technical Report No. 110. Stanford University, Department of Statistics.

Fuller, W. A. (2002): "Regression Estimation for Survey Samples." In: *Survey Methodology 28 (1)*, pp. 5–24.

Fuller, W. A. (2009): *Sampling Statistics*. Hoboken: John Wiley & Sons.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., Rossi, F. and Ulerich, R. (2009): *GNU Scientific Library Reference Manual*. 3rd ed. Network Theory Ltd.

Galton, F. (1886): "Regression Towards Mediocrity in Hereditary Stature." In: *The Journal of the Anthropological Institute of Great Britain and Ireland 15*, pp. 246–263.

Ganesh, N., Pineau, V., Chakraborty, A. and Dennis, J. M. (2017): "Combining Probability and Non-Probability Samples Using Small Area Estimation." In: *JSM Proceedings, Survey Research Methods Section*. Alexandria: American Statistical Association, pp. 1657–1667.

Gauss, C. F. (1809): *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium Auctore Carolo Friderico Gauss*. Hamburg: sumtibus Frid. Perthes et IH Besser.

Geiger, C. and Kanzow, C. (2002): *Theorie Und Numerik Restringierter Optimierungsaufgaben*. Berlin: Springer.

Gelman, A. (2007): "Struggles with Survey Weighting and Regression Modeling." In: *Statistical Science 22 (2)*, pp. 153–164.

Gelman, A., Goel, S., Rivers, D. and Rothschild, D. (2016a): "The Mythical Swing Voter." In: *Quarterly Journal of Political Science 11 (1)*, pp. 103–130.

Gelman, A., Goel, S., Rothschild, D. and Wang, W. (2016b): "High-frequency Polling with Non-representative Data." In: *Political Communication in Real Time. Theoretical and Applied Research Approaches*. Ed. by D. Schill, R. Kirk and A. E. Jasperson. New York: Routledge, pp. 89–105.

Gelman, A. and Little, T. C. (1997): "Poststratification into Many Categories Using Hierarchical Logistic Regression." In: *Survey Methodology 23 (2)*, pp. 127–135.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C. and Villa-Vialaneix, N. (2017): "Random Forests for Big Data." In: *Big Data Research 9*, pp. 28–46.

Genz, A. and Bretz, F. (2009): *Computation of Multivariate Normal and T Probabilities*. Lecture Notes in Statistics. Heidelberg: Springer.

Ghitza, Y. and Gelman, A. (2013): "Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups." In: *American Journal of Political Science 57 (3)*, pp. 762–776.

Ghosh, M., Natarajan, K., Stroud, T. and Carlin, B. P. (1998): "Generalized Linear Models for Small-area Estimation." In: *Journal of the American Statistical Association 93 (441)*, pp. 273–282.

Giles, D. E. (2001): "A Saddlepoint Approximation to the Distribution Function of the Anderson-Darling Test Statistic." In: *Communications in Statistics-simulation and Computation 30 (4)*, pp. 899–905.

Gill, P. E. and Murray, W. (1978): "Numerically Stable Methods for Quadratic Programming." In: *Mathematical Programming 14 (1)*, pp. 349–372.

Gill, P. E., Murray, W. and Wright, M. H. (1981): *Practical Optimization*. London: Academic press.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L. (2009): "Detecting Influenza Epidemics Using Search Engine Query Data." In: *Nature 457 (7232)*, p. 1012.

Goldfarb, D. (1970): "A Family of Variable-metric Methods Derived by Variational Means." In: *Mathematics of Computation 24 (109)*, pp. 23–26.

Goller, C. and Kuchler, A. (1996): *Learning Task-dependent Distributed Representations by Backpropagation through Structure*. Technical Report Report AR-95-02. Technische Universität München, Fakultät für Informatik.

Golub, G. H., Heath, M. and Wahba, G. (1979): "Generalized Cross-validation As a Method for Choosing a Good Ridge Parameter." In: *Technometrics 21 (2)*, pp. 215–223.

Golub, H. and Van Loan, C. F. (1996): *Matrix Computations, Johns Hopkins Uni*. 3rd ed. Baltimore and London: Johns Hopkins University Press.

Govindaraju, V., Raghavan, V. and Rao, C. R. (2015): *Handbook of Statistics 33: Big Data Analytics*. Amsterdam: Elsevier.

Graf, M., Alfons, A., Bruch, C., Filzmoser, P., Hulliger, B., Lehtonen, R., Meindl, B., Münnich, R., Schoch, T., Templ, M., Valaste, M., Wenger, A. and Zins, S. (2011): *State-of-the-art of Laeken Indicators*. Research Project Report WP1 – D1.1. FP7-SSH-2007-217322 AMELI.

Green, P. J. (1984): "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives." In: *Journal of the Royal Statistical Society: Series B (methodological) 46 (2)*, pp. 149–170.

Green, P. and Silverman, B. (1994): *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*. London: Chapman & Hall.

Greenacre, Z. A. (2016): "The Importance of Selection Bias in Internet Surveys." In: *Open Journal of Statistics 6 (03)*, pp. 397–404.

Greene, W. H. (2008): *Econometric Analysis*. 6th ed. Upper Saddle River: Pearson Education India.

Greenstadt, J. (1970): "Variations on Variable-metric Methods." In: *Mathematics of Computation 24 (109)*, pp. 1–22.

Groves, R. M. (1989): *Survey Errors and Survey Costs*. Hoboken: John Wiley & Sons.

Groves, R. M. (2011): "Three Eras of Survey Research." In: *Public Opinion Quarterly 75 (5)*, pp. 861–871.

Groves, R. M. and Couper, M. P. (1998): *Nonresponse in Household Interview Surveys*. New York et al.: John Wiley & Sons.

Groves, R. M., Presser, S. and Dipko, S. (2004): "The Role of Topic Interest in Survey Participation Decisions." In: *Public Opinion Quarterly 68 (1)*, pp. 2–31.

Gu, C. (1990): "Adaptive Spline Smoothing in Non-Gaussian Regression Models." In: *Journal of the American Statistical Association 85 (411)*, pp. 801–807.

Gu, C. (1992): "Cross-validating Non-Gaussian Data." In: *Journal of Computational and Graphical Statistics 1 (2)*, pp. 169–179.

Guarnieri, S., Piazza, F. and Uncini, A. (1999): "Multilayer Feedforward Networks with Adaptive Spline Activation Function." In: *Ieee Transactions on Neural Networks 10 (3)*, pp. 672–683.

Guarte, J. M. and Barrios, E. B. (2006): "Estimation under Purposive Sampling." In: *Communications in Statistics–simulation and Computation 35 (2)*, pp. 277–284.

Guennebaud, G., Jacob, B. et al. (2010): *Eigen V3*. URL: `http://eigen.tuxfamily.org` *(Retrieved 03.06.2019)*.

Guggemos, F. and Tillé, Y. (2010): "Penalized Calibration in Survey Sampling: Design-based Estimation Assisted by Mixed Models." In: *Journal of Statistical Planning and Inference 140 (11)*, pp. 3199–3212.

Hackbusch, W. (1994): *Iterative Solution of Large Sparse Systems of Equations*. Berlin: Springer.

Hagan, M. T., Demuth, H. B., Beale, M. H. and De Jesús, O. (1996): *Neural Network Design*. 2nd ed. Boston: Pws Pub.

Hager, W. W. (1989): "Updating the Inverse of a Matrix." In: *Siam Review 31 (2)*, pp. 221–239.

Hahn, E. D. and Soyer, R. (2005): *Probit and Logit Models: Differences in the Multivariate Realm*.

Han, S.-P. (1977): "A Globally Convergent Method for Nonlinear Programming." In: *Journal of Optimization Theory and Applications 22 (3)*, pp. 297–309.

Hansen, M. H. and Hurwitz, W. N. (1943): "On the Theory of Sampling from Finite Populations." In: *The Annals of Mathematical Statistics 14 (4)*, pp. 333–362.

Harville, D. A. (1977): "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems." In: *Journal of the American Statistical Association 72 (358)*, pp. 320–338.

Hastie, T. and Tibshirani, R. (1998): "Classification by Pairwise Coupling." In: *The Annals of Statistics 26 (2)*, pp. 451–471.

Hastie, T. J. and Tibshirani, R. J. (1984): *Generalized Additive Models*. Technical Report No. 98. Standord University, Division of Biostatistics, pp. 371–386.

Hastie, T. J. and Tibshirani, R. J. (1986): "Generalized Additive Models." In: *Statistical Science 1 (3)*, pp. 297–318.

Hastie, T. J. and Tibshirani, R. J. (1990): *Generalized Additive Models*. London: Chapman & Hall.

Hastie, T. J., Tibshirani, R. J. and Friedman, J. (2008): *The Elements of Statistical Learning*. 2nd ed. New York: Springer.

Hawkins, D. M. (1981): "A New Test for Multivariate Normality and Homoscedasticity." In: *Technometrics 23 (1)*, pp. 105–110.

He, T. (2011): "Lasso and General L1-regularized Regression under Linear Equality and Inequality Constraints." Doctoral dissertation. Purdue University.

Heckathorn, D. D. (2002): "Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations." In: *Social Problems 49 (1)*, pp. 11–34.

Heckman, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." In: *Annals of Economic and Social Measurement 5 (4)*, pp. 475–492.

Heckman, J. J. (1979): "Sample Selection Bias As a Specification Error." In: *Econometrica 47 (1)*, pp. 153–161.

Heckman, J. J., Ichimura, H., Smith, J. and Todd, P. (1998): "Characterizing Selection Bias Using Experimental Data." In: *Econometrica 66 (5)*, pp. 1017–1098.

Heij, V. de, Schouten, B. and Shlomo, N. (2010): *RISQ Manual, Tools in SAS and R for the Computation of R-indicators and Partial R-indicators*. Research Project Deliverable 12.1. RISQ.

Henderson, C. R. (1950): "Estimation of Genetic Parameters." In: *The Annals of Mathematical Statistics 21 (2)*, pp. 309–310.

