PH.D. THESIS

On Robust Small Area Estimation

SUBMITTED BY

Tobias Schoch
Viktoriastrasse 41
CH-3013 Bern
born January 9, 1980, Zürich

TO THE

Fachbereich IV, Volkswirtschaftslehre
Lehrstuhl für Wirtschafts- und Sozialstatistik
Universität Trier

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor rerum politicarum

SUPERVISORS

Univ.-Prof. Dr. Ralf Thomas Münnich

Prof. Dr. Beat Hulliger

Oral examination: June 8, 2018

Acknowledgments

I would like to express my deep gratitude to my supervisors Univ.-Prof. Dr. Ralf Thomas Münnich and Prof. Dr. Beat Hulliger. I am also grateful for the interesting and stimulating discussions with all the friends and colleagues that I met at conferences or with whom I worked in projects (in particlar the AMELI project, 2008-2011). Finally, I would like to express my sincere gratitude to my parents who always supported me openheartedly in what I considered to be my way.

Contents

Zusammenfassung

Traditionell werden Zufallsstichprobenerhebungen so geplant, dass nationale Statistiken zuverlässig mit einer adäquaten Präzision geschätzt werden können. Hierbei kommen vorrangig designbasierte, Modell-unterstützte (engl. model assisted) Schätzmethoden zur Anwendung, die überwiegend auf asymptotischen Eigenschaften beruhen. Für kleinere Stichprobenumfänge, wie man sie für Small Areas (Domains bzw. Subpopulationen) antrifft, eignen sich diese Schätzmethoden eher nicht, weswegen für diese Anwendung spezielle modellbasierte Small Area-Schätzverfahren entwickelt wurden. Letztere können zwar Verzerrungen aufweisen, besitzen jedoch häufig einen kleineren mittleren quadratischen Fehler der Schätzung als dies für designbasierte Schätzer der Fall ist.

Den Modell-unterstützten und modellbasierten Methoden ist gemeinsam, dass sie auf statistischen Modellen beruhen; allerdings in unterschiedlichem Ausmass. Modell-unterstützte Verfahren sind in der Regel so konstruiert, dass der Beitrag des Modells bei sehr grossen Stichprobenumfängen gering ist (bei einer Grenzwertbetrachtung sogar wegfällt). Bei modellbasierten Methoden nimmt das Modell immer eine tragende Rolle ein, unabhängig vom Stichprobenumfang. Diese Überlegungen veranschaulichen, dass das unterstellte Modell, präizser formuliert, die Güte der Modellierung für die Qualität der Small Area-Statistik von massgeblicher Bedeutung ist. Wenn es nicht gelingt, die empirischen Daten durch ein passendes Modell zu beschreiben und mit den entsprechenden Methoden zu schätzen, dann können massive Verzerrungen und / oder ineffiziente Schätzungen resultieren.

Die vorliegende Arbeit beschäftigt sich mit der zentralen Frage der Robustheit von Small Area-Schätzverfahren. Als robust werden statistische Methoden dann bezeichnet, wenn sie eine beschränkte Einflussfunktion und einen möglichst hohen Bruchpunkt haben. Vereinfacht gesprochen zeichnen sich robuste Verfahren dadurch aus, dass sie nur unwesentlich durch Ausreisser und andere Anomalien in den Daten beeinflusst werden. Die Untersuchung zur Robustheit konzentriert sich auf die folgenden Modelle bzw. Schätzmethoden:

(i) modellbasierte Schätzer für das Fay-Herriot-Modell (Fay und Herrot, 1979, J. Amer. Statist. Assoc.) und das elementare Unit-Level-Modell (vgl. Battese et al., 1988, J. Amer. Statist. Assoc.).

(ii) direkte, Modell-unterstützte Schätzer unter der Annahme eines linearen Regressionsmodells.

Das Unit-Level-Modell zur Mittelwertschätzung beruht auf einem linearen gemischten Gauss'schen Modell (engl. mixed linear model, MLM) mit blockdiagonaler Kovarianzmatrix. Im Gegensatz zu bspw. einem multiplen linearen Regressionsmodell, besitzen MLM-Modelle keine nennenswerten Invarianzeigenschaften, so dass eine Kontamination der abhängigen Variable unvermeidbar zu verzerrten Parameterschätzungen führt. Für die Maximum-Likelihood-Methode kann die resultierende Verzerrung nahezu beliebig groß werden. Aus diesem Grund haben Richardson und Welsh (1995, Biometrics) die robusten Schätzmethoden RML 1 und RML 2 entwickelt, die bei kontaminierten Daten nur eine geringe Verzerrung aufweisen und wesentlich effizienter sind als die Maximum-Likelihood-Methode. Eine Abwandlung von Methode RML 2 wurde Sinha und Rao (2009, Canad. J. Statist.) für die robuste Schätzung von Unit-Level-Modellen vorgeschlagen. Allerdings erweisen sich die gebräuchlichen numerischen Verfahren zur Berechnung der RML-2-Methode (dies gilt auch für den Vorschlag von Sinha und Rao) als notorisch unzuverlässig. In dieser Arbeit werden zuerst die Konvergenzprobleme der bestehenden Verfahren erörtert und anschließend ein numerisches Verfahren vorgeschlagen, das sich durch wesentlich bessere numerische Eigenschaften auszeichnet. Schließlich wird das vorgeschlagene Schätzverfahren im Rahmen einer Simulationsstudie untersucht und anhand eines empirischen Beispiels zur Schätzung von oberirdischer Biomasse in norwegischen Kommunen illustriert.

Das Modell von Fay-Herriot kann als Spezialfall eines MLM mit blockdiagonaler Kovarianzmatrix aufgefasst werden, obwohl die Varianzen des Zufallseffekts für die Small Areas nicht geschätzt werden müssen, sondern als bereits bekannte Größen betrachtet werden. Diese Eigenschaft kann man sich nun zunutze machen, um die von Sinha und Rao (2009) vorgeschlagene Robustifizierung des Unit-Level-Modells direkt auf das Fay-Herriot Model zu übertragen. In der vorliegenden Arbeit wird jedoch ein alternativer Vorschlag erarbeitet, der von der folgenden Beobachtung ausgeht: Fay und Herriot (1979) haben ihr Modell als Verallgemeinerung des James-Stein-Schätzers motiviert, wobei sie sich einen empirischen Bayes-Ansatz zunutze machen. Wir greifen diese Motivation des Problems auf und formulieren ein analoges robustes Bayes'sches Verfahren. Wählt man nun in der robusten Bayes'schen Problemformulierung die ungünstigste Verteilung (engl. least favorable distribution) von Huber (1964, Ann. Math. Statist.) als A-priori-Verteilung für die Lokationswerte der Small Areas, dann resultiert als Bayes-Schätzer [=Schätzer mit dem kleinsten Bayes-Risk] die Limited-Translation-Rule (LTR) von Efron und Morris (1971, J. Amer. Statist. Assoc.). Im Kontext der frequentistischen Statistik kann die Limited-Translation-Rule nicht verwendet werden, weil sie (als Bayes-Schätzer) auf unbekannten Parametern beruht. Die unbekannten Parameter können jedoch nach dem empirischen Bayes-Ansatz an der Randverteilung der abhängigen Variable geschätzt werden. Hierbei gilt es zu beachten (und dies wurde in der Litera-

tur vernachlässigt), dass die Randverteilung unter der ungünstigsten A-priori-Verteilung nicht einer Normalverteilung entspricht, sondern durch die ungünstigste Verteilung nach Huber (1964) beschrieben wird. Es ist nun nicht weiter erstaunlich, dass es sich bei den Maximum-Likelihood-Schätzern von Regressionskoeffizienten und Modellvarianz unter der Randverteilung um M-Schätzer mit der Huber'schen psi-Funktion handelt.

Unsere theoriegeleitete Herleitung von robusten Schätzern zum Fay-Herriot-Modell zeigt auf, dass bei kontaminierten Daten die geschätzte LTR (mit Parameterschätzungen nach der M-Schätzermethodik) optimal ist und, dass die LTR ein integraler Bestandteil der Schätzmethodik ist (und nicht als "Zusatz" o.Ä. zu betrachten ist, wie dies andernorts getan wird). Die vorgeschlagenen M-Schätzer sind robust bei Vorliegen von atypischen Small Areas (Ausreissern), wie dies auch die Simulations- und Fallstudien zeigen. Um auch Robustheit bei Vorkommen von einflussreichen Beobachtungen in den unabhängigen Variablen zu erzielen, wurden verallgemeinerte M-Schätzer (engl. generalized M-estimator) für das Fay-Herriot-Modell entwickelt.

List of Mathematical Symbols

I. General mathematical notation

$(a, b),\ [a, b]$	open and closed interval in $\mathbb{R}$ ($a < b$)
$\mathbb{R},\ \mathbb{N}$	real and natural numbers
$\inf,\ \sup$	(pointwise) infimum and supremum
$\min,\ \max$	minimum and maximum
$\operatorname{sgn},\ \exp,\ \log$	sign(um), exponential, and (natural) logarithm function
λ_n	Lebesgue measure on $\mathbb{R}^n$
a.e., μ-a.e.	almost everywhere (w.r.t. measure μ)
$L^1(\Omega, \mu)$	space (equivalence class) of absolutely μ-integrable functions on Ω
$\operatorname{Lip}(U),\ \operatorname{AC}(U)$	class of Lipschitz continuous and absolutely continuous functions on $U \subset \mathbb{R}$

II. Statistics and probability-related notation

r.v.	random variable
$X \sim F$	r.v. X is distributed according to the law F
$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	p-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (in the $p = 1$ dimensional case, the subscript p is omitted)
$\Phi,\ \phi$	cumulative distribution function (c.d.f.) and probability density function (p.d.f.) of the standard normal distribution
$\xrightarrow{\text{p}},\ \xrightarrow{\text{a.s.}}$	convergence in probability and almost surely
$\mathcal{O},\ \mathcal{O}_p$	big oh notation
$\mathbb{E}_\xi,\ \mathbb{V}_\xi,\ \operatorname{Cov}_\xi$	expectation, variance and covariance (operator) w.r.t. model ξ
$\mathbb{E}_p,\ \mathbb{V}_p,\ \operatorname{Cov}_p$	expectation, variance and covariance (operator) w.r.t. the sampling design
$U,\ s$	universe (population) and sample
$p(s)$	sampling design

$o,\ o_p$	little oh notation
$N,\ n$	population and sample size
$\pi_i,\ \pi_{ij}$	first and second order sample inclusion probabilities of the elements i and ij

III. Vector and matrix notation

$\circ,\ \otimes$	elementwise (Hadamard) matrix product and Kronecker product
$\|\cdot\|_p$	for finite vector spaces: Euclidean norm ($p = 2$), maximal norm ($p = \infty$), spectral matrix norm ($p = S$), and Frobenius matrix norm ($p = F$); norm on the space of Lebesgue integrable function ($p = L^1$)
$\operatorname{tr} \mathbf{A}$	trace of matrix $\mathbf{A}$
$\mathbf{A}^T$	transpose of matrix $\mathbf{A}$
∇f	gradient of the function/ map f
$\mathbf{I}_n,\ \mathbf{J}_n$	identity matrix and matrix of ones of size $(n \times n)$
$\mathbb{1},\ \mathbf{1}_n$	indicator function and n-vector of ones
$\mathbf{A}^{-1}$	matrix inverse of the square matrix $\mathbf{A}$
$\mathbf{A}^{+}$	pseudoinverse of matrix $\mathbf{A}$ (e.g., Moore–Penrose inverse)
$\operatorname{diag}(\mathbf{a})$	diagonal matrix along the elements $a_1, \ldots, a_p$
$\operatorname{blockdiag}(\mathbf{A}, \mathbf{B})$	block diagonal matrix with matrices $\mathbf{A}$ and $\mathbf{B}$ stacked along the main diagonal (the remaining elements of the matrix are equal to zero)
$\underline{\lambda}(\mathbf{A}),\ \overline{\lambda}(\mathbf{A})$	minimum and maximum eigenvalue of matrix $\mathbf{A}$

List of Tables

List of Figures

1. Introduction

The demand for reliable statistics has been growing over the past decades as more and more decision makers (in businesses and politics) request evidence-based knowledge. In this context, sample surveys have long been used as cost-effective means for data collection. Such data have been effectively used to provide suitable statistics not only for the population targeted by the survey, but also for a variety of subpopulations, often called domains or areas. Domains may be geographical areas such as Bundesländer, Gemeinden, or socio-demographic groups (e.g., female in the age bracket [75–80] years) or any other kind of subpopulation. We will use the terms domain and area interchangeably. With respect to size, a domain is considered large if the domain-specific sample size is sufficiently large to provide a direct estimate[1] of the domain characteristic (e.g., mean) with *adequate* precision. Vice versa, a domain is regarded small if the domain-specific sample is not large enough to produce a direct estimate with reliable precision.

Growing demand for small-area estimates

A *universally* agreed measure that indicates adequate precision does not exist. Whether the achieved precision meets expectations or requirements must be decided on a case-by-case basis. For the German census 2011, authorities in charge formulated the goal that the precision, in terms of relative root mean squared error, for area-level estimates (i.e., communities and urban districts) has to be smaller than 0.5%; see Münnich et al. (2012, chap. 2.1.2). In order to reach such precision goals, explicit small-area estimation (SAE) methods are inevitable as direct estimators virtually never comply with the requirements. Hence, there is a growing demand for indirect small-area estimates, particularly in central Europe. Governments in countries outside central Europe have passed laws requiring regular production of reliable small area estimates already in late 1980s and early 1990s. For example, the U.S. Congress has passed a law requiring the Secretary of Commerce beginning in 1996 to publish poverty estimates for states, counties, and – to the extent feasible – even for local jurisdictions of government and school districts; see the compiled manual by Schaible (1996). In the European Union, the AMELI[2] research project

[1] A direct domain estimator does not include sample data other than from the domain under study [this notion will be made precise later].

[2] AMELI: Advanced Methods for European Laeken Indicators: research project funded by the European Commission in the Seventh Framework Programme for Research.

made important contributions to the analysis of income inequality and poverty measurement for small areas; see Münnich et al. (2011).

In contrast to direct estimators, indirect estimating methods of small area characteristics borrow strength either from other small areas (cross-sectional estimates), over time (time series) or from both (cross-sectional time series); see e.g. Datta (2009, 252). The latter two methods require a system of surveys which is repeated regularly over time. The focus in this thesis is put exclusively on cross-sectional estimates.

Robustness

Small area estimation relies *explicitly* on prior assumptions and beliefs about the data generating process, which are commonly summarized as "the model". Therefore, SAE methods are commonly referred to as model-based estimating methods. Most of the traditional direct estimators, on the other hand, are merely model-assisted estimators in the sense that they are asymptotically model independent. Though, neither class of estimators is completely model free; hence, both classes may suffer from model misspecification (although to varying degrees). A major cause of model misspecification relates to the choice of the underlying distributional assumption. Therefore, robustness considerations become of vital importance as classical estimators can be severely biased (and / or have inflated variance) in the presence of atypical or outlying observations. It is commonly agreed that every robust statistical procedure should satisfy the following two principles; see Huber (1981, 5).

Principle 1 *(Qualitative robustness)*. A *small* change in the underlying distribution should cause only a *small* change in the performance of a statistical procedure.

Principle 2 *(Quantiative robustness)*. A *somewhat larger deviation* from the core model should not cause a *catastrophe*.

Principle 1 originates from F. Hampel's ground breaking ideas related to his notion of the influence curve (later called influence functions); see Hampel (1968). Principle 2 is complementary to the first principle. It is closely tied to the notion of *breakdown point*, which is – loosely speaking – the maximum number of "bad points" an estimator can withstand before breaking down. A breakdown is observed when an estimator attains values of no "relevant value" for the problem at hand (e.g., infinity or zero, depending on the context). The concept of breakdown point was also introduced by Hampel (1968) and further developed by Donoho and Huber (1983) and others.

Throughout the thesis the two principles will play a major role. We will study

estimators that comply with both principles.

1.1. Goal and contribution

There is an *enormous body of literature*, both concerning robust statistics and small area estimation, which renders any attempt to really grasp the subject in all aspects ridiculous. For the sake of illustration, we give a rough "estimate" of how productive the scientific community in the field of SAE was. To this end, we counted the references listed in the 2nd edition of J.N.K. Rao's book [with co-author Isabel Molina, 2015; for the first edition, see Rao, 2003]. Overall, we counted 487 references.

Despite the vast body of literature, we spotted some "niches", where we can contribute to the existing research with new and interesting results. Our contributions relate to both, the model-assisted / randomization-assisted and the model-based strand of research.

Randomization-assisted inference

The ratio estimator and asymptotic robustness (see Chapter 3)

Suppose a finite population U consisting of N elements y_i, $i = 1, \dots, N$. The y_i's are regarded as fixed real-valued constants and are assumed unknown. The goal is to estimate the population y-mean, where the population is assumed to satisfy the model $y_i = x_i\beta$ (plus some stochastic component), $i = 1, \dots, N$; the x_i's are known constants (auxiliary variable), and $\beta \in \mathbb{R}^+$. It is well-known that the generalized difference (GD) estimator of the y-mean is a design unbiased statistic (see e.g. Cassel, Särndal, and Wretman, 1976, 616). However, the GD estimator is not applicable in practice as it depends on the unknown β. Therefore, Cassel et al. (1976, 617) suggested to replace the unknown β by an estimate of β; the resulting estimator is called generalized regression estimator (GREG). The GREG estimator is not design consistent; however, this is no reason to worry as exact design-unbiasedness is of doubtful virtue (Särndal, 1980, 641). Under mild assumptions, the design bias of the GREG estimator vanishes asymptotically as the sample size grows without bounds. Therefore, the GREG estimator is said to be asymptotically design consistent (ADC and asymptotically design unbiased, ADU). More importantly, the contribution of the assisting model to the estimator vanishes as the sample size grows (irrespective whether the model holds or not). This property is an intrinsic characteristic of ADU estimators and has become the cornerstone of the model-assisted sampling paradigm. C.E. Särndal and some his coauthors argue that randomization can therefore be seen as a source of *robustness* against model failure; see Särndal (1980, 641) and in a similar vein, see Cassel, Särndal, and Wretman (1977,

chap. 7). Though, it must be pointed out that robustness in the sense of C.E. Särndal and coauthors takes only effect when the sample size becomes large.

Clearly, this type of robustness differs from the classical notion of robustness related to Principles 1 and 2, but is of particular interest to the field of sampling. When we do have firm beliefs that the variable of interest underlies a data generating mechanism with heavy tails or if it is supposed to have an asymmetric distribution, then the argument that the contribution of the assisting model vanishes as $N \to \infty$ is not helpful—or may even be misleading. For sufficiently large N, our estimator will indeed become indistinguishable from the sample mean (either of type Horvitz–Thompson or Hajek, depending on our formulation of the GREG estimator). But, the sample mean is not an adequate measure of central tendency in the presence of heavy tails or an asymmetric distribution. The sample mean either results in a biased or inefficient estimate, or both. Hedlin, Falvey, Chambers, and Kokic (2001, 543) assess the situation as follows: "[i]t is just not true that GREG estimators are relatively robust to model choice".

For the moment, let us suppose that the sample data are relatively "well-behaved" and not outlier-prone. In such situations, design consistency may indeed play an important role. We shall thus pick up the notion of robustness / consistency due to Särndal (and his coauthors) and introduce the idea of *strongly* design consistent estimators. Surprisingly, we require only mild regularity assumptions on the design and the behavior of the x_i's and, this is important, the additional assumption that the study variable y_i is nonnegative. The restriction to nonnegative y_i's is quite natural in the context of ratio estimation of the mean or total. Our results contribute to the understanding of ratio estimators and the class of QR-estimators due to Wright (1983).

Robust estimation under the linear assisting model (see Chapter 4)

One may come to the conclusion that under the linear assisting model (almost) "every" conceivable robust method of some importance has already been developed, discussed at large and applied in practical applications. There is indeed a large body of literature which supports this impression. It is neither our goal nor is it truly possible to do full justice to such a large subject in such a limited space. Nonetheless, we have elaborated – in a rather encyclopedic attempt – a compilation of robust domain estimators under the linear assisting model. The compilation of estimators distinguishes whether a population model (treating all domains the same) or domain-specific model is assumed, whether a domain-specific or population-level estimator and / or predictor is applied, whether the Horvitz–Thompson or the Hajek estimator builds the base estimator, etc. Our list of estimators are motivated by the work of Hidiroglou and Patak (2004), who came up with a similar list of non-robust estimators. The compiled list of esti-

mators (and the respective nomenclatura) may help others to figure out what estimator they need.

Model-based inference

In this part, we turn to explicit SAE models for estimating small area means that make specific allowance for between-area variation. The two most commonly used best known classes of models are the area-level and (basic) unit-level model (Rao, 2003, chap. 5). The major difference between the two model classes is the level of aggregation: the area-level model relates area-specific (i.e. aggregated) auxiliary data to the small area means, whereas the unit-level model – as its name implies – depends on within-area units. The best known proposal of an area-level model is due to Fay and Herriot (1979); the basic unit-level model (a.k.a. nested error regression model) has been introduced by Battese, Harter, and Fuller (1988). The key to understanding the models is that they can be regarded as special cases of the broad family of mixed linear models (MLM) with block-diagonal covariance matrix. This then opens the possibility to express optimal estimators of the area-specific means via application of the empirically best linear unbiased predictor (EBLUP) theory; see (Rao, 2003, chap. 7.1).

However, MLM's with block-diagonal covariance matrix have – unlike location-scale or regression models – no nice invariance structure. This implies that the model parameters cannot be estimated consistently in the presence of contamination—there is an unavoidable asymptotic bias (see e.g. Welsh and Richardson, 1997). In the presence of contamination, any method estimates the parameter at the core model plus an unknown bias. In the case of the maximum likelihood estimator, the bias can be arbitrarily large which renders these estimators extremely inefficient. Therefore, Richardson and Welsh (1995) and Welsh and Richardson (1997) suggested robust estimators (called RML 1 and 2) for MLM's with a block-diagonal covariance matrix. Though, it took quite some time until the research community in SAE adopted the work of A.M. Richardson and A.H. Welsh.[3] This eventually was the case when S.K. Sinha and J.N.K. Rao published their remarkable robustification of the EBLUP method; see Sinha and Rao (2009). Technically, their M-estimator is an approximation to the RML 2 method of Richardson and Welsh (1995).

Robust estimation under the basic unit-level model (see Chapter 5)

The approximation suggested by Sinha and Rao (2009) is – unfortunately – not sufficient to ensure numerically sound solutions to the notoriously instable esti-

[3] Instead of directly adopting the work of Richardson and Welsh (1995), researchers around R.L. Chambers studied an alternative approach: M-quantile models for small area estimation; see e.g. Chambers and Tzavidis (2006).

mating problem in robust MLMs. Convergence issues are not only experienced in the Sinha–Rao setup, but are also reported by Richardson (1995, chap. 6.5) and Chaubey and Venkateswarlu (2002). Our experiences show that the major source of convergence failure is due to the Newton–Raphson method (which has a quadratic rate of convergence within some neighborhood of the root, but can go horribly wrong outside the neighborhood) and the parametrization of the model; see Schoch (2011a). Our contributions to solve the outlined problems are:

- a robust method for computing the parameters under the basic unit-level model which is equivalent to the RML 2 approach, but uses a different parametrization. Our method bypasses some of the notorious instabilities experienced in different setups, and it has superior numerical properties;
- a robust prediction method that is much simpler than the one suggested by Sinha and Rao (2009);
- an algorithm for computing robust estimates of the area-specific means that is numerically stable and computationally efficient.

Robust estimation under the Fay–Herriot model (see Chapter 6)

The Fay–Herriot (FH) model and some of the more sophisticated area-level models can be regarded as special cases of an MLM—although a degenerate MLM, pretending certain variance parameters are known. Therefore, the robustification can be tackled using the M-estimator theory developed by Richardson and Welsh (1995). So, the robustification under the FH model is an immediate consequence of applying the method of Sinha and Rao (2009) to the area-level model. The details of this approach have been written up in the Ph.D. thesis of Warnholz (2016b, chap. 3.3).

We have chosen a different approach to obtain robust estimators and predictors under the Fay–Herriot model. To this end, we go back to the roots of the FH model, namely: James–Stein estimation. In their seminal paper, Fay and Herriot (1979) motivate their SAE method as an extension of the James–Stein estimator that obtains via an empirical Bayes argument, which has been popularized by Efron and Morris (1972, 1973b). We pick up this line of argument, adopt a robust Bayes view and derive from there a robust empirical Bayes method that creates a direct link to the seminal work of Huber (1964), but under a slightly more general model. Technically, the resulting M-estimator under the FH model coincides largely with the Sinha–Rao proposal applied to the FH model. Nevertheless, our major contribution is to put the derivation of M-estimators under the area-level model onto a formal theoretical foundation, instead of, what Stahel and Welsh (1997) called, merely “huberizing suitable

quantities". Via this theoretical motivation, we can view the existing estimators and their behavior from a new perspective. In addition to M-estimators, we introduce the class of generalized regression M-estimator (GM). The GM-estimators are an indispensable tool in the presence of influential observations in the design space of the model and outlying areas.

1.2. Organization of the thesis

The chapters are to a large extent self-contained. However, some of the introductory material is presented in Chapter 2 (e.g., introduction of notation and some preliminary stuff).

We discuss (i) the major research question, (ii) relevant literature, (iii) and our line of attack at the beginning of every part / chapter separately. Furthermore, the chapters include their own very specific simulation study and conclude with a specific summary. The final conclusion (see Chapter 7) draws together the findings from all chapters and closes with a general discussion on robust small area estimation.

2. Preliminaries, basic concepts, and notation

In the subsequent sections, we introduce mathematical notation and concepts that will be relevant to all (or almost all) chapters of the thesis. The discussion of those concepts that are required in only one chapter will be deferred for the moment and addressed in the respective chapter.

2.1. Domain estimation

Throughout the thesis, we use the terms domain of interest and area interchangeably (in the literature, the latter term typically refers to geographically determined entities). Consider the population U of size $N > 0$ and suppose a sample s of size $0 < n < N$ has been drawn according to some sampling design $p(s)$ (the notion of a sampling design will be made explicit, see below). For the moment, it does not matter whether U is regarded as a finite or infinite population. Associated with each element in U is the real-valued variable Y_i, $i = 1, \ldots, N$. Depending on the context, the Y_i's will be regarded as random variables (r.v.) or constants. For the moment, this differentiation does not matter. We denote by y_i constants that are supposed known for all sampled elements, but are regarded unknown for all elements in the set $\{i \in U : U \setminus s\}$.

Domain structure

The population U is supposed to be divided into $d = 1, \ldots, D$ domains, and the objective is to estimate characteristics for each domains on grounds of the sampled data. To this end, we we shall define the *domain structure* (Lehtonen and Veijanen, 2009, 222) as the partitioning of U into subsets $U_1, \ldots, U_D$, which is made rigorous in the following definition.

Definition 2.1 (Domain structure)**.** *The population U is supposed to be structured into $d = 1, \ldots, D$ mutually exclusive and exhaustive domains spanning the whole population such that*

$$U = \bigcup_{d=1}^{D} U_d,$$

where U_d denotes the set of elements that fall into domain d.

In view of the definition, it is obvious that the sample s features a corresponding partitioning $s = \cup_{d=1}^{D} s_d$, where $s_d = U_d \cap s$ is the part of s that falls into domain d ($d = 1, \ldots, D$).

Remarks. (i) The division of U into domains given by Definition 2.1 represents a special type of partitioning into subsets and is technically regarded as a stratification. The definition implies that $N = \sum_{d=1}^{D} N_d$ and $n = \sum_{d=1}^{D} n_d$, where N_d and n_d are, respectively, the size of U_d and s_d.

(ii) In practice, domain structures typically cut across design strata or clusters (or, rarely, coincide with a set of primary sampling units (PSU) of a multistage design). This has important implications whether the n_d are fixed known constants or random quantities. In this respect, we distinguish two types of domains (cf. Lehtonen and Veijanen, 2009, 222–23):

- *planned* (or primary) domains and
- *unplanned* (or secondary) domains.

Unplanned domains (and thus unplanned domain structures) occur if the information of domain membership is *not incorporated* into the sampling design. Hence, the domain structure may cut across a partitioning induced by the sample design. An immediate consequence is that the n_d's will be random. The random nature of the n_d has implications for variance estimation, but apart from this, it does not pose any further difficulties. As a rule, the random nature of the n_d's tends to increase the variance of the estimators.

(iii) The implied partitioning of s in to $s_1, \ldots, s_D$ may be such that the sample domain structure is sparsely populated; it may even be the case that certain s_d's are empty (i.e., $s_d = \varnothing$). In such a case, we set $n_d = 0$.

Some further comments are in order. If the elements in U_d can be identified (and listed) beforehand, the domain is typically designated as a planed domain or stratum. However, even if U_d could be singled out as a stratum, one does not always choose to do so in practice. In particular, when the number of domains D is large, a high cost is associated with controlled selection in every domain (as well as other impracticalities). Also, in order to avoid confusion, we use the term strata (or cluster) only to mean a partitioning of U other than the domain structure.

It is convenient to work with domain indicators (functions). Let $\mathbb{1}\{i \in U_d\}$ denote the domain indicator which equals one if element i satisfies $i \in U_d$ and zero otherwise. The indicators may be thought of as being arranged over all domains $d = 1, \ldots, D$ in the $(D \times 1)$ vector $\mathbf{1}_i = (\mathbb{1}\{i \in U_1\}, \ldots, \mathbb{1}\{i \in U_D\})^T$, which consists of $D-1$ entries of zero and a single entry 1 indicating the domain of element $i \in U$. The domain totals are defined as

$$T_{y,d} = \sum_{i \in U} y_i \mathbb{1}\{i \in U_d\} = \sum_{i \in U_d} y_i \quad \text{for all } d = 1, \ldots, D, \tag{2.1}$$

where the expression $y_i \mathbb{1}\{i \in U_d\}$ in (2.1) is called *extended domain variable* by

H.O. Hartely and others; see Lehtonen and Veijanen (2009, 223).

Direct and indirect estimators

Domain-specific estimators that rely exclusively on data of the respective domain are called *direct* estimators. In contrast, *indirect* estimators exploit also data from other (e.g. neighboring) domains. Direct estimators are easily seen to be a special case of indirect estimators. In practice, direct estimators are typically used for planned domain structures whereas indirect estimators are a natural application for unplanned domains (Lehtonen and Veijanen, 2009, 225). The distinction between direct and indirect estimators will become important when we study model-assisted estimators. Moreover, estimators that can be represented in terms of an extended domain variable are *additive* in the sense that the sum over all domains equals the population total (Lehtonen and Veijanen, 2009, 223).

2.2. Survey sampling and randomization inference

Consider the survey population U introduced above. We shall now assume that U is a finite population that consists of N, $1 \leq N < \infty$, fixed values $y_1, \ldots, y_N$. For ease of disposition, we shall temporarily assume that N is a known quantity, and that the finite population U is regarded as a set of labels $\{i : i = 1, \ldots, N\}$. The ith element of U will be referred to by its label i. The values y_i, $i \in U$ are supposed unknown throughout our discussion. The population characteristics of the study variables (possibly vector-valued) have to be estimated from sampled data.

Often samples are obtained from drawing units successively from the population according to some randomization scheme. Denote by $\mathscr{S}$ the set of all sets s, given a predetermined randomization scheme. It is useful to state the probability of selecting a particular sample s. Consequently, we think of s as a realization of the set-valued random variable S taking values $s \in \mathscr{S}$. Define the set-valued function $p : \mathscr{S} \to \mathbb{R}^+$, called *sampling design*, such that $p(S = s)$ gives the probability of selecting s under the sampling scheme in use. Since $p(s)$ is a probability distribution on $\mathscr{S}$, we note that $p(s)$ satisfies $p(s) \geq 0$, for all $s \in \mathscr{S}$, and $\sum_{s \in \mathscr{S}} p(s) = 1$; see Cassel et al. (1977, 9).

The inclusion of a given element $i \in U$ in a sample s is a random event indicated by the zero-one sample indicator $\mathbb{1}^p\{i \in s\}$ which is equal to one if element i is in the sample and zero otherwise. Under the design-based inference paradigm, the logical status of $\mathbb{1}_i^p$ is that of a random variable; the superscript p will be omitted whenever no confusion can arise. Let the first-order sample

inclusion probability π_i be defined as

$$\pi_i = \sum_{s \ni i} p(s) \quad \text{for all } i \in U,$$

where $s \ni i$ means that the sum is over those samples s that contain element i. Likewise, we obtain the second-order sample inclusion probability, $\pi_{ij} = \sum_{s \ni (i,j)} p(s)$, where $\pi_{ii} = \pi_i$ for all $i, j \in U$; see e.g. Särndal, Swensson, and Wretman (1992, 28).

The space of eligible sampling designs which satisfy the assumptions discussed so far is to big to be useful. We shall therefore restrict the set of designs as follows.

Definition 2.2. *A sampling design $p(s)$ is such that*

(i) all labels $i \in U$ are identifiable,

(ii) it is possible to observe and measure without error the variables of interest for each sampled element $i \in s$,

(iii) the randomization scheme is non-informative,

(iv) $p(s) > 0$ for all $s \in \mathscr{S}$,

(v) $\pi_i > 0$ for all $i \in U$ and $\pi_{ij} > 0$ for all $i, j \in U$, $i \neq j$.

Remarks. (i) The first assumption in Definition 2.2 states that each element $i \in U$ is unambiguously identifiable, meaning that each $i \in U$ can be uniquely labeled and the label of each unit is known; see Cassel et al. (1977, 4). Assumption (ii) implies *absence of nonresponse* and that the variables of interest can be obtained without measurement error.

(ii) A randomization scheme is said *non-informative* if the sample inclusion probabilities are independent of the study variables (Cassel et al., 1977, 12); the π_i's are allowed to depend on some auxiliary variables. The assumption of a non-informative design is not very restrictive; also, non-informative designs are the rule in practical (multi-purpose survey) applications since they do require only one sampling design instead of considering a design for each variable of interest. Assumption (iv) restricts attention to the (sub-) class of *possible samples*, i.e. the set of samples that can actually be obtained; see e.g. Särndal et al. (1992, 28).

(iii) Under the assumption that the π_i's satisfy $\pi_i > 0$ for all $i \in U$ [see (v) in Definition 2.2], every population element has a chance to be sampled. This is a necessary condition for a design $p(s)$ to be called a probability

sampling design (Särndal et al., 1992, 32). Moreover, it is a necessary and sufficient condition for obtaining design unbiased estimators (Cassel et al., 1977, 68). In practice, one nevertheless uses designs where this assumption is not met. For instance in business surveys, enterprises are drawn by cut-off sampling. That is, very small enterprises are given an inclusion probability of zero since their contribution to e.g. the estimate of the total turnover value is deemed negligible (also compared to the cost of sampling, surveying and maintaining the database of a large number of very small firms); see e.g., Hidiroglou and Lavallée (2009). In this context, it is still possible to obtain an unbiased estimator of the total for the subset of U for which $\pi_i \neq 0$ (cf. Särndal et al., 1992, 527–28). If both assumptions in (v) hold, a sampling design is called (design) *measurable* since it allows the calculations of valid variance estimates. More precisely, assumption (v) is a sufficient condition for the existence of unbiased variance estimators (Särndal et al., 1992, 33).

(iv) Definition 2.2 encompasses element-sampling and rather complex multi-stage designs.

Under assumptions (i) and (ii) of Definition 2.2, it can be shown by sufficiency arguments [see Cassel et al. (1977, chap. 2.2) or Thompson (1997, 10)] that all that is required for any inference about the study variable of interest is the reduced data

$$\mathscr{D}^* = \{(i, y_i) : \ i \in s\}. \tag{2.2}$$

The definition of $\mathscr{D}^*$ is still more general than we typically will require. In fact, it suffices for our purposes to work with the set of unlabeled measurements

$$\mathscr{D} = \{y_i : \ i \in s\}, \tag{2.3}$$

which contains all the information relevant for inference. That is, for a large class of estimators, the labels (e.g., drawing order, etc.) are non-informative in the inference-making process.

Some frequently used estimators require additional information in order to be computed, such as the values x_i of an auxiliary variable, or the probabilities π_i, say, with which the various units $i \in U$ are selected when sampling is with unequal probabilities. We shall therefore assume that such information is available in a complete list for the population. Hence, once a sample has been selected, we can go back an consult the list to obtain the values of x_i or π_i for each $i \in U$. Such estimators are also called label-dependent (Cassel et al., 1977, 18). From here on, we regard either $\mathscr{D}^*$ or $\mathscr{D}$ as embodying our sample data.

2.2.1. Randomization-based inference

Consider the finite population U and let $p(s)$ be a sampling design in accordance with Definition 2.2. In close analogy with classical statistics, we may define the design probability space $(\mathscr{S}, \mathcal{B}, p)$, where $\mathscr{S}$ is the sample space, $\mathcal{B}$ denotes the sets of all (measurable) subsets of $\mathscr{S}$ and $p(\cdot)$ is regarded as a probability measure; see Rubin-Bleuer and Schiopu-Kratina (2005).

Suppose a r.v. D taking values in the sample space $\mathcal{D}$, which shall be defined as $\mathcal{D} = \{\mathscr{D} : s \in \mathscr{S}, \mathbf{y} \in \Omega\}$; in general we take $\Omega = \mathbb{R}^N$. With this, we define the notion of a *statistic* $\theta(D)$, which is an estimable function θ on $\mathcal{D}$ such that for any given $s \in \mathscr{S}$, θ depends on $\mathbf{y}$ only through the elements y_i, $i \in s$ (Cassel et al., 1977, 20). The key is that the y_i's are variables in the sense of taking (possibly) different values for the elements $i \in U$ but they *are not treated as random variables*. The random nature of $\theta(D)$ stems solely from the fact that θ takes values in $\mathcal{D}$ (Särndal et al., 1992, 34–5). Once a sample s has been realized given a sampling design $p(s)$, an estimator of θ shall be given by $\hat{\theta} = \theta(s)$, where s is a realization of the r.v. S defined on the possible set of sample $\mathscr{S}$.

In what follows, we suppress the dependency of an estimator $\hat{\theta}(s)$ on s and use the abbreviated notation $\hat{\theta}$ instead.

A strategy

In finite-population sampling, both sampling design and estimator are under the control of the statistician and therefore closely related. A design–estimator pair (p, θ) is often referred to as *strategy*; see Chaudhuri and Stenger (2005, chs. 2 and 3). The standard paradigm under which randomization inference (a.k.a. design-based inference) operates requires us to find estimators of a population characteristic (e.g., total) that are design unbiased (at least approximatively or nearly design unbiased); see below. Expectation with respect to (w.r.t.) the design is denoted by the operator $\mathbb{E}_p$ and is defined as

$$\mathbb{E}_p\big(\hat{\theta}\big) = \sum_{s \in \mathscr{S}} p(s)\hat{\theta}(s),$$

likewise we write $\mathbb{V}_p$ to mean the variance operator,

$$\mathbb{V}_p\big(\hat{\theta}\big) = \sum_{s \in \mathscr{S}} p(s)\big(\hat{\theta}(s) - \mathbb{E}_p\big[\hat{\theta}(s)\big]\big)^2.$$

The operator cov_p shall be defined analogously.

Basic (direct) estimators

It will prove useful to introduce some basic estimators of the most common characteristics of interest, population total and arithmetic mean. These basic estimators will serve as benchmark or reference line in comparison with more elab-

orate estimators. For any design $p(s)$ that satisfies Definition 2.2, the (Narain-) Horvitz–Thompson (HT) estimator,

$$\hat{T}_y^{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

is an unbiased estimator of the y-total $T_y = \sum_{i \in U} y_i$ with variance (cf. Horvitz and Thompson, 1952)

$$v_{HT} = \mathbb{V}\big[\hat{T}_y^{HT}\big] = \sum_{i \in U} \frac{y_i^2}{\pi_i}(1 - \pi_i) + \sum_{i,j \in U,\, i \neq j} \sum \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}(\pi_{ij} - \pi_i \pi_j).$$

Note: Let A denote a set of indices (e.g., the set of population indices, U). The double sum notation

$$\sum_{i,j \in A} \sum y_{ij}$$

is our shorthand notation for

$$\sum_{i \in A} \sum_{j \in A} y_{ij} = \sum_{i \in A} y_{ii} + \sum_{i,j \in A,\, i \neq j} \sum y_{ij}.$$

An unbiased estimator of v_{HT} is given by

$$\hat{v}_{HT} = \sum_{i \in s} \frac{y_i^2}{\pi_i} \frac{(1 - \pi_i)}{\pi_i} + \sum_{i,j \in s,\, i \neq j} \sum \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}}.$$

The HT estimator is the only unbiased homogeneous linear estimator whose weights $w_i = 1/\pi_i$ do not depend on the sample. It performs poorly when n is not of fixed size (viz. unplanned domains). When n is fixed, the variance of the HT estimator features the following representation

$$v_{HT,2} = \sum_{i,j \in U,\, i < j} \sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

which is the basis of the unbiased variance estimator due to Sen, Yates and Grundy (SYG) [not shown]. In particular, the SYG-estimator is uniformly non-negative if $\pi_i \pi_j \geq \pi_{ij}$ for all $i \neq j$; the HT variance estimator, on the other hand, can take negative values. Despite its popularity, the HT estimator has some undesirable features. The estimator is scale invariant but not location invariant. In particular, the lack of location invariance restricts the number of practical situations in which the HT estimator is useful. If on the other hand it holds that $\pi_i \propto y_i$, the variance of the HT estimator is minimized. The attempt to relate the π_i's directly to the y_i's is ruled out because $p(s)$ is required to be non-informative (see Definition 2.2). If there exists an auxiliary variable x_i such that $x_i > 0$, $i \in U$, which is known to be positively correlated with y_i, we may take $\pi_i \propto x_i$ for all $i \in U$ (Särndal et al., 1992, chap. 2.8). Such a strategy proves to be favorable in this particular case. For different designs, estimators

other than the HT estimator are superior; see Chaudhuri and Stenger (2005, chs. 2 and 3).

Another important characteristic is the population y-mean, denoted by $\bar{y}_U$. If N is supposed known, then an obvious estimator of the population y-mean is given by the HT type estimator

$$\hat{\bar{y}}_{HT} = \frac{\hat{T}_y^{HT}}{N} = \frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i}.$$

Another candidate estimator for the population y-mean is the Hajek (HJ) estimator,

$$\hat{\bar{y}}_{HJ} = \frac{\hat{T}_y^{HT}}{\sum_{i\in s} 1/\pi_i} = \frac{\hat{T}_y^{HT}}{\hat{N}}, \quad \text{where } \hat{N} = \sum_{i\in s} 1/\pi_i.$$

The HJ estimator is often superior to the HT estimator in terms of variance even when N is known (although $\hat{\bar{y}}_{HJ}$ forgoes the information about N if N is in fact known; see Särndal et al., 1992, 182–83). Since the Hajek estimator is a ratio of two unknown totals, it is a non-linear statistic. Consequently, it is only approximately design unbiased and exact expressions for its variance do not exist (except in trivial cases). Therefore, one has to resort to approximate variance estimation; see Särndal et al. (1992, Result 5.7.1).

In line with estimation of the y-total at the population level, we shall denote the HT estimator of the y-total in domain d by

$$\hat{T}_{y,d}^{HT} = \sum_{i\in s_d}\frac{y_i}{\pi_i}, \qquad d = 1, \ldots, D.$$

Variance estimators for $\hat{T}_{y,d}^{HT}$ are similar to those for the population total and are therefore not discussed. The domain-specific (arithmetic) mean is defined as $\bar{y}_{U_d} = \sum_{i\in U_d} y_i/N_d$, where $N_d = \sum_{i\in U} \mathbb{1}\{i \in U_d\}$ is the size of domain d. The Hajek type domain-specific estimator is given by

$$\hat{\bar{y}}_d^{HJ} = \frac{1}{\hat{N}_d}\sum_{i\in s_d}\frac{y_i}{\pi},$$

where $\hat{N}_d = \sum_{i\in s_d} 1/\pi_i$ is an unbiased estimator of N_d. For variance estimation of the domain-specific Hajek estimator, approximate variance estimators are needed.

2.2.2. Finite-sample asymptotics

Exact distributional results exist only for a very limited number of strategies. For most estimator–design configurations, approximations based on large-sample theory are needed. In what follows, we assume that $p(s)$ satisfies Definition 2.2 and that certain additional condition are met (see below).

Our asymptotic framework is that of Isaki and Fuller (1982); see also Fuller

(2009, chap. 1.3). We define sequences that will permit us to establish large-sample properties of the estimators. Let U_t denote the tth population of size N_t and consider the nested sequence $\{U_t, t \geq 1\}$, with $t = 1, 2, \ldots$ such that $U_1 \subset U_2 \subset \ldots$. Associated with the ith element of the tth population is a tuple of vectors $(\mathbf{y}_{i,t},\ (\mathbf{x}_{1i,t}, \ldots, \mathbf{x}_{pi,t})^T)$, where $\mathbf{y}_{i,t}$ is the (vector-valued) variable of interest and the $\mathbf{x}_{1i,t}, \ldots, \mathbf{x}_{pi,t}$ are known, fixed numbers (auxiliary data).

Consider a sequence of samples s_t (of size n_t) is drawn from U_t according to some sequence of sampling designs $\{p(s_t), t \geq 1\}$. The first- and second-order sample inclusion probabilities are denoted by, respectively, $\pi_{i,t}$ and $\pi_{ij,t}$.

Assumption 2.1. *Let $p(s_t)$ be a sampling design that satisfies Definition 2.2 for all $t = 1, 2, \ldots$. In addition, the sampling design $p(s_t)$ must be such that*

(i) the samples are of fixed size n_t, $0 < n_t < N_t$ for all $t = 1, 2, \ldots$;

(ii) the samples are drawn using some without-replacement sampling mechanism.

The restriction to fixed-size sample designs is not consequential, but it simplifies notation considerably. Moreover, we assume that n_t grows as fast as N_t.

Assumption 2.2. *The (asymptotic) sampling fraction f satisfies*

$$\lim_{t\to\infty} \frac{n_t}{N_t} =: f \quad \textit{such that} \quad f \in (0, 1).$$

The sample sizes n_t form a sequence $\{n_t, t \geq 1\}$. Under Assumption 2.2, we have that $n_t \to \infty$ as $t \to \infty$ [cf. Assumption A5 in Breidt and Opsomer (2000) or the setup of Isaki and Fuller (1982)]. This implies that n_{N_t} is of the same (limiting) order as N_t, where $n_{N_t} = \lfloor f_{N_t} \rfloor$, i.e. largest integer less than or equal to fN_t; For ease of notation, we write n_t instead of n_{N_t}. Notable asymptotic setups which do not require n_t to grow as fast as N_t can be found in Robinson and Särndal (1983) and Fuller (2009, Thm. 1.3.6).

All *limiting processes* are taken as $t \to \infty$. It will prove useful to impose some regularity conditions on the asymptotic behavior of the sampling designs. To this end, we introduce different *sets* of assumptions.

Assumption 2.3 (Minimal)**.**

(i) There exist constants L_π and U_π independent of $\pi_{i,t}$ such that the first-order sample inclusion probabilities satisfy, for all t,

$$0 < \underline{\pi} \leq \pi_{i,t} \leq \overline{\pi} < 1 \qquad \forall i \in U_t,$$

(ii) and the second-order sample inclusion probabilities obey the inequality, for

all t,

$$\pi_{ij,t} \leq \pi_{i,t}\pi_{j,t} \quad \textit{for all } i,j \in U,\ i \neq j.$$

Assumption 2.4 (Weak; cf. Särndal and Robinson, 1983)**.**

(i) The first-order sample inclusion probabilities are such that

$$\lim_{t\to\infty} N_t \min_{1\leq i\leq N_t} \pi_{i,t} = \infty \quad \forall i \in U_t,$$

(ii) and the second-order sample inclusion probabilities are such that

$$\lim_{t\to\infty} \max_{1\leq i\neq j\leq N_t} \left| \frac{\pi_{ij,t}}{\pi_{i,t}\pi_{j,t}} - 1 \right| = 0 \quad \forall i,j \in U_t.$$

Assumption 2.5 (Strong; cf. Särndal and Robinson, 1983)**.**

(i) There exist positive real-valued constants P_1 and P_2 such that the first-order sample inclusion probabilities satisfy

$$\lim_{t\to\infty} \max_{1\leq i\leq N_t} \left(\frac{n_t}{N_t\pi_i} \right) \leq P_1 < \infty \quad \forall i \in U_t,$$

(ii) and the second-order sample inclusion probabilities are such that

$$\lim_{t\to\infty} n_t \max_{1\leq i\neq j\leq N_t} |\pi_{ij} - \pi_i\pi_j| \leq P_2 < \infty \quad \forall i,j \in U_t.$$

Remarks. (i) Assumption 2.4 can be regarded as a "standard" assumption in the literature; cf. Robinson and Särndal (1983), Breidt and Opsomer (2000) or – with slight modifications – Wu (2003). This Assumptions holds for both simple and simple stratified element (single-stage) sampling designs; cf. Breidt and Opsomer (2000). If in addition Assumption $\pi_{ij,t} \leq \pi_{i,t}\pi_{j,t}$ holds for all $i,j \in U_t$, $i \neq j$, (which is not uncommon, see below), then the asymptotic setup is appropriate for Poisson sampling, Midzuno's or Sampfort's drawing method and many other unequal probability designs; see Chauvet (2014) for further details.

(ii) The crucial "ingredient" of Assumption 2.4 is the hypothesis of part (ii), which is an asymptotic uncorrelatedness (negligibility) assumption. This is easily seen from the fact that, for any proper design, we have

$$\operatorname{cov}_p[\mathbb{1}_{i\in s_t}, \mathbb{1}_{j\in s_t}] = \pi_{ij,t} - \pi_{i,t}\pi_{j,t}, \qquad \text{for all } i,j \in U_t,\ i \neq j. \tag{2.4}$$

(iii) Assumption 2.3 is considerably weaker than Assumption 2.4 insofar that it does not impose asymptotic uncorrelatedness. Part (ii) of Assumption 2.3 requires only that the inequality $\pi_{ij,t} \leq \pi_{i,t}\pi_{j,t}$, $i \neq j$, is maintained as

$t \to \infty$. In fact this, together with our condition of fixed-sized sampling designs (cf. Assumption 2.1), implies that $\mathrm{cov}_p[\mathbb{1}_{i\in s_t}, \mathbb{1}_{j\in s_t}] \leq 0$ in (2.4) and is thus sufficient to ensure nonnegativeness of the variance of a total (cf. Sen–Yates–Grundy variance representation; Särndal et al., 1992, 47).

Let $\{X_n, n \geq 1\}$ denote a sequence of random variables. We say that the sequence of r.v. $\{X_n, n \geq 1\}$ converges *in probability* to the r.v. X, formally

$$X_n \xrightarrow{\text{p}} X, \tag{2.5}$$

if for every $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\big\{|X_n - X| > \epsilon\big\} = 0.$$

The we say that $\{X_n, n \geq 1\}$ converges *almost surely* (a.k.a. with probability one or almost everywhere) to the r.v. X, written as

$$X_n \xrightarrow{\text{a.s.}} X, \tag{2.6}$$

meaning that

$$\mathbb{P}\Big\{ \lim_{n\to\infty} X_n = X \Big\} = 1.$$

Note that the notions of convergence in (2.5) and (2.6) refer to, depending on the context, the sampling design or a model (or both). In order to avoid confusion, we will always indicate whether we refer to the design, the model or both.

Next, we introduce the notions of asymptotic design unbiasedness and design consistency for a sequence of estimators. To this end, suppose the asymptotic framework introduced above. Let $\{U_t, t \geq 1\}$ be the sequence of populations, where U_t is of size N_t. The associated limiting processes are obtained as $t \to \infty$ which implies $N_t \to \infty$. Let T_{N_t} be the y-total (we write T_{y,N_t} if we wish to highlight the dependency on y) and consider the sequence $\{T_{N_t}, t \geq 1\}$; likewise, denote by $\{\hat{T}_{n_t}, t \geq 1\}$ the sequence of estimators of the population y-total. With this, we define the following.

Definition 2.3. *An estimator $\hat{T}_{n_t}$ of the population total T_{N_t} (more precisely, a sequence $\{\hat{T}_{n_t}, t \geq 1\}$ of estimators and an associated sequence of sampling designs) is said to be*

(1) asymptotically design unbiased (ADU) for T_{N_t} if

$$\lim_{t\to\infty} \mathbb{E}_p \left[\frac{\hat{T}_{n_t} - T_{N_t}}{N_t} \right] = 0,$$

(2) (weakly) design consistent for T_{N_t} if

$$\lim_{t\to\infty} \mathbb{P}_p\big\{|\hat{T}_{n_t} - T_{N_t}| > \epsilon N_t\big\} = 0$$

for every $\epsilon > 0$;

(3) strongly design consistent for T_{N_t} if

$$\mathbb{P}_p\left\{\lim_{t\to\infty}\frac{\hat{T}_{n_t}}{N_t}=\frac{T_{N_t}}{N_t}\right\}=1.$$

In all three cases, it is assumed that $n_t \to \infty$ and $N_t \to \infty$ as $t \to \infty$ under the asymptotic framework.

Remark. The first part of the definition can be found in Fuller (2009, chap. 1.3.1). Robinson and Särndal (1983) and Särndal and Wright (1984) use similar but slightly different definitions of ADU-ness. In general, the sampling literature focuses on design consistency not strong consistency. In fact, the latter mode of consistency is mentioned in the handbook articles of Sen (1988) and Prášková and Sen (2009) but has not been worked out there. It is briefly illustrated in Sen and Singer (1993, Example 2.4.9) for simple random sampling without replacement (using martingale theory).

2.3. Robustness

In Chapter 1, we introduced the two major principles of robustness. Estimators and estimating strategies which satisfy these principles are considered robust with regard to deviations from the assumed core model. We will discuss the characteristics of such estimators in more detail for the class of M-estimators. Since M-estimators can be seen as a generalization of maximum likelihood estimators (m.l.e.), they form a natural class of estimators in the predominantly parametric estimation context of SAE models.[1]

Besides M-estimators, other classes of robust estimators are known. In the context of regression estimation, the best known estimator classes are L-estimators [linear functions of order statistics; see e.g. Rousseeuw and Leroy (1987, chap. 3), Jurečková and Sen (1996, chap. 4), and David and Nagaraja (2003, chap. 8)], R-estimators [rank-based statistics; see e.g. Hettmansperger and McKean (1998, chap. 3) or Jurečková and Sen (1996, chap. 6)], S-estimators [estimators on the basis of a robust scale; Rousseeuw and Yohai (1984)] etc. These classes will not be considered.

Subsequently, we introduce location M-estimators and discuss their robustness properties in the light of two important tools: influence function and breakdown point. We then broaden the scope of discussion and treat location-scale models and finally the regression case.

[1] Exceptions to parametric estimation are the papers of e.g. Chambers, Dorfman, and Werly (1993) on kernel-smoothing calibration estimators or Breidt and Opsomer (2000) on local polynomial regression estimation in survey sampling.

2.3.1. M-estimators

The roots of M-estimators and their defining ψ-functions lie in the parametric setup of Huber (1964). Denote by X_1 a random variable (r.v.) taking values in $\mathbb{R}$ with an everywhere positive, unimodal, and symmetric probability density function (p.d.f.) f_0 on $\mathbb{R}$. The corresponding cumulative distribution function (c.d.f.) is denoted by F_0. Let $\rho_0 = -\log f_0$, then the ψ-function is defined as

$$\psi_0(u) = \rho_0'(u) = \frac{\mathrm{d}}{\mathrm{d}u}\rho_0(u).$$

For the time being, we shall assume that ρ_0 is continuously differentiable for all $u \in \mathbb{R}$. For location estimators θ, it is sufficient – by equivariance considerations – to restrict attention to functions of the form $\psi(u, \theta) = \psi(u - \theta)$ (Godambe, 1960); see also Huber (1981, 45).

Suppose a sample of $n \geq 1$ independent and identically distributed (i.i.d.) random variables X_i, $i = 1, \ldots, n$, with p.d.f. $f_0 \equiv \phi$, ϕ denoting the p.d.f. of the standard normal distribution; the c.d.f. will be denoted by Φ. The realizations of the r.v. X_i are denoted by x_i. Note that under the Gaussian model, we have $\psi_0(u) = u$. If the X_i (for all $i = 1, \ldots, n$) are truly from $\mathcal{N}(0, 1)$, then the most efficient estimator of location is the maximum likelihood estimator, defined implicitly as the solution θ to the estimating equation

$$\sum_{i=1}^{n} \psi_0(x_i - \theta) = 0. \tag{2.7}$$

It follows immediately from (2.7) that θ is the (arithmetic) mean. However, the mean ceases to be the most efficient estimator if the p.d.f. of X_i has even slightly fatter tails than ϕ [but is still considered to be symmetric around the origin]. In such cases, alternative estimators posses far better efficiency and are considered superior. We illustrate this remarkable fact on the basis of the contaminated standard normal d.f. using the following contamination neighborhood or mixture distribution.

Definition 2.4 (Symmetry preserving ϵ-contamination neighborhood)**.** *Let $0 < \epsilon < 1$ be fixed and denote by H an arbitrary symmetric c.d.f. taking values in $\mathbb{R}$. The ϵ-contamination neighborhood of the standard Gaussian distribution is defined as*

$$F_\epsilon(x) = (1 - \epsilon)\Phi(x) + \epsilon H(x). \tag{2.8}$$

The constant ϵ captures the importance of contamination relative to the true distribution Φ. Note that the c.d.f. H is considered to be symmetric (i.e. $F(-x) = 1 - F(x)$ for all $x \geq 0$). A candiate c.d.f. H is e.g.

$$H(x) = \frac{1}{2}\delta_x + \frac{1}{2}\delta_{-x},$$

δ_x denoting the Dirac probability measure at location x (see e.g. Collins, 1977, 647). The above contamination neighborhood of the parametric model is called "gross-error model" in Huber (1964). The contamination model could be easily generalized to much more general neighborhoods in an abstract setting; see Huber (1981, chap. 2). The symmetry-preserving nature of the contamination model is of crucial importance as it implies that estimators of the form (2.7) with a monotone and odd ψ-function satisfy $\int \psi(x-\theta)\mathrm{d}F_\epsilon(x) = 0$ (see Fisher consistency, below), i.e. such estimators are asymptotically unbiased. Therefore, the optimal estimator can be picked among competing estimators on grounds of (asymptotic) variance considerations only. It is important to remark that M-estimator theory in general is not constrained to symmetric distributions; see e.g. Collins (1976) and in a more general context, Hampel, Ronchetti, Rousseeuw, and Stahel (1986, chap. 8.2a).

For the next step, it is fundamental to bear in mind that the amount of contamination ϵ is completely *unknown* to the statistician. If ϵ were known, we could infer an optimal estimator directly from the mixture distribution in (2.7). Huber (1964), however, proposed a versatile and generally applicable approach to the problem of not knowing ϵ. In place of deriving the ψ-function directly from the d.f. of the core model (as we did above with ψ_0), he suggested to turn towards an abstract class of candidate ψ-functions. In his setup for location estimators with known scale, Huber (1964, 76) assumes that the ρ-function is convex and therefore ψ is monotone non-decreasing. Location M-estimators with known scale, σ, (hence, we can take w.l.o.g. $\sigma = 1$) are defined like θ in (2.7) but with ψ_0 replaced by the particular ψ-function.