Henderson, C. R. (1953): "Estimation of Variance and Covariance Components." In: *Biometrics 9 (2)*, pp. 226–252.

Henderson, C. R. (1963): "Selection Index and Expected Genetic Advance." In: *Statisitical Genetics and Plant Breeding 982*, pp. 141–163.

Henderson, C. R., Kempthorne, O., Searle, S. R. and Von Krosigk, C. (1959): "The Estimation of Environmental and Genetic Trends from Records Subject to Culling." In: *Biometrics 15 (2)*, pp. 192–218.

Hestenes, M. R. and Stiefel, E. (1952): "Methods of Conjugate Gradients for Solving Linear Systems." In: *Journal of Research of the National Bureau of Standards 49 (6)*, pp. 409–436.

Hinkins, S., Oh, H. L. and Scheuren, F. (1997): "Inverse Sampling Design Algorithms." In: *Survey Methodology 23 (1)*, pp. 11–22.

Hinkley, D. and Shi, S. (1989): "Importance Sampling and the Nested Bootstrap." In: *Biometrika 76 (3)*, pp. 435–446.

Hirano, K., Imbens, G. W. and Ridder, G. (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." In: *Econometrica 71 (4)*, pp. 1161–1189.

Hitchman, S. C., Brose, L. S., Brown, J., Robson, D. and McNeill, A. (2015): "Associations between E-cigarette Type, Frequency of Use, and Quitting Smoking: Findings from a Longitudinal Online Panel Survey in Great Britain." In: *Nicotine & Tobacco Research 17 (10)*, pp. 1187–1194.

Hoerl, A. E. and Kennard, R. W. (1970): "Ridge Regression: Biased Estimation for Nonorthogonal Problems." In: *Technometrics 12 (1)*, pp. 55–67.

Holm, S. (1979): "A Simple Sequentially Rejective Multiple Test Procedure." In: *Scandinavian Journal of Statistics 6 (2)*, pp. 65–70.

Holt, D. T. (2007): "The Official Statistics Olympic Challenge: Wider, Deeper, Quicker, Better, Cheaper." In: *The American Statistician 61 (1)*, pp. 1–8.

Hong, S. and Nadler, D. (2012): "Which Candidates Do the Public Discuss Online in an Election Campaign?: The Use of Social Media by 2012 Presidential Candidates and Its Impact on Candidate Salience." In: *Government Information Quarterly 29 (4)*, pp. 455–461.

Hong, X. and Chen, S. (2011): "Modeling of Complex-valued Wiener Systems Using B-spline Neural Network." In: *Ieee Transactions on Neural Networks 22 (5)*, pp. 818–825.

Hornik, K., Stinchcombe, M., White, H. et al. (1989): "Multilayer Feedforward Networks Are Universal Approximators." In: *Neural Networks 2 (5)*, pp. 359–366.

Horvitz, D. G. and Thompson, D. J. (1952): "A Generalization of Sampling without Replacement from a Finite Universe." In: *Journal of the American Statistical Association 47 (260)*, pp. 663–685.

Iacus, S. M., King, G. and Porro, G. (2009): "CEM: Software for Coarsened Exact Matching." In: *Journal of Statistical Sof 30 (9)*, pp. 1–27.

Iacus, S. M., King, G. and Porro, G. (2011): "Multivariate Matching Methods That Are Monotonic Imbalance Bounding." In: *Journal of the American Statistical Association 106 (493)*, pp. 345–361.

Iacus, S. M., King, G. and Porro, G. (2012): "Causal Inference without Balance Checking: Coarsened Exact Matching." In: *Political Analysis 20 (1)*, pp. 1–24.

Iman, R. L. and Davenport, J. M. (1976): "New Approximations to the Exact Distribution of the Kruskal-Wallis Test Statistic." In: *Communications in Statistics-theory and Methods 5 (14)*, pp. 1335–1348.

Imbens, G. W. (2000): "The Role of the Propensity Score in Estimating Dose-response Functions." In: *Biometrika 87 (3)*, pp. 706–710.

Insua, D. R. and Müller, P. (1998): "Feedforward Neural Networks for Nonparametric Regression." In: *Practical Nonparametric and Semiparametric Bayesian Statistics*. Ed. by D. Dey, P. Müller and D. Sinha. New York: Springer, pp. 181–193.

Isaksson, A. and Lee, S. (2005): "Simple Approaches to Estimating the Variance of the Propensity Score Weighted Estimator Applied on Volunteer Panel Web Survey Data – a Comparative Study." In: *JSM Proceedings, Survey Research Methods Section*. Alexandria: American Statistical Association, pp. 3143–3149.

Jacoby, J. and Handlin, A. H. (1991): "Non-probability Sampling Designs for Litigation Surveys." In: *Trademark Reporter 81 (2)*, p. 169.

Jahn, J. (2011): *Vector Optimization*. Berlin: Springer.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013): *An Introduction to Statistical Learning*. New York et al.: Springer.

Jamshidian, M. and Jalal, S. (2010): "Tests of Homoscedasticity, Normality, and Missing Completely at Random for Incomplete Multivariate Data." In: *Psychometrika 75 (4)*, pp. 649–674.

Jamshidian, M., Jalal, S. J. and Jansen, C. (2014): "MissMech: An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR)." In: *Journal of Statistical Software 56 (6)*, pp. 1–31.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. and Usher, A. (2015): "Big Data in Survey Research: AAPOR Task Force Report." In: *Public Opinion Quarterly 79 (4)*, pp. 839–880.

Jarre, F. and Stoer, J. (2004): *Optimierung*. Berlin: Springer.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016): "Combining Satellite Imagery and Machine Learning to Predict Poverty." In: *Science 353 (6301)*, pp. 790–794.

Johnson, N. L. and Kotz, S. (1972): *Distributions in Statistics: Continuous Multivariate Distributions*. New York et al.: Wiley.

Johnson, S. G. (n.d.): *The NLopt Nonlinear-optimization Package*. URL: `http://github.com/stevengj/nlopt` *(Retrieved 27.06.2019)*.

Jørgensen, B. (1984): "The Delta Algorithm and GLIM." In: *International Statistical Review 52 (3)*, pp. 283–300.

Justel, A., Peña, D. and Zamar, R. (1997): "A Multivariate Kolmogorov-Smirnov Test of Goodness of Fit." In: *Statistics & Probability Letters 35 (3)*, pp. 251–259.

Kale, B. (1961): "On the Solution of the Likelihood Equation by Iteration Processes." In: *Biometrika 48 (3/4)*, pp. 452–456.

Kalton, G. (1983): *Introduction to Survey Sampling*. Newbury Park: SAGE Publications.

Karush, W. (1939): "Minima of Functions of Several Variables with Inequalities As Side Constraints." Master's thesis. University of Chicago. Quoted in Kjeldsen (2000).

Kendler, K. S., Myers, J., Potter, J. and Opalesky, J. (2009): "A Web-based Study of Personality, Psychopathology and Substance Use in Twin, Other Relative and Relationship Pairs." In: *Twin Research and Human Genetics 12 (2)*, pp. 137–141.

Kern, C., Klausch, T. and Kreuter, F. (2019): "Tree-based Machine Learning Methods for Survey Research." In: *Survey Research Methods 13 (1)*, pp. 73–93.

Kessy, A., Lewin, A. and Strimmer, K. (2018): "Optimal Whitening and Decorrelation." In: *The American Statistician 72 (4)*, pp. 309–314.

Kiær, A. N. (1895): "Observations Et Expériences Concernant Les Dénombrements Représentatifs." In: *Bulletin of the International Statistical Institute 9 (2)*, pp. 176–183.

Kiær, A. N. (1897): *Den Repraesentative Undersøgelsesmethode*. Quoted in Kruskal and Mosteller (1980).

Kim, J. K., Kwon, Y. and Paik, M. C. (2016): "Calibrated Propensity Score Method for Survey Nonresponse in Cluster Sampling." In: *Biometrika 103 (2)*, pp. 461–473.

Kim, J. K. and Park, M. (2010): "Calibration Estimation in Survey Sampling." In: *International Statistical Review 78 (1)*, pp. 21–39.

Kim, J. K., Park, S., Chen, Y. and Wu, C. (2018): "Combining Non-probability and Probability Survey Samples through Mass Imputation." In: *Arxiv Preprint Arxiv:1812.10694*.

Kim, J. K. and Wang, Z. (2019): "Sampling Techniques for Big Data Analysis." In: *International Statistical Review 87 (S1)*, pp. 177–191.

King, G. and Nielsen, R. (2019): "Why Propensity Scores Should Not Be Used for Matching." In: *Political Analysis 27 (4)*, pp. 435–454.

Kirkwood, T. and Westlake, W. J. (1981): "Bioequivalence Testing–a Need to Rethink." In: *Biometrics 37 (3)*, pp. 589–594.

Kish, L. (1965): *Survey Sampling*. New York: John Wiley & Sons.

Kjeldsen, T. H. (2000): "A Contextualized Historical Analysis of the Kuhn–Tucker Theorem in Nonlinear Programming: The Impact of World War II." In: *Historia Mathematica 27 (4)*, pp. 331–361.

Klenke, A. (2013): *Probability Theory: A Comprehensive Course*. London: Springer Science & Business Media.

Kloft, M., Brefeld, U., Sonnenburg, S. and Zien, A. (2011): "Lp-norm Multiple Kernel Learning." In: *Journal of Machine Learning Research 12 (Mar)*, pp. 953–997.

Kolb, J.-P. (2012): "Methoden Zur Erzeugung Synthetischer Simulationsgesamtheiten." Doctoral dissertation. University of Trier.

Kolmogorov, A. (1933): "Sulla Determinazione Empirica Di Una Legge Di Distribuzione." In: *Giornale Dell'istituto Italiano Degli Attua 4*, pp. 83–91.

Kott, P. S. (2003): "A Practical Use for Instrumental-variable Calibration." In: *Journal of Official Statistics 19 (3)*, p. 265.

Kott, P. S. (2006): "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors." In: *Survey Methodology 32 (2)*, pp. 133–142.