In order to restrict the influence of outliers, the candidate ψ-functions must be *bounded* functions $\psi : \mathbb{R} \to \mathbb{R}$. Boundedness is a necessary condition that follows from Principle 1 (qualitative robustness). This can be seen from the fact that Principle 1 is essentially a continuity principle in the following sense: The location estimator θ expressed as a functional $\theta(F)$ of the c.d.f. F, defined by

$$\int \psi\big(x - \theta(F)\big)\mathrm{d}F(x) = 0,$$

is weakly continuous [w.r.t. the weak(-star) topology in the space of probability measures] at F_0 for our target $\theta(F_0) = 0$ if and only if ψ is bounded (Huber, 1981, Thm. 2.6).

Influence function and qualitative robustness

Whether an estimator satisfies Principle 1, can be studied by means of the *influence function* (IF, Hampel, 1974, originally called influence curve). Let $T(F)$ be the estimator under consideration [this can be any estimator expressible as a functional, not only the location estimators considered so far], then the IF describes the effect of an *infinitesimal contamination* at the point x_0 on the es-

timate; x_0 being an element of the probability space under study.

Definition 2.5 (Influence function)**.** *Let $T(F)$ be any estimator expressible as a functional of the c.d.f. F. The influence function is defined (if it exits) as*

$$IF(T, x_0, F) = \lim_{\epsilon \downarrow 0} \frac{T\big((1-\epsilon)F + \epsilon\delta_{x_0}\big) - T(F)}{\epsilon}, \tag{2.9}$$

where δ_{x_0} is a point mass at x_0 which is assumed to be defined on the same probability space as F.

Under appropriate regularity conditions, the IF can be obtained as a Gâteaux derivative of T (Hampel et al., 1986, 84–5). Though, the IF is commonly used as a heuristic tool without bothering about regularity conditions. If the IF of $T(F)$ is a bounded function, then the estimator is said to be *qualitatively robust* and therefore satisfies Principle 1. For M-estimators of location, the IF is *proportional* to its ψ-function (Huber, 1981, 45). The proportionality often holds for more complex M-estimators (e.g. estimators of multivariate location and scatter, see Hampel et al., 1986, 283).

The influence function can be considered as a "limit version" of the sensitivity curve (SC) (Maronna, Martin, and Yohai, 2006, 55). The latter displays the (standardized) effect on an estimator based on the sample $\{x_1, \ldots, x_n\}$ when we add the additional observation x_0 to the sample. The SC is useful when studying finite-sample behavior. Hulliger (1995) extended the notion of the SC to the finite-population sampling context; see also Hulliger (1991, chap. 4.4.2) for more details. A related but different representation of the finite-sampling influence function is given by Deville (1999).

Asymptotic minimax optimality

The great achievement of Huber (1964) was to prove that the M-estimator defined by the Huber ψ-function ψ_k (see below) is *optimal in an asymptotic minimax sense*. That is, he showed that the location M-estimator with ψ-function given by (where k is a constant such that $k \geq 0$)

$$\psi_k(u) = \begin{cases} u & \text{if } |u| \leq k, \\ k \operatorname{sgn}(u) & \text{otherwise}, \end{cases}$$

minimizes the maximal asymptotic variance of the estimator over the family of *symmetric* ϵ-contamination neighborhoods F_ϵ, where F_ϵ is specified in Definition 2.4. Figure 2.1 shows a display of the Huber ψ-function together with the corresponding ψ'- and w-functions, where $w(u) = \psi(u)/u$. The function ψ_k is directly tied to the least favorable distribution F^* (i.e. the d.f. minimizing the Fisher information of the location estimator T over all candidate d.f. in F_ϵ), which is a smooth blend of the Gaussian and the Laplace (a.k.a. double exponential) dis-

tribution. The negative logarithm of the p.d.f. of the least favorable distribution is usually defined as the ρ-function, given by

$$\rho_k(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \leq k, \\ k|u| - \frac{1}{2}k^2 & \text{otherwise.} \end{cases}$$

Huber's original argument is in fact more general than the one we presented. He studied symmetric contamination neighborhoods for general (symmetric) core models with c.d.f., say G (with p.d.f. g) instead of Φ, the c.d.f of the standardized Gaussian. Under his generalized assumptions, the ψ-function is given by (omitting the index k)

$$\psi_g(u) = \begin{cases} -g'(u)/g(u) & \text{if } |u| \leq k, \\ k\,\mathrm{sgn}(u) & \text{otherwise.} \end{cases}$$

A considerably less well known part of Huber's seminal 1964 paper is concerned with asymptotic bias; see Huber (1964, sec. 7). There, Huber proves that among all translation equivariant estimators of location, the sample median minimizes the maximum asymptotic bias over the general contamination neighborhood $(1-\epsilon)\Phi(x) + \epsilon H(x)$, where H is an arbitrary (not necessarily symmetric) c.d.f.

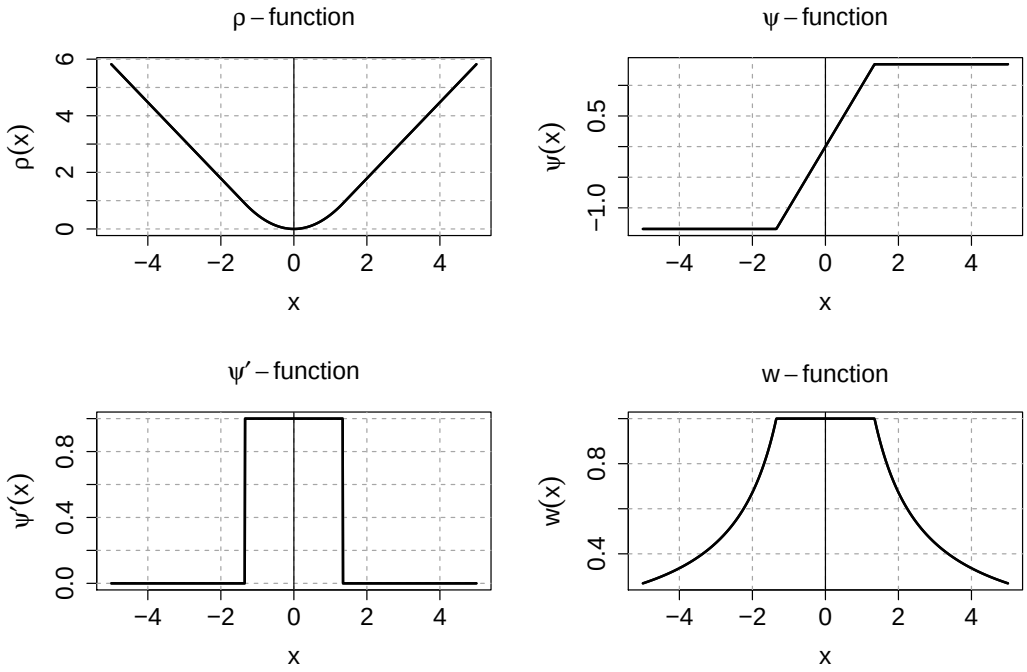


Figure 2.1.: Huber ψ-, ρ-, ψ'-, and w-function with tuning constant $k = 1.345$.

The location-scale family

Suppose a parametric model $\{F_\theta, \theta \in \Theta\}$ indexed by the parameter θ (for ease

of discussion, the parameter space Θ shall be taken equal to $\mathbb{R}$). An estimator T of θ is called Fisher consistent for θ if, at an exact model distribution F_θ, T gives always the correct answer θ. Formally, we have.

Definition 2.6 (Fisher consistency)**.** *An estimator T of θ is said to be Fisher consistent at F_θ if $T(F_\theta) = \theta$ for all $\theta \in \Theta$.*

So far, we have considered only location estimators and assumed that the scale $\sigma = 1$ is known. M-estimators of location are *not scale equivariant* (Hampel et al., 1986, 105), which implies that the estimators may depend heavily on measurement units. This can be fixed by considering

$$\sum_{i=1}^{n} \psi_k \left(\frac{x_i - \theta}{\sigma} \right) = 0$$

instead of (2.7); the term $(x_i - \theta)/\sigma$ is called studentized residual. When σ is unknown, we may apply the estimator of Proposal 2 in Huber (1964, sec. 11), i.e. compute θ and σ simultaneously, where σ is defined by an M-estimator of scale, which is the solution $\hat{\sigma}$ to

$$\sum_{i=1}^{n} \psi_k^2 \left(\frac{x_i - \theta}{\sigma} \right) = n\kappa,$$

where κ is a correction term to ensure that the estimator of σ is a Fisher consistent estimator at the (Gaussian) core model; κ is determined by $\kappa = \mathbb{E}\psi_k^2(Z)$, where $Z \sim \mathcal{N}(0,1)$ and $\mathbb{E}$ denotes expectation w.r.t. $\mathcal{N}(0,1)$. The simultaneous estimation of θ and σ is not imperative; on the contrary, Hampel et al. (1986, 105) provide evidence that the location M-estimator with predetermined scale is often superior to the joint solution for (θ, σ). Andrews, Bickel, Hampel, Huber, Rogers, and Tukey (1972, 239) recommend (as result of their large simulation study) the median absolute deviation of the median (MAD; with Fisher-consistency correction) as a preliminary scale estimator. Alternative methods are discussed in Maronna et al. (2006, chap. 2.6).

It is well known that the m.l.e. of σ^2 at the Gaussian model, defined as $\hat{\sigma}^2_{mle} = (1/n)\sum_{i \le n}(x_i - \bar{x}_i)^2$, where $\bar{x}$ is the sample mean, has a bias of $-\sigma^2/n$. Consequently, a bias corrected estimator obtains when the normalization $1/n$ is replaced by $1/(n-1)$. What is perhaps less well known is that for the Gaussian distribution, the m.l.e. of scale, $\hat{\sigma} = \sqrt{\hat{\sigma}^2_{mle}}$, has a bias of $\mathbb{E}[\sigma - \hat{\sigma}] = -3/4n$. A bias-corrected estimator is thus $n/(n - 3/4)\hat{\sigma}$. Since the robust estimator of σ suffers from the same problem, Clarke and Milne (2004) derive a first-order small sample bias correction for the scale estimator in Huber's Proposal 2.

Breakdown point and quantitative robustness

Principle 2 states that a somewhat larger deviation from the model should not

cause a catastrophe, i.e. the estimator should not break down. Hampel (1968) introduced the concept of breakdown point (BP) for an estimator of parameter $\theta \in \Theta$, which is – roughly speaking – the largest amount of contamination (i.e., proportion of atypical points) that the data may contain such that the estimator still gives some information about θ.

A crucial strength of the BP is that it is defined without recourse to a probabilistic model. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a sample of fixed size n taking values in some Euclidean space. We may think of contaminating the sample in many ways; Donoho and Huber (1983, 160) discuss three important cases. For highly structured problems e.g., multilevel level models), we focus on the *ϵ-replacement BP* which is defined by replacing an arbitrary subset of size $m < n$ of the sample by arbitrary values $x_1^\circ, \ldots, x_m^\circ$.[2] The fraction of "bad" values in the corrupted sample $\mathcal{X}^c$ is $\epsilon = m/n$, which implies that $\epsilon \in [0, 1)$.

We shall consider estimation of location first. Let $\theta = \{\hat{\theta}\}_{n=1,2,\ldots}$ be an estimator of location (possibly multivariate), and denote by $\hat{\theta}(\mathcal{X})$ its value at the sample $\mathcal{X}$. Define the maximum bias that can be caused by ε-contamination as

$$\text{bias}(\varepsilon; \mathcal{X}, \hat{\theta}) = \sup \left| \hat{\theta}(\mathcal{X}^c) - \hat{\theta}(\mathcal{X}) \right|,$$

where the supremum is taken over the set of all ε-corrupted samples. Donoho and Huber (1983, 161–2) define the finite sample BP (FBP), $\epsilon^*(\mathcal{X}, \hat{\theta})$, of an estimator of location $\hat{\theta}$ at $\mathcal{X}$ as follows.

Definition 2.7 (Finite sample replacement BP of an estimator of location)**.** *The finite sample ϵ-replacement breakdown point of the estimator of location $\hat{\theta}$ at $\mathcal{X}$ is defined as*

$$\varepsilon^*(\mathcal{X}, \hat{\theta}) = \inf \left\{ \varepsilon : \text{bias}(\varepsilon; \mathcal{X}, \hat{\theta}) = \infty \right\},$$

where ε^* is the smallest ε for which the estimator when applied to the ε-corrupted sample $\mathcal{X}^c$ can take values arbitrary far from $\hat{\theta}(\mathcal{X})$.

Definition 2.7 makes sense only for estimators $\hat{\theta}$ taking values in unbounded Euclidean space. Notably, for scale estimators $\hat{\sigma}$, which are inherently nonnegative, it makes sense to say that $\hat{\sigma}$ has broken down if contamination may either drive it to 0 or ∞ (called "implosion" and "explosion" in Huber, 1981, chap. 6.6). For scale estimators, Definition 2.7 must therefore be slightly altered to be meaningful. Insofar, the FBP shall be defined as the least fraction of contamination that can send the value of the estimator to the boundary of the parameter space (see Donoho and Huber, 1983, 161–2 and 168), which is formally defined (making use of $\log$) as follows.

[2] When studying the BP of location estimators, one may also consider ϵ-contamination (instead of ϵ-replacement), i.e., the case a certain proportion of the data has been contaminated. By the invariance of location estimators under arbitrary permutations of the (unstructured) observations, ϵ-contamination does not pose any difficulties. However, as Donoho and Huber (1983, 165) point out, ϵ-contamination is "distinctly awkward in structured problems".

Definition 2.8 (Finite sample replacement BP of an estimator of scale)**.** *The finite sample ϵ-replacement breakdown point of the estimator of scale $\hat{\sigma}$ at $\mathcal{X}$ is defined as*

$$\varepsilon^*(\mathcal{X}, \hat{\sigma}) = \inf\left\{\varepsilon : \left|\log\left(\hat{\sigma}(\mathcal{X}^c)\right)\right| = \infty\right\},$$

where ε^* is the smallest ε for which the estimator when applied to the ε-corrupted sample $\mathcal{X}^c$, can take values outside the parameter space $(0, \infty)$.

For M-estimators of location with monotone, bounded ψ-functions and known scale, the FBP equals $(n-1)/2n$, which is roughly 0.5 for large n (Maronna et al., 2006, 61). In view of an FBP of almost 50%, we may conclude that the notion of breakdown point is barely relevant for all practical reasons since empirical data virtually never consists of 50% bad points. However, for more complex models, the FBP can become dangerously low. This is demonstrated for the one-way random effect model by Wellmann (1994, 2000), which is of particular relevance since this type of model is a workhorse in SAE. For more complex models, breakdown point consideration are much more involved than in the location or regression case (cf. Davies and Gather, 2005, 2007).

Redescending ψ-functions

M-estimators with monotone ψ-functions have the advantage that they are uniquely defined. Also, by monotonicity of ψ, any outlying observation contributes to the estimator, irrespective of how far from the bulk of data such an observation lies. This property can be disadvantageous in the case of very fat-tailed distributions and it would instead be beneficial in terms of efficiency to discard extreme outlying observations from the computation completely or at least to reduce their impact considerably. This is in fact equivalent to how the ψ-function associated with the Student t-distribution behaves, that is $\psi(u) = 0$ if $|u| \leq k$ for some $k < \infty$, which is said to redescend. In this spirit, F. Hampel defined the three-part *redescending* ψ-function (see Andrews et al., 1972, sec. 2C3), which vanishes outside $[-r, r]$, where $r \in \mathbb{R}^+$ can be chosen. Another redescender is the Tukey bisquare [or biweight] function (see Figure 2.2), whose ρ-function is (see Hampel et al., 1986, chap. 2.6)

$$\rho_b(u) = \begin{cases} 1 - \left(1 - (u/b)^2\right)^3 & \text{if } |u| \leq b, \\ 1 & \text{otherwise}, \end{cases}$$

and taking the derivative w.r.t. u, we get

$$\psi_b(u) = \begin{cases} x\left(1 - (u/b)^2\right)^2 & \text{if } |u| \leq b, \\ 0 & \text{otherwise}. \end{cases}$$

Note that ψ_b is everywhere differentiable and vanishes outside of $[-b, b]$. The tuning constant b for attaining 95% efficiency of the estimator of location at

the Gaussian model is equal to 4.685. By their non-monotone nature, estimators based on redescending ψ-functions are not uniquely defined and may have multiple solutions to the estimating equations. Also, redescenders are in general much more sensitive to (wrong) scaling than monotone ψ-functions (Huber, 1981, 102–3).

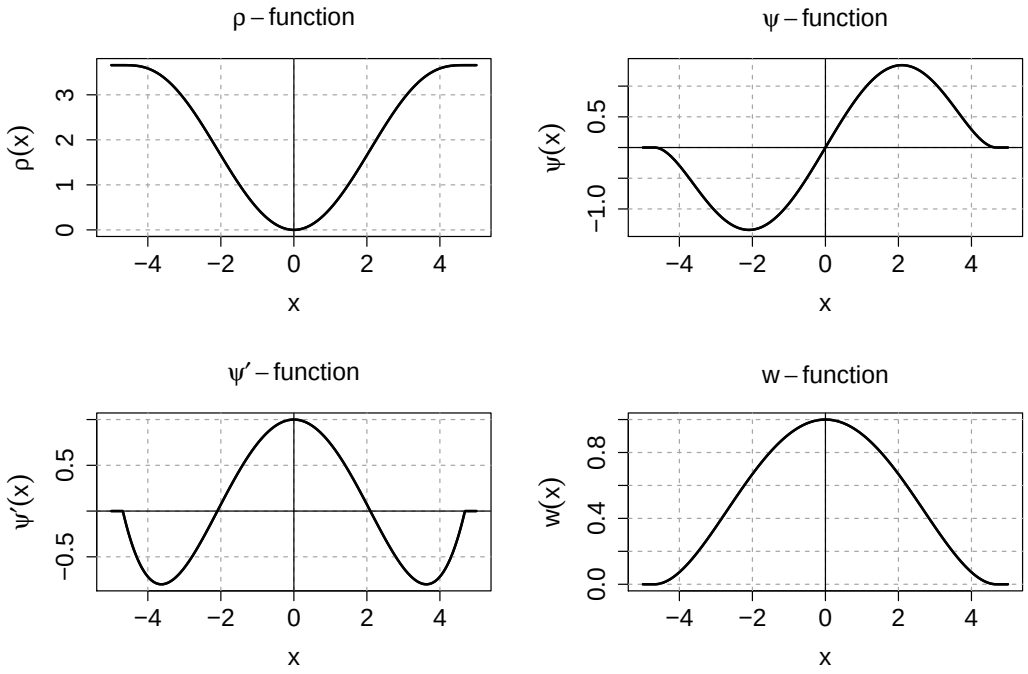


Figure 2.2.: Tukey bisquare ψ-, ρ-, ψ'-, and w-function with tuning constant $b = 4.685$.

Regression M-estimators

Let $\mathbf{x}_i$, $i = 1, \ldots, n$, denote p-vectors of regressors and define the matrix $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$, which is assumed to be of full rank; the study variable y_i is supported on $\mathbb{R}$ with realizations y_i, $i = 1, \ldots, n$. Suppose the linear model $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is unknown and the e_i are i.i.d. errors with expectation zero and constant but unknown standard deviation σ. The d.f. of the e_i is supposed to be fat-tailed. Consequently, the goal is to estimate $\boldsymbol{\beta}$ and σ robustly. For the further course of discussion, we have to distinguish two cases depending on the nature of the $\mathbf{x}_i$'s: either the $\mathbf{x}_i$'s are regarded (i) as fixed regression carriers / constants (cf. Huber, 1983) or (ii) as random variables.

- When the $\mathbf{x}_i$'s are regarded as constants, the regression M-estimator with monotone ψ-function is a straightforward extension of the M-estimators we studied in the location-scale case; see Maronna et al. (2006, chap. 4).
- In case (ii), outliers may occur also in $\mathbf{x}_i$ (in addition to Y_i) and operate as *leverage points* h_i of $\mathbf{X}$, where $\{h_1, \ldots, h_p\} = \operatorname{diag}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$. In

particular, a small number of atypical $(\mathbf{x}_i, y_i)$ pairs with high leverage can exert a heavy influence on any monotone M-estimator and completely distort the estimate. There are essentially two classes of M-estimators that are resistant to such outlying pairs: (1) M-estimators with redescending ψ-functions (see Maronna et al., 2006, chap. 5.3) and (2) *generalized M-estimators* (GM). We shall only look at the latter class of estimators.[3]

Generalized M-estimators follow actually the simplest way to fix the weakness of monotone M-estimators when outlying $(\mathbf{x}_i, y_i)$ pairs occur (Maronna et al., 2006, chap. 5.11). They simply downweight the influential $\mathbf{x}_i$ and prevent that such $\mathbf{x}_i$ dominate the estimating equation. More specifically, define a weight function $\omega : \mathbb{R}^+ \to \mathbb{R}^+$; then, the estimating equation for $\boldsymbol{\beta}$ writes

$$\sum_{i=1}^{n} \omega\big(d(\mathbf{x}_i)\big)\psi\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\mathbf{x}_i = \mathbf{0}, \tag{2.10}$$

where $d : \mathbb{R}^p \to \mathbb{R}^+$ is a distance function (e.g. robust Mahalanobis distance). The unknown scale σ can be estimated in a similar fashion as proposed in the location-scale model.

We may let the weight ω depend on the residuals, $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$, as well as the predictor variables. To this end, suppose a function $\eta : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}$ given by $\eta(\sigma, r)$ such that $\eta(\sigma, \cdot)$ is continuous in σ on $\mathbb{R}^+$ and, for all $\sigma \in \mathbb{R}^+$, $\eta(\cdot, r)$ is a nondecreasing and bounded function in r. Estimators on the basis of η are called *generalized M-estimators* (GM) for $\boldsymbol{\beta}$ and take the form (Hampel et al., 1986, chap. 6.3)

$$\sum_{i=1}^{n} \eta\left(d(\mathbf{x}_i),\ \frac{r_i(\boldsymbol{\beta})}{\sigma}\right)\mathbf{x}_i = \mathbf{0}.$$

Two particular forms of GM-estimators have been of primary interest in the literature:

(i) Mallows-type, where $\eta(\sigma, r) = \omega(\sigma)\psi(r)$ and

(ii) Hampel–Krasker–Welsch-type with $\eta(\sigma, r) = \psi(\sigma r)/\sigma$.

The Mallows-type coincides with (2.10) and is far more common in practical applications. GM-estimators have some attractive properties (cf. Maronna et al., 2006, chap. 5.11), namely, (i) their influence function is bounded if $\eta(\sigma, r) \cdot \sigma$ is bounded. (ii) GM-estimators have an asymptotic normal distribution under mild regularity conditions. (iii) GM-estimators can be computed like ordinary monotone M-estimates. However, they also have some drawbacks. (a) Their efficiency depends on the distribution of the $\mathbf{x}_i$. In case of fat-tailed distributions, GM-estimators cannot be simultaneously robust and very efficient. (b)

[3] Besides redescending and generalized M-estimators, also MM-estimators can deal with leverage points; see Maronna et al. (see e.g. 2006, chap. 5).

The FBP is less than 0.5 and will be quite low for large p; see Maronna et al. (2006, 149–50) for a more detailed discussion.

2.3.2. Robustness in finite population sampling

Robustness in the context of finite-population sampling differs from our discussion of robustness so far. The reason is that the *variance–bias* trade-off achievable with robust estimators is of much greater importance in finite-population sampling (and is also of different nature than encountered in "classical" statistics). This will become apparent once we have introduced the notions of *representative* and *non-representative* outliers due to Chambers (1986). To this end, we shall suppose a sample of elements $\{y_1, \ldots, y_n\}$ from a finite population U. Some of the y_i's are recognized as being exceedingly large compared with the bulk of data. We shall call those observations outliers. The question is whether these outliers are correct observations or not, and whether we should use them in inference on the population or not. Now, either an outlier may be a correct (but influential) observation from the target population (called representative outlier) or it may be an incorrect observation, for instance, due to coding errors or because it is an element from outside the target population (called non-representative). Clearly, keeping a non-representative outlier quasi "untreated" in the sample and apply non-robust methods will lead to biased estimates and an inflated variance of the estimator. The case of representative outliers is more intricate. Discarding a representative outlier (i.e. a correct but atypical observation) leads to biased estimates. Keeping it, however, "untreated" will inflate the estimators' variance since – as a rule – the outlier would show up only in a few of the possible samples. This reasoning implies trade-off a between bias and variance, and the trade-off is particularly accentuated under asymmetric or heavy- tailed distributions; see also Hulliger (1995).

Robust procedures (e.g. M-estimators) may yield efficient estimates (in terms of mean squared error, MSE) in the presence of representative outliers. However, this efficiency is obtained through the introduction of some bias (via using an inconsistent estimator). When the sample size is small and whence, as a rule, the variance is the dominant factor in the MSE, small biases introduced through robustification can be worthwhile if the variance can in turn be significantly reduced; see e.g. Chambers (1986), Lee (1991) and Hulliger (1991, 1995). However, in some cases the bias can be substantial and may render the robust procedure grossly inefficient. This happens all the more as the sample size grows because the variance decreases, but the bias typically does not. As a

result, the bias tends to dominate the MSE. Such troublesome situations have been of great concern to advocates of robust methods. Therefore, Chambers (1986) proposed a robust ratio M-estimator in which the incurred bias is estimated by a robust technique and then "added back" to some extent to the robust estimator; the resulting estimator is called bias-corrected; see also Welsh and Ronchetti (1998) who suggested a modification of Chamber's bias-corrected estimator. Hulliger (1991, 1995) suggested another solution to the inconsistency problem, insofar that he considers a set of eligible estimators which includes the non-robust, but consistent estimator (i.e. the Horvitz–Thompson estimator); the method is called minimum estimated risk estimator and is an adaptive procedure. The key to his proposal is the allowance for the consistent estimator as this choice ensures that the overall procedure is consistent.

2.4. Bayesian statistics

In this section, we shall focus on the family of parametric models $\mathcal{F} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, specified in terms of the c.d.f. F (we write $F_{\boldsymbol{\theta}}$ or $F(\mathbf{y} \,|\, \boldsymbol{\theta})$ if we wish to highlight the dependency on $\boldsymbol{\theta}$), where $\boldsymbol{\theta} \in \Theta$ is a parameter and Θ will be a subset of $\mathbb{R}^p$ [the sample space is a subset of $\mathbb{R}^n$]. The family $\mathcal{F}$ is assumed to consist of a set of absolutely continuous c.d.f.'s F; this implies that any $F \in \mathcal{F}$ has a density f [respectively, $f_{\boldsymbol{\theta}}$ or $f(\mathbf{y} \,|\, \boldsymbol{\theta})$] w.r.t. the Lebesgue measure. We write $(\Omega, \mathcal{B}, F_{\boldsymbol{\theta}})$ to mean the probability space under consideration, where Ω is the sample space and $\mathcal{B}$ is a sigma algebra over over the set Ω. We use the qualifier "sampling" together with the term model (i.e., "sampling model") when we wish to point out that "model" refers to the data generating model/ mechanism, not the prior distribution (model).

We shall adopt a Bayesian point of view and introduce to this end the notion of a decision rule. In very general terms, a decision rule is a function which maps an observation to an appropriate action. An estimator is regarded as a decision rule used for estimating a parameter. More formally, the notion of a (deterministic) decision rule is given as follows (Berger, 1985, chap. 1.3.2).

Definition 2.9 (Decision rule)**.** *Let* $\mathbf{Y}$ *denote a r.v. defined on* $(\Omega, \mathcal{B}, F_{\boldsymbol{\theta}})$*,* $\boldsymbol{\theta} \in \Theta$*. Let* A *be a set of possible actions, then a decision rule* $\boldsymbol{\delta}$ *is a function* $\boldsymbol{\delta} : \Omega \to A$*.*

Remarks. (i) The full generality of the definition will typically not be exploited. For our purposes it will prove sufficient to take A equal to the parameter space Θ as our major interest lies in estimating the unknown $\boldsymbol{\theta}$. Moreover,

if $\mathbf{y}$ is an observed realization of the r.v. $\mathbf{Y}$, then $\boldsymbol{\delta}(\mathbf{y})$ is the action that will take place, i.e. the estimator under consideration.

(ii) We will always assume that the rules (functions) are appropriately measurable. We write δ and $\boldsymbol{\delta}$ to mean, respectively, a univariate and a n-variate decision rule.

(iii) It will (later) proof useful to study whole classes, say $\mathcal{D}$, of eligible rules. For instance, it may be sensible to define a class of rules such that all its elements δ have finite expected loss, i.e. $R(\theta, \delta) < \infty$ for all $\theta \in \mathbb{R}$ [see below].

A frequentist decision-theorist seeks to evaluate, for each $\boldsymbol{\theta} \in \Theta$, how much s/he would "expect" to lose if s/he had used $\boldsymbol{\delta}(\mathbf{Y})$ repeatedly with varying $\mathbf{Y}$; the evaluation tool to do so is called risk function. In order to give the definition of the risk function, we first have to introduce the notion of a loss function. Let $\mathcal{D}$ be a sufficiently well-behaved class of eligible decision rules (without going into the details). A loss function for estimating $\boldsymbol{\theta} \in \Theta$ by rule $\boldsymbol{\delta} \in \mathcal{D}$ is a real-valued non-negative function $L : \Theta \times \mathcal{D} \to [0, \infty)$ that satisfies the condition $L\big(\boldsymbol{\theta}, g(\boldsymbol{\theta})\big) = 0$ for all $\boldsymbol{\theta} \in \Theta$ and any real-valued function $g : \mathbb{R}^n \to \mathbb{R}^n$. The value of $L\big(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{y})\big)$ is the "cost" of action $\boldsymbol{\delta}(\mathbf{y})$ under parameter $\boldsymbol{\theta}$. A common loss function is the sum of squared error loss given by

$$L(\boldsymbol{\theta}, u) = \|\boldsymbol{\theta} - u\|_2^2,$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Alternatively, we may define *average* squared error loss. The accuracy, or rather inaccuracy, of a rule $\boldsymbol{\delta}$ is measured by the (frequentist) risk function (see e.g. Berger, 1985, chap. 1.3).

Definition 2.10 (Frequentist risk / expected loss)**.** *The risk (function) of decision rule $\boldsymbol{\delta} \in \mathcal{D}$ for estimating $\boldsymbol{\theta}$ under sampling model $F(\mathbf{y} \,|\, \boldsymbol{\theta})$ and loss function $L(\boldsymbol{\theta}, \cdot)$ is defined as*

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) = \mathbb{E}_{\boldsymbol{\theta}}\big\{L\big(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{y})\big)\big\} \equiv \int_{\Omega} L\big(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{y})\big) \mathrm{d}F(\mathbf{y} \mid \boldsymbol{\theta}).$$

Clearly, since $F_{\boldsymbol{\theta}}$ is supposed to have a p.d.f., we may define $R(\boldsymbol{\theta}, \boldsymbol{\delta})$ in terms of the p.d.f. $f_{\boldsymbol{\theta}}$. Note that $\mathbb{E}_{\boldsymbol{\theta}}$ means expectation w.r.t. sampling model $F_{\boldsymbol{\theta}}$ conditional on $\boldsymbol{\theta}$. Our definition of the risk function is not mathematically rigorous, but we stick with this definition as we are primarily interested in the underlying ideas, not regularity conditions. We shall assume that the class of rules $\mathcal{D}$ and the parametric model $F_{\boldsymbol{\theta}}$ are such that $R(\boldsymbol{\theta}, \boldsymbol{\delta})$ is well-defined.

Of course, one would like to find a $\boldsymbol{\delta}$ in class $\mathcal{D}$ that minimizes the risk value for all $\boldsymbol{\theta} \in \Theta$. Though, except in trivial cases, there exists no uniformly best estimator, i.e. no estimator which simultaneously minimizes the risk for all values of $\boldsymbol{\theta}$ (Lehmann and Casella, 1998, 5). One potential way to deal with the non-existence of such optimal estimator is to restrict the class of estimators (i.e., enforce certain conditions, e.g. require unbiasedness). Another type of restriction that could be imposed on the class of estimators are equivalence relations or conditions, which are quite natural when symmetries are present in the problem; see Lehmann and Casella (1998, chap. 3). However, in many important situations the application of unbiasedness or equivalence restrictions is rather limited. There exists an alternative approach: Instead of seeking an uniformly minimum risk estimator, one may require that the risk has to be low in *some overall sense*. Two natural global measures of the size of the risk are the average,

$$\int_{\Theta} R(\boldsymbol{\theta}, \boldsymbol{\delta}) \omega(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \tag{2.11}$$

for some "weight" function $\omega(\cdot)$, and the maximum of the risk function,

$$\sup_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \boldsymbol{\delta}). \tag{2.12}$$

The estimator minimizing (2.11) leads to the *Bayes estimator* [see below] if $\omega(\cdot)$ is taken to be equal to the *prior* probability density function (p.d.f.) $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{\eta})$ of the r.v. $\boldsymbol{\Theta}$ (π is indexed by the parameter $\boldsymbol{\eta} \in \mathbb{R}^m$). Observe that $\boldsymbol{\Theta}$ is regarded as a r.v. with realization $\boldsymbol{\theta}$. In passing, we note that minimization of (2.12) leads to the *minimax* estimator. Under prior p.d.f. π, we define the Bayes risk function as follows.

Definition 2.11 (Bayes risk)**.** *The Bayes risk of a decision rule $\boldsymbol{\delta}$ for estimating $\boldsymbol{\theta} \in \Theta$ versus prior $\pi(\boldsymbol{\theta} \mid \boldsymbol{\eta})$ is defined as*

$$r(\boldsymbol{\delta}, \pi) = \mathbb{E}_{\pi}\big\{R(\boldsymbol{\theta}, \boldsymbol{\delta})\big\} \equiv \int_{\Theta} R(\boldsymbol{\theta}, \boldsymbol{\delta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \mathrm{d}\boldsymbol{\theta}.$$

We have implicitly assumed that $\boldsymbol{\delta}$, π, and R are such that the Bayes risk is well-defined; for a discussion of regularity conditions, we refer to Berger (1985, chap. 1). Notice that $\mathbb{E}_{\pi}$ denotes expectation w.r.t. the prior p.d.f. π. Equipped with this definition, we obtain a formal definition of the Bayes estimator (see e.g. Lehmann and Casella, 1998, 225).

Definition 2.12 (Bayes estimator)**.** *A Bayes estimator $\boldsymbol{\delta}^*$ versus prior p.d.f. π is the minimizer of the Bayes risk $r(\cdot, \pi)$.*

From this definition it is clear that a Bayes estimator is specific to the choice of prior distribution. Moreover, it is sometimes easier to work with the *posterior* p.d.f. ϱ, which is given by

$$\varrho(\boldsymbol{\theta} \,|\, \mathbf{y}, \boldsymbol{\eta}) = \frac{f(\mathbf{y} \,|\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \,|\, \boldsymbol{\eta})}{m(\mathbf{y} \,|\, \boldsymbol{\eta})}$$

where

$$m(\mathbf{y} \,|\, \boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{\theta}}\big\{ f(\mathbf{y} \,|\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \,|\, \boldsymbol{\eta}) \big\} \equiv \int_{\Theta} f(\mathbf{y} \,|\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \,|\, \boldsymbol{\eta}) \mathrm{d}\boldsymbol{\theta}$$

is the *marginal* p.d.f. The level of mathematical rigor taken in this disposition is sufficient for our purposes; the interested reader will find technically more rigorous approaches to Bayesian statistics in the cited literature.

2.5. Mixed linear models

A large class of models in small-area estimation can be regarded as a mixed linear model (MLM) with block-diagonal covariance matrix (also known as linear mixed model, see e.g., Searle, Casella, and McCulloch, 1992; Demidenko, 2004) of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{t=1}^{c-1} \mathbf{Z}_t \mathbf{v}_t + \mathbf{e},$$

where

(i) $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_g)^T$ is a $(n \times 1)$ vector of observations, where the $\mathbf{y}_i$'s ($i = 1, \ldots, g$) denote group- or area-specific subvectors,

(ii) $\mathbf{X}$ and $\mathbf{Z}_t$ are known $(n \times q)$ and $(n \times p_t)$ matrices of full rank, respectively, $t = 1, \ldots, c-1$ ($\mathbf{X}$ and the $\mathbf{Z}_t$'s are also composed of the $i = 1, \ldots, g$ area-specific submatrices),

(iii) $\boldsymbol{\beta}$ is a q-vector of unknown parameters (fixed effects),

(iv) each of the $\mathbf{v}_t$ denotes a $(1 \times p_t)$ vector of unobserved random effects, $t = 1, \ldots, c-1$,

(v) $\mathbf{e} = (\mathbf{e}_1, \ldots, \mathbf{e}_g)^T$ is a $(n \times 1)$ vector of unobserved errors.

In addition, it is assumed that the p_t levels of each random effect $\mathbf{v}_t$ are independently distributed with zero mean and (constant) variance σ_t^2; also, $\mathbf{v}_1, \ldots, \mathbf{v}_t$ and $\mathbf{e}$ are assumed to be independent. Hence, it follows that

$$\mathbb{E}[\mathbf{y} \mid \mathbf{X}] = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \mathbb{V}[\mathbf{y} \mid \mathbf{X}] = \mathbf{V} = \theta_c \mathbf{I}_n + \sum_{t=1}^{t-1} \boldsymbol{\theta}_t \mathbf{Z}_t \mathbf{Z}_t^T,$$

where $\boldsymbol{\theta} = (\sigma_1^2, \ldots, \sigma_c^2)^T$ and $\mathbf{I}_n$ is the $(n \times n)$ identity matrix. A commonly used alternative expression of the above model can be obtained (cf. Searle et al., 1992,

p. 171) if we put $\mathbf{Z}_c = \mathbf{I}_n$ and $\mathbf{v}_c = \mathbf{e}$; hence,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{t=1}^{c} \mathbf{Z}_t \mathbf{v}_t, \quad \text{and} \quad \mathbf{V} = \sum_{t=1}^{c} \sigma_t^2 \mathbf{Z}_t \mathbf{Z}_t^T.$$

The above model description provides only a rough summary; for a discussion of the more intricate details, we refer the reader to e.g. Searle et al. (1992, ch. 6.1). The significance of MLM's in the SAE context is the fact that the basic area- and unit-level models (and several extensions of those models) can be seen as MLM's with block-diagonal covariance structure. More importantly, estimators under these models are obtained from the theory of (empirical) best unbiased prediction (E)BLUP; see Rao (2003, ch. 6).

2.6. Notation

In this section, we introduce some basic mathematical notation that will be required in several chapters. Consider the n-dimensional Euclidean space $\mathbb{R}^n$ ($1 \leq n \leq \infty$) and let $\mathbf{x} \in \mathbb{R}^n$. We always regard $\mathbf{x} \in \mathbb{R}^n$ as a n-dimensional *column* vector. There is one single exception to this rule, namely, when $\mathbf{x}$ is the argument of a function $f : \mathbb{R}^n \to \mathbb{R}$. In this particular case, the argument $\mathbf{x}$ is regarded as a row vector such that we have $f(\mathbf{x}) \equiv f(x_1, \ldots, x_n)$.

For a fixed number $q \geq 1$, we define the q-norm by $\|\mathbf{x}\|_q = (\sum_{i \leq n} |x_i|^q)^{1/q}$. If we want to refer to *any* norm, irrespective of what value q is, we write $\|\cdot\|$.

Vector-valued functions. Let I be equal to either $\mathbb{R}$ or $\mathbb{R}^n$. Vector-valued functions are denoted by bold symbols, e.g. $\boldsymbol{f} : I \to \mathbb{R}^n$ where map $\boldsymbol{f}(\mathbf{x})$ is understood as a column vector. Such functions are sometimes easier to handle in a component-wise manner, i.e. $\boldsymbol{f}(\mathbf{x}) = \big(f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})\big)^T$, where each $f_i : I \to \mathbb{R}$ for all $i = 1, \ldots, n$.

With regard to vector calculus, confusion can easily arise on how to layout derivatives. In order to avoid any confusion, we choose the

Numerator layout (also known as Jacobian layout); that is,

$$\text{(i)} \quad \frac{\partial y}{\partial \mathbf{x}} \quad \text{is a row vector,} \qquad \text{(ii)} \quad \frac{\partial \mathbf{y}}{\partial x} \quad \text{is a column vector,}$$

(where y is a function of the real n-vector $\mathbf{x}$ in case (i) and $\mathbf{y}$ is function of the real scalar x in case (ii). Also note that when the definition in case (i) is replaced by $\partial y / \partial \mathbf{x}^T$, both cases represent column vectors.

Gradient. Consider the function $f : X \subset \mathbb{R}^n \to \mathbb{R}$ (where X is open) and let $\boldsymbol{\xi} \in X$. If $f(\mathbf{x})$ is everywhere differentiable in $\mathbf{x}$ at $\boldsymbol{\xi}$, the gradient of f, i.e.

a p-dimensional row vector, is given by

$$f'(\boldsymbol{\xi}) = \nabla f = \frac{\partial}{\partial \mathbf{x}} f = \left(\frac{\partial}{\partial x_1} f(\boldsymbol{\xi}), \ldots, \frac{\partial}{\partial x_n} f(\boldsymbol{\xi}) \right),$$

and for each $i = 1, \ldots, n$, the partial derivative of f w.r.t. x_i on X, denoted by $\partial/(\partial x_i) f$, is the function

$$(x_1, \ldots, x_n) \mapsto \frac{\partial}{\partial x_i} f(x_1, \ldots, x_n).$$

Jacobian (matrix). Let $\boldsymbol{f} : G \subset \mathbb{R}^n \to \mathbb{R}^q$ (where G is open) and assume that it is everywhere differentiable in G. Then we have for $\xi \in G$ and $\mathbf{h} \in \mathbb{R}^n$

$$\boldsymbol{f}(\boldsymbol{\xi} + \mathbf{h}) - \boldsymbol{f}(\boldsymbol{\xi}) = \boldsymbol{D}\boldsymbol{f}(\boldsymbol{\xi})\mathbf{h} + \mathbf{r}(\mathbf{h}) \quad \text{with} \quad \lim_{\mathbf{h} \to \mathbf{0}} \frac{\mathbf{r}(\mathbf{h})}{\|\mathbf{h}\|} = \mathbf{0},$$

where $\mathbf{r}(\mathbf{h})$ is the q-vector of component-wise remainder terms r_i ($i = 1, \ldots, q$), $\| \cdot \|$ denotes any norm that is compatible with the norms on G and $\mathbb{R}^q$; and $\boldsymbol{D}\boldsymbol{f}(\boldsymbol{\xi})$ is the $(q \times n)$ *Jacobian* matrix of $\boldsymbol{f}$.

3. Asymptotic robustness: Strong consistency of the ratio estimator

3.1. Introduction

Let U denote a finite population of units labeled $i = 1, \ldots, N$. Associated with the ith element in U is a realization y_i of the unknown study variable Y_i. We assume that a sample s of fixed size n is selected (without replacement) from U according to sampling design $p(s)$, which is supposed to satisfy Definition 2.2. The first- and second-order sample inclusion probabilities are denoted by π_i and π_{ij}, $i, j \in U$, respectively. Expectation and variance w.r.t. the randomization distribution are denoted by $\mathbb{E}_p$ and $\mathbb{V}_p$, respectively; cov_p shall be defined in the same way.

The problem is to estimate / predict the population mean,

$$\bar{y}_U = \sum_{i \in U} y_i / N,$$

taking advantage of superpopulation model ξ. Model ξ relates Y_i to the *auxiliary* variable x_i, where $x_i > 0$ and x_i is known for all $i \in U$ prior to sampling. Under model ξ, the Y_i's are supposed to be r.v.'s taking values in $[0, \infty)$,

$$\xi: \quad Y_i = x_i \beta + E_i \quad \text{for all } i \in U, \tag{3.1}$$

where

$$\mathbb{E}_\xi E_i = 0 \quad \forall i \in U, \quad \text{and} \quad \mathbb{E}[E_i E_j] = \begin{cases} \sigma^2 v(x_i) & \text{if } i = j,\ i, j \in U, \\ 0 & \text{otherwise.} \end{cases}$$

The parameters $\beta \in \mathbb{R}^+$ and $\sigma \in \mathbb{R}^+$ are unknown; the variance function $v(\cdot)$ is defined as $v : \mathbb{R}^+ \to \mathbb{R}^+$ and is supposed known. We restrict attention to the regression through the origin (RTO) model. The extension to the multiple regression superpopulation model is straightforward and does neither pose any difficulties nor require notable modifications of the underlying theory.

Let μ_i, $i \in U$, denote *arbitrary* known real numbers. The *generalized difference predictor* (GD) due to Cassel et al. (1976, 616) is given by

$$\bar{Y}_{GD} = \sum_{i \in s} \frac{y_i - \mu_i}{\pi_i N} + \bar{\mu}_U, \tag{3.2}$$

where

$$\bar{\mu}_U = \sum_{i \in U} \mu_i / N.$$

It is easy to see that $\bar{Y}_{GD}$ is p-unbiased. Moreover, it is ξ-unbiased for any model if $\mu_i = \mathbb{E}_\xi Y_i$ for all $i \in U$. In the case that $\mu_i \equiv 0$, $\bar{Y}_{GD}$ reduces to the Horvitz–Thompson (HT) estimator of the mean; this is also the case (but rather artificial) if $\mu_i \propto \pi_i$ and $p(s)$ is a fixed-sized design (Cassel et al., 1977, 95).

Now, in the context of model ξ, we may choose $\mu_i = x_i\beta$ for all $i \in U$, where β is supposed known. This configuration has some remarkable optimality properties under the model, which shall be illustrated for the particular case when the variance function is specified as $v(x_i) = x_i$ for all $i \in U$. Denote by $\mathrm{pps}(\sqrt{x})$ the PPS design with size measure $\sqrt{x_i}$. Let the mean squared error be defined as

$$\mathrm{MSE}_{\xi p} = \mathbb{E}_\xi \mathbb{E}_p\big((\bar{Y} - \bar{y}_N)^2\big),$$

where $\bar{Y}$ is any p-unbiased linear estimator/ predictor of $\bar{y}_U$. Let $p(s)$ denote an arbitrary fixed-size design, then the strategy $\big(\mathrm{pps}(\sqrt{x}), \bar{Y}_{GD}\big)$ is optimal among all fixed-size strategies $(p, \bar{Y})$ because $\mathrm{MSE}_{\xi p}$ attains its minimum. [Remark: Strictly speaking, we must choose

$$\pi_i = \frac{n\sqrt{x_i}}{\sum_{i \in U} \sqrt{x_i}}$$

in order to maintain our assumption of fixed-size designs; see Cassel et al. (1976, chap. 3)].

When β is not known, Cassel et al. (1976, 617–8) propose the *generalized regression estimator* (GREG),

$$\hat{\bar{y}}_{GR} = \sum_{i \in s} \frac{y_i - x_i\hat{\beta}}{\pi_i N} + \sum_{i \in U} \frac{x_i\hat{\beta}}{N}, \tag{3.3}$$

where $\hat{\beta}$ denotes a sample-based estimator of β under model ξ. For large samples, the choice of $\hat{\beta}$ hardly matters in terms efficiency. In particular, $\hat{\bar{y}}_{GR}$ based on either the π-weighted estimator of β or the best linear unbiased estimator (BLUE) under model ξ are equally efficient with regard to a first-order approximation to $\mathrm{MSE}_{\xi p}$ (Särndal, 1980, 645).

The estimator $\hat{\bar{y}}_{GR}$ is not p-unbiased for $\bar{y}_U$ which, however, is no reason to worry. In the eyes of Särndal (1980, 641), exact design unbiasedness is of doubtful virtue (a point of view shared with J. Hájek). The price for exact p-unbiasedness can be high in terms of unexploited efficiency gains; and the focus on exact p-unbiasedness shrinks the class of candidate estimators considerably.

Asymptotic robustness – model independence

As the design bias of $\hat{\bar{y}}_{GR}$ vanishes asymptotically under mild assumptions, $\hat{\bar{y}}_{GR}$ is asymptotically design unbiased (ADU) and design consistent (ADC) whether

the model holds or not. Furthermore, the additional contribution to $\mathrm{MSE}_{\xi p}$ arising from the fact that β has to estimated by $\hat{\beta}$ is small compared to the leading term (Särndal, 1980, 641). Hence, randomization provides a source of *robustness against model failure* (in large samples; Särndal, 1980, 641); a property that Särndal (1980, 643) calls "model independence"; see also Cassel et al. (1977, chap. 7).

The "robustness" of $\hat{\bar{y}}_{GR}$ or any other ADU estimator in terms of (potential) model failure to which Cassel et al. (1977) refer to, is of asymptotic nature. That is, the contribution of any assisting model to the estimator vanishes as the sample size grows (irrespective of whether the model holds or not). This property is an intrinsic characteristic of ADU estimators and has become a cornerstone of the model-assisted sampling paradigm. Consequently ADUness of an estimator is commonly regarded as a minimum requirement that any candidate estimator is required to satisfy in the context of model-assisted sampling. Therefore, our discussion will focus on ADUness / ADCness and robustness will refer to asymptotic robustness against model failure. This notion of robustness will be elaborated in more detail in what follows. Clearly, in small samples, ADUness does not provide any protection against model failure at all.

Contribution

ADCness or ADUness have been studied for different estimators and under various asymptotic frameworks; see Brewer (1979), Särndal (1980), Isaki and Fuller (1982), Robinson and Särndal (1983), Wright (1983), or Särndal and Wright (1984) among others. All papers on ADUness / ADCness have, as far as we know, in common that they focus on weak consistency. Our approach aims at *strong consistency* instead (using a strong law of large numbers, SLLN). To this end, we restrict attention to nonnegative r.v.'s that are not necessarily independently distributed from each other. The restriction to nonnegative r.v.'s is quite natural in the context of the ratio model.

The key property for the application of a particular type of SLLN is the fact that under mild assumptions (see below) the designs $p(s)$ are such that $\pi_{ij} \leq \pi_i\pi_j$ for all $i, j \in U$, $i \neq j$, which implies $\mathrm{cov}_p(\mathbb{1}_{i\in s}, \mathbb{1}_{j\in s}) \leq 0$; the sample inclusion indicator $\mathbb{1}_{i\in s}$ equals one (zero) in the presence (absence) of element i in the sample s. Now, the nonpositive covariance plays a crucial role as it implies

$$\mathbb{V}_p(\hat{T}_y) \leq \sum_{i\in U} \mathbb{V}_p(y_i \mathbb{1}_{i\in s}/\pi_i),$$

where $\hat{T}_y = \sum_{i\in U} \mathbb{1}_{i\in s} y_i/\pi_i$ is the HT estimator of the population y-total. This property simplifies and weakens the set of assumptions that have to be imposed in the asymptotic framework.

The opinions whether strong consistency (more general the SLLN) should

be preferred over weak consistency (and the weak law of large numbers) are divided. We argue in line of W. Feller that "[the weak law of large numbers] is of very limited interest and should be replaced by the more precise and more useful strong law of large numbers" (Feller, 1968, 152).[1]

Outline of the chapter

The remainder of this chapter is organized as follows. In Section 3.2, we provide more details on populations, samples, and the asymptotic framework under study. In Section 3.3, we present the strong consistency results for the estimators: Horvitz–Thompson estimator, ratio estimator, and the class of QR estimators. Finally, we draw together the major findings (see Section 3.4).

3.2. Population, sample, and asymptotic framework

The finite population U defined in the preceding paragraph will be equipped with a subscript t; hence we have U_t which is of size N_t, $t = 1, 2, \ldots$ (the index t is unimportant for the moment). This applies also to the sample s_t, the sample size n_t, the first- and second-order sampling inclusion probabilities $\pi_{i,t}$ and $\pi_{ij,t}$, and our notation of the sample indicator variable $\mathbb{1}_{i \in s_t}$. In the light of the new naming convention, we define the (unknown) y-population mean to be

$$\bar{y}_{U_t} = \frac{1}{N_t} \sum_{i \in U_t} y_i,$$

the x-mean, $\bar{x}_{U_t}$, is a known quantity and shall be defined in the same way.

3.2.1. Sampling

We shall assume that a sample s_t is drawn form U_t. The sampling design $p(s_t)$ is required to adhere to the conditions formulated in the introductory chapter. For ease of reading, we summarize the key points of our assumptions. First, we suppose that $p(s_t)$ satisfies Definition 2.2, i.e. it holds that

(i) all labels $i \in U_t$ are identifiable,

(ii) it is possible to observe and measure without error the variable(s) of interest Y_i for each sampled element $i \in s_t$,

(iii) the randomization scheme is non-informative,

(iv) $p(s_t) > 0$ for all $s_t \in \mathscr{S}_t$,

[1] A contrary view is taken by B.L. van der Waerden who writes "[d]aneben gibt es auch noch das 'starke Gesetz der grossen Zahlen', das aber in der mathematischen Statistik keine grosse Rolle spielt."; see B.L. Van der Waerden (1971) Mathematische Statistik, 3rd ed. Springer: Berlin, p. 98.

(v) $\pi_{i,t} > 0$ for all $i \in U_t$ and $\pi_{ij,t} > 0$ for all $i, j \in U_t$, $i \neq j$.

Moreover, we assume that Assumptions 2.1, 2.2, and 2.3 hold which are repeated here for the sake of disposition.

Assumption 2.2. *The (asymptotic) sampling fraction f satisfies*

$$\lim_{t\to\infty} \frac{n_t}{N_t} =: f \quad \textit{such that} \quad f \in (0, 1).$$

Assumption 2.1. *Let $p(s_t)$ be a sampling design that satisfies Definition 2.2 for all $t = 1, 2, \ldots$. In addition, the sampling design $p(s_t)$ must be such that*

(i) the samples are of fixed size n_t, $0 < n_t < N_t$ for all $t = 1, 2, \ldots$;

(ii) the samples are drawn using some without-replacement sampling mechanism.

Assumption 2.3 (Minimal)**.**

(i) There exist constants L_π and U_π independent of $\pi_{i,t}$ such that the first-order sample inclusion probabilities satisfy, for all t,

$$0 < \underline{\pi} \leq \pi_{i,t} \leq \overline{\pi} < 1 \qquad \forall i \in U_t,$$

(ii) and the second-order sample inclusion probabilities obey the inequality, for all t,

$$\pi_{ij,t} \leq \pi_{i,t}\pi_{j,t} \quad \textit{for all } i, j \in U,\ i \neq j.$$

For any $t = 1, 2, \ldots$, the *covariance* of the sample-inclusion indicators will be denoted by

$$\Delta_{ij,t} = \mathrm{Cov}_p[\mathbb{1}_{i\in s_t}, \mathbb{1}_{j\in s_t}] = \pi_{ij,t} - \pi_{i,t}\pi_{j,t}. \tag{3.4}$$

In view of (3.4), Part ii) of Assumption 2.3 can equivalently be expressed as $\Delta_{ij,t} \leq 0$, $i, j \in U_t$, $i \neq j$. Following Robinson (1982, 237), we say that two observations $i, j \in U_t$, $i \neq j$, are "tied" if $\Delta_{ij,t} > 0$. It has been shown by Robinson (1982) for the case of the Horvitz–Thompson estimator that tying can lessen the rate of convergence. Fortunately, the vast majority of single-stage without-replacement sampling designs satisfies $\Delta_{ij,t} \leq 0$. However, this property usually does not hold for one-stage cluster sampling and multistage designs (due to positive correlation within clusters).

Moreover, by the hypothesis of fixed-size sampling designs, the identities

$$\sum_{i=1}^{N_t} \pi_{i,t} = n_t, \quad \text{and} \quad \sum_{j=1,\, j\neq i}^{N_t} \pi_{ij,t} = (n_t - 1)\pi_{i,t} \quad \text{for all } i \in U_t, \tag{3.5}$$

obtain (cf. Särndal et al., 1992, Result 2.6.2), which imply

$$\sum_{j=1,\, j\neq i}^{N_t} \Delta_{ij,t} = \pi_{i,t}(\pi_{i,t} - 1) \quad \text{for all } i \in U_t. \tag{3.6}$$

3.2.2. Asymptotic robustness

Our asymptotic framework is that of Isaki and Fuller (1982); see also Fuller (2009, chap. 1.3) and our introductory Chapter 2.2.2. Let $\{U_t, t \geq 1\}$ denoted the nested sequence of populations U_t of size N_t; the samples s_t form an analogous but not necessarily nested sequence. All limiting processes will be taken as $t \to \infty$. We assume in addition that our Assumptions on the sampling design are maintained as $t \to \infty$ and that $n_t \to \infty$ as $N_t \to \infty$.

Let $\hat{\bar{y}}$ denote a generic estimator of $\bar{y}_{U_t}$. In the introduction to this chapter, we argued, referring to Särndal (1980), that randomization provides a source of *robustness against model failure* in the sense that ADU estimators are asymptotically independent of the model. From this perspective, $\hat{\bar{y}}$ is regarded as robust if it is exactly p-unbiased or if it is ADU. Tam (1988) calls such estimators "weakly robust". In this respect, ADUness and weak robustness of an estimator are equivalent concepts. Strong robustness on the other hand obtains if (Tam, 1988, 224)

(i) $\hat{\bar{y}}$ is a weakly robust estimator

(ii) and if its expected variance under the model attains the minimum of the Godambe–Joshi lower bound asymptotically.

The requirement that an estimator $\hat{\bar{y}}$ must attain the minimum of the Godambe–Joshi lower bound in order to be regarded as a strongly robust strategy has been proposed by Godambe (1982) for exactly p-unbiased estimators; Tam (1988) extends this condition to ADU estimators (though, ADU estimators attain the bound only *asymptotically*). Suppose $\hat{\bar{y}}$ is p-biased but ADU; then, the asymptotic expected variance of $\hat{\bar{y}}$ under model ξ, given by

$$AV = \lim_{t\to\infty} \mathbb{E}_p \mathbb{E}_\xi \big((\hat{\bar{y}} - \bar{y}_{U_t})^2 \big),$$

is supposed to attain the Godambe–Joshi lower bound,

$$\frac{1}{N_t^2} \sum_{i \in U_t} \left(\frac{1}{\pi_i} - 1 \right) v(x_i)^2. \tag{3.7}$$

The asymptotic expected variance, AV, is typically sought to be minimized by a suitable choice of design (more precisely, sampling strategy) at the planning stage of a survey.

The condition that the asymptotic expected variance attains the Godambe–Joshi lower bound is rather bulky to work with. Fortunately, there exists a

sufficient condition for strong robustness which facilitates the analysis considerably. Let $\hat{\bar{y}}$ be ADU for $\bar{y}_{U_t}$. If $\hat{\bar{y}}$ is also a ξ-unbiased estimator, then Theorem 3.1. in Tam (1988) implies that $\hat{\bar{y}}$ is strongly robust for $\bar{y}_{U_t}$. Hence, all we need to ensure for an ADU estimator to be strongly robust is that it is also ξ-unbiased.

In fact, all estimators studied in this chapter (namely, HT-, ratio-, and QR-estimator) are ξ-unbiased by construction. As a consequence, all estimators are strongly robust once we have proved that they are ADU (respectively, asymptotically design consistent or strongly design consistent). Therefore, we shall simplify our discussion insofar that robustness will *always* mean strong robustness in what follows.