Kott, P. S. and Liao, D. (2017): "Calibration Weighting for Nonresponse That Is Not Missing at Random: Allowing More Calibration Than Response-Model Variables." In: *Journal of Survey Statistics and Methodology 5 (2)*, pp. 159–174.

Kovar, J., Rao, J. and Wu, C. (1988): "Bootstrap and Other Methods to Measure Errors in Survey Estimates." In: *Canadian Journal of Statistics 16 (S1)*, pp. 25–45.

Kraft, D. (1988): *A Software Package for Sequential Quadratic Programming*. Forschungsbericht 88-28. Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt.

Kraft, D. (1994): "Algorithm 733: TOMP–Fortran Modules for Optimal Control Calculations." In: *Acm Transactions on Mathematical Software (toms) 20 (3)*, pp. 262–281.

Kreuter, F. and Olson, K. (2013): "Paradata for Nonresponse Error Investigation." In: *Improving Surveys with Paradata: Analytic Uses of Process Information*. Ed. by F. Kreuter. Hoboken: Wiley & Sons, pp. 13–42.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M. and Raghunathan, T. E. (2010): "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-response: Examples from Multiple Surveys." In: *Journal of the Royal Statistical Society: Series a (statistics in Society) 173 (2)*, pp. 389–407.

Kruskal, W. and Mosteller, F. (1979a): "Representative Sampling, I: Non-scientific Literature." In: *International Statistical Review 47 (1)*, pp. 13–24.

Kruskal, W. and Mosteller, F. (1979b): "Representative Sampling, II: Scientific Literature, Excluding Statistics." In: *International Statistical Review 47 (2)*, pp. 111–127.

Kruskal, W. and Mosteller, F. (1979c): "Representative Sampling, III: The Current Statistical Literature." In: *International Statistical Review 47 (4)*, pp. 245–265.

Kruskal, W. and Mosteller, F. (1980): "Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939." In: *International Statistical Review 48 (2)*, pp. 169–195.

Kruskal, W. H. (1952): "A Nonparametric Test for the Several Sample Problem." In: *The Annals of Mathematical Statistics 23 (4)*, pp. 525–540.

Kruskal, W. H. and Wallis, W. A. (1952): "Use of Ranks in One-criterion Variance Analysis." In: *Journal of the American Statistical Association 47 (260)*, pp. 583–621.

Kuhn, H. W. and Tucker, A. W. (1951): "Nonlinear Programming." In: *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by J. Neyman. Berkeley: University of California Press.

Kuk, A. Y. (1989): "Double Bootstrap Estimation of Variance under Systematic Sampling with Probability Proportional to Size." In: *Journal of Statistical Computation and Simulation 31 (2)*, pp. 73–82.

Kullback, S. and Leibler, R. A. (1951): "On Information and Sufficiency." In: *The Annals of Mathematical Statistics 22 (1)*, pp. 79–86.

Lakens, D. (2017): "Equivalence Tests: A Practical Primer for T Tests, Correlations, and Meta-analyses." In: *Social Psychological and Personality Science 8 (4)*, pp. 355–362.

Laney, D. (2001): "3D Data Management: Controlling Data Volume, Velocity and Variety." In: *Meta Group Research Note 6 (70)*, p. 1.

Laube, P., Franz, M. O. and Umlauf, G. (2018): "Deep Learning Parametrization for B-Spline Curve Approximation." In: *Arxiv Preprint Arxiv:1807.08304v1*.

Lawson, C. L. and Hanson, R. J. (1995): *Solving Least Squares Problems*. Philadelphia, PA: SIAM.

Lax, J. R. and Phillips, J. H. (2009): "Gay Rights in the States: Public Opinion and Policy Responsiveness." In: *American Political Science Review 103 (3)*, pp. 367–386.

Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014): "The Parable of Google Flu: Traps in Big Data Analysis." In: *Science 343 (6176)*, pp. 1203–1205.

Ledwina, T. (1994): "Data-driven Version of Neyman's Smooth Test of Fit." In: *Journal of the American Statistical Association 89 (427)*, pp. 1000–1005.

Lee, B. K., Lessler, J. and Stuart, E. A. (2010): "Improving Propensity Score Weighting Using Machine Learning." In: *Statistics in Medicine 29 (3)*, pp. 337–346.

Lee, S. (2004): "Statistical Estimation Methods in Volunteer Panel Web Surveys." Doctoral dissertation. University of Maryland.

Lee, S. and Valliant, R. (2009): "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment." In: *Sociological Methods & Research 37 (3)*, pp. 319–343.

Lee, Y. and Nelder, J. A. (1996): "Hierarchical Generalized Linear Models." In: *Journal of the Royal Statistical Society: Series B (methodological) 58 (4)*, pp. 619–656.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006): *Generalized Linear Models with Random Effects: Unified Analysis Via H-likelihood*. Boca Raton: CRC Press.

Legendre, A. M. (1805): *Nouvelles Méthodes Pour La Détermination Des Orbites Des Comètes*. Paris: F. Didot.

Lehdonvirta, V., Oksanen, A., Räsänen, P. and Blank, G. (2020): "Social Media, Web, and Panel Surveys: Using Non-Probability Samples in Social and Policy Research." In: *Policy & Internet 13 (1)*, pp. 134–155.

Lenard, M. L. (1979): "A Computational Study of Active Set Strategies in Nonlinear Programming with Linear Constraints." In: *Mathematical Programming 16 (1)*, pp. 81–97.

Lenau, S. (2020): *Sqp: (Sequential) Quadratic Programming*. R package version 0.5. URL: `https://CRAN.R-project.org/package=sqp` *(Retrieved 05.05.2021)*.

Lenau, S. and Münnich, R. (2017): "Use of New Data & Non–Probability Sampling." In: *InGRID Deliverable 23.2: Future Needs in Statistics*. Ed. by C. Articus, Y. G. Berger, G. Betti, J. P. Burgard, A. Byrne, T. Chandola, A. D'Agostino, F. Gagliardi, C. Giusti, H. Goldstein, S. Lenau, S. Marchetti, R. Münnich, A. Potsi, M. Pratesi, N. Shlomo and V. Verma. Trier: InGRID project, pp. 59–75. *(Retrieved 27.09.2018)*.

Lerman, R. I. and Yitzhaki, S. (1985): "Income Inequality Effects by Income Source: A New Approach and Applications to the United States." In: *The Review of Economics and Statistics 67 (1)*, pp. 151–156.

Lewis, P. A. (1961): “Distribution of the Anderson-Darling Statistic.” In: *The Annals of Mathematical Statistics 32 (4)*, pp. 1118–1124.

Li, X. S. (2005): “An Overview of SuperLU: Algorithms, Implementation, and User Interface.” In: *Acm Transactions on Mathematical Software 31 (3)*, pp. 302–325.

Lilliefors, H. W. (1967): “On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown.” In: *Journal of the American Statistical Association 62 (318)*, pp. 399–402.

Limentani, G. B., Ringo, M. C., Ye, F., Bergquist, M. L. and McSorley, E. O. (2005): “Beyond the T-test: Statistical Equivalence Testing.” In: *Analytical Chemistry 77 (11)*, pp. 221–226.

Lin, Z., Reay, D. S., Williams, B. W. and He, X. (2006): “Torque Ripple Reduction in Switched Reluctance Motor Drives Using B-spline Neural Networks.” In: *Ieee Transactions on Industry Applications 42 (6)*, pp. 1445–1453.

Lindeberg, J. W. (1922): “Eine Neue Herleitung Des Exponentialgesetzes in Der Wahrscheinlichkeitsrechnung.” In: *Mathematische Zeitschrift 15 (1)*, pp. 211–225.

Little, R. J. (1986): “Survey Nonresponse Adjustments for Estimates of Means.” In: *International Statistical Review 54 (2)*, pp. 139–157.

Little, R. J. (1988a): “A Test of Missing Completely at Random for Multivariate Data with Missing Values.” In: *Journal of the American Statistical Association 83 (404)*, pp. 1198–1202.

Little, R. J. (1988b): “Missing-data Adjustments in Large Surveys.” In: *Journal of Business & Economic Statistics 6 (3)*, pp. 287–296.

Little, R. J. (1993): “Post-stratification: A Modeler’s Perspective.” In: *Journal of the American Statistical Association 88 (423)*, pp. 1001–1012.

Little, R. J. and Rubin, D. B. (2019): *Statistical Analysis with Missing Data*. Hoboken: John Wiley & Sons.

Little, R. J., West, B. T., Boonstra, P. S. and Hu, J. (2020): “Measures of the Degree of Departure from Ignorable Sample Selection.” In: *Journal of Survey Statistics and Methodology 8 (5)*, pp. 932–964.

Lohr, S. (2010): *Sampling: Design and Analysis*. Boston: Brooks/Cole Cenage Learning.

Lohr, S. L. and Raghunathan, T. E. (2017): “Combining Survey Data with Other Data Sources.” In: *Statistical Science 32 (2)*, pp. 293–312.

Loosveldt, G. and Sonck, N. (2008): “An Evaluation of the Weighting Procedures for an Online Access Panel Survey.” In: *Survey Research Methods 2 (2)*, pp. 93–105.

Luhmann, M. (2017): “Using Big Data to Study Subjective Well–being.” In: *Current Opinion in Behavioral Sciences 18*, pp. 28–33.

Lumley, T. (2004): “Analysis of Complex Survey Samples.” In: *Journal of Statistical Software 9 (1)*, pp. 1–19.

Lumley, T. and Scott, A. (2017): “Fitting Regression Models to Survey Data.” In: *Statistical Science 32 (2)*, pp. 265–278.

Lunceford, J. K. and Davidian, M. (2004): "Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study." In: *Statistics in Medicine 23 (19)*, pp. 2937–2960.

Luts, J., Molenberghs, G., Verbeke, G., Van Huffel, S. and Suykens, J. A. (2012): "A Mixed Effects Least Squares Support Vector Machine Model for Classification of Longitudinal Data." In: *Computational Statistics & Data Analysis 56 (3)*, pp. 611–628.