3.2.3. Law of large numbers

Consider the sequence $\{X_i, i \geq 1\}$ of square integrable random variables that take values in $\mathbb{R}$. Put $S_n = \sum_{i \leq n} X_i$, and denote expectation and variance by, respectively, $\mathbb{E}X_i$ and $\mathbb{V}X_i$ for all i. A. A. Markov showed that $\mathbb{V}S_n = o(n^2)$ is a sufficient condition for the *weak* law of large numbers to hold, i.e. that $S_n/n \xrightarrow{\text{p}} 0$ as $n \to \infty$ (cf. Chung, 2001, chap. 5.1). Consider

$$\begin{aligned}\mathbb{E}(S_n^2) &= \mathbb{E}\bigg(\bigg(\sum_{i=1}^{n} X_i\bigg)^2\bigg) = \mathbb{E}\bigg(\sum_{i=1}^{n} X_i^2 + 2 \sum_{1 \leq i < j \leq n} X_i X_j\bigg) \\ &= \sum_{i=1}^{n} \mathbb{E}(X_i^2) + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}(X_i X_j) \qquad (3.8)\end{aligned}$$

and observe that there are n^2 terms in (3.8). Suppose that all terms in (3.8) are bounded by some constant. Then, we have $\mathbb{E}(S_n^2) = \mathcal{O}(n^2)$, which, however, falls critically short of *Markov's condition*. In order to ensure that the variance of S_n is $o(n^2)$, we are required to impose certain assumptions on the dependence structure of the X_i in order to "cause enough cancellation among the mixed terms" (Chung, 2001, 107). Obviously, the easiest case obtains when the X_i's are (totally or pairwise) independent.

In what follows, we shall be interested in the strong law of large numbers (SLLN). We say that the sequence $\{X_i, i \geq 1\}$ satisfies the SLLN if $(S_n - \mathbb{E}S_n)/n \to 0$ almost surely (a.s.) as $n \to \infty$.

Independent random variables

Before we continue our discussion, we focus on independent square integrable r.v.'s X_i with bounded $\mathbb{E}(S_n)$ for all $n = 1, 2, \ldots$ Besides Markov's condition, A.N. Kolmogorov and V.V. Petrov established other sufficient conditions. We say that the X_i's obey *Kolmogorov*'s condition if

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}(X_i^2)}{i^2} < \infty,$$

which then implies an SLLN by Kolmogorov (cf. Petrov, 1975, Thm. 14). Before we give Petrov's condition, we introduce the following class of functions.

Definition 3.1. *Denote by* $\Gamma = \{\gamma : \mathbb{R} \to \mathbb{R}^+\}$ *a class of functions such that*

(i) γ *is nondecreasing in the domain of* $x > x_0$ *for some real* x_0 *and*

(ii) the series

$$\sum_{n=1}^{\infty} \frac{1}{n\gamma(n)}$$

is convergent.

The value x_0 in the definition is not assumed to be the same for different candidate functions in Γ. Examples of functions in Γ are x^δ and $(\log x)^{1+\delta}$ for any $\delta > 0$; see Petrov (1975, chap. 9). The condition due to *Petrov* is given by (see Petrov, 1969)

$$\mathbb{V}S_n = \mathcal{O}\left(\frac{n^2}{\gamma(n)}\right) \quad \text{for any } \gamma \in \Gamma.$$

Petrov's condition implies Kolmogorov's condition (Korchevsky, 2010). Further, it is easily seen that Petrov's condition can be regarded as a strengthened version of Markov's condition.

It is interesting to ask under what additional hypotheses a Kolmogorov- or Petrov-type SLLN holds if we drop the independence assumption. It turns out that nice results obtain when the X_i's are nonnegative. To see this, observe that nonnegative r.v.'s have the peculiar property that the summand of the covariance terms in (3.8) is negative provided that the X_i's are negatively correlated. As a consequence, it is sufficient to "control" the term $\sum_{i \le n} \mathbb{E}X_i^2$.

SLLN for nonnegative not necessarily independent random variables

Unless otherwise stated, we adhere to the following notation. Denote by

$$\{X_i, i \ge 1\} \text{ a seq. of nonnegative r.v.'s with finite second moment,} \tag{3.9}$$

$$\{w_i, i \ge 1\} \text{ a seq. of nonnegative real numbers,} \tag{3.10}$$

$$\{a_i, i \ge 1\} \text{ a monotone seq. of positive real numbers,} \tag{3.11}$$

where $0 < a_i \uparrow \infty$ as $i \to \infty$.

There exists a considerably large body of SLLN for nonnegative r.v.'s. Two strands of research are noteworthy, and are related to the following authors:

(i) Etemadi (1983a,b) [see also Csörgö, Tandori, and Totik (1983)] who assumes Kolmogorov's condition and requires in addition that $\mathbb{E}(X_i)$ is uniformly bounded and

$$\mathbb{E}(X_i Y_j) \le \mathbb{E}(X_i)(\mathbb{E}X_j) \quad \text{for all } j > i,$$

(ii) Petrov (2009) who utilizes Petrov's [i.e. "his"] condition and imposes the hypothesis that $\mathbb{E}(S_{n,m}) \leq C_1(n-m)$ with $S_{n,m} = \sum_{i=m+1}^{n} X_i$ for sufficiently large n and m, where C_1 is a constant independent of n and m; see also Kuczmaszewska (2016).

Chen and Sung (2016) establish a theorem which unifies both strands.

Etemadi's SLLN for nonnegative r.v.'s

In what follows, we pursue the approach of Etemadi (1983a), which can be inferred from his theorem.

Theorem 3.1 (Etemadi, 1983a, Thm. 1)**.** *Let $\{X_i, i \geq 1\}$ be a sequence of nonnegative r.v.'s with finite second moment such that*

(i) $\sup_{i\geq 1} \mathbb{E}X_i < \infty$*,*

(ii) $\mathbb{E}[X_i X_j] \leq \mathbb{E}X_i \mathbb{E}X_j$ *for all* $j > i$*, and*

(iii) $\sum_{i=1}^{\infty} \mathbb{V}X_i/i^2 < \infty$*.*

Then as $n \to \infty$*,*

$$\frac{S_n - \mathbb{E}S_n}{n} \to 0 \text{ a.s.}$$

Proof. Let $a > 0$ and consider the subsequence $\{\lfloor a^n \rfloor, n \geq 1\}$ of $\{n\}$. Define $S_{\lfloor a^n \rfloor} = \sum_{i \leq \lfloor a^n \rfloor} X_i$. By the Chebyshev inequality, for every $\epsilon > 0$

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathbb{P}\Big\{ \big| S_{\lfloor a^n \rfloor} - \mathbb{E}(S_{\lfloor a^n \rfloor}) \big| > \epsilon \lfloor a^n \rfloor \Big\} &\leq \frac{C_1}{\epsilon^2} \sum_{n=1}^{\infty} \mathbb{V}S_{\lfloor a^n \rfloor} / \lfloor a^n \rfloor^2 \\
&= \frac{C_1}{\epsilon^2} \sum_{n=1}^{\infty} \frac{1}{\lfloor a^n \rfloor^2} \Big(\sum_{j=1}^{\lfloor a^n \rfloor} \mathbb{V}X_j \Big) = \frac{C_1}{\epsilon^2} \sum_{j=1}^{\infty} \mathbb{V}X_j \sum_{n:\lfloor a^n \rfloor \geq j}^{\lfloor a^n \rfloor} \frac{1}{\lfloor a^n \rfloor^2} \qquad \text{(A.1)} \\
&\leq \frac{C_1}{\epsilon^2} \sum_{j=1}^{\infty} \mathbb{V}X_j / j^2 < \infty, \qquad \text{(A.2)}
\end{aligned}$$

where C_1 is a generic (and unimportant) constant that differs from line to line. Since ϵ is arbitrary, by the Borel–Cantelli lemma [see Lemma A.5], as $n \to \infty$,

$$\mathbb{P}\big[|S_{\lfloor a^n \rfloor}| > \lfloor a^n \rfloor \epsilon \text{ i.o.} \big] = 0,$$

which implies

$$\frac{S_{\lfloor a^n \rfloor} - \mathbb{E}(S_{\lfloor a^n \rfloor})}{\lfloor a^n \rfloor} \xrightarrow{\text{a.s.}} 0.$$

To handle the intermediate values, we have for any given $j \in \mathbb{N}$ s.t. $\lfloor a^n \rfloor \leq j < \lfloor a^{n+1} \rfloor$

$$\frac{S_j - \mathbb{E}(S_j)}{j} \leq \left| \frac{S_{\lfloor a^{n+1} \rfloor} - \mathbb{E}(S_{\lfloor a^{n+1} \rfloor})}{\lfloor a^{n+1} \rfloor} \right| \cdot \frac{\lfloor a^{n+1} \rfloor}{\lfloor a^n \rfloor} + \frac{\mathbb{E}(S_{\lfloor a^{n+1} \rfloor}) - \mathbb{E}(S_{\lfloor a^n \rfloor})}{\lfloor a^n \rfloor} \qquad \text{(C)}$$

and an analogous estimate from below. Thus by hypothesis (i), (C), and the fact that $\lfloor a^{n+1} \rfloor / \lfloor a^n \rfloor \to a$ (cf. Durrett, 2010, 75), we have

$$\frac{1}{a} \sup_{i \geq 1} \mathbb{E}(X_i) \leq \liminf_{n \to \infty} \frac{S_m - \mathbb{E}(S_m)}{m} \leq \limsup_{n \to \infty} \frac{S_j - \mathbb{E}(S_j)}{j} \leq a \sup_{i \geq 1} \mathbb{E} X_i,$$

for every $a > 1$ which completes the proof. ■

Remarks. (i) Walk (2005, Thm. 1) gave a generalization of Theorem 3.1, having replaced hypothesis (ii) and (iii) by the weaker assumption $\sum_{i=1}^{\infty} \mathbb{V}(X_1 + \cdots + X_i)/i^3 < \infty$.

(ii) The proof of Theorem 1 in Etemadi (1983a) is kept very short. We added some references.

Arbitrary normalization

Etamadi's SLLN uses the normalization in terms of the sequence of positive integers $n = 1, 2, \ldots$ For many applications, it is useful to consider an arbitrary norming sequence in place of the classical one.

Consider the arbitrary monotone sequence of positive real number $\{a_n, n \geq 1\}$ defined in (3.11). Korchevsky (2015) formulated the *generalized* Petrov condition,

$$\mathbb{E}|S_n - \mathbb{E}S_n|^p = \mathcal{O}\left(\frac{a_n^p}{\gamma(a_n)}\right) \quad \text{for some } \gamma \in \Gamma \text{ and } p \geq 1. \tag{3.12}$$

Under the assumptions in (3.12) and the additional hypothesis $\mathbb{E}S_n = \mathcal{O}(a_n)$, Korchevsky (2015, Thm. 1) proved

$$\frac{S_n - \mathbb{E}S_n}{a_n} \to 0 \text{ a.s.} \qquad \text{as } n \to \infty. \tag{3.13}$$

Main results

We provide an extension of Theorem 3.1 to an arbitrary norming sequence in place of the classical one—similar to the Petrov-type result in (3.13). Our theorem provides two interesting corollaries: an SLLN for weighted sums and a result on the strong stability of sums of nonnegative r.v.'s. In addition, we give a generalization of Wu's (1981) lemma (see Lemma 3.1) for nonnegative but not necessarily independent r.v.'s.

Theorem 3.2. *Consider* $\{X_i, i \geq 1\}$ *and* $\{a_i, i \geq 1\}$ *in, respectively, (3.9) and (3.11). Suppose the hypotheses,*

(i) $\mathbb{E}S_n = \mathcal{O}(a_n)$*,*

(ii) $\mathbb{E}[X_i X_j] \leq \mathbb{E}X_i \mathbb{E}X_j$ *for all* $j > i$*,*

(iii) $\sum_{i=1}^{\infty} \mathbb{V}X_i / a_i^2 < \infty$*,*

then as $n \to \infty$*,*

$$\frac{S_n - \mathbb{E}S_n}{a_n} \to 0 \text{ a.s.}$$

Proof. Let $\beta > 1$ be a real number and $m, n \in \mathbb{N}$. For all $m \geq 1$, define

$$n_m = \inf\left\{n : a_n \geq \beta^m\right\}. \tag{3.14}$$

Note that $\{n_m, m \geq 1\}$ is a monotone sequence of positive integers, $0 < n_1 \leq n_2 \leq \cdots \leq n_m \uparrow \infty$ as $m \to \infty$ since $\{a_n, n \geq 1\}$ is monotone and $0 < a_n \uparrow \infty$ as $n \to \infty$. By Chebyshev's inequality and hypothesis (ii), for any $\epsilon > 0$,

$$\begin{aligned}\epsilon^2 \sum_{m=1}^{\infty} \mathbb{P}\left\{\left|S_{n_m} - \mathbb{E}S_{n_m}\right| > \epsilon \cdot a_{n_m}\right\} &\leq \sum_{m=1}^{\infty} \frac{\mathbb{V}S_{n_m}}{a_{n_m}^2} \\ &\leq \sum_{m=1}^{\infty} \frac{1}{a_{n_m}^2} \sum_{i=1}^{n_m} \mathbb{V}X_i = \sum_{i=1}^{\infty} \mathbb{V}[X_i] t_i, \end{aligned} \tag{3.15}$$

where

$$t_i = \sum_{m \in M} \frac{1}{a_{n_m}^2} \qquad \text{and} \qquad M_i = \{m : n_m \geq i\}. \tag{3.16}$$

Next, we use an argument similar to the one in Etemadi (1983b). Since $a_n \uparrow \infty$ as $n \to \infty$, it follows that $a_{n_m} \sim \beta^m$ for all large m. Thus for some constant $C_1 > 0$ and every $i = 1, 2, 3, \ldots$,

$$M_i = \left\{m : n_m \geq i\right\} \subset \left\{m : a_{n_m} \geq a_i\right\} \subset \left\{m : C_1 \beta^m \geq a_i\right\} =: M_i^*, \text{ say,} \tag{3.17}$$

since $n_m \geq i$ and monotonicity of $\{a_n, n \geq 1\}$ imply $a_{n_m} \geq a_i$. By this and (3.14), the geometric series on the far left in (3.16) obtains for all $i \geq 1$

$$t_i = \sum_{m \in M_i} a_{n_m}^{-2} \leq \sum_{m \in M_i} \beta^{-2m} \leq \sum_{m \in M_i^*} \beta^{-2m} = \frac{C_\beta}{\beta^{2m_i}}, \tag{3.18}$$

where $m_i = \inf M_i^*$ and $C_\beta = (1 - 1/\beta^2)^{-1}$ is a constant. From (3.17) and the fact that $m_i \in M_i^*$, it is easy to see that $C_1^2/a_i^2 \geq \beta^{-2m_i}$. This together with hypothesis (iii) and (3.15) implies

$$\epsilon^2 \sum_{m=1}^{\infty} \mathbb{P}\left\{\left|S_{n_m} - \mathbb{E}S_{n_m}\right| > \epsilon \cdot a_{n_m}\right\} \leq C_1^2 C_\beta \sum_{i=1}^{\infty} \frac{\mathbb{V}X_i}{a_i^2} < \infty. \tag{3.19}$$

Since $\epsilon > 0$ is arbitrary, (3.19) and the Borel–Cantelli lemma [see Lemma A.5] imply $\mathbb{P}\{|S_{n_m} - \mathbb{E}S_{n_m}| > \epsilon \cdot a_{n_m} \text{ i.o.}\} = 0$ (where i.o. stands for infinitely often), therefore as $m \to \infty$

$$\frac{S_{n_m} - \mathbb{E}S_{n_m}}{a_{n_m}} \to 0 \text{ a.s.} \tag{3.20}$$

Thus we showed the desired result for the subsequence. This result can be extended to the whole sequence. Let $m \in [n_m, n_{m+1})$. By monotonicity of S_m, we

have

$$\frac{|S_m - \mathbb{E}S_m|}{a_m} \leq \frac{|S_{n_{m+1}} - \mathbb{E}S_{n_m}|}{a_{n_m}} \leq \frac{a_{n_{m+1}}}{a_{n_m}} \frac{|S_{n_{m+1}} - \mathbb{E}S_{n_{m+1}}|}{a_{n_{m+1}}} + \frac{|\mathbb{E}S_{n_{m+1}} - \mathbb{E}S_{n_m}|}{a_{n_m}}. \tag{3.21}$$

By (3.20) and since $a_{n_{m+1}}/a_{n_m} \to \beta$ (as $m \to \infty$), the first summand on the far right of (3.21) converges a.s. to zero as $m \to \infty$. By hypothesis (i), and for m large enough, there is a constant C_2 such that $\mathbb{E}S_m \leq C_2 a_m$. Hence, for large m, we have obtain for the second summand in (3.21) that

$$\frac{|\mathbb{E}S_{n_{m+1}} - \mathbb{E}S_{n_m}|}{a_{n_m}} \leq C_2 \frac{|a_{n_{m+1}} - a_{n_m}|}{a_{n_m}} \to C_2(\beta - 1) \qquad (\text{as } m \to \infty),$$

thus,

$$\limsup_{m\to\infty} \left| \frac{S_m - \mathbb{E}S_m}{a_m} \right| \leq C_2(\beta - 1).$$

Likewise we obtain a lower bound of the left-hand side in (3.21), which then provides that $\liminf_{m\to\infty} |(S_m - \mathbb{E}S_m)/a_m| \geq C_2(1 - 1/\beta)$. Therefore,

$$\left(1 - \frac{1}{\beta}\right) C_2 \leq \liminf_{m\to\infty} \left| \frac{S_m - \mathbb{E}S_m}{a_m} \right| \leq \limsup_{m\to\infty} \left| \frac{S_m - \mathbb{E}S_m}{a_m} \right| \leq (\beta - 1)C_2,$$

and as $\beta > 1$ is arbitrary, $|(S_m - \mathbb{E}S_m)/a_m| \xrightarrow{\beta\downarrow 1} 0$ a.s. ■

Remark. The above theorem can be seen as a special case of the more general result in Chandra and Goswami (1992, Thm. 1). They require the existence of a double sequence $\{\rho_{ij}\}$ of real values such that $\mathbb{V}S_n \leq \sum_{i=1}^n \sum_{j=1}^n \rho_{ij}$ for all $n \geq 1$ with $\sum_{i=1}^\infty \sum_{j=1}^\infty \rho_{ij}/a^2_{\max(i,j)} < \infty$ instead of our hypotheses (ii) and (iii).

An immediate corollary of Theorem 3.2 is the following SLLN for weighted sums, which is similar (but not identical) to a result in Etemadi (1983b). Consider $\{w_i, i \geq 1\}$ in (3.10), put

$$W_n = \sum_{i=1}^n w_i, \qquad T_n = \sum_{i=1}^n w_i X_i, \tag{3.22}$$

and let $\{w_i, i \geq 1\}$ be such that

$$\frac{w_n}{W_n} \to 0 \quad \text{and} \quad W_n \to \infty \quad \text{as } n \to \infty. \tag{3.23}$$

Corollary 3.1. *Let $\{X_i, i \geq 1\}$ be given in (3.9), and let $\{w_i, i \geq 1\}$ in (3.10) satisfy (3.23). Consider T_n and W_n in (3.22). If*

(i) $\mathbb{E}T_n = \mathcal{O}(W_n)$,

(ii) $\mathbb{E}[X_i X_j] \leq \mathbb{E}X_i \mathbb{E}X_j$ *for all* $j > i$,

(iii) $\sum_{i=1}^{\infty} w_i^2 \mathbb{V}[X_i]/W_i^2 < \infty$,

then as $n \to \infty$,

$$\frac{T_n - \mathbb{E}T_n}{W_n} \to 0 \text{ a.s.}$$

Proof. The proof is straightforward. Put $a_i \equiv W_i$ and apply Theorem 3.2 using $\{w_i X_i, i \geq 1\}$ and $\{W_i, i \geq 1\}$ in place of $\{X_i, i \geq 1\}$ and $\{a_i, i \geq 1\}$, hence the assertion obtains. ■

Another interesting corollary of Theorem 3.2 refers to the (strong) stability of sums of r.v.'s. Consider $\{X_i, i \geq 1\}$ in (3.9) and let $S_n = \sum_{i \leq n} X_i$ be such that

$$\mathbb{E}S_n \to \infty \quad \text{and} \quad \frac{\mathbb{E}X_n}{\mathbb{E}S_n} \to 0 \quad \text{as } n \to \infty. \tag{3.24}$$

Corollary 3.2. *Let* $\{X_i, i \geq 1\}$ *in (3.9) satisfy (3.24). If*

(i) $\mathbb{E}[X_i X_j] \leq \mathbb{E}X_i \mathbb{E}X_j$ *for all* $j > i$,

(ii) $\sum_{n=1}^{\infty} \mathbb{V}[X_n]/(\mathbb{E}S_n)^2 < \infty$,

then as $n \to \infty$,

$$\frac{S_n}{\mathbb{E}S_n} \to 1 \text{ a.s.}$$

Proof. The proof obtains from Theorem 3.2. Define $Y_i = X_i/\mathbb{E}X_i$ (where we assume without loss of generality that $\mathbb{E}X_i > 0$ for all i), and put $w_i = \mathbb{E}X_i$. Note that $T_n = \sum_{i \leq n} w_i Y_i = \sum_{i \leq n} X_i$ and $W_n = \sum_{i \leq n} w_i = \mathbb{E}S_n$, hence the assertion follows by application of Corollary 3.1. ■

The next theorem provides an SLLN under a suitable normalization when the sum of the variances of the elements in the partial sum S_n,

$$B_n = \sum_{i=1}^{n} \mathbb{V}X_i,$$

grows without bounds (i.e. $B_n \to \infty$ as $n \to \infty$).

Theorem 3.3. *Let* $\{X_i, i \geq 1\}$ *be defined as in (3.9). Suppose that the hypotheses (i) and (ii) of Theorem 3.2 hold. Let* $B_n \to \infty$ *as* $n \to \infty$, *then*

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{B_n \gamma(B_n)}} \to 0 \text{ a.s.} \qquad \text{for any } \gamma \in \Gamma.$$

Proof. Put $b_n = \sqrt{B_n \gamma(B_n)}$, where $B_n = \sum_{i \leq n} \mathbb{V}X_i$, and note that $b_n \uparrow \infty$ as $n \to \infty$. Application of the generalized Dini–Abel lemma [see Lemma A.7] implies hypothesis (iii) of Theorem 3.2, hence the assertion follows by Kronecker's lemma [see Lemma A.6]. ■

Remark. Theorem 3.3 is comparable with Theorem 1 in Petrov (1969) except that it does not require the r.v.'s to be independently distributed.

As a corollary of Theorem 3.3 we obtain the next result, which can be regarded as a generalization of Lemma 2 in Wu (1981) – see Lemma 3.1 – under the hypothesis of nonnegative but not necessarily independent r.v.'s.

Lemma 3.1 (Wu, 1981)**.** *Let* $\{X_i, i \geq 1\}$ *be a sequence of independent r.v.'s with* $\mathbb{E}X_i = 0$ *and* $\mathbb{V}X_i = \sigma_i^2 < \infty$*. Suppose a sequence of positive real numbers* $\{A_n, n \geq 1\}$ *such that*

$$A_n \to \infty, \qquad \limsup_{n\to\infty} \frac{\left(\sum_{i\leq n} \sigma_i^2\right)^{1/2+\delta}}{A_n} < \infty \qquad \textit{for some } \delta > 0. \tag{3.25}$$

Then as $n \to \infty$*,*

$$\frac{S_n}{A_n} \to 0 \textit{ a.s.}$$

Proof. See Lemma 2 in Wu (1981). ■

Under the hypothesis of independence, Lemma 3.1 proved to be a very popular and valuable tool in a large number of papers; see e.g. Fahrmeir and Kaufmann (1985) in the context of the generalized linear model. In view of the wide applicability of Lemma 3.1, our Corollary 3.3 can be useful in its own right.

Corollary 3.3. *Suppose that the hypotheses of Theorem 3.3 hold. Let* $\{w_n, n \geq 1\}$ *be a sequence of positive real numbers such that*

$$w_n \to \infty, \qquad \textit{and} \qquad \limsup_{n\to\infty} \frac{B_n^{1/2+\delta}}{w_n} < \infty \qquad \textit{for some } \delta > 0.$$

Then as $n \to \infty$*,*

$$\frac{S_n - \mathbb{E}S_n}{w_n} \to 0 \textit{ a.s.}$$

3.3. Strong consistency results

Our discussion of strong consistency is organized in increasing complexity, starting with the Horvitz–Thompson estimator (Section 3.3.1), followed by the ratio estimator (Section 3.3.2) and eventually reaching the class of QR predictors/ estimators (Section 3.3.3).

3.3.1. Horvitz–Thompson estimator

For any $t = 1, 2, \ldots,$, define the Horvitz–Thompson (HT) estimator of the population y-total $T_{y,t} = \sum_{i \in U_t} y_i$ as

$$\hat{T}_{y,t} = \sum_{i \in U_t} \frac{y_i}{\pi_{i,t}} \mathbb{1}_{i \in s_t}.$$

The HT estimator of the x-total, $T_{x,t}$, shall be defined in the same way. An obvious estimator of the population y-mean, $\bar{y}_{U_t}$, is the HT-type estimator given by

$$\hat{\bar{y}}_{HT,t} = \frac{\hat{T}_{y,t}}{N_t}.$$

Robinson (1982) studied mean square (weak) consistency of this estimator. For our purposes, it is instructive to present Robinson's main arguments (see his Theorem 1 and 2) in condensed form. To this end, define the sets, for all $i = 1, \ldots, N_t$ and $t = 1, 2, \ldots,$

$$D_i^+ = \big\{ j \in U_t : \ \Delta_{ij,t} > 0 \big\}, \quad \text{and} \quad D_i^- = \big\{ j \in U_t : \ \Delta_{ij,t} \leq 0 \big\}, \tag{3.26}$$

having suppressed the index t in the notation of D_i^+ and D_i^- for the sake of simplicity. Note that these definitions enable us to "separate" tied form untied observations. Put

$$D^+ = \cup_{i=1}^{N_t} D_i^+, \qquad D^- = \cup_{i=1}^{N_t} D_i^- .$$

Under the hypothesis of fixed-size designs, the variance of $\hat{\bar{y}}_{HT,t}$ is given by (cf. Särndal et al., 1992, Result 2.8.2)

$$N_t^2 \mathbb{V}_p\big[\hat{\bar{y}}_{HT,t}\big] = -\frac{1}{2} \sum \sum_{i \neq j} \Delta_{ij,t} \left(\frac{y_i}{\pi_{i,t}} - \frac{y_j}{\pi_{j,t}} \right)^2 \leq -\frac{1}{2} \sum \sum_{i \in D^-} \Delta_{ij,t} \left(\frac{y_i}{\pi_{i,t}} - \frac{y_j}{\pi_{j,t}} \right)^2$$

(since the double sum over the index set D^- is non-positive), moreover,

$$\begin{aligned} &\leq -\sum \sum_{i \in D^-} \Delta_{ij,t} \frac{y_i^2}{\pi_{i,t}^2} = -\sum_{i=1}^{N_t} \frac{y_i^2}{\pi_{i,t}^2} \sum_{D_i^-} \Delta_{ij,t} \\ &= -\sum_{i=1}^{N_t} \frac{y_i^2}{\pi_{i,t}^2} \Bigg(\sum_{j=1,\, j \neq i}^{N_t} \Delta_{ij,t} - \sum_{D_i^+} \Delta_{ij,t} \Bigg), \end{aligned}$$

which then, together with (3.5) and (3.6), implies

$$= \sum_{i=1}^{N_t} \frac{y_i^2}{\pi_{i,t}^2} \left(\pi_{i,t}(1 - \pi_{i,t}) + \sum_{D_i^+} \Delta_{ij,t} \right) \leq \sum_{i=1}^{N_t} \frac{y_i^2}{\pi_{i,t}^2} \left(\pi_{i,t} + \sum_{D_i^+} \Delta_{ij,t} \right),$$

hence, for N_t large, we have (see Robinson, 1982, 236)

$$= \mathcal{O}(N_t)\big(\min_{i\in U_t}\pi_{i,t}\big)^{-1}\left(1+\big(\min_{i\in U_t}\pi_{i,t}\big)^{-1}\max_{i\in U_t}\sum_{D_i^+}\Delta_{ij,t}\right). \tag{3.27}$$

Before addressing the implications of (3.27), we shall impose an assumption concerning the behavior of the sequence $\{y_i, i\geq 1\}$ of real numbers.

Assumption 3.1. *Let $\{y_i, i\geq 1\}$ be such that*

$$\lim_{t\to\infty}\frac{1}{N_t}\sum_{i=1}^{N_t} y_i^2 < \infty.$$

Under Assumption 3.1, Theorem 2 of Robinson (1982) establishes that

$$\hat{\bar{y}}_{HT,t} - \bar{y}_{U_t} = \mathcal{O}(N_t^{-1/2}\delta^{-1/2}) + \mathcal{O}(N_t^{-1/2}\delta^{-1}\zeta),$$

where (as a consequence of Eq. 3.27)

$$\delta = \min_{i\in U_t}\pi_{i,t}, \qquad \text{and} \qquad \zeta = \max_{i\in U_t}\sum_{D_i^+}\Delta_{ij,t}.$$

Hence, sufficient conditions of (weak) consistency of $\hat{\bar{y}}_{HT,t}$ for the population mean are $\delta N_t \to \infty$ and $\zeta = o(\delta {N_t}^{1/2})$ as $t\to\infty$. Note that the latter condition means that the observations are not strongly tied. Moreover, as Robinson (1982, 237) points out, it is desirable if $\Delta_{ij,t}\leq 0$, for then the sets $D_i^+ \equiv \varnothing$, i.e. the empty set for all $i\in U_t$, hence $\zeta = 0$. If in addition the $\pi_{i,t}N_t/n_t$ are bounded away from zero it follows that $\hat{\bar{y}}_{HT,t} - \bar{y}_{U_t}$ is $\mathcal{O}({n_t}^{-1/2})$, implying weak consistency as $n_t\to\infty$ ($t\to\infty$).

The next theorem establishes strong consistency of the Horvitz–Thompson type estimator $\hat{\bar{y}}_{HT,t}$ under hypotheses similar to the ones used by Robinson (1982).

Theorem 3.4. *Suppose Assumptions 2.1, 2.2, and 2.3. Let $\{y_i, i\geq 1\}$ be a sequence of nonnegative real numbers that satisfies Assumption 3.1. Then, as $t\to\infty$,*

$$\hat{\bar{y}}_{HT,t} - \bar{y}_{U_t} \to 0 \quad \textit{a.s.} \quad \textit{(w.r.t. p-distr.)}.$$

Proof. The proof follows by application of Theorem 3.2 (with $a_i \equiv n$); hence, it is sufficient to check that the hypotheses of Theorem 3.2 are satisfied. Define the r.v. $Y_i = y_i\mathbb{1}_{i\in s_t}/\pi_{i,t}$, for all $i\in U_t$, let $\hat{T}_{N_t} = \sum_{i\leq N_t} Y_i$, and observe that by Assumption 3.1 there is a constant C such that

$$\mathbb{E}_p[\hat{T}_{N_t}] = \sum_{i=1}^{N_t}\mathbb{E}_p[Y_i] = \sum_{i=1}^{N_t} y_i \leq \sum_{i=1}^{N_t} y_i^2 = N_t\underbrace{\frac{1}{N_t}\sum_{i=1}^{N_t} y_i^2}_{\leq C} = \mathcal{O}(N_t),$$

hence, hypothesis (i) is satisfied. By Assumption 2.3, we have (ii) $\mathbb{E}_p[Y_iY_j] \leq \mathbb{E}_p[Y_i]\mathbb{E}_p[Y_j]$ for all $i, j \in U_t$. Finally, we have to check hypothesis (iii) of Theorem 3.2. To this end, note that by Assumptions 2.1, 2.2, and 3.1,

$$\begin{aligned}\frac{1}{N_t^2}\sum_{i=1}^{N_t}\mathbb{V}_p[Y_i] &= \frac{1}{N_t^2}\sum_{i=1}^{N_t} y_i^2\left(\frac{1-\pi_{i,t}}{\pi_{i,t}}\right) \leq \sum_{i=1}^{N_t}\frac{y_i^2}{\pi_{i,t}} \\ &\leq \Bigg[\underbrace{\min_{1\leq i\leq N_t}(\pi_{i,t})}_{\geq\lambda}\Bigg]^{-1} \underbrace{\frac{n_t}{N_t}}_{\to f}\frac{1}{n_t}\underbrace{\left(\frac{1}{N_t}\sum_{i=1}^{N_t} y_i^2\right)}_{\leq C} = \mathcal{O}(n_t^{-1}),\end{aligned} \tag{3.28}$$

hence, all hypothesis of Theorem 3.2 are satisfied, which implies the result. ■

Remark. In contrast to Theorem 2 in Robinson (1982), Theorem 3.4 requires $\{y_i, i \geq 1\}$ to be a sequence of *nonnegative* real numbers. With this additional assumption, however, we obtain strong instead of weak consistency. The restriction to non-negativity does not limit the applicability of the result in any noticeable manner as population totals are only meaningful population characteristics in conjunction with nonnegative data.

3.3.2. Ratio estimator

Let $\{x_i, i \geq 1\}$ denote a sequence of nonnegative real numbers that are known for all $i \in U_t$; hence, the population x-mean, $\bar{x}_{U_t}$, is a known quantity. We shall assume that the population-level relationship between y_i and x_i can be *approximated* by the heteroscedastic model ξ: $y_i = x_i\beta + e_i$, $i \in U_t$, where $\mathbb{E}_\xi[e_i \mid x_i] = 0$ and

$$\mathbb{E}_\xi[e_ie_j \mid x_i, x_j] = \begin{cases} x_i\sigma^2 & \text{if } i = j \\ 0 & \text{otherwise.}\end{cases} \tag{3.29}$$

The parameters $\beta \in \mathbb{R}^+$ and $\sigma^2 \in \mathbb{R}^+$ are supposed unknown. Note that model ξ is merely motivated as an *assisting model*. In this context, the ratio estimator of the population y-mean,

$$\bar{y}_{\text{rat},t} = \bar{x}_{U_t}\hat{\beta}_t, \quad \text{where} \quad \hat{\beta}_t = \hat{T}_{y,t}/\hat{T}_{x,t}, \tag{3.30}$$

is among survey statisticians' preferred estimators (although not optimal in terms of efficiency when model ξ is the true data generating mechanism). However, $\bar{y}_{\text{rat},t}$ proves to be more efficient as an estimator of $\bar{y}_{U_t}$ than the HT-type estimator, $\hat{T}_{y,t}/N_t$, in a large number of applications.

Now, it is important to note that both model ξ and the sampling design induce separate stochastic behavior. It is thus natural to study strong consistency of the estimator $\hat{\beta}_t$ for β (more precisely, the sequence $\{\hat{\beta}_t, t \geq 1\}$) under the *compound* design–model distribution (abbreviated by ξp-distr.). Further, we in-

troduce another assumption.

Assumption 3.2. *Let $\{x_i, i \geq 1\}$ be a sequence of positive real numbers such that*

$$T_{x,t} = \sum_{i=1}^{N_t} x_i \to \infty \quad \textit{as} \quad t \to \infty \qquad \textit{and} \qquad \sum_{t=1}^{\infty} \left(\frac{x_t}{T_{x,t}} \right)^2 < \infty.$$

Remark. The "critical" part of Assumption 3.2 is whether $(x_t/T_{x,t})^2$ is summable. Suppose that the x_i's are bounded by some constant $c > 0$; then it is easily seen from

$$\begin{aligned} \sum_{t=1}^{\infty} \frac{x_t^2}{\left(\sum_{i=1}^{t} x_i\right)^2} \leq c \sum_{t=1}^{\infty} \frac{1}{\left(\sum_{i=1}^{t} x_i\right)^2} \leq c \sum_{t=1}^{\infty} \frac{1}{\left(t \min_{1 \leq i \leq t}(x_i)\right)^2} \\ \leq \frac{c}{\min_{i \geq 1}(x_i^2)} \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty \end{aligned}$$

that the Assumption holds. However, the assumption is also satisfied in case the x_i's are allowed to grow with $i = 1, 2, \ldots$ (although not too fast).

It is straightforward to show that $\bar{y}_{\text{rat},t}$ is a strongly consistent estimator of $\bar{y}_{U_t}$ in the above asymptotic framework. Indeed, Theorem 3.5 extends some of the results in Robinson and Särndal (1983) under the hypothesis of nonnegative r.v.'s. Notably, our assumptions are weaker and in particular we do not require their asymptotic uncorrelatedness assumption (see Robinson and Särndal, 1983, Assumption A5),

$$\lim_{t \to \infty} \max_{1 \leq i \neq j \leq N_t} |\pi_{ij,t}/(\pi_{i,t}\pi_{j,t}) - 1| = 0.$$

Theorem 3.5. *Suppose Assumptions 2.1, 2.2, and 2.3. Let the sequences of populations and samples be as described. Let $\{x_i, i \geq 1\}$ denote a sequence of nonnegative real numbers that satisfies Assumption 3.2.*

(i) Suppose model ξ, let $\hat{\beta}$ be defined in (3.30), and let $\boldsymbol{x}_{N_t} = \{x_1, \ldots, x_{N_t}\}$. Then, we have conditional on $\boldsymbol{x}_{N_t}$, as $t \to \infty$,

$$\left(\hat{\beta} - \beta\right) \mid \boldsymbol{x}_{N_t} \to 0 \textit{ a.s.},$$

hence,

$$\hat{\bar{y}}_{rat,t} \to \bar{y}_{U_t} \quad \textit{a.s.}$$

(ii) If $\{y_i, i \geq 1\}$ is a sequence of nonnegative real numbers that satisfies Assumption 3.1, then unconditionally (i.e., regardless of whether model ξ holds),

$$\hat{\bar{y}}_{rat,t} - \bar{y}_{U_t} \to 0 \quad \textit{a.s.} \quad \textit{for} \quad t \to \infty.$$

Proof. Part (i). Define the r.v. $X_i = x_i \mathbb{1}_{i \in s_t}/\pi_{i,t}$, where the x_i's are positive real numbers. Let $\hat{T}_{x,k} = \sum_{i \leq k} X_i$ and observe that $\mathbb{E}_p[\hat{T}_{x,k}] = T_{x,k} = \sum_{i \leq k} x_i$. Also note that $\mathbb{V}_p[X_i] = (1/\pi_{i,t} - 1)\, x_i^2$ for all $i \in U_t$. By Assumptions 2.1, 2.2, 2.3 [Part i], and 3.2, we have

$$\lim_{t\to\infty} \sum_{k=1}^{N_t} \frac{\mathbb{V}_p[X_k]}{(\mathbb{E}_p \hat{T}_{x,k})^2} = \lim_{t\to\infty} \sum_{k=1}^{N_t} \left(\frac{1}{\pi_{k,t}} - 1\right)\left(\frac{x_k}{T_{x,k}}\right)^2 \leq \lim_{t\to\infty} \sum_{k=1}^{N_t} \frac{1}{\pi_{k,t}}\left(\frac{x_k}{T_{x,k}}\right)^2$$
$$\leq \underbrace{\left[\lim_{t\to\infty} \min_{i\in U_t}(\pi_{i,t})\right]^{-1}}_{\geq \lambda} \underbrace{\lim_{t\to\infty} \sum_{k=1}^{N_t} \left(\frac{x_k}{T_{x,k}}\right)^2}_{<\infty} < \infty. \tag{3.31}$$

Moreover, Assumption 2.3 [Part ii] ensures that $\mathrm{Cov}_p[X_i, X_j] \leq 0$ for all $i, j \in U_t$, $i \neq j$. Hence, this together with (3.31) implies the hypotheses of Corollary 3.2, from which we conclude that (as $t \to \infty$)

$$\frac{\hat{T}_{x,t}}{T_{x,t}} \to 1 \quad \text{a.s.} \qquad \text{(w.r.t. } p\text{-distribution)}. \tag{3.32}$$

In view of (3.32), and observing that

$$\hat{\bar{y}}_{rat,t} = \hat{\bar{y}}_{HT,t}\left(\frac{T_{x,t}}{\hat{T}_{x,t}}\right),$$

it is easy to see that $\hat{\bar{y}}_{rat,t} - \bar{y}_{U_t} \to 0$ a.s. as $t \to \infty$ provided that $\hat{\bar{y}}_{HT,t}$ is a strongly consistent estimator for $\bar{y}_{U_t}$. This is indeed the case under the hypotheses of Theorem 3.4; hence, this concludes the proof of Part (i). The proof of Part (ii) is analogous to the proof of the first part; hence, it is omitted. ■

3.3.3. Class of QR predictors

Under the superpopulation model ξ and the additional assumption that the variance specification takes the form $v(x_i) = x_i$, we obtained the ratio estimator. However, the model provides the basis for other candidate estimators of the y-mean. Of particular importance is the class of ***QR predictors*** due to Wright (1983). *Note:* we write "predictor" when we refer to the device to compute the y-mean when the underlying parameters, e.g. β, are known; the term "estimator", on the other hand, is used to mean the estimation device based on sample-based parameter estimates.

Let $r_i \geq 0$ and $q_i > 0$ be known real numbers for all $i \in U_t$. The class of ***QR predictors*** of the y-mean is defined in terms of the choice of the tuple (r_i, q_i), and is given by

$$\bar{Y}_{QR} = \frac{1}{N} \sum_{i\in s_t} r_i(y_i - x_i\beta) + \sum_{i\in U_t} x_i\beta, \tag{3.33}$$

where

$$\beta = \left(\sum_{i \in U_t} \pi_i q_i x_i^2 \right)^{-1} \sum_{i \in U_t} \pi_i q_i x_i Y_i. \tag{3.34}$$

The predictor $\bar{Y}_{QR}$ can be seen as generalization of $\bar{Y}_{GD}$ in (3.2). The two predictors are equal if $r_i = 1/\pi_i$ for all $i \in U_t$. Note that $\bar{Y}_{QR}$ is not a statistic as it depends on the population-level parameter β which is unknown (since the y_i, $i \in U_t \setminus s_t$, are unobserved). $\bar{Y}_{QR}$ will thus be regarded as a r.v. w.r.t. the ξp-distribution. If we replace the unknown β in (3.33) by a sample-based estimator

$$\hat{\beta} = \left(\sum_{i \in s_t} q_i x_i^2 \right)^{-1} \sum_{i \in s_t} q_i x_i Y_i, \tag{3.35}$$

we obtain the ***QR estimator*** of the population y-mean,

$$\hat{\bar{y}}_{QR} = \frac{1}{N} \sum_{i \in s_t} r_i (y_i - x_i \hat{\beta}) + \sum_{i \in U_t} x_i \hat{\beta}. \tag{3.36}$$

The class of QR estimators under the ratio model ξ includes a large number of estimators / predictors that have been proposed in the literature, ranging from purely model-based predictors ($\hat{\bar{y}}_{BLU}$; BLU: best linear unbiased predictor) to the classical ratio estimator; see Table 3.1.

Table 3.1.: Estimators under the ratio model ξ

Estimator / predictor	q_i	r_i	proposed by
$\hat{\bar{y}}_{HTR} = \hat{\beta}_R \bar{x}_{U_t}$	$1/(\pi_i x_i)$	0	Hajek (1971)
$\hat{\bar{y}}_{CR} = \bar{y}_\pi + \hat{\beta}_{CR}(\bar{x}_{U_t} - \bar{x}_\pi)$	$1/\pi_i$	$1/\pi_i$	cf. Cochran (1977)
$\hat{\bar{y}}_{GR} = \bar{y}_\pi + \hat{\beta}_{BLU}(\bar{x}_{U_t} - \bar{x}_\pi)$	$1/v(x_i)$	1	Cassel et al. (1976)
$\hat{\bar{y}}_{BLU} = f\bar{y}_{s_t} + (1-f)\hat{\beta}_{BLU}\bar{x}_{\not s}$	$1/v(x_i)$	1	Royall (1970)
$\hat{\bar{y}}_{BR} = f\bar{y}_{s_t} + (1-f)\hat{\beta}_{BLU}\bar{x}_{\not s}$	$(1-\pi_i)/(\pi_i x_i)$	1	Brewer (1979)

where

$\hat{\beta}_R = \bar{y}_\pi \,/\, \bar{x}_\pi$,

$\hat{\beta}_{CR} = \sum_{i \in s_t} x_i Y_i/\pi_i \,/\, \sum_{i \in s_t} x_i^2/\pi_i$,

$\hat{\beta}_{BLU} = \sum_{i \in s_t} Y_i x_i / v(x_i) \,/\, \sum_{i \in s_t} x_i^2 / v(x_i)$,

$\hat{\beta}_{BR} = \sum_{i \in s_t} a_i x_i Y_i \,/\, \sum_{i \in s_t} a_i x_i^2$,

and

$a_i = (1-\pi_i)/(\pi_i x_i)$, $\quad (\bar{y}_\pi, \bar{x}_\pi) = \sum_{i \in s_t} (y_i, x_i)/(\pi_i N)$, $\qquad \bar{x}_{\not s} = \sum_{i \in U_t \setminus s_t} x_i/(N-n)$,
$\bar{y}_{s_t} = \sum_{i \in s_t} y_i/N$, $\quad f = n/N$.

See Wright (1983, 880).

ADUness of the QR estimator

Consider the asymptotic framework laid out above. Define the random vector under model ξ by $\mathbf{Y}_{N_t} = (Y_1, \ldots, Y_{N_t})^T$; the realizations of which are denoted by $\mathbf{y}_{N_t} = (y_1, \ldots, y_{N_t})^T$. We follow the argument of Wright (1983, 880–1), who provides sufficient conditions for ADUness of $\hat{\bar{y}}_{QR}$ under the additional assumption that $\bar{Y}_{QR}$ is exactly p-unbiased, i.e.

$$\mathbb{E}_p\big(\bar{Y}_{QR} - \bar{y}_{U_t} \,|\, \mathbf{Y}_{N_t} = \mathbf{y}_{N_t}\big) = 0 \quad \forall t = 1, 2, \ldots. \tag{3.37}$$

This additional p-unbiasedness assumption on $\bar{Y}_{QR}$ is not necessary but simplifies matters (without any relevant loss of generality). The following lemma, which derives from Theorem 1 in Wright (1983) provides sufficient conditions under which (3.37) holds. Since this result will play a role in the further course of our disposition, we formulated it as a lemma for the sake for referencing.

Lemma 3.2 (implied by Wright, 1983)**.** *Consider $\bar{Y}_{QR}$ in (3.33). If either*

(i) $r_i = 1/\pi_i$ *for all* $i \in U_t$*, or*

(ii) the choice (r_i, q_i) *obeys, for all* $i \in U_t$*,*

$$\frac{1 - \pi_i r_i}{\pi_i q_i} \propto x_i,$$

then $\bar{Y}_{QR}$ is p-unbiased for $\bar{y}_{U_t}$.

Proof. Observe that

$$\mathbb{E}_p(T_{QR}) = \frac{1}{N_t}\left[\sum_{i\in U_t} \pi_i r_i Y_i + \sum_{i\in U_t}(1 - \pi_i r_i) x_i \left(\sum_{i\in U_t} \pi_i q_i x_i^2\right)^{-1} \sum_{i\in U_t} \pi_i q_i x_i Y_i\right], \tag{3.38}$$

which must be equal to $\bar{y}_N$ for p-unbiasedness. This is achieved for any choice of (r_i, q_i) provided that the identity

$$\frac{1}{N_t}\pi_i r_i + \frac{1}{N}\sum_{i\in U_t}(1 - \pi_i r_i) x_i \left(\sum_{i\in U_t} \pi_i q_i x_i^2\right)^{-1} \pi_i q_i x_i = \frac{1}{N} \tag{3.39}$$

holds for all $i \in U_t$. From this, the assertion follows. ∎

Remark. Under hypothesis (i) of the above lemma, $\bar{Y}_{QR}$ is p-unbiased regardless of the choice of q_i. The choice of q_i, however, may matter in small sample applications. The corresponding QR predictor is then equal to the generalized regression (GREG) predictor. Conversely, every GREG-type predictor is ADU by construction. Note that we do distinguish hypothesis (i) from (ii) for didactic reasons, although this is not necessary as (i) is a special case of (ii).

Next, suppose that the above lemma holds for $\bar{Y}_{QR}$, i.e. that $\bar{Y}_{QR}$ is p-unbiased for $\bar{y}_{U_t}$. Hence, if we manage to show that

$$\lim_{t\to\infty} \mathbb{E}_p\big(\hat{\bar{y}}_{QR} - \bar{Y}_{QR}\big) = 0, \tag{3.40}$$

then $\hat{\bar{y}}_{QR}$ is ADU for $\bar{y}_{U_t}$ [Wright (1983) uses the asymptotic framework of Brewer (1979)]. Now, observe that

$$\hat{\bar{y}}_{QR} - \bar{Y}_{QR} = \underbrace{\left(\bar{x}_{U_t} - \frac{1}{N}\sum_{i\in s_t} r_i x_i\right)}_{=A,\text{ say}} \big(\hat{\beta} - \beta\big) \tag{3.41}$$

from which we see that condition (3.40) obtains if $\hat{\beta}-\beta \xrightarrow{\xi\text{p}} 0$ [respectively, almost surely] as $t \to \infty$ and provided that $|A|$ is bounded away from infinity. This approach ensures that (3.40) holds without having to impose any restrictions on A (and notably on the choice of r_i). The superscript "ξp" in the probability limit of $\hat{\beta}$ refers to convergence "in probability" w.r.t. the compound ξp-distribution. Wright (1983, 880) argues that mild restrictions on the sampling design and the moments of x_i and Y_i ensure this probability limit. Therefore ADUness of $\hat{\bar{y}}_{QR}$ essentially depends on the choice of (r_i, q_i).

Moreover, Wright (1983) proves in his Theorem 2 that any QR estimator which is ADU is identical to the GREG estimator that uses the same set of q_i's. This remarkable result implies that nothing is gained from any choice other than $r_i = 1/\pi_i$ except possibly for computational reasons. The latter point has been raised by Särndal and Wright (1984) who discusses estimators in "cosmetic form" (e.g. simple projection vs. linear prediction form; see also Brewer, 1979). More important, Theorem 2 of Wright (1983) implies that variance estimation for any QR estimator which is ADU can be obtained – in large samples – simply in the same way it is recommended for generalized regression estimation; see Särndal and Wright (1984, sec. 4) and Särndal et al. (1992, ch 6.6).

Main result

Under model ξ, the estimators $\hat{\bar{y}}_{HTR}$, $\hat{\bar{y}}_{CR}$, $\hat{\bar{y}}_{GR}$, and $\hat{\bar{y}}_{BR}$ (see Table 3.1) each satisfy the ADU condition for any sampling design and any x_i. However, $\hat{\bar{y}}_{BLU}$ is ADU if and only if $v(x_i)(1/\pi_i - 1)$ is a multiple of x_i.

The focus of Wright (1983) is on ADUness and this implies that he mainly builds on (3.40). However, as he points out, emphasis of ADUness is somewhat arbitrary and can be replaced by design consistency. Indeed, instead of (3.40) we may require that

$$\lim_{t\to\infty} \mathbb{P}_p\big\{\bar{Y}_{QR} - \hat{\bar{y}}_{QR}\big\} = 0 \tag{3.42}$$

holds, which obtains from the argument we used in (3.41), and provided that $\hat{\beta} - \beta \xrightarrow{\xi\text{p}} 0$ as $t \to \infty$. Eventually, design consistency of $\hat{\bar{y}}_{QR}$ for $\bar{y}_{U_t}$ follows from

Lemma 3.2. Since our interest is on strong consistency, we focus on

$$\mathbb{P}_p\big\{ \lim_{t\to\infty} \hat{\bar{y}}_{QR} = \bar{y}_{U_t} \big\} = 1 \tag{3.43}$$

instead of (3.42). The rest of our argument, however, will follow the lines of Wright (1983). To this end we restrict attention to situations where Lemma 3.2 applies. This will provide us with sufficient conditions (though not the most general conditions) for strong design consistency to obtain. Before giving the result, we do have to discuss some technical details. Observe that the class of QR estimators offers a certain degree of choice in terms of the tuple (r_i, q_i). In order to ensure that our results apply for all candidate QR-estimators, we impose some conditions on the behavior of the sequence of nonnegative real numbers, $\{x_i, i \geq 1\}$, given a particular choice of q_i. Now, suppose a sequence of random variables w.r.t. p-distribution $\{X_i, i \geq 1\}$, specified as

$$X_i = \mathbb{1}_{i\in s_t} q_i x_i^2, \tag{3.44}$$

where $q_i > 0$ and $x_i > 0$ are real numbers. It is easily seen that $\mathbb{E}_p(X_i) = \pi_{i,t} q_i x_i^2$ for all $i \in U_t$. Likewise we obtain, for $i \in U_t$,

$$\mathbb{V}_p(X_i) = \mathbb{E}_p\big(X_i^2\big) - \mathbb{E}_p\big(X_i\big)^2 = \pi_{i,t}\big(1 - \pi_{i,t}\big) q_i^2 x_i^4, \tag{3.45}$$

and for $i, j \in U_t$, $i \neq j$,

$$\operatorname{cov}_p\big(X_i, X_j\big) = \mathbb{E}_p\big(X_i X_j\big) - \mathbb{E}_p(X_i)\mathbb{E}_p(X_j) = \big(\pi_{ij,t} - \pi_{i,t}\pi_{j,t}\big) q_i q_j x_i^2 x_j^2. \tag{3.46}$$

By Assumption 2.3 (ii) and the hypotheses that $x_i > 0$ and $q_i > 0$, $\operatorname{cov}_p\big(X_i, X_j\big)$ is nonpositive for all $i, j \in U_t$ and all $t = 1, 2, \ldots$. With regard to our definition of $\{X_i, i \geq 1\}$ in (3.44), Assumption 3.3 can be expressed as

Assumption 3.3. *Consider the sequence* $\{X_i, i \geq 1\}$ *defined in (3.44), where* $q_i > 0$ *and* $x_i > 0$ *are real numbers and* $0 < x_i \uparrow \infty$*, such that*

$$\sum_{i\leq N_t} \mathbb{E}_p(X_i) \to \infty \qquad \textit{and} \qquad \frac{\mathbb{E}_p(X_{N_t})}{\sum_{i\leq N_t} \mathbb{E}_p(X_i)} \to 0 \qquad \textit{as} \quad t \to \infty \tag{3.47}$$

Remarks. In general, q_i can be chosen to take any nonnegative real value as long as Assumption 3.3 is satisfied. For several of the QR estimators, Assumption 3.3 simplifies to some extent, such that we get more familiar expressions in (3.47). For instance, in case of the ratio estimator, we have $q_i = 1/(\pi_{i,t} x_i)$ for all $i \in U_t$, hence $\{x_i, i \geq 1\}$ must satisfy $\sum_{i\in U_t} x_i \to \infty$ and $x_i / \sum_{i\in U_t} x_i \to 0$ as $t \to \infty$.

In what follows, we adhere to the notation

$$\begin{aligned} S_{q_i x_i^2} &= \sum_{i \in U_t} \pi_i q_i x_i^2, & \hat{S}_{q_i x_i^2} &= \sum_{i \in U_t} \mathbb{1}_{i \in s_t} q_i x_i^2, \\ S_{q_i x_i Y_i} &= \sum_{i \in U_t} \pi_i q_i x_i Y_i, & \hat{S}_{q_i x_i Y_i} &= \sum_{i \in U_t} \mathbb{1}_{i \in s_t} q_i x_i Y_i. \end{aligned} \tag{3.48}$$

With this, $\hat{\beta}$ in (3.35) writes as a ratio of sample totals,

$$\hat{\beta} = \hat{S}_{q_i x_i Y_i} / \hat{S}_{q_i x_i^2},$$

By a first-order Taylor series approximation of this ratio around $(S_{q_i x_i Y_i}, S_{q_i x_i^2})$, we have

$$\hat{\beta} - \beta = \frac{1}{S_{q_i x_i^2}} \left(\hat{S}_{q_i x_i Y_i} - \beta \hat{S}_{q_i x_i^2} \right) + R, \tag{3.49}$$

with

$$R = \left(\hat{\beta} - \beta \right) \left(1 - \frac{\hat{S}_{q_i x_i^2}}{S_{q_i x_i^2}} \right). \tag{3.50}$$

From (3.49), it is straightforward to show that $\hat{\beta} - \beta \xrightarrow{\xi \mathrm{p}} 0$ as $t \to \infty$ [respectively, almost survey]; this implies (3.43); hence, strong consistency of the QR estimator obtains.

3.4. Summary and discussion

It is well-known that most of the model-assisted estimators (e.g., GREG estimator) are not design consistent; however, this is no reason to worry as exact design-unbiasedness is of doubtful virtue (Särndal, 1980, 641). Under mild assumptions, the design bias of such estimator vanishes asymptotically as the sample size grows. Therefore, these estimators are said to be asymptotically design consistent (or asymptotically design unbiased, ADU). The contribution of the assisting model to the estimator vanishes as the sample size grows (irrespective whether the model holds or not). This property is an intrinsic characteristic of ADU estimators and has become the cornerstone of the model-assisted sampling paradigm. C.E. Särndal and some of his coauthors argue that randomization can be seen as a source of *robustness* against model failure; see Särndal (1980, 641). Clearly, this notion of robustness is only relevant when the sample size becomes large.

We have discussed the notion of robustness / consistency due to C.E. Särndal (and his coauthors) and introduced the idea of *strongly* design consistent estimators. Surprisingly, we require only mild regularity assumptions on the design and the behavior of the auxiliary variables x_i and, this is important, the additional assumption that the study variable y_i is nonnegative. The restriction to nonnegative y_i's is quite natural in the context of ratio estimation of the mean

or total. Our results contribute to the understanding of ratio estimators and the class of QR-estimators due to Wright (1983).

4. Robust model-assisted estimation under the linear model

4.1. Introduction

One may come to the conclusion that under the linear assisting model (almost) "every" conceivable robust method of some importance has already been developed, discussed and applied in practical applications. There is indeed a huge body of literature which supports this impression. It is neither our goal nor is it truly possible to do full justice to such a huge subject in such a limited space. For these reasons, we limit ourselves to review only the most important publications.

To fix notation, let U denote a finite population of size $N > 0$; s is a sample of size $0 < n < N$ that has been drawn from U using some sampling design $p(s)$ [these notions will be made more precise later]. The goal is to estimate the population total or mean of variable y_i whose values are assumed unknown for $i \in U$.

Review of the literature

Early outlier resistant finite-population estimators of the y-total did not consider the incorporation of auxiliary information at the estimation stage; see e.g. Searls (1966) or Fuller (1991). We do not discuss such early procedures; instead, we refer the reader to the review article of Beaumont and Rivest (2009).

For ease of discussion, we consider only a scalar-valued auxiliary variable $x_i \in \mathbb{R}^+$ which is known for all population elements $i \in U$, and is available at the design stage. Some estimators incorporate the auxiliary information already at the design stage; other estimators use the auxiliary information only at the estimation stage.

Under the ratio superpopulation model,

$$\xi : \begin{cases} y_i = x_i\beta + e_i & \text{for all } i \in U,\ \beta \in \mathbb{R}^+, \\ \mathbb{E}_\xi(e_i) = 0 \qquad \text{and} \quad \mathbb{V}_\xi(e_i) = \sigma^2 h(x_i) & \text{for all } i \in U,\ \sigma \in \mathbb{R}^+, \end{cases}$$

where $h : \mathbb{R}^+ \to \mathbb{R}^+$ is a known function, and the additional assumption that

$$\mathbb{E}_\xi(e_i e_j) = 0 \qquad \text{for all } i \neq j, \quad i, j \in U,$$

Chambers's (1986) robust bias-calibrated (or bias-corrected) estimator is given

by

$$\hat{Y}^C = \sum_{i \in s} y_i + \sum_{i \in U \setminus s} x_i \hat{\beta} + \underbrace{\left[\sum_{i \in U \setminus s} \sqrt{h(x_i)} \right] \frac{1}{n} \sum_{i \in s} \hat{\sigma} \psi_k \left(\frac{y_i - x_i \hat{\beta}}{\hat{\sigma} \sqrt{h(x_i)}} \right)}_{\text{bias correction term}}$$

where $\hat{\beta}$ and $\hat{\sigma}$ denote, respectively, an outlier resistant estimator of β and σ; the expression of $\hat{Y}^C$ is due to Welsh and Ronchetti (1998, 423). Estimator $\hat{Y}^C$ is by construction a *model-based estimator*. As a consequence, the reliability of the estimator depends heavily on the validity of model ξ. Chambers (1986) also derived formulas of the asymptotic variance and asymptotic bias of $\hat{Y}^C$ under an outlier-prone, but symmetric superpopulation model. The tuning constant k of the Huber ψ-function ψ_k (see bias correction term) does not have to be related to the computation of the outlier resistant estimate $\hat{\beta}$. However, one typically chooses a large value for k when the population y-values are assumed to have a relatively high variability (i.e., contain a substantive amount of representative outliers).