Lynch, C. (2008): "How Do Your Data Grow?" In: *Nature 455 (7209)*, pp. 28–29.

Lynn, P. (2014): "Longer Interviews May Not Affect Subsequent Survey Participation Propensity." In: *Public Opinion Quarterly 78 (2)*, pp. 500–509.

Ma, P. and Sun, X. (2015): "Leveraging for Big Data Regression." In: *Wiley Interdisciplinary Reviews: Computational Statistics 7 (1)*, pp. 70–76.

Maddala, G. (1983): *Limited-dependent and Qualitative Variables in Economics*. New York: Cambridge University Press.

Magnussen, S. (2015): "Arguments for a Model-dependent Inference?" In: *Forestry 88 (3)*, pp. 317–325.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L. and Gabrielli, L. (2015): "Small Area Model-based Estimators Using Big Data Sources." In: *Journal of Official Statistics 31 (2)*, pp. 263–281.

Marler, R. T. and Arora, J. S. (2004): "Survey of Multi-objective Optimization Methods for Engineering." In: *Structural and Multidisciplinary Optimization 26 (6)*, pp. 369–395.

Marler, R. T. and Arora, J. S. (2010): "The Weighted Sum Method for Multi-objective Optimization: New Insights." In: *Structural and Multidisciplinary Optimization 41 (6)*, pp. 853–862.

Marsaglia, G. and Marsaglia, J. (2004): "Evaluating the Anderson–darling Distribution." In: *Journal of Statistical Software 9 (2)*, pp. 1–5.

Martin, R., Peters, G. and Wilkinson, J. (1965): "Symmetric Decomposition of a Positive Definite Matrix." In: *Numerische Mathematik 7 (5)*, pp. 362–383.

Martin, R., Peters, G. and Wilkinson, J. (1966): "Iterative Refinement of the Solution of a Positive Definite System of Equations." In: *Numerische Mathematik 8 (3)*, pp. 203–216.

Masch, L. and Gabriel, O. W. (2020): "How Emotional Displays of Political Leaders Shape Citizen Attitudes: The Case of German Chancellor Angela Merkel." In: *German Politics 29 (2)*, pp. 158–179.

Massey, F. J. (1950): "A Note on the Estimation of a Distribution Function by Confidence Limits." In: *The Annals of Mathematical Statistics 21 (1)*, pp. 116–119.

McCabe, S. E. (2008): "Misperceptions of Non-medical Prescription Drug Use: A Web Survey of College Students." In: *Addictive Behaviors 33 (5)*, pp. 713–724.

McCornack, R. L. (1956): "A Criticism of Studies Comparing Item-weighting Methods." In: *Journal of Applied Psychology 40 (5)*, p. 343.

McCullagh, P. and Nelder, J. A. (1989): *Generalized Linear Models*. 2nd ed. London, New York: CRC press.

McLachlan, G. J. (2004): *Discriminant Analysis and Statistical Pattern Recognition*. Hoboken: John Wiley & Sons.

Melnik, M. and Pusev, R. (2015): *Uniftest: Tests for Uniformity*. R package version 1.1. URL: `https://CRAN.R-project.org/package=uniftest` *(Retrieved 05.05.2021)*.

Meng, X.-L. (2018): "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election." In: *The Annals of Applied Statistics 12 (2)*, pp. 685–726.

Meng, X., Duan, N., Chen, C. and Alegria, M. (2009): "Power-shrinkage: An Alternative Method for Dealing with Excessive Weights." In: *JSM Proceedings, Survey Research Methods Section*. Alexandria: American Statistical Association. *(Retrieved 15.08.2018)*.

Mercer, A. W., Kreuter, F., Keeter, S. and Stuart, E. A. (2017): "Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference." In: *Public Opinion Quarterly 81 (S1)*, pp. 250–271.

Merkle, H., Burgard, J. P. and Münnich, R. (2016): "The AMELIA Dataset - A Synthetic Universe for Reproducible Research." In: *InGRID Deliverable 23.1: Case Studies*. Ed. by Y. G. Berger, J. P. Burgard, A. Byrne, A. Cernat, C. Giusti, P. Koksel, S. Lenau, S. Marchetti, H. Merkle, R. Münnich, I. Permanyer, M. Pratesi, N. Salvati, N. Shlomo, D. Smith and N. Tzavidis. Trier: InGRID project, pp. 59–75. *(Retrieved 27.09.2018)*.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F. (2019): *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-3. URL: `https://CRAN.R-project.org/package=e1071` *(Retrieved 05.05.2021)*.

Milborrow, S. (2019): *Earth: Multivariate Adaptive Regression Splines*. Derived from mda:mars by Hastie, T. and Tibshirani, R. `R` package version 5.1.1. URL: `https://CRAN.R-project.org/package=earth` *(Retrieved 05.05.2021)*.

Mills, J. P. (1926): "Table of the Ratio: Area to Bounding Ordinate, for Any Portion of Normal Curve." In: *Biometrika 18 (3–4)*, pp. 395–400.

Mirabbasi, R., Kisi, O., Sanikhani, H. and Meshram, S. G. (2019): "Monthly Long-term Rainfall Estimation in Central India Using M5Tree, MARS, LSSVR, ANN and GEP Models." In: *Neural Computing and Applications 31 (10)*, pp. 6843–6862.

Moser, B. K., Stevens, G. R. and Watts, C. L. (1989): "The Two-sample T Test Versus Satterthwaite's Approximate F Test." In: *Communications in Statistics-theory and Methods 18 (11)*, pp. 3963–3975.

Mozer, M. C. (1995): "A Focused Backpropagation Algorithm for Temporal Pattern Recognition." In: *Complex Systems 3*, pp. 349–381.

Münnich, R., Schürle, J., Bihler, W., Boonstra, H., Knotterus, P., Nieuwenbroek, N., Haslinger, A., Laaksonen, S., Eckmair, D., Quatember, A. et al. (2003): *Monte Carlo Simulation Study of European Surveys*. Research Project Report WP3 - D 3.1/3.2. IST–2000–26057–DACSEIS.

Münnich, R., Burgard, J. P. and Vogt, M. (2013): "Small Area-statistik: Methoden Und Anwendungen." In: *Asta Wirtschafts-und Sozialstatistisches Archiv 6 (3-4)*, pp. 149–191.

Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P. and Kolb, J.-P. (2012): *Stichprobenoptimierung Und Schätzung Im Zensus 2011*. Statistik und Wissenschaft Bd. 21.

Münnich, R. and Lenau, S. (2019): "Bias Correction for Non-probability Samples." In: *Proceedings of the Statistische Woche 2019*. Trier: German Statistical Society.

Münnich, R., Sachs, E. W. and Wagner, M. (2012): "Calibration of Estimator-weights Via Semismooth Newton Method." In: *Journal of Global Optimization 52 (3)*, pp. 471–485.

Münnich, R., Wagner, J., Hill, J., Stoffels, J., Buddenbaum, H. and Udelhoven, T. (2016): "Schätzung Von Holzvorräten Unter Verwendung Von Fernerkundungsdaten." In: *Asta Wirtschafts-und Sozialstatistisches Archiv 10 (2-3)*, pp. 95–112.

Münnich, R. and Zins, S. (2011): *Variance Estimation for Indicators of Poverty and Social Exclusion*. Research Project Report WP3 – D3.2. FP7-SSH-2007-217322 AMELI.

Münnich, R. T. and Zwick, M. (2016): "Big Data Und Was Nun? Neue Datenbestände Und Ihre Auswirkungen. Vorwort Der Herausgeber." In: *Asta Wirtschafts-und Sozialstatistisches Archiv 10*, pp. 73–77.

Nassimbeni, G. (2001): "Technology, Innovation Capacity, and the Export Attitude of Small Manufacturing Firms: A Logit/tobit Model." In: *Research Policy 30 (2)*, pp. 245–262.

National Research Council of the United States (2013): *Frontiers in Massive Data Analysis*. Washington, D.C.: National Academies Press.

Nelder, J. A. and Wedderburn, R. W. (1972): "Generalized Linear Models." In: *Journal of the Royal Statistical Society: Series a (general) 135 (3)*, pp. 370–384.

Nesterov, Y. E. (2004): *Introductory Lectures on Convex Optimization: A Basic Course*. New York: Springer Science & Business Media.

Neyman, J. (1934): "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." In: *Journal of the Royal Statistical Society 97 (4)*, pp. 558–625.

Neyman, J. (1937): "» Smooth Test» for Goodness of Fit." In: *Scandinavian Actuarial Journal 1937 (3-4)*, pp. 149–199.

Nielsen, M., Haun, D., Kärtner, J. and Legare, C. H. (2017): "The Persistent Sampling Bias in Developmental Psychology: A Call to Action." In: *Journal of Experimental Child Psychology 162*, pp. 31–38.

Nocedal, J. and Wright, S. J. (1999): *Numerical Optimization*. New York: Springer.

Noh, M., Wu, L. and Lee, Y. (2012): "Hierarchical Likelihood Methods for Nonlinear and Generalized Linear Mixed Models with Missing Data and Measurement Errors in Covariates." In: *Journal of Multivariate Analysis 109*, pp. 42–51.

Nsoesie, E. O., Kluberg, S. A. and Brownstein, J. S. (2014): "Online Reports of Foodborne Illness Capture Foods Implicated in Official Foodborne Outbreak Reports." In: *Preventive Medicine 67*, pp. 264–269.

O'Sullivan, F., Yandell, B. S. and Raynor, W. J. (1986): "Automatic Smoothing of Regression Functions in Generalized Linear Models." In: *Journal of the American Statistical Association 81 (393)*, pp. 96–103.

Okner, B. (1972): "Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File." In: *Annals of Economic and Social Measurement 1 (3)*. Ed. by S. V. Berg, pp. 325–362.

Olson, K. and Parkhurst, B. (2013): "Collecting Paradata for Measurement Error Evaluations." In: *Improving Surveys with Paradata: Analytic Uses of Process Information*. Ed. by F. Kreuter. Hoboken: Wiley & Sons, pp. 43–72.