Under model ξ [but without the orthogonality assumption, $\mathbb{E}_\xi(e_i e_j) = 0$], Gwet and Rivest (1992) proposed a robust *model-assisted* ratio estimator of the population y-mean for simple random sampling. Their estimator [under the assumption that $h(x_i) \equiv x_i$] is given by

$$\hat{\bar{y}} = \hat{\beta} \bar{x},$$

where $\bar{x} = (1/N) \sum_{i \in U} x_i$, and $\hat{\beta}$ is implicitly defined by the estimating equation

$$\sum_{i \in s} \sqrt{x_i} \psi_k \left(\frac{y_i - x_i \hat{\beta}}{\hat{\sigma} \sqrt{x_i}} \right) = 0, \tag{4.1}$$

$\hat{\sigma}$ denoting a robust estimate of the superpopulation parameter σ [in fact, Gwet and Rivest (1992, 1175) have suggested a Schweppe-type GM-estimator, not the M-estimator in (4.1)].

Gwet and Rivest (1992) present evidence that their robust estimator is more efficient than the ordinary ratio estimator in the presence of representative outliers. Notably, when the distribution of the error terms e_i (see model ξ) is symmetric about zero and has fatter tails than the Gaussian distribution, the robust estimator will be – as a rule – more efficient (irrespective of the sample size). Under the assumption of a skewed population distribution of the y_i's, the situation is more intricate. For small sample sizes, the design mean squared error (MSE) of the estimator is dominated by the variance, and robustification can lead to considerable efficiency gains over the ordinary ratio estimator. However, the gains in efficiency decrease as the sample size grows because the MSE tends to become dominated by the estimator's bias.

Gwet and Rivest (1992) studied robust ratio estimation only for simple random sampling. Hulliger (1995) considered the robustification problem related to model ξ when the auxiliary variable x_i is utilized at the design stage and is

also incorporated into a model-assisted estimation strategy. In particular, he investigated the PPS design, where the size variable is x_i; the resulting estimator $X\hat{\beta}$, X denoting the population x-total, is a robust Horvitz–Thompson (RHT) estimator. The estimate $\hat{\beta}$ in the proposal of Hulliger (1995, 81) is a Mallows type GM-estimate; and the (normalized) median of the absolute deviations of the residuals about zero is used to estimate the superpopulation parameter σ. Hulliger (1995) developed formulae for the asymptotic variance and asymptotic bias of the RHT estimator.

In relatively large samples, the bias of estimator RHT (more generally, any robust estimator) can be substantial and may thus render the robust procedure grossly inefficient. As a remedy, Hulliger (1995) suggested a solution to the inconsistency problem of robust estimators insofar that he considers a set of eligible estimators which includes the non-robust, but consistent estimator (i.e. the Horvitz–Thompson estimator). The method is called minimum estimated risk (MER) estimator and is an adaptive procedure. The key to his proposal is the allowance for the consistent estimator as this choice ensures that the overall procedure is consistent.

Gwet (1997, chap. 3.3) has generalized the M- and GM-estimators under the ratio model to the full regression framework (i.e. generalized regression estimator, GREG). In addition, he introduced two new classes of estimators for skewed populations that deal with the bias incurred through robustification [see his chap. 3.4]. The last contribution is due to Beaumont and Alavi (2004); it is to a great extent a survey of existing robust methods.

Open questions and contribution

As we have pointed out, M- and GM-estimators have been studied in the context of finite-population sampling for quite some time. Robust GREG type estimation *for domains*, on the other hand, has not been thoroughly investigated. Therefore, we shall study robust Horvitz–Thompson and Hajek type GREG estimators. Our contribution is motivated by the work of Hidiroglou and Patak (2004) who came up with a list of (non-robust) GREG estimators for domains.

In this chapter, little emphasis is placed on asymptotic and other mathematical properties of the proposed methods. Existence results for robust ratio and robust GREG-type estimators, including the estimator's asymptotic behavior under a great variety of regularity conditions, are discussed in the Ph.D. theses of Hulliger (1991) and Gwet (1997). Their results are also valid (with minor adaptations) for our robust estimators.

Outline of the chapter

The remainder of this chapter is organized as follows. In Sections 4.2 and 4.3, we define population, sampling design, domain structure, and give a short re-

view of GREG estimation. The suggested robust domain estimators are discussed in Section 4.4, followed by short note on variance estimation (see Section 4.5). Finally, the major findings are summarized in Section 4.6.

4.2. Preliminaries, assumptions and notation

We shall assume that a sample s is drawn from the finite population U. The sampling design $p(s)$ is required to adhere to the conditions formulated in the introductory chapter (see Definition 2.2). For ease of referencing, we summarize the key points of our assumptions.

(i) all labels $i \in U$ are identifiable,

(ii) it is possible to observe and measure without error the variable(s) of interest Y_i for each sampled element $i \in s$,

(iii) the randomization scheme is non-informative,

(iv) $p(s) > 0$ for all $s \in \mathscr{S}$ (see Definition 2.2),

(v) $\pi_i > 0$ for all $i \in U$ and $\pi_{ij} > 0$ for all $i, j \in U$, $i \neq j$.

Domain structure

The population U is supposed to be partitioned into $d = 1, \ldots, D$ mutually exclusive and exhaustive domains spanning the whole population (see Definition 2.1) such that

$$U = \bigcup_{d=1}^{D} U_d,$$

where U_d denotes the set of elements that fall into domain d. The sample s features a corresponding partitioning $s = \cup_{d=1}^{D} s_d$, where $s_d = U_d \cap s$ is the part of s that falls into domain d ($d = 1, \ldots, D$). The definition implies that $N = \sum_{d=1}^{D} N_d$ and $n = \sum_{d=1}^{D} n_d$, where N_d and n_d are, respectively, the size of U_d and s_d. Following Lehtonen and Veijanen (2009, 222–23), we distinguish two types of domains :

- *planned* (or primary) domains and
- *unplanned* (or secondary) domains.

Unplanned domains (and thus unplanned domain structures) occur if the information of domain membership is *not incorporated* into the sampling design. Hence, the domain structure may cut across a partitioning induced by the sample design. An immediate consequence is that the n_d's will be random for all $d = 1, \ldots, D$. The random nature of the n_d's has implications for variance estimation, but apart from this it does not pose any further difficulties. As a rule, the random nature of the n_d's tends to increase the variance of the estimators.

Domain totals

The population and domain-specific x-totals are defined as, respectively,

$$\mathbf{X} = \sum_{i \in U} \mathbf{x}_i, \qquad \mathbf{X}_d = \sum_{i \in U_d} \mathbf{x}_i, \qquad d = 1, \ldots, D,$$

and the corresponding Horvitz–Thompson estimators are given by, respectively,

$$\hat{\mathbf{X}}^{HT} = \sum_{i \in s} w_i \mathbf{x}_i, \qquad \hat{\mathbf{X}}_d^{HT} = \sum_{i \in s_d} w_i \mathbf{x}_i,$$

where $w_i = 1/\pi_i$; likewise, we write $\hat{Y}_d^{HT}$ to mean the HT estimator of the population y-total in domain d. The Hajek (HJ) estimators of the population and domain x-total are given by, respectively,

$$\hat{\mathbf{X}}^{HJ} = \frac{N}{\hat{N}} \sum_{i \in s} w_i \mathbf{x}_i, \qquad \hat{\mathbf{X}}_d^{HJ} = \frac{N_d}{\hat{N}_d} \sum_{i \in s_d} w_i \mathbf{x}_i, \qquad d = 1, \ldots, D,$$

where

$$\hat{N} = \sum_{i \in s} w_i \qquad \hat{N}_d = \sum_{i \in s_d} w_i.$$

The Hajek y-totals are defined in the same manner. Since the HJ estimator forms a ratio of two random variables (w.r.t. the randomization distribution), it is (in general) only approximately unbiased. As a consequence, we cannot derive exact variance expressions (except for some special cases); see Särndal et al. (1992, Result 5.7.1). Nevertheless, the HJ estimator proves to be superior in terms of variability even when N (resp. N_d) is known. In particular, under variable-size sampling designs or poor correlation between the y_i's and the π_i's, the HJ estimation strategy tends to be better in terms of variance than the HT estimator; see Särndal et al. (1992, 183–84).

4.3. Review of GREG estimation theory

Before we begin with the discussion of robust domain-specific GREG estimators, it will prove useful to review the most important properties of ordinary GREG estimation theory.

Following Estevao, Hidiroglou, and Särndal (1995), GREG estimation is best understood in terms of the three concepts *model level*, *model group*, and *model type* which shall be introduced subsequently.

Model level

The notion of model level relates to the *type of unit* used in the model formulation. We say a model is defined at the elemental level if it is expressed in terms of auxiliary data on individual or within-domain elements (Estevao et al., 1995,

185). For single-stage element designs, a model is necessarily specified at the elemental level. If one considers single-stage cluster designs, the model may be formulated for elements or for clusters of elements. In case of multistage designs, even more choices are sensible (e.g., models at the level of the primary, secondary, etc. sampling units). However, we shall restrict attention to single-stage designs.

Model group

Consider the population-level models

$$\xi_P : \begin{cases} y_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i, & i = 1, \ldots, N \\ \mathbb{E}_\xi[e_i] = 0 \quad \text{and} \quad \mathbb{V}_\xi[e_i] = \sigma^2 c_i, & i = 1, \ldots, N \end{cases}$$

and

$$\xi_D : \begin{cases} y_i = \mathbf{x}_i^T \boldsymbol{\beta}_d + e_i, & i = 1, \ldots, N_d \\ \mathbb{E}_\xi[e_i] = 0 \quad \text{and} \quad \mathbb{V}_\xi[e_i] = \sigma_d^2 c_i, & i = 1, \ldots, N_d \end{cases}$$

where $\boldsymbol{\beta}, \boldsymbol{\beta}_d \in \mathbb{R}^p$ and $\sigma, \sigma_d \in \mathbb{R}^+$ are unknown parameters; the $\mathbf{x}_i$'s denote p-vectors of known auxiliary variables. The c_i's satisfy $c_i > 0$ and are defined in relation to the variance structure; for instance, let $h : \mathbb{R}^+ \to \mathbb{R}^+$ be a known function, then one may take $c_i \equiv h(\mathbf{x}_i)$.

Estimators under model ξ_P attempt to "borrow strength" from domains other than the domain of interest (Lehtonen and Veijanen, 2009, 224); such estimators are usually called indirect estimators; likewise, estimators under model ξ_D are called direct. The models ξ_P and ξ_D represent two extremes: either a single set of parameters $(\boldsymbol{\beta}, \sigma^2)$ is assumed to describe all domains, or each domain is modelled via its own set of parameters. In fact, there is a whole spectrum of models between these two extrema. Intermediate models usually provide a more parsimonious parametrization than model ξ_D; yet, they offer much more flexibility than model ξ_P.

The notion of *model groups* is due to Särndal et al. (1992, chap. 10.7) and has been more elaborated in Estevao et al. (1995, Sect. 4). Let $\{U_j, j = 1, \ldots, J\}$ be a set of non-overlapping subsets of the population such that $U = \cup_{j=1}^J U_j$. For each *model group* U_j, $j = 1, \ldots, J$, a separate model $y_i = \mathbf{x}_i^T \boldsymbol{\beta}_j + e_i$, $i \in U_j$, is specified. Usually, the type of auxiliary information used in the model group's model are the same among groups, but this is not mandatory; the auxiliary variables included in $\mathbf{x}_i$ may in fact differ from group to group. Also, when the subsets $\{U_j, j = 1, \ldots, J\}$ coincide with the partitioning $\cup_{d=1}^D U_d$ induced by the domain structure, the resulting estimator is direct. The boundaries of the sets U_j do not have to agree with domain boundaries. The minimum requirement for a partitioning to qualify as a model group is the existence of one or more auxiliary variables with known totals. Model groups may or may not be directly related to the domain of interest. By way of example, it may sometimes make

sense to use a cross-classification of age and occupation categories to define a model group for the estimation of a regionally determined domain of interest. Furthermore, the domain of interest is allowed to intersect with one or more model groups (Estevao et al., 1995, 191). This may happen, for instance, in a population of business establishments when the classification of major economic activity of a given business establishment changes over time or may only be discovered during the survey.

The notion of models groups is extremely flexible and enables the statistician to specify highly elaborate model configurations. The implementation of a model-group estimation strategy is achieved via the choice of variables present in $\mathbf{x}_i$ and the specification of the index sets $\{d = 1, \ldots, D\}$ (domains) and $\{j = 1, \ldots, J\}$ (model groups). Beyond this, model groups do not pose any further difficulties. For the sake of simplicity, we shall assume throughout the further course of discussion that the partitioning induced by the domain structure falls together with the one generated by the model groups.

Model type

The auxiliary variables determine the model type (Estevao et al., 1995, 185). We distinguish the following special types (for $i \in U$):

(i) common mean model, where $\mathbf{x}_i \equiv 1$ and $c_i \equiv 1$;

(ii) ratio model, where $\mathbf{x}_i \equiv x_i$ with $x_i > 0$ and $c_i \equiv x_i$;

(iii) simple regression model, where $\mathbf{x}_i \equiv (1, x_i)^T$ and $c_i \equiv 1$.

It is well known that for fixed-size sampling designs, the HT estimator of the domain mean, $(1/N_d) \sum_{i \in s_d} y_i/\pi_i$, obtains as a special case of the ratio model if we take $\mathbf{x}_i \equiv \pi_i$ and $c_i \equiv \pi_i$.

Remarks.

(i) By the linearity of the model structure, it is sufficient to know only the x-totals and x-values of the sampled elements. We do not require the $\mathbf{x}_i$'s $i \in U \setminus s$, respectively, $i \in U \setminus s_d$.

(ii) The notion of (linear) GREG estimation has been generalized in several directions. One way to generalize the linear GREG is to consider a transformation of the response variable; see e.g. Chambers and Dorfman (2003) or Karlberg (2000). Another form of generalization is to allow for study variables of different types (e.g., continuous, polytomous, binary, count), which is achieved by specifying an assisting model in the class of generalized linear models (e.g., linear, logistic, Poisson, etc.; see Lehtonen and Veijanen, 2009). Lehtonen and Veijanen (1998) were among the earliest to use

a logistic model. An even more general approach obtains when we chose our assisting models from the class of generalized mixed linear model; see Lehtonen, Särndal, and Veijanen (2003), Lehtonen, Särndal, and Veijanen (2005) and Lehtonen (2011). Yet, the design principles underlying the construction of the estimators are still those of the (linear) GREG estimator of the population y-total, namely,

$$\sum_{i \in U} \hat{y}_i + \underbrace{\sum_{i \in s} w_i (y_i - \hat{y}_i)}_{\text{correction term}}, \tag{4.2}$$

where $\hat{y}_i$ denotes the fitted values under the model. Consequently, estimators of the form (4.2) are called "extended GREG" estimators (Lehtonen and Veijanen, 2009; Lehtonen, 2011).

(iii) The correction term in (4.2), which is characteristic for GREG estimation, ensures that the estimator is asymptotically design unbiased (and asymptotically design consistent); see also our discussion in Section 3.1. Without the correction term, the estimator reduces to the synthetic or projection estimator. A sufficient condition that the correction term is zero by construction is $c_j = \boldsymbol{\lambda}^T \mathbf{x}_i$ for all $i \in U$ (resp. $i \in U_d$), where $\boldsymbol{\lambda} \in \mathbb{R}^p$ is a known constant (Särndal et al., 1992, Result 6.5.1). Two special cases when the condition holds are: (i) the variance is constant over the observations or (ii) the $\mathbf{x}_i$'s contain a constant (regression intercept).

(iv) The correction term in (4.2) may also be viewed as a nonparametric adjustment for bias caused by potential model misspecification error (Chambers et al., 1993). The size of the absolute value of the correction term relative to the size of the projection term, $\sum_{i \in U} \hat{y}_i$, can be thought of as a measure of model misspecification (Hedlin et al., 2001, 530). In view of this, the correction term is usually said to account for the potential design bias through explicitly estimating the residuals. However, the problem of model failure and bias in rather small samples is much more involved and cannot be argued away by referring to large-sample arguments such as asymptotic design unbiasedness. In the words of Hedlin et al. (2001, 543): "[i]t is just not true that GREG estimators are relatively robust to model choice".

Properties of the GREG estimator

The Horvitz-Thompson type GREG estimator of the population y-total due to Cassel et al. (1976) shall be defined as

$$\hat{Y}^{HT} = \mathbf{X}^T \hat{\boldsymbol{\beta}} + \sum_{i \in s} w_i (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}}$ is a sample-based estimate of the population fit coefficient. Estimator $\hat{Y}^{HT}$ can be expressed as a g-weighted estimator (see e.g. Särndal et al., 1992, 232),

$$\hat{Y}^{HT} = \sum_{i \in s} w_i g_i y_i,$$

where the g-weights are defined as

$$g_i = 1 + \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i \in s} \frac{w_i \mathbf{x}_i \mathbf{x}_i^T}{c_i} \right)^{-1} \frac{\mathbf{x}_i}{c_i}.$$

The g-weights satisfy the calibration property

$$\sum_{i \in s} w_i g_i \mathbf{x}_i = \mathbf{X}, \tag{4.3}$$

and play an important role for variance estimation of $\hat{Y}^{HT}$ [see below]. The product of the design weight w_i and the g-weight g_i is known as regression or calibration weight. For the HT estimator of the domain total Y_d, we have the following analogous representation,

$$\hat{Y}_d^{HT} = \sum_{i \in s} w_i g_i y_i, \qquad i = 1, \ldots, D.$$

Observe that the g-weighted sum is defined over the set of *all sampled elements* (not $i \in s_d$).

The model's raison d'être

The question of whether a model is of any use to survey sampling has been a long standing controversy (Brewer, 2013). Early textbooks (namely, the books of R. Yates, W.E. Deming, or W.G. Cochran) had in common that models did not play any relevant role. In case of the influential book due Hansen, Hurwitz, and Madow (1953), models are literally absent; the analysis of Brewer (2013, 255) shows that the words "model" and "models" do not occur in text, although Hansen and his co-authors' treatment includes a chapter on regression estimation. From today's perspective, this is rather strange; in the words of Brewer (2013, 255): "I don't see how one can have a regression estimator without a regression model, at least in the back of one's mind".

The model-free orthodoxy remained the dominant paradigm in survey statistics up until the early 1970s. The change is marked by a series of papers published by R.M. Royall (and co-authors) who reinstated purposive sampling and prediction-based inference. In particular, the paper of Royall (1970) came – in view of Brewer (2013, 256) – "as a considerable shock to the finite-population sampling establishment". Later, Royall has withdrawn the most extreme of his recommendations (see also Brewer, 1999). In the years to follow, a third

position, which combines the merits from randomization-based and prediction-based inference, received a lot of attention. Eventually, the debate has lost much of its former explosiveness and most survey statisticians now agree with the quote of Brewer (2005, 390): "[e]ach approach has its merits, and there are advantages in using both together"; see also Särndal (2011).

In robust statistics, a model is key to estimation. This is self-evident for Chambers's (1986) proposal of robust methods in finite population estimation as he refers to prediction-based inference. Yet, the model is equally important when one considers robust estimators under the randomization-based or randomization-assisted paradigm. The reason is that (parametric) robustness theory establishes an abstract notion of neighborhood of the true model and studies estimators in that neighborhood. In order to utilize this theory in a finite-population estimation context, we cannot do without explicitly formulating a model and fixing some form of neighborhood. A *prototypical* robustification approach is discussed in Hulliger (1995) and works as follows. First, B. Hulliger "uncovers" the implicit (superpopulation) model that underlies the HT estimator; then, he expresses the estimator under the model as a functional of the empirical distribution function (taking the complex nature of the sampling design into account). This intermediate step establishes a direct link to classical robust statistics and eventually leads to an M-estimator of the HT estimation strategy (see Hulliger, 1995, sec. 2).

4.4. Robust domain GREG estimators

In what follows, we do have rather strong beliefs in the (superpopulation) models (e.g. ξ_D), but are not willing to entirely rely on such assumptions. As a consequence, we derive M-estimators that are robust in the presence of outliers in the response variable. No attempt is made to limit the impact of influential observations or outliers in the auxiliary variables.

Let $\mathbf{r} = (r_1, \ldots, r_n)^T \in \mathbb{R}^n$. We define the normalized weighted median of the absolute residuals about zero by

$$\mathrm{MAD}_{\mathrm{w}}(\mathbf{r}) = 1.4826 \times \mathrm{median}_{\mathrm{w}}\big(|r_i|,\ i = 1, \ldots, n\big),$$

where $\mathrm{median}_{\mathrm{w}}$ denotes the weighted median.

By function ψ, we shall mean the Huber ψ-function or any other monotone, odd, and almost everywhere differentiable function $\mathbb{R} \to \mathbb{R}$. We write ψ_k if we wish to highlight the dependency of ψ on tuning constant k.

4.4.1. Horvitz–Thompson type estimators

Subsequently, we introduce three robust Horvitz–Thompson type GREG domain estimators of the domain y-total Y_d ($d = 1, \ldots, D$).

Estimator $\hat{Y}_{d,1}^{HT}$ is defined in terms of the g-weights (for $i \in s$)

$$g_i = v_i + \left(\mathbf{X} - \sum_{i \in s} v_i w_i \mathbf{x}_i\right)^T \left(\sum_{i \in s} \frac{u_i w_i \mathbf{x}_i \mathbf{x}_i^T}{c_i}\right)^{-1} \frac{u_i \mathbf{x}_i}{c_i},$$

with

$$u_i = \frac{\psi_{k_1}(r_i)}{r_i}, \qquad v_i = \frac{\psi_{k_2}(r_i)}{r_i}, \qquad r_i = \frac{y_i \mathbb{1}\{i \in s_d\} - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,1}}{\hat{\sigma}_{d,1}\sqrt{c_i}},$$

where $\hat{\boldsymbol{\beta}}_{d,1}$ is the solution to the estimating equation

$$\sum_{i \in s} w_i \psi_{k_1}\left(\frac{y_i \mathbb{1}\{i \in s_d\} - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,1}}{\hat{\sigma}_{d,1}\sqrt{c_i}}\right) \frac{\mathbf{x}_i}{\sqrt{c_i}} = \mathbf{0}$$

and $\hat{\sigma}_{d,1} = \mathrm{MAD}_w(r_i,\, i \in s)$.

Remarks *(general)***.** The following remarks concern all robust domain estimators to be discussed subsequently. (For the sake of readability, the remarks are not repeated.)

(i) Estimator $\hat{Y}_{d,1}^{HT}$ and all subsequent estimators can be written in the form

$$\hat{Y}_{d,1}^{HT} = \mathbf{X}^T \hat{\boldsymbol{\beta}}_{d,1} + \hat{\sigma}_{d,1} \sum_{i \in s} w_i \sqrt{c_i} \psi_{k_2}\left(\frac{y_i \mathbb{1}\{i \in s_d\} - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,1}}{\hat{\sigma}_{d,1}\sqrt{c_i}}\right),$$

or equivalently as (using the definition of the v_i's)

$$\hat{Y}_{d,1}^{HT} = \mathbf{X}^T \hat{\boldsymbol{\beta}}_{d,1} + \sum_{i \in s} w_i v_i \big(y_i \mathbb{1}\{i \in s_d\} - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,1}\big).$$

From the last formula, we recognize that the v_i's play a fundamental role in order to control the degree of robustness. The u_i's on the other hand (not shown) obtain as a "byproduct" of the estimation of $\hat{\boldsymbol{\beta}}_{d,1}$. Now, if we take $v_i \equiv u_i$ (i.e., $k_1 = k_2$), then the resulting estimator $\hat{Y}_{d,1}^{HT}$ is quite robust with respect to non-representative outliers (in the sense of Chambers, 1986); however, this choice may not be appropriate in the presence of representative outliers. Choosing the tuning constant k_2 (which is associated with the v_i's) somewhat larger than k_1, the resulting estimator is a bias-corrected estimator in the sense of Chambers (1986), respectively, Welsh and Ronchetti (1998). We also note that the synthetic estimator is obtained when we put $v_i \equiv 0$.

(ii) The g-weights of estimator $\hat{Y}_{d,1}^{HT}$ (and all subsequent estimators) are such that the calibration property [see also Formula (4.3)] holds,

$$\sum_{i \in s} w_i g_i \mathbf{x}_i = \mathbf{X}$$

for all choices of the tuning constants k_1 and k_2. Unlike the g-weights of the ordinary GREG, the g-weights of the robust estimator $\hat{Y}_{d,1}^{HT}$ (and all subsequent estimators) depend on variable y_i (except when $v_i \equiv u_i \equiv 1$).

(iii) Our choice of $\hat{\sigma}_{d,1} = \mathrm{MAD}_w(r_i, i \in s)$ together with the regression M-estimator of $\boldsymbol{\beta}$ achieves, as a rule, a higher degree of robustness, compared with the joint M-estimate of $(\boldsymbol{\beta}, \sigma)$. This claim is substantiated by an argument in Hampel et al. (1986, 329).

(iv) The proposed M-estimator of $\boldsymbol{\beta}$ is robust against outliers in the response variable y_i. It is not robust with respect to influential observations in the model's design space; see the discussion of GM-estimators in Section 2.3.1. In the context of population-level GREG estimation, Beaumont and Alavi (2004) have studied Schweppe-type GM-estimators; see also Gwet and Rivest (1992) or Hulliger (1995).

(v) Another channel for non-robustness to "creep in" besides the variables $(y_i, \mathbf{x}_i)$, is the design weight. Outlying design weights or design weights that are highly variable can influence the robustness of estimator $\hat{Y}_{d,1}^{HT}$ badly. As a countermeasure, several authors have proposed some form of weight trimming; see e.g. Hulliger (1999), Duchesne (1999), or Beaumont and Alavi (2004). Their methods are also applicable in the context of domain estimation.

(vi) We have already pointed out in the introduction that robust estimators tend to be biased when the distribution of the y_i's is noticeably skewed. In large samples, the bias tends to overshadow the variance reduction obtainable via robustification such that one might end up with a larger MSE compared with the ordinary GREG. A solution to this problem is to generalized the notion of Hulliger's (1995) minimum estimated risk (MER) estimators to domain-specific estimators.

Remarks *(concerning estimator $\hat{Y}_{d,1}^{HT}$)*.

(i) Let us for the moment consider the ordinary GREG domain estimator which obtains when we specify $k_1 = k_2 = \infty$ (hence, $u_i \equiv v_i \equiv 1$). This estimator takes the form

$$= \hat{Y}_d^{HT} + \underbrace{(\mathbf{X} - \hat{\mathbf{X}})^T \left(\sum_{i \in s} \frac{w_i \mathbf{x}_i \mathbf{x}_i^T}{c_i} \right)^{-1}}_{\substack{\text{domain independent} \\ (1 \times p) \text{ matrix}}} \underbrace{\sum_{i \in s_d} \frac{w_i \mathbf{x}_i y_i}{c_i}}_{\substack{\text{domain dependent} \\ (p \times 1) \text{ matrix}}}$$

and was first suggested by M.A. Hidiroglou; see Estevao et al. (1995). From the above formula, we recognize the peculiar characteristic of the estima-

tor. In terms of auxiliary variables, the estimator builds on all elements $i \in s$ whereas with respect to computing the regression coefficient, it uses only y-values from domain d.

(ii) Following Hidiroglou and Patak (2004), the estimator is called *not* domain dependent.

The next two estimators use domain-level auxiliary data.

Estimator $\hat{Y}_{d,2}^{HT}$ is defined in terms of the g-weights (for $i \in s$)

$$g_i = \mathbb{1}_{i \in s_d} \left[v_i + \left(\mathbf{X}_d - \sum_{i \in s_d} v_i w_i \mathbf{x}_i \right)^T \left(\sum_{i \in s_d} \frac{u_i w_i \mathbf{x}_i \mathbf{x}_i^T}{c_i} \right)^{-1} \frac{u_i \mathbf{x}_i}{c_i} \right]$$

with

$$u_i = \frac{\psi_{k_1}(r_i)}{r_i}, \qquad v_i = \frac{\psi_{k_2}(r_i)}{r_i}, \qquad r_i = \frac{\mathbb{1}\{i \in s_d\} y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,2}}{\hat{\sigma}_{d,2} \sqrt{c_i}},$$

where $\hat{\boldsymbol{\beta}}_{d,2}$ is the solution to the estimating equation

$$\sum_{i \in s_d} w_i \psi_{k_1} \left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,2}}{\hat{\sigma}_{d,2} \sqrt{c_i}} \right) \frac{\mathbf{x}_i}{\sqrt{c_i}} = \mathbf{0}$$

and $\hat{\sigma}_{d,2} = \mathrm{MAD}_w(r_i,\, i \in s_d)$.

Estimator $\hat{Y}_{d,3}^{HT}$ is defined in terms of the g-weights (for $i \in s$)

$$g_i = v_i \mathbb{1}_{i \in s_d} + \left(\mathbf{X}_d - \sum_{i \in s_d} v_i w_i \mathbf{x}_i \right)^T \left(\sum_{i \in s} \frac{u_i w_i \mathbf{x}_i \mathbf{x}_i^T}{c_i} \right)^{-1} \frac{u_i \mathbf{x}_i}{c_i}$$

with

$$u_i = \frac{\psi_{k_1}(r_i)}{r_i}, \qquad v_i = \frac{\psi_{k_2}(r_i)}{r_i}, \qquad r_i = \frac{\mathbb{1}\{i \in s_d\} y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,3}}{\hat{\sigma}_{d,3} \sqrt{c_i}},$$

where $\hat{\boldsymbol{\beta}}_{d,3}$ is the solution to the estimating equation

$$\sum_{i \in s} w_i \psi_{k_1} \left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,3}}{\hat{\sigma}_{d,3} \sqrt{c_i}} \right) \frac{\mathbf{x}_i}{\sqrt{c_i}} = \mathbf{0}$$

and $\hat{\sigma}_{d,3}\mathrm{MAD}_w(r_i,\, i \in s_d)$.