Opsomer, J. D. (2009): "Introduction to Part 4." In: *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*. Ed. by D. Pfeffermann and C. R. Rao. Amsterdam: Elsevier, pp. 11–31.

Orchard, T. and Woodbury, M. (1972): "A Missing Information Principle: Theory and Applications." In: *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics*. Berkeley: University of California Press, pp. 697–715.

Osborne, M. R. (1992): "Fisher's Method of Scoring." In: *International Statistical Review 60 (1)*, pp. 99–117.

Park, D. K., Gelman, A. and Bafumi, J. (2004): "Bayesian Multilevel Estimation with Poststratification: State-level Estimates from National Polls." In: *Political Analysis 12 (4)*, pp. 375–385.

Park, M. and Fuller, W. A. (2005): "Towards Nonnegative Regression Weights for Survey Samples." In: *Survey Methodology 31 (1)*, pp. 85–93.

Pasek, J. (2016): "When Will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies As Criteria for Correspondence." In: *International Journal of Public Opinion Research 28 (2)*, pp. 269–291.

Patterson, H. D. and Thompson, R. (1971): "Recovery of Inter-block Information When Block Sizes Are Unequal." In: *Biometrika 58 (3)*, pp. 545–554.

Pearl, J. (2010): "The Foundations of Causal Inference." In: *Sociological Methodology 40 (1)*, pp. 75–149.

Pedraza, P. de, Guzi, M. and Tijdens, K. (2020): "Life Satisfaction of Employees, Labour Market Tightness and Matching Efficiency." In: *International Journal of Manpower 42 (3)*, pp. 341–355.

Pedraza, P. de, Tijdens, K., Bustillo, R. M. de and Steinmetz, S. (2010): "A Spanish Continuous Volunteer Web Survey: Sample Bias, Weighting and Efficiency." In: *Revista Española De Investigaciones Sociológicas 131 (1)*, pp. 109–130.

Pedraza, P. de, Tijdens, K. G. and Bustillo, R. M. de (2007): *Sample Bias, Weights and Efficiency of Weights in a Continuous Web Voluntary Survey*. Working paper 2007-60. AIAS.

Petrucci, A. and Rocco, E. (2019): "A Proposal to Assess the Representativeness of Nonprobability Surveys." In: *Proceedings of the Statistische Woche 2019*. Trier: German Statistical Society.

Pfeffermann, D. (2011): "Modelling of Complex Survey Data: Why Is It a Problem? How Should We Approach It?" In: *Survey Methodology 37 (2)*, pp. 115–136.

Pfeffermann, D. (2015): "Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture." In: *Journal of Survey Statistics and Methodology 3 (4)*, pp. 425–483.

Pfeffermann, D. and Rao, C. R. (2009): *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*. Amsterdam: Elsevier.

Pfeffermann, D. and Sikov, A. (2011): "Imputation and Estimation under Nonignorable Nonresponse for Household Surveys with Missing Covariate Information." In: *Journal of Official Statistics 27 (2)*, pp. 181–209.

Pfeffermann, D. and Sverchkov, M. (1999): "Parametric and Semi-parametric Estimation of Regression Models Fitted to Survey Data." In: *Sankhyā: The Indian Journal of Statistics, Series B 61 (1)*, pp. 166–186.

Piazza, F., Smerilli, S., Uncini, A., Griffo, M. and Zunino, R. (1997): "Fast Spline Neural Networks for Image Compression." In: *Proceedings of the 8th Italian Workshop on Neural Nets, 23–25 May 1996, Vietri Sul Mare, Salerno, Italy*. Ed. by M. Marinaro and R. Tagliaferri. Salerno: Springer, pp. 223–230.

Piegl, L. and Tiller, W. (1997): *The NURBS Book*. New York: Springer.

Piegl, L. A. and Tiller, W. (1998): "Computing the Derivative of NURBS with Respect to a Knot." In: *Computer Aided Geometric Design 15 (9)*, pp. 925–934.

Pinheiro, J. and Bates, D. (2000): *Mixed-effects Models in S and S-PLUS*. New York: Springer Science & Business Media.

Posawang, P., Phosaard, S., Polnigongit, W. and Pattara-Atikom, W. (2010): "Perception-based Road Traffic Congestion Classification Using Neural Networks and Decision Tree." In: *Electronic Engineering and Computing Technology*. Ed. by S.-I. Ao and L. Gelman. Lecture Notes in Electrical Engineering. Dordrecht: Springer, pp. 237–248.

Posner, M. A. and Ash, A. (2012): *Comparing Weighting Methods in Propensity Score Analysis*. Working paper.

Powell, M. J. (1978): "A Fast Algorithm for Nonlinearly Constrained Optimization Calculations." In: *Numerical Analysis. Proceedings of the Biennial Conference, Dundee*. Ed. by G. A. Watson. Berlin: Springer, pp. 144–157.

Pratesi, M., Manfreda, K. L., Biffignandi, S. and Vehovar, V. (2004): "List-based Web Surveys: Quality, Timeliness, and Nonresponse in the Steps of the Participation Flow." In: *Journal of Official Statistics 20 (3)*, p. 451.

Preston, J. (2009): "Rescaled Bootstrap for Stratified Multistage Sampling." In: *Survey Methodology 35 (2)*, pp. 227–234.

Procházková, J. and Procházka, D. (2007): "Implementation of NURBS Curve Derivatives in Engineering Practice." In: *Proceedings of the 15th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision. Posters Proceedings*. Ed. by V. Skala. Plzen: University of West Bohemia.

R Core Team (2018): *R: A Language and Environment for Statistical Computing*. URL: `https://www.R-project.org/` *(Retrieved 05.05.2021)*.

Rafei, A., Flannagan, C. A. and Elliott, M. R. (2020): "Big Data for Finite Population Inference: Applying Quasi-Random Approaches to Naturalistic Driving Data Using Bayesian Additive Regression Trees." In: *Journal of Survey Statistics and Methodology 8 (1)*, pp. 148–180.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M. and Mozetič, I. (2015): "The Effects of Twitter Sentiment on Stock Price Returns." In: *Plos One 10 (9)*, e0138441.

Rao, C. R. (1952): *Advanced Statistical Methods in Biometric Research.* New York: Hafner Press.

Rao, J., Scott, A. and Benhin, E. (2003): "Undoing Complex Survey Data Structures: Some Theory and Applications of Inverse Sampling." In: *Survey Methodology 29 (2)*, pp. 107–118.

Rao, J., Wu, C. and Yue, K. (1992): "Some Recent Work on Resampling Methods for Complex Surveys." In: *Survey Methodology 18 (2)*, pp. 209–217.

Rao, J. N. K. and Molina, I. (2015): *Small Area Estimation.* New York et al.: John Wiley & Sons.

Rao, J. N. (2003): *Small Area Estimation.* Hoboken: Wiley.

Rao, J. N. and Wu, C. (1988): "Resampling Inference with Complex Survey Data." In: *Journal of the American Statistical Association 83 (401)*, pp. 231–241.

Raphson, J. (1690): *Analysis Aequationum Universalis.* London: Quoted in Cajori (1911).

Raya-Armenta, J. M., Lozano-Garcia, J. M. and Avina-Cervantes, J. G. (2018): "B-spline Neural Network for Real and Reactive Power Control of a Wind Turbine." In: *Electrical Engineering 100 (4)*, pp. 2799–2813.

Reams, R. (1999): "Hadamard Inverses, Square Roots and Products of Almost Semidefinite Matrices." In: *Linear Algebra and Its Applications 288 (1)*, pp. 35–43.

Reinsch, C. H. (1967): "Smoothing by Spline Functions." In: *Numerische Mathematik 10 (3)*, pp. 177–183.

Rendtel, U. and Schimpl-Neimanns, B. (2001): "Die Berechnung Der Varianz Von Populationsschätzern Im Scientific Use File Des Mikrozensus Ab 1996." In: *Zuma Nachrichten 25 (48)*, pp. 85–116.

Ripley, B. and Venables, W. (2016): *Package 'nnet'.* R package version 7.3-12, pp. 3–12. URL: `https://cran.r-project.org/package=nnet` *(Retrieved 05.05.2021).*

Ripley, B. D. (1996): *Pattern Recognition and Neural Networks.* Cambridge et al.: Cambridge university press.

Rivers, D. (2007): "Sampling for Web Surveys." In: *JSM Proceedings, Survey Research Methods Section.* Alexandria: American Statistical Association. *(Retrieved 06.03.2019).*

Robinson, G. K. et al. (1991): "That BLUP Is a Good Thing: The Estimation of Random Effects." In: *Statistical Science 6 (1)*, pp. 15–32.

Rodríguez-Mazahua, L., Rodríguez-Enríquez, C.-A., Sánchez-Cervantes, J. L., Cervantes, J., García-Alcaraz, J. L. and Alor-Hernández, G. (2016): "A General Perspective of Big Data: Applications, Tools, Challenges and Trends." In: *The Journal of Supercomputing 72 (8)*, pp. 3073–3113.

Rosenbaum, P. R. (2010): *Design of Observational Studies.* New York: Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects." In: *Biometrika 70 (1)*, pp. 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." In: *The American Statistician 39 (1)*, pp. 33–38.

Rosenblatt, M. (1952): "Remarks on a Multivariate Transformation." In: *The Annals of Mathematical Statistics 23 (3)*, pp. 470–472.

Roster, C. A., Rogers, R. D., Albaum, G. and Klein, D. (2004): "A Comparison of Response Characteristics from Web and Telephone Surveys." In: *International Journal of Market Research 46 (3)*, pp. 359–373.

Royall, R. M. (1970): "On Finite Population Sampling Theory under Certain Linear Regression Models." In: *Biometrika 57 (2)*, pp. 377–387.

Royall, R. M. (1992): "Robustness and Optimal Design under Prediction Models for Finite Populations." In: *Survey Methodology 18 (2)*, pp. 179–185.

Rubin, D. B. (1973): "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies." In: *Biometrics 29 (1)*, pp. 185–203.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." In: *Journal of Educational Psychology 66 (5)*, pp. 688–701.

Rubin, D. B. (1976): "Inference and Missing Data." In: *Biometrika 63 (3)*, pp. 581–592.

Rubin, D. B. (1979): "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." In: *Journal of the American Statistical Association 74 (366)*, pp. 318–328.

Rubin, D. B. (1987): *Multiple Imputation for Nonresponse in Surveys*. New York et al.: John Wiley & Sons.

Rubin, D. B. (2006): *Matched Sampling for Causal Effects*. Cambridge et al.: Cambridge University Press.

Rubin, D. B. and Thomas, N. (1996): "Matching Using Estimated Propensity Scores: Relating Theory to Practice." In: *Biometrics 52 (1)*, pp. 249–264.

Rubin, D. B. and Thomas, N. (2000): "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." In: *Journal of the American Statistical Association 95 (450)*, pp. 573–585.

Rupp, K., Tillet, P., Rudolf, F., Weinbub, J., Morhammer, A., Grasser, T., Jüngel, A. and Selberherr, S. (2016): "ViennaCL – Linear Algebra Library for Multi-and Many-Core Architectures." In: *Siam Journal on Scientific Computing 38 (5)*, S412–S439.

Rupp, M. (2018): "Optimization for Multivariate and Multi-domain Methods in Survey Statistics." Doctoral dissertation. University of Trier.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003): *Semiparametric Regression*. Cambridge et al.: Cambridge university press.

Ryzin, G. G. van (2008): "Validity of an On-line Panel Approach to Citizen Surveys." In: *Public Performance & Management Review 32 (2)*, pp. 236–262.

Saad, Y. and Schultz, M. H. (1986): "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems." In: *Siam Journal on Scientific and Statistical Computing 7 (3)*, pp. 856–869.

Saad, Y. (2003): *Iterative Methods for Sparse Linear Systems*. Philadelphia, PA: SIAM.

Salganik, M. J. and Heckathorn, D. D. (2004): "Sampling and Estimation in Hidden Populations Using Respondent-driven Sampling." In: *Sociological Methodology 34 (1)*, pp. 193–240.

Samuels, M. L. (1991): "Statistical Reversion toward the Mean: More Universal Than Regression toward the Mean." In: *The American Statistician 45 (4)*, pp. 344–346.

Sanders, D., Clarke, H. D., Stewart, M. C. and Whiteley, P. (2007): "Does Mode Matter for Modeling Political Choice? Evidence from the 2005 British Election Study." In: *Political Analysis 15 (3)*, pp. 257–285.

Sanderson, C. and Curtin, R. (2016): "Armadillo: A Template-based C++ Library for Linear Algebra." In: *Journal of Open Source Software 1 (2)*, pp. 26–33.

Sanderson, C. and Curtin, R. (2018): "A User-friendly Hybrid Sparse Matrix Class in C++." In: *International Congress on Mathematical Software*. Ed. by J. H. Davenport, M. Kauers, G. Labahn and J. Urban. Vol. 10931. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 422–430.

Santos Coelho, L. dos and Guerra, F. A. (2008): "B-spline Neural Network Design Using Improved Differential Evolution for Identification of an Experimental Nonlinear Process." In: *Applied Soft Computing 8 (4)*, pp. 1513–1522.

Särndal, C. E. and Lundström, S. (2005): *Estimation in Surveys with Nonresponse*. Chichester: Wiley.

Särndal, C.-E. (1978): "Design-based and Model-based Inference in Survey Sampling." In: *Scandinavian Journal of Statistics 5 (1)*, pp. 27–42.

Särndal, C.-E. (2007): "The Calibration Approach in Survey Theory and Practice." In: *Survey Methodology 33 (2)*, pp. 99–119.

Särndal, C.-E. (2011): "The 2010 Morris Hansen Lecture. Dealing with Survey Nonresponse in Data Collection, in Estimation." In: *Journal of Official Statistics 27 (1)*, p. 1.

Särndal, C., Swensson, B. and Wretman, J. (1992): *Model Assisted Survey Sampling*. New York: Springer.

Satterthwaite, F. E. (1946): "An Approximate Distribution of Estimates of Variance Components." In: *Biometrics Bulletin 2 (6)*, pp. 110–114.

Schaid, D. J., Tong, X., Batzler, A., Sinnwell, J. P., Qing, J. and Biernacka, J. M. (2019): "Multivariate Generalized Linear Model for Genetic Pleiotropy." In: *Biostatistics 20 (1)*, pp. 111–128.

Schall, R. (1991): "Estimation in Generalized Linear Models with Random Effects." In: *Biometrika 78 (4)*, pp. 719–727.

Schillewaert, N. and Meulemeester, P. (2005): "Comparing Response Distributions of Offline and Online Data Collection Methods." In: *International Journal of Market Research 47 (2)*, pp. 163–178.

Schimpl-Neimanns, B. (2011): "Schätzung Des Stichprobenfehlers in Mikrozensus Scientific Use Files Ab 2005." In: *Asta Wirtschafts-und Sozialstatistisches Archiv 5 (1)*, pp. 19–38.

Schittkowski, K. (1981): "The Nonlinear Programming Method of Wilson, Han, and Powell with an Augmented Lagrangian Type Line Search Function." In: *Numerische Mathematik 38 (1)*, pp. 83–114.

Schmidt-Hieber, J. et al. (2020): "Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function." In: *Annals of Statistics 48 (4)*, pp. 1875–1897.

Scholz, F. W. and Stephens, M. A. (1987): "K-sample Anderson–Darling Tests." In: *Journal of the American Statistical Association 82 (399)*, pp. 918–924.

Schonlau, M., Van Soest, A. and Kapteyn, A. (2007): *Are 'Webographic' or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?* Working Paper WR-506. RAND.

Schonlau, M., Van Soest, A., Kapteyn, A. and Couper, M. (2009): "Selection Bias in Web Surveys and the Use of Propensity Scores." In: *Sociological Methods & Research 37 (3)*, pp. 291–318.

Schouten, B. (2007): "A Selection Strategy for Weighting Variables under a Not-missing-at-random Assumption." In: *Journal of Official Statistics 23 (1)*, p. 51.

Schouten, B. (2018): "Statistical Inference Based on Randomly Generated Auxiliary Variables." In: *Journal of the Royal Statistical Society: Series B (statistical Methodology) 80 (1)*, pp. 33–56.

Schouten, B. and Bethlehem, J. (2009): *Representativeness Indicators for Measuring and Enhancing the Composition of Survey Response*. Research Project Deliverable 9. RISQ.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. and Skinner, C. (2012): "Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response through R-indicators and Partial R-indicators." In: *International Statistical Review 80 (3)*, pp. 382–399.

Schouten, B., Cobben, F. and Bethlehem, J. (2009): "Indicators for the Representativeness of Survey Response." In: *Survey Methodology 35 (1)*, pp. 101–113.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2016): "Does More Balanced Survey Response Imply Less Non-response Bias?" In: *Journal of the Royal Statistical Society: Series a (statistics in Society) 179 (3)*, pp. 727–748.

Schouten, B., Shlomo, N. and Skinner, C. (2009): *How to Use R-indicators*. Research Project Deliverable 3. RISQ.

Schouten, B., Shlomo, N. and Skinner, C. (2011): "Indicators for Monitoring and Improving Representativeness of Response." In: *Journal of Official Statistics 27 (2)*, pp. 231–253.

Schroedter, J. H., Lechert, Y. and Lüttinger, P. (2006): *Die Umsetzung Der Bildungsskala ISCED-1997 Für Die Volkszählung 1970, Die Mikrozensus-Zusatzerhebung 1971 Und Die Mikrozensen 1976-2004 (Version 1)*. Methodenbericht 2006/08. ZUMA.

Schuirmann, D. J. (1987): "A Comparison of the Two One-sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability." In: *Journal of Pharmacokinetics and Biopharmaceutics 15 (6)*, pp. 657–680.

Searle, S. R., Casella, G. and McCulloch, C. E. (2006): *Variance Components*. Hoboken: John Wiley & Sons.

Sen, A. R. (1953): "On the Estimate of the Variance in Sampling with Varying Probabilities." In: *Journal of the Indian Society of Agricultural Statistics 5 (1194)*, p. 127.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J. and Cook, E. F. (2008): "Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study." In: *Pharmacoepidemiology and Drug Safety 17 (6)*, pp. 546–555.

Shanno, D. F. (1970): "Conditioning of Quasi-Newton Methods for Function Minimization." In: *Mathematics of Computation 24 (111)*, pp. 647–656.

Sherman, J. and Morrison, W. J. (1950): "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix." In: *The Annals of Mathematical Statistics 21 (1)*, pp. 124–127.

Shlomo, N., Skinner, C., Schouten, B., Heij, V. de, Bethlehem, J. and Ouwehand, P. (2009a): *Indicators for Representative Response Based on Population Totals*. Research Project Deliverable 2.2. RISQ.

Shlomo, N. and Goldstein, H. (2015): "Editorial: Big Data in Social Research." In: *Journal of the Royal Statistical Society: Series a (statistics in Society) 178 (4)*, pp. 787–790.

Shlomo, N., Skinner, C. and Schouten, B. (2012): "Estimation of an Indicator of the Representativeness of Survey Response." In: *Journal of Statistical Planning and Inference 142 (1)*, pp. 201–211.

Shlomo, N., Skinner, C., Schouten, B., Bethlehem, J. and Zhang, L.-C. (2009b): *Statistical Properties of R-indicators*. Research Project Deliverable 2.1. RISQ.

Signorini, A., Segre, A. M. and Polgreen, P. M. (2011): "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the US during the Influenza A H1N1 Pandemic." In: *Plos One 6 (5)*, e19467.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011): "Regularization Paths for Cox's Proportional Hazards Model Via Coordinate Descent." In: *Journal of Statistical Software 39 (5)*, pp. 1–13.

Simonoff, J. S. (2003): *Analyzing Categorical Data*. New York: Springer.

Sinclair, C. and Spurr, B. (1988): "Approximations to the Distribution Function of the Anderson–darling Test Statistic." In: *Journal of the American Statistical Association 83 (404)*, pp. 1190–1191.

Skinner, C. (2009): "Statistical Disclosure Control for Survey Data." In: *Handbook of Statistics 29A: Sample Surveys: Design, Methods and Applications*. Ed. by D. Pfeffermann and C. R. Rao. Amsterdam: Elsevier, pp. 381–396.

Skinner, C., Shlomo, N., Schouten, B., Zhang, L.-C. and Bethlehem, J. (2009): "Measuring Survey Quality through Representativeness Indicators Using Sample and Population Based Information." In: *Proceedings of the NTTS 2009*. Brussels: Eurostat.

Smirnov, N. (1936): "Sur La Distribution De ω^2." In: *Comptes Rendus De L'acadáemie Des Sciences, Paris 202 (44)*, pp. 449–452.

Smith, T. (1983): "On the Validity of Inferences from Non-random Samples." In: *Journal of the Royal Statistical Society: Series a (general) 146 (4)*, pp. 394–403.

Smola, A. J. and Schölkopf, B. (2004): "A Tutorial on Support Vector Regression." In: *Statistics and Computing 14 (3)*, pp. 199–222.

Smyk, M., Tyrowicz, J. and Van der Velde, L. (2021): "A Cautionary Note on the Reliability of the Online Survey Data: The Case of Wage Indicator." In: *Sociological Methods & Research 50 (1)*, pp. 429–464.

Specht, D. F. et al. (1991): "A General Regression Neural Network." In: *Ieee Transactions on Neural Networks 2 (6)*, pp. 568–576.

Spijkerman, R., Knibbe, R., Knoops, K., Van De Mheen, D. and Van Den Eijnden, R. (2009): "The Utility of Online Panel Surveys Versus Computer-assisted Interviews in Obtaining Substance-use Prevalence Estimates in the Netherlands." In: *Addiction 104 (10)*, pp. 1641–1645.

Sra, S., Nowozin, S. and Wright, S. J. (2012): *Optimization for Machine Learning*. Cambridge: MIT Press.

Statistisches Bundesamt (2013): *Qualitätsbericht Mikrozensus 2012*. Statistisches Bundesamt (DESTATIS), Wiesbaden. URL: `https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bevoelkerung/mikrozensus-2012.pdf?__blob=publicationFile` *(Retrieved 02.05.2021)*.

Statistisches Bundesamt (2017): *Datenhandbuch Zum Mikrozensus Scientific Use File 2012*. Statistisches Bundesamt (DESTATIS), Wiesbaden. URL: `https://www.forschungsdatenzentrum.de/sites/default/files/mz_2012_suf_dhb.pdf` *(Retrieved 02.05.2021)*.

Statistisches Bundesamt (2020): *Ihr Nutzen. Unser Auftrag.* Statistisches Bundesamt (DESTATIS), Wiesbaden. URL: `https://www.bmi.bund.de/SharedDocs/downloads/DE/behoerden/destatis-stba-ihr-nutzen-unser-auftrag.pdf?__blob=publicationFile&v=8` *(Retrieved 02.05.2021)*.

Steinmetz, S., Bianchi, A., Tijdens, K. and Biffignandi, S. (2014): "Improving Web Survey Quality. Potentials and Constraints of Propensity Score Adjustments." In: *Online Panel Research: A Data Quality Perspective*. Ed. by M. Callegaro, R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick and P. J. Lavrakas. New York: John Wiley & Sons, pp. 273–298.

Steinmetz, S., Tijdens, K. and Pedraza, P. de (2009): *Comparing Different Weighting Procedures for Volunteer Web Surveys: Lessons to Be Learned from German and Dutch WageIndicator Data*. Working paper 09-76. AIAS.

Steinmetz, S. and Tijdens, K. (2009): "Can Weighting Improve the Representativeness of Volunteer Online Panels? Insights from the German Wage Indicator Data." In: *Concepts & Methods 5 (1)*, pp. 7–11.

Stephens, M. A. (1976): "Asymptotic Results for Goodness-of-fit Statistics with Unknown Parameters." In: *The Annals of Statistics 4 (2)*, pp. 357–369.

Stone, M. (1974): "Cross-validatory Choice and Assessment of Statistical Predictions." In: *Journal of the Royal Statistical Society: Series B (methodological) 36 (2)*, pp. 111–133.

Stone, M. (1977): "Asymptotics for and against Cross-validation." In: *Biometrika 64 (1)*, pp. 29–35.

Storey, J. D. (2002): "A Direct Approach to False Discovery Rates." In: *Journal of the Royal Statistical Society: Series B (statistical Methodology) 64 (3)*, pp. 479–498.

Stuart, E. A. (2010): "Matching Methods for Causal Inference: A Review and a Look Forward." In: *Statistical Science: A Review Journal of the Institute of Mathematical Statistics 25 (1)*, p. 1.

Stürmer, T., Schneeweiss, S., Rothman, K. J., Avorn, J. and Glynn, R. J. (2007): "Performance of Propensity Score Calibration–a Simulation Study." In: *American Journal of Epidemiology 165 (10)*, pp. 1110–1118.

Sue, V. M. and Ritter, L. A. (2012): *Conducting Online Surveys*. Los Angeles et al.: Sage Publications.

Sverchkov, M. and Pfeffermann, D. (2004): "Prediction of Finite Population Totals Based on the Sample Distribution." In: *Survey Methodology 30 (1)*, pp. 79–92.

Tam, S.-M. and Clarke, F. (2015): "Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics." In: *International Statistical Review 83 (3)*, pp. 436–448.

Therneau, T. and Atkinson, B. (2018): *Rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13. URL: `https://CRAN.R-project.org/package=rpart` *(Retrieved 05.05.2021)*.

Thompson, W. A. et al. (1962): "The Problem of Negative Estimates of Variance Components." In: *The Annals of Mathematical Statistics 33 (1)*, pp. 273–289.

Tibshirani, R. (1994): *A Proposal for Variable Selection in the Cox Model*. Technical Report. University of Toronto, Department of Statistics.

Tibshirani, R. (1996): "Regression Shrinkage and Selection Via the Lasso." In: *Journal of the Royal Statistical Society: Series B (methodological) 58 (1)*, pp. 267–288.

Tibshirani, R. (1997): "The Lasso Method for Variable Selection in the Cox Model." In: *Statistics in Medicine 16 (4)*, pp. 385–395.

Tijdens, K. (2008): "The WageIndicator Web Survey." In: *Access Panels and Online Research, Panacea or Pitfall? Proceedings of the DANS Symposium*. Ed. by I. Stoop and M. Wittenberg. The Hague: DANS Symposium Publications, pp. 91–99.

Tijdens, K. (2014): "Dropout Rates and Response Times of an Occupation Search Tree in a Web Survey." In: *Journal of Official Statistics 30 (1)*, pp. 23–43.

Tijdens, K. and Steinmetz, S. (2016): "Is the Web a Promising Tool for Data Collection in Developing Countries? An Analysis of the Sample Bias of 10 Web and Face-to-face Surveys from Africa, Asia, and South America." In: *International Journal of Social Research Methodology 19 (4)*, pp. 461–479.

Tijdens, K., Van Klaveren, M., Bispinck, R., Dribbusch, H. and Öz, F. (2014): "Wage and Workforce Adjustments in the Economic Crisis in Germany and the Netherlands." In: *European Journal of Industrial Relations 20 (2)*, pp. 165–183.

Tijdens, K., van Zijl, S., Hughie-Williams, M., Klaveren, M. van and Steinmetz, S. (2010): *Codebook and Explanatory Note on the WageIndicator Dataset: A Worldwide, Continuous, Multilingual Web-survey on Work and Wages with Paper Supplements*. Working paper 10-102. AIAS.

Tijdens, K. (2020): *Managing Surveys: Ten Lessons Learned from Web-surveys*. WageIndicator Report April 2020. WageIndicator Foundation, Amsterdam.

Tillé, Y. (2006): *Sampling Algorithms*. New York: Springer.

Tillé, Y., Wilhelm, M. et al. (2017): "Probability Sampling Designs: Principles for Choice of Design and Balancing." In: *Statistical Science 32 (2)*, pp. 176–189.

Tobin, J. (1958): "Estimation of Relationships for Limited Dependent Variables." In: *Econometrica: Journal of the Econometric Society 26 (1)*, pp. 24–36.

Tongco, M. D. C. (2007): "Purposive Sampling As a Tool for Informant Selection." In: *Ethnobotany Research and Applications 5*, pp. 147–158.

Toomet, O. and Henningsen, A. (2008): "Sample Selection Models in R: Package SampleSelection." In: *Journal of Statistical Software 27 (7)*, pp. 1–23.

Tran, M.-N., Nguyen, N., Nott, D. and Kohn, R. (2020): "Bayesian Deep Net GLM and GLMM." In: *Journal of Computational and Graphical Statistics 29 (1)*, pp. 97–113.

Trefethen, L. N. and Bau, D. (1997): *Numerical Linear Algebra*. Philadelphia, PA: SIAM.

Tsong, Y., Dong, X. and Shen, M. (2017): "Development of Statistical Methods for Analytical Similarity Assessment." In: *Journal of Biopharmaceutical Statistics 27 (2)*, pp. 197–205.

UNESCO (2006): *International Standard Classification of Education: ISCED 1997*. Montreal: United Nations Educational, Scientific and Cultural Organization (UNESCO).

UNESCO (2012): *International Standard Classification of Education: ISCED 2011*. Montreal: United Nations Educational, Scientific and Cultural Organization (UNESCO), Institute for Statistics.

United Nations (2013): *Guidelines on Integrated Economic Statistics*. United Nations, Department of Economic and Social Affairs, ST/ESA/STAT/SER.F/108. URL: https://unstats.un.org/unsd/nationalaccount/docs/ies-guidelines-e.pdf *(Retrieved 09.08.2019)*.

United Nations (2014): *Report of the Global Working Group on Big Data for Official Statistics*. United Nations, Economic and Social Council, E/CN.3/2015/4. URL: https://unstats.un.org/unsd/statcom/doc15/2015-4-BigData-E.pdf *(Retrieved 09.08.2019)*.

Valliant, R. (2009): "Model-based Prediction of Finite Population Totals." In: *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*. Ed. by D. Pfeffermann and C. R. Rao. Amsterdam: Elsevier, pp. 11–31.

Valliant, R. and Dever, J. A. (2011): "Estimating Propensity Adjustments for Volunteer Web Surveys." In: *Sociological Methods & Research 40 (1)*, pp. 105–137.

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000): *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

van Buuren, S. (2018): *Flexible Imputation of Missing Data*. Boca Raton et al.: CRC press.

van den Brakel, J., Söhler, E., Daas, P. and Buelens, B. (2017): "Social Media As a Data Source for Official Statistics; the Dutch Consumer Confidence Index." In: *Survey Methodology 43 (2)*, pp. 183–210.

Van To, T. and Kositviwat, T. (2005): "Using Rational B-spline Neural Networks for Curve Approximation." In: *Proceedings of the 7th WSEAS International Conference on Mathematical Methods and Computational Techniques In Electrical Engineering*. Sofia: World Scientific, Engineering Academy and Society (WSEAS), pp. 42–50.

Varian, H. R. (2014): "Big Data: New Tricks for Econometrics." In: *Journal of Economic Perspectives 28 (2)*, pp. 3–28.

Venables, W. N. and Ripley, B. D. (2002): *Modern Applied Statistics with S*. New York: Springer.

Visintin, S., Tijdens, K., Steinmetz, S. and Pedraza, P. de (2015): "Task Implementation Heterogeneity and Wage Dispersion." In: *Iza Journal of Labor Economics 4 (1)*, pp. 1–24.

Vosen, S. and Schmidt, T. (2011): "Forecasting Private Consumption: Survey-based Indicators Vs. Google Trends." In: *Journal of Forecasting 30 (6)*, pp. 565–578.

WageIndicator Foundation (2011): *Weights for the WageIndicator Dataset*. Distributed as part of the WageIndicator data documentation. URL: `https://datasets.iza.org/dataset/59/wageindicator-survey` *(Retrieved 23.10.2018)*.

Wagner, M. (2013): "Numerical Optimization in Survey Statistics." Doctoral dissertation. University of Trier.

Wallis, J. (1685): *A Treatise of Algebra, Both Historical and Practical*. London: Royal Society of London. Quoted in Cajori (1911).

Wang, C., Chen, M.-H., Schifano, E., Wu, J. and Yan, J. (2016): "Statistical Methods and Computing for Big Data." In: *Statistics and Its Interface 9 (4)*, p. 399.

Wang, H., Yang, M. and Stufken, J. (2019): "Information-based Optimal Subdata Selection for Big Data Linear Regression." In: *Journal of the American Statistical Association 114 (525)*, pp. 393–405.

Wang, K. and Lei, B. (2001): "Using B-spline Neural Network to Extract Fuzzy Rules for a Centrifugal Pump Monitoring." In: *Journal of Intelligent Manufacturing 12 (1)*, pp. 5–11.

Wang, Q., Zhang, X., Zhang, Y. and Yi, Q. (2013): "AUGEM: Automatically Generate High Performance Dense Linear Algebra Kernels on X86 CPUs." In: *SC'13: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. New York: IEEE, pp. 1–12.

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015): "Forecasting Elections with Non-representative Polls." In: *International Journal of Forecasting 31 (3)*, pp. 980–991.

Ward, C. and Ronald, K. (2008): *Numerical Mathematics and Computing*. 6th ed. Belmont: Thomson Brooks/Cole.

Weisberg, H. F. (2005): *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago and London: The University of Chicago Press.

Welch, B. L. (1947): "The Generalization of Student's' Problem When Several Different Population Variances Are Involved." In: *Biometrika 34 (1/2)*, pp. 28–35.

Welch, B. L. (1951): "On the Comparison of Several Mean Values: An Alternative Approach." In: *Biometrika 38 (3/4)*, pp. 330–336.

West, B. T., Little, R. J., Andridge, R. R., Boonstra, P. S., Ware, E. B., Pandit, A. and Alvarado-Leiton, F. (2021): "Assessing Selection Bias in Regression Coefficients Estimated from Non-Probability Samples, with Applications to Genetics and Demographic Surveys." In: *Arxiv Preprint Arxiv:2004.06139*.

Westlake, W. J. (1976): "Symmetrical Confidence Intervals for Bioequivalence Trials." In: *Biometrics 32 (4)*, pp. 741–744.

Willenborg, L. and De Waal, T. (2001): *Elements of Statistical Disclosure Control*. New York: Springer Science & Business Media.

Wilson, R. B. (1963): "A Simplicial Algorithm for Concave Programming." Doctoral dissertation. Harvard University.

Witting, H. (1985): *Mathematische Statistik I: Parametrische Verfahren Bei Festem Stichprobenumfang*. Wiesbaden: Springer.

Wolfe, P. (1969): "Convergence Conditions for Ascent Methods." In: *Siam Review 11 (2)*, pp. 226–235.

Wolfinger, R. and O'connell, M. (1993): "Generalized Linear Mixed Models a Pseudo-likelihood Approach." In: *Journal of Statistical Computation and Simulation 48 (3-4)*, pp. 233–243.

Wolter, K. (2007): *Introduction to Variance Estimation*. New York: Springer Science & Business Media.

Wood, S. N. (2011): "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." In: *Journal of the Royal Statistical Society (b) 73 (1)*, pp. 3–36.

Wood, S. N. (2017): *Generalized Additive Models: An Introduction with R*. 2nd ed. Boca Raton: CRC press.

Wood, S. N., Pya, N. and Säfken, B. (2016): "Supplementary Material: Smoothing Parameter and Model Selection for General Smooth Models." In: *Journal of the American Statistical Association 111 (516)*.

Wooldridge, J. M. (2012): *Introductory Econometrics: A Modern Approach*. 5th ed. Mason: Cengage Learning.

Wu, C. and Sitter, R. R. (2001): "A Model-calibration Approach to Using Complete Auxiliary Information from Survey Data." In: *Journal of the American Statistical Association 96 (453)*, pp. 185–193.

Xiong, Y., Kim, H. J. and Singh, V. (2019): "Mixed Effects Neural Networks (menets) with Applications to Gaze Estimation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New York: IEEE, pp. 7743–7752.

Xu, W., Li, Z., Cheng, C. and Zheng, T. (2013): "Data Mining for Unemployment Rate Prediction Using Search Engine Query Data." In: *Service Oriented Computing and Applications 7 (1)*, pp. 33–42.

Xu, X., Wong, S. and Choi, K. (2014): "A Two-stage Bivariate Logistic-Tobit Model for the Safety Analysis of Signalized Intersections." In: *Analytic Methods in Accident Research 3–4*, pp. 1–10.

Yan, T. and Tourangeau, R. (2008): "Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times." In: *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition 22 (1)*, pp. 51–68.

Yang, S. and Kim, J. K. (2018): "Integration of Survey Data and Big Observational Data for Finite Population Inference Using Mass Imputation." In: *Arxiv Preprint Arxiv:1807.02817*.

Yates, F. and Grundy, P. M. (1953): "Selection without Replacement from within Strata with Probability Proportional to Size." In: *Journal of the Royal Statistical Society: Series B (methodological) 15 (2)*, pp. 253–261.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A. and Wang, R. (2011): "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-probability Samples." In: *Public Opinion Quarterly 75 (4)*, pp. 709–747.

Yee, T. W. and Wild, C. (1996): "Vector Generalized Additive Models." In: *Journal of the Royal Statistical Society: Series B (methodological) 58 (3)*, pp. 481–493.

Yiu, K. F. C., Wang, S., Teo, K. L. and Tsoi, A. C. (2001): "Nonlinear System Modeling Via Knot-optimizing B-spline Networks." In: *Ieee Transactions on Neural Networks 12 (5)*, pp. 1013–1022.

Yule, G. U. (1922): *An Introduction to the Theory of Statistics*. 6th ed. London: Charles Griffin & Co.

Zhang, L.-C. (2000): "Post-stratification and Calibration–a Synthesis." In: *The American Statistician 54 (3)*, pp. 178–184.

Zhang, L.-C., Thomsen, I. and Kleven, Ø. (2013): "On the Use of Auxiliary and Paradata for Dealing With Non-sampling Errors in Household Surveys." In: *International Statistical Review 81 (2)*, pp. 270–288.

Zhang, W. and Goh, A. T. (2016): "Multivariate Adaptive Regression Splines and Neural Network Models for Prediction of Pile Drivability." In: *Geoscience Frontiers 7 (1)*, pp. 45–52.

Zhang, X., Zhao, Y., Guo, K., Li, G. and Deng, N. (2017): "An Adaptive B-spline Neural Network and Its Application in Terminal Sliding Mode Control for a Mobile Satcom Antenna Inertially Stabilized Platform." In: *Sensors 17 (5)*, p. 978.

Zou, H. and Hastie, T. (2005): "Regularization and Variable Selection Via the Elastic Net." In: *Journal of the Royal Statistical Society: Series B (statistical Methodology) 67 (2)*, pp. 301–320.

Simon Lenau – Education & Academic Career

Academic Career

10/2015 – 10/2021	**Research assistant** *Trier University,* *Economic and Social Statistics Department*
12/2012 – 10/2015	**Student assistant** *Trier University,* *Economic and Social Statistics Department*
11/2013 – 10/2014	**Student assistant** *Trier University,* *Chair in Comparative Politics*
04/2012 – 09/2012	**Student assistant** *Technical University of Kaiserslautern,* *Departement of Educational Science*

Education and Studies

10/2015 – 10/2021	**Doctoral student** *Trier University,* *Economic and Social Statistics Department*
10/2012 – 09/2015	**M.Sc. Survey Statistics** *Trier University*
10/2008 – 03/2012	**B.A. Integrative Sozialwissenschaft** *Technical University of Kaiserslautern*
07/2000 – 03/2008	**Abitur** *Burggymnasium Kaiserslautern*