Remark. The regression parameter $\hat{\boldsymbol{\beta}}$ in estimator $\hat{Y}_{d,3}^{HT}$ is computed by regressing y_i on $\mathbf{x}_i$ using all sampled elements.

4.4.2. Hajek type estimators

In this paragraph, we introduce the Hajek (HJ) type GREG estimators of the domain y-total Y_d ($d = 1, \ldots, D$).

Estimator $\hat{Y}_{d,1}^{HJ}$ is defined in terms of the g-weights (for $i \in s$)

$$g_i = \frac{N_d}{\hat{N}_d} v_i \mathbb{1}\{i \in s_d\} + \left(\mathbf{X} - \frac{N}{\hat{N}} \sum_{i \in s} v_i w_i \mathbf{x}_i\right)^T \left(\sum_{i \in s} \frac{u_i w_i \mathbf{x}_i \mathbf{x}_i^T}{c_i}\right)^{-1} \frac{u_i \mathbf{x}_i}{c_i},$$

with

$$u_i = \frac{\psi_{k_1}(r_i)}{r_i}, \qquad v_i = \frac{\psi_{k_2}(r_i)}{r_i}, \qquad r_i = \frac{y_i \mathbb{1}\{i \in s_d\} - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,1}}{\hat{\sigma}_{d,1} \sqrt{c_i}},$$

where $\hat{\boldsymbol{\beta}}_{d,1}$ is the solution to the estimating equation

$$\sum_{i \in s} w_i \psi_{k_1}\left(\frac{y_i \mathbb{1}\{i \in s_d\} - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,1}}{\hat{\sigma}_{d,1} \sqrt{c_i}}\right) \frac{\mathbf{x}_i}{\sqrt{c_i}} = \mathbf{0}$$

and $\hat{\sigma}_{d,1} = \mathrm{MAD}_w(r_i,\, i \in s)$.

Estimator $\hat{Y}_{d,2}^{HJ}$ is defined in terms of the g-weights (for $i \in s$)

$$g_i = \mathbb{1}_{i \in s_d}\left[\frac{N_d}{\hat{N}_d} v_i + \left(\mathbf{X}_d - \frac{N_d}{\hat{N}_d} \sum_{i \in s_d} v_i w_i \mathbf{x}_i\right)^T \left(\sum_{i \in s_d} \frac{u_i w_i \mathbf{x}_i \mathbf{x}_i^T}{c_i}\right)^{-1} \frac{u_i \mathbf{x}_i}{c_i}\right]$$

with

$$u_i = \frac{\psi_{k_1}(r_i)}{r_i}, \qquad v_i = \frac{\psi_{k_2}(r_i)}{r_i}, \qquad r_i = \frac{\mathbb{1}\{i \in s_d\} y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,2}}{\hat{\sigma}_{d,2} \sqrt{c_i}},$$

where $\hat{\boldsymbol{\beta}}_{d,2}$ is the solution to the estimating equation

$$\sum_{i \in s_d} w_i \psi_{k_1}\left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,2}}{\hat{\sigma}_{d,2} \sqrt{c_i}}\right) \frac{\mathbf{x}_i}{\sqrt{c_i}} = \mathbf{0}$$

and $\hat{\sigma}_{d,2} = \mathrm{MAD}_w(r_i,\, i \in s_d)$.

Estimator $\hat{Y}_{d,3}^{HJ}$ is defined in terms of the g-weights (for $i \in s$)

$$g_i = \frac{N_d}{\hat{N}_d} v_i \mathbb{1}_{i \in s_d} + \left(\mathbf{X}_d - \frac{N_d}{\hat{N}_d} \sum_{i \in s_d} v_i w_i \mathbf{x}_i\right)^T \left(\sum_{i \in s} \frac{u_i w_i \mathbf{x}_i \mathbf{x}_i^T}{c_i}\right)^{-1} \frac{u_i \mathbf{x}_i}{c_i}$$

with

$$u_i = \frac{\psi_{k_1}(r_i)}{r_i}, \qquad v_i = \frac{\psi_{k_2}(r_i)}{r_i}, \qquad r_i = \frac{\mathbb{1}\{i \in s_d\} y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,3}}{\hat{\sigma}_{d,3} \sqrt{c_i}},$$

where $\hat{\boldsymbol{\beta}}_{d,3}$ is the solution to the estimating equation

$$\sum_{i \in s} w_i \psi_{k_1}\left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{d,3}}{\hat{\sigma}_{d,3} \sqrt{c_i}}\right) \frac{\mathbf{x}_i}{\sqrt{c_i}} = \mathbf{0}$$

and $\hat{\sigma}_{d,3}\mathrm{MAD}_w(r_i,\, i \in s_d)$.

Remarks. (i) The M-estimators of $\boldsymbol{\beta}$ contained in the Hajek type estimators

correspond exactly to their Horvitz-Thompson counterparts.

(ii) Clearly, when it holds that $\hat{N} = N$ (respectively, $\hat{N}_d = N_d$), the HJ and HT estimators coincide.

(iii) Consider the domain-specific HT type estimators $\hat{Y}_{d,1}^{HT}$ and $\hat{Y}_{d,3}^{HT}$ with $k_1 = k_2 = \infty$ (i.e., non-robust estimators). These estimators satisfy the *additivity* or benchmarking property,

$$\sum_{d=1}^{D} \hat{Y}_{d,1}^{HT} = \sum_{d=1}^{D} \hat{Y}_{d,3}^{HT} = \hat{Y}^{HT} + \left(\mathbf{X} - \hat{\mathbf{X}}^{HT}\right)^T \hat{\boldsymbol{\beta}}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in s} \frac{w_i \mathbf{x}_i \mathbf{x}_i^T}{c_i} \right)^{-1} \sum_{i \in s} \frac{w_i \mathbf{x}_i y_i}{c_i}.$$

The non-robust Hajek-type estimators $\hat{Y}_{d,1}^{HJ}$, $\hat{Y}_{d,2}^{HJ}$, and $\hat{Y}_{d,3}^{HJ}$ do not satisfy the additivity property (except under special circumstances), neither does $\hat{Y}_{d,2}^{HT}$ (Hidiroglou and Patak, 2004, Remark 3.1). None of the robust estimators we have proposed satisfies the additivity property (except perhaps in special but uninteresting cases).

(iv) The non-robust Hajek type estimators are *nearly conditionally* (conditional on the sample size) unbiased whereas the HT type estimators are not (Hidiroglou and Patak, 2004, 69). This property also holds for the robust estimators.

4.5. Variance estimation

To fix ideas, we consider the linear unbiased estimator

$$\hat{Y}_b = \sum_{i \in s} b_i y_i$$

of the population y-total, where the b_i's are known real numbers which do not depend on the y_i's. The design variance of $\hat{Y}_b$ is (Thompson, 1997, 19)

$$\mathbb{V}\big(\hat{Y}_b\big) = \sum_{i \in U} a_i y_i^2 + \sum_{i,j \in U,\, i \neq j} \sum a_{ij} y_i y_j,$$

where

$$a_i = \mathbb{V}\big(b_i \mathbb{1}\{i \in s\}\big) \qquad \text{and} \qquad a_{ij} = \operatorname{cov}\big(b_i \mathbb{1}\{i \in s\},\, b_j \mathbb{1}\{j \in s\}\big),$$

$\mathbb{1}\{i \in s\}$ denoting the sample inclusion indicator variable. An unbiased Horvitz–Thompson type estimator of $\mathbb{V}(\hat{Y}_b)$ is (provided that all $\pi_{ij} > 0$; see Definition

2.2)

$$\hat{\mathbb{V}}(\hat{Y}_b) = \sum_{i \in s} \frac{a_i}{\pi_i} y_i^2 + \sum_{i,j \in s,\, i \neq j} \sum \frac{a_{ij}}{\pi_{ij}} y_i y_j.$$

If the b_i's in the definition of $\hat{Y}_b$ are independent of the $\mathbb{1}\{i \in s\}$-variables, the expressions of the variance and variance estimator simplify; for the latter we have,

$$\hat{\mathbb{V}}(\hat{Y}_b) = \sum_{i \in s} (1 - \pi_i) \left(\frac{b_i y_i}{\pi_i} \right)^2 + \sum_{i,j \in s,\, i \neq j} \sum \frac{\Delta_{ij}}{\pi_{ij}} \frac{b_i y_i}{\pi_i} \frac{b_j y_j}{\pi_j}, \tag{4.4}$$

where $\Delta_{ij} = \pi_{ij} - \pi_i \pi_i$ for all $i, j \in U$. Alternatively, if the sampling design is of fixed size, we might consider the variance estimator due to Sen–Yates–Grundy (see Section 2.2). Clearly, the same variance estimation technique can also be applied for the estimation of a domain y-total Y_d (provided the domain structure does not contain unplanned designs; see below).

Model-assisted estimators

Let us come back to model-assisted domain estimators. For illustration purposes, we restrict attention to estimator $\hat{Y}_{d,2}^{HT}$; yet, the variance estimation technique also holds for every other estimator we have proposed.

The key to variance estimation is that estimator $\hat{Y}_{d,2}^{HT}$ is expressible as a calibration estimator $\sum w_i g_i y_i$, where summation is over the index set $\{i \in s\}$, $w_i = 1/\pi_i$, and g_i denotes the g-weight. Then, under the (superpopulation-) model, the domain y-total Y_d can be written as $\mathbf{X}_d^T \boldsymbol{\beta}$. Now, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, we have (cf. Särndal et al., 1992, 402)

$$\begin{aligned} \hat{Y}_{d,2}^{HT} - Y_d &= \sum_{i \in s_d} w_i g_i y_i - \sum_{i \in U_d} y_i \\ &= \sum_{i \in s_d} w_i g_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) - \sum_{i \in U_d} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \end{aligned}$$

$\big[$letting $e_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^T \boldsymbol{\beta}\big]$

$$= \sum_{i \in s_d} w_i g_i e_i(\boldsymbol{\beta}) - \sum_{i \in U_d} e_i(\boldsymbol{\beta}), \tag{4.5}$$

where the second line follows from the calibration property of the g-weights,

$$\sum_{i \in s_d} w_i g_i \mathbf{x}_i^T \boldsymbol{\beta} = \mathbf{X}_d^T \boldsymbol{\beta} = \sum_{i \in U_d} \mathbf{x}_i^T \boldsymbol{\beta}.$$

Since $\hat{Y}_{d,2}^{HT}$ is approximately design-unbiased, we can obtain an estimate of the design mean squared error (MSE) for any $\boldsymbol{\beta} \in \mathbb{R}^p$ using (4.5) together with

$$\mathbb{E}\Big(\hat{Y}_{d,2}^{HT} - Y_d \Big)^2. \tag{4.6}$$

The sum of the $e_i(\boldsymbol{\beta})$'s over the index set $\{i \in U_d\}$ in (4.5) does not contribute to the design MSE. If the term $\mathbf{X}_d^T\boldsymbol{\beta}$ were of interest of its own and given that we do have strong beliefs in the superpopulation model, then we might consider the compound MSE, i.e. the MSE resulting from the joint stochastic behavior induced by the model and the sampling design (see Thompson, 1997, chap. 6.2.2). Here, we focus on the design MSE.

Special case

For the next step, we shall assume the special case when $k_1 = k_2 = \infty$; hence, $\hat{Y}_{d,2}^{HT}$ coincides with the ordinary domain GREG estimator. For this special case, the g-weights do not depend on the y_i's. Moreover, we may in view of (4.6) utilize the variance estimator defined in (4.4) with g_i in place of b_i and having substituted $e_i(\boldsymbol{\beta})$ for y_i. Obviously, the so derived variance estimator is of little practical value as $\boldsymbol{\beta}$ is unknown. However, we may consider the approximate variance estimator with $e_i(\hat{\boldsymbol{\beta}})$ in place of $e_i(\boldsymbol{\beta})$, where $\hat{\boldsymbol{\beta}}$ is a sample-based estimate of $\boldsymbol{\beta}$. The resulting variance estimator is the well-known g-weighted approximate variance estimator of the domain GREG estimator (see Särndal et al., 1992, Result 10.5.1). A formal derivation of the variance estimator via a first-order Taylor series expansion is given [for the GREG estimator of the population y-total] in Result 6.6.1 of Särndal et al. (1992); see also Binder (1983) and Rao (1994).

Variance of the robust domain GREG estimators

So far we have investigated variance estimators for the special case when $k_1 = k_2 = \infty$. In the general case, the g-weights are not necessarily independent of the y_i's. This implies that, in principle, we cannot use the variance estimator in (4.4) with the tuple (g_i, e_i) in place of (b_i, y_i). However, as an *approximation*, we suggest the following estimator

$$\hat{\mathbb{V}}\big(\hat{Y}_{d,2}^{HT}\big) \approx \sum_{i\in s}(1-\pi_i)\left(\frac{g_i e_i(\hat{\boldsymbol{\beta}})}{\pi_i}\right)^2 + \sum_{i,j\in s,\, i\neq j}\sum \frac{\Delta_{ij}}{\pi_{ij}}\frac{g_i e_i(\hat{\boldsymbol{\beta}})}{\pi_i}\frac{g_j e_j(\hat{\boldsymbol{\beta}})}{\pi_j}, \tag{4.7}$$

where $e_i(\hat{\boldsymbol{\beta}}) = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$. Alternatively, if the sampling design is of fixed size, we might consider the variance estimator due to Sen–Yates–Grundy. As we have indicated, variance estimators for the HT-type estimators $\hat{Y}_{d,1}^{HT}$ and $\hat{Y}_{d,3}^{HT}$ can be obtained in the same manner.

For HJ-type estimators, an approximate unconditional variance is discussed in Särndal et al. (1992, Result 10.5.1) or Hidiroglou and Patak (2004, 69). The approximate estimator of the unconditional variance for HJ-type estimators is the same as for HT-type estimators and thus defined like the estimator in (4.7).

Caution. It is absolutely crucial to point out that the variance estimators are only meaningful when the r.v.'s e_i have roughly a symmetric distribution. Put another way, if the distribution of the y_i's is strongly skewed, the estimators (e.g. $\hat{Y}_{d,2}^{HT}$) are biased estimators of the domain y-total, and the variance estimate alone does not appropriately describe the estimated design MSE in (4.6). In such situations, it may be advisable to consider a domain-specific minimum estimated risk (MER) estimator of the total; i.e., to apply the methods in Hulliger (1995) [and presumably to consider a resampling variance estimator].

Remarks. (i) Our derivation of the variance estimator via (4.5) shows that estimators like the one defined in (4.7) are sufficient to ensure "robust" variance estimates; notably, we do not have to "huberize" the $e_i(\hat{\boldsymbol{\beta}})$'s in the variance formulae. Moreover, our variance estimator when utilized for estimating the population y-total (instead of a domain total) coincides with the estimator studied in Beaumont and Alavi (2004, 11).

(ii) The literature on GREG variance estimation for the population y-total contains some aspects that are also applicable to robust domain estimation, namely: The "traditional" GREG variance estimator [which corresponds to taking $g_i \equiv 1$ in the variance formula] is known to produce slight underestimation for small sample sizes (Estevao et al., 1995, 185). The inclusion of the g-weights reduces the degree of underestimation. Another reasons to prefer the g-weighted variance estimator over the traditional one is because the former has better conditional inference properties (Särndal et al., 1992, Remark 7.10.4); see also Hidiroglou and Patak (2004).

(iii) Note that the variance estimator depends on the second-order inclusion probabilities, π_{ij}, which are typically not made available by national statistical institutes. Therefore, alternative variance estimators that do not require the π_{ij}'s are often desired. Fortunately, the variance expressions simplify considerably for many sampling designs, and estimation can be done using standard statistical software.

(iv) For unplanned domains, variance estimators like the one defined in (4.7) tend to underestimate the true variance as they do not account for the "randomness" of the domain membership. We can account for this either via:

   (a) fitting the extended domain variable $y_{d,i} = \mathbb{1}\{i \in U_d\} y_i$ (Estevao et al., 1995) or

   (b) replacing residuals r_i in the variance estimator by "extended residuals" $\mathbb{1}\{i \in U_d\} y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$; this proposal is due to C.E. Särndal (Lehtonen and Veijanen, 2009, 234).

(v) Alternatives to the approximate analytic variance estimators have been studied in the context of robust estimation of the population y-total by Hulliger (1991, chap. 4.2.5) [bootstrap] and Gwet (1997, chap. 4) [Jackknife]. Obviously, these resampling variance methods are also applicable for domain estimation.

4.6. Summary and discussion

We have suggested three robust Horvitz–Thompson and three robust Hajek type GREG domain estimators. What can be said about their performance (without having done a simulation study)? First, the three estimators (either of type HT or HJ) *differ substantially* with regard to the assumed assisting model, the required form of auxiliary information, etc. Hence, it would be completely misleading to compare the competing estimators. Second, robust estimators like the ones we proposed can be significantly more efficient compared with classical estimators if the distribution of the study variable y_i is not too skewed but contains representative outliers (and provided the model is otherwise properly specified). This claim is substantiated in the literature on robust estimation for finite-population sampling (see e.g. Gwet and Rivest, 1992; Hulliger, 1995; Duchesne, 1999; Beaumont and Alavi, 2004; Beaumont and Rivest, 2009). What if the distribution of the y_i's is strongly skewed? Then, the robust estimators can be severely biased and the design mean squared error of the estimators is (ceteris paribus) dominated by the bias. This is especially accentuated when the sample size is relatively large. In such situations, robust methods can be even less efficient than classical GREG estimators. However, much of the efficiency loss encountered with biased robust estimators can be avoided in practice by considering a symmetrizing transformation prior to estimation.

What type of estimator is to be preferred, the Horvitz–Thompson or Hajek type estimator? We have a strong preference in favor of Hajek-type GREG estimators for the following reasons (which are also valid in the case of robust domain estimation):

- The Hajek-type total of the estimated residuals tends to be more stable with varying sample sizes n_d compared with the Horvitz–Thompson estimator (Lehtonen and Veijanen, 2009, 234).

- The bias of the Hajek-type estimator is nearly zero *conditional* on the realized sample sizes; this is not the case for Horvitz–Thompson type estimators (Hidiroglou and Patak, 2004).

- The *conditional* coverage rates of confidence intervals for Hajek-type estimators are very close to the nominal rate (Hidiroglou and Patak, 2004, 69);

this property also holds for moderately small sample sizes (e.g., $n_d = 20$) as the simulation of Hidiroglou and Patak (2004, sec. 4) reveals.

Two out of three reasons that speak for Hajek-type GREG estimators are conditional in nature. But this does not necessarily have to be a disadvantage. On the contrary, conditional properties are important in practice since practitioners are interested in estimates given their sample rather than in quantities averaged over all possible samples.

5. Robust estimation under the basic unit-level model

5.1. Introduction

Consider a finite survey population U. The population U is assumed to be partitioned into g small areas $U_1, \ldots, U_g$. Let N_i be the size of U_i and assume that $U = \bigcup_{i \leq g} U_i$, then we have $N = \sum_{i \leq g} N_i$; see domain structure, Definition 2.1. A sample s_i of $n_i \geq 1$ units is assumed to be drawn from U_i, $i = 1, \ldots, g$, according to a specified sampling design. If the implied partitioning of s by the domain structure may result in empty areas (i.e. when $\{j : j \in s_i\} \equiv \varnothing$), we put $n_i = 0$; this does do no harm to the estimators discussed in this Chapter.

We make the crucial *assumption* that sampling is *ignorable* in the sense of Sugden and Smith (1984); sometimes it may be justified to impose a slightly less restrictive assumption, e.g. requiring the sampling design to be self-weighting. However, we stick with the assumption of ignorable sampling. We refer the reader to the papers of Münnich and Burgard (2012) and Burgard, Münnich, and Zimmermann (2014), where the impact of non-ignorable sampling designs on model-based estimators is studied; see also You and Rao (2002), who proposed a pseudo-EBLUP method that incorporates design weights.

Associated with the jth unit in area i is the variable of interest y_{ij}, which is supposed known for all $j \in s_i$ ($i = 1 \ldots, g$). In addition, we assume that a p-vector of auxiliary variables $\mathbf{x}_{ij}$ has been observed for all $j \in U_i$ and all areas $i = 1, \ldots, g$. Unlike the area-level or Fay–Herriot model, the data are assumed to be available at the unit- or individual level. Moreover, it is assumed that the relationship between y_{ij} and $\mathbf{x}_{ij}$ is given at the population level by the basic unit-level model through a nested-error linear regression (Battese et al., 1988) of the form

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}, \quad j = 1, \ldots, N_i, \quad i = 1, \ldots, g, \tag{5.1}$$

where

$$u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2), \qquad i = 1, \ldots, g,$$

and

$$e_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_e^2), \qquad j = 1, \ldots, N_i, \quad i = 1, \ldots, g,$$

The area-specific random effects u_i and the unit-level errors e_{ij} are assumed to be stochastically independent from each other. The model in (5.1) is a random-

intercept mixed linear model (MLM) with block diagonal structure; see Section 2.5. The parameters β, σ_e^2, and σ_u^2 of the population model (5.1) are supposed unknown [a rigorous definition of the parameter space will be discussed below]. The goal is to estimate the y-mean for each of the $i = 1, \ldots, g$ areas.

Motivation and goal

Under model (5.1) and in particular, the Gaussian distributional assumption on the r.v.'s u_i and e_{ij}, Battese et al. (1988) show that optimal estimators of the area-specific y-means can be obtained by appealing to the theory of empirical best linear unbiased predictor (EBLUP) [this is made more precise; see below]. The EBLUP method, however, can be highly influenced in the presence of outliers. Since MLM's with block diagonal covariance matrix have, unlike location-scale or regression models, no nice invariance structure, the parameters cannot be estimated consistently in the presence of contamination—there is an unavoidable asymptotic bias (Welsh and Richardson, 1997, 349). In the presence of contamination, any method estimates the parameters at the core model plus an unknown bias. In the case of the maximum likelihood estimator (m.l.e.), the bias can be arbitrarily large which renders these estimators extremely inefficient.

A large number of robust methods to estimate the parameters of an MLM have been proposed (see discussion in Section 5.4, below). Among the most promising approaches are M-estimators under the MLM (called RML 1 and 2; see Richardson and Welsh, 1995; Welsh and Richardson, 1997). In the context of SAE, the proposal of Sinha and Rao (2009) is by far the best known robustification and has yet inspired further developments; see e.g., Schmid and Münnich (2014); Chambers, Chandra, Salvati, and Tzavidis (2014); Jiongo, Haziza, and Duchesne (2013) (to name only a few). From a methodological point of view, the proposal of Sinha and Rao (2009) is in fact an *approximation* to the RML 2 method. This approach seeks to approximate the square root of the inverse covariance matrix $\mathbf{\Omega}_i^{-1}$ in area i (which is a function of σ_e^2 and σ_u^2) required to make the M-estimates scale equivariant. Instead of using a properly defined square root of $\mathbf{\Omega}_i^{-1}$ under the model, which would ensure a *proper decorrelation* of the residuals, Sinha and Rao (2009) suggest to take a diagonal matrix with elements equal to $(\mathrm{diag}(\mathbf{\Omega}_i))^{-1/2}$. Clearly, their proposal is easier to deal with since it neglects the off-diagonal structure of $\mathbf{\Omega}_i$. Yet, their approximation to the problem is not sufficient to ensure a numerically sound solutions to the notoriously instable estimating problem in robust MLM's. Moreover, the improper decorrelation resulting from their proposal makes it hard to choose appropriate robustness tuning constants.

Convergence issues are not only experienced in the setting of Sinha and Rao (2009), but are also reported by Richardson (1995, chap. 6.5) and Chaubey and

Venkateswarlu (2002), who did not resort to the above mentioned approximation. The latter authors report that almost 25% of their Monte Carlo replications did not converge. Our experiences (see Schoch, 2011a) show that the major source of convergence failure is due to the Newton–Raphson method (which has quadratic convergence rate within some neighborhood of the root, but it can go horribly wrong outside the neighborhood) and the parametrization of the model. With regard to the open problems, our contributions are:

- a robust method for computing the parameters under the basic unit-level model in (5.1) which is equivalent to the RML 2 approach, but uses a different parametrization (thus, bypasses some of the notorious instabilities experienced in different setups) and has superior numerical properties (in particular, it does not suffer from failure of convergence, an issue encountered far too often with other implementations);
- a robust prediction method that is much simpler than the one suggested by Sinha and Rao (2009);
- an algorithm for computing robust estimates of the area-specific y-means that is numerically stable and computationally efficient.

Outline of this chapter

In Sections 5.2 and 5.3, we define the model and study the EBLUP method. Section 5.4 is devoted to the development of the robust EBLUP method and in Section 5.5, we suggest a parametric bootstrap method for the estimation of the mean square prediction error. Section 5.6 studies the computational details of the proposed robustification (e.g., choice of starting values for the iterative algorithm, etc.), followed by a model-based simulation study (see Section 5.7). In Section 5.8, the methods are applied to empirical data on average aboveground forest biomass in Norwegian municipalities (see case study). Finally, Section 5.9 draws together the major findings.

Further remarks

Little emphasis is placed on asymptotics and other mathematical properties of the proposed method. Since our proposal is equivalent to the RML 2 approach, the technical results of Welsh and Richardson (1997) apply. Large parts of this chapter are based on the papers Schoch (2012), Schoch (2011a), and Schoch (2011b). The proposed robust EBLUP method has been implemented in the `R` statistical software and published in the `R`-package `rsae`; see Schoch (2011c).

5.2. Definitions and assumptions

The specification of the model in (5.1) shall be made rigorous in what follows.

Definition 5.1. *The basic unit-level model entails*

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega}_i(\boldsymbol{\theta})), \quad i = 1, \ldots, g, \tag{5.2}$$

where the $\mathbf{y}_i$*'s* ($i = 1, \ldots, g$) *are independent* $(n_i \times 1)$ *random vectors with*

$$\boldsymbol{\theta} = (\sigma_e^2, \sigma_u^2)^T \qquad \textit{and} \qquad \boldsymbol{\Omega}_i(\boldsymbol{\theta}) = \sigma_e^2\mathbf{I}_i + \sigma_u^2\mathbf{1}_i\mathbf{1}_i^T, \tag{5.3}$$

where $\mathbf{I}_i$ *is the* $(n_i \times n_i)$ *identity matrix and* $\mathbf{1}_i$ *the* n_i*-vector of ones.*

Whenever no confusion can arise, we suppress the functional dependence of $\boldsymbol{\Omega}_i(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$. Note that the area-specific response vectors $\mathbf{y}_i$ ($i = 1, \ldots, g$) can be of different length (i.e., unbalanced data). For notational simplicity, it is sometimes useful to work with the stacked vectors and matrices, given by

$$\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_g^T)^T, \quad \mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_g^T)^T, \quad \boldsymbol{\Omega} = \text{blockdiag}(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_g),$$

where the operator "blockdiag" stacks the matrices along the diagonal; the off-diagonal blocks are set equal to zero. An alternative representation of $\boldsymbol{\Omega}$ is given by means of the direct sum $\boldsymbol{\Omega} = \boldsymbol{\Omega}_1 \oplus \cdots \oplus \boldsymbol{\Omega}_g$. In addition, we impose the following assumptions (restrictions) on the definition of model (5.2).

Assumption 5.1. *The parameter space of model (5.2) is* $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{\boldsymbol{\beta}} \times \boldsymbol{\Theta}_{\boldsymbol{\theta}}$ *with* $\boldsymbol{\Theta}_{\boldsymbol{\beta}} = \{\boldsymbol{\beta} \in \mathbb{R}^p\}$ *and* $\boldsymbol{\Theta}_{\boldsymbol{\theta}} = \{\boldsymbol{\theta} = (\sigma_e^2, \sigma_u^2)^T : \sigma_e^2 > 0,\ \sigma_u^2 \geq 0\}$.

Note that we allow for the possibility that the random-effect variance, σ_u^2, can be zero. In this case, the model reduces to an ordinary linear regression model. The next assumption is concerned with the design matrix.

Assumption 5.2. *The* $i = 1, \ldots, g$ *design matrices* $\mathbf{X}_i$ *of size* $(n_i \times p)$ *are supposed to have*

(i) full column rank p*,*

(ii) a column of n_i *ones as the first column.*

Remarks. (i) All results in this chapter remain valid if Assumption 5.2(i) does not hold. The exclusion of rank deficient design matrices is not consequential, but simplifies the discussion of the proposed algorithm considerably.

(ii) Definition 5.1 points out that we shall only be concerned with independently distributed r.v.'s $\mathbf{y}_i$ under the assumed model. The assumption of

a Gaussian distribution has been introduced for pedagogical reasons and will be relaxed in favor of more general distributional specifications. Yet, we stick with the independence assumption; see e.g. Schmid and Münnich (2014) or Schmid, Tzavidis, Münnich, and Chambers (2016) who consider robust estimation in the presence of spatial correlation among the areas.

5.3. Best linear unbiased predictor

Suppose for the time being that $\boldsymbol{\beta}$ and u_i $(i = 1, \ldots, g)$ in model (5.1) are known. For N_i large and the sampling fraction $f_i = n_i/N_i$ small (for all i), it is easily seen that the area-specific population means $\bar{y}_i = (1/N_i) \sum_{j \in U_i} y_j$ can be predicted by

$$\bar{y}_i \approx \mu_i = \bar{\mathbf{x}}^T \boldsymbol{\beta} + u_i, \quad i = 1, \ldots, g, \tag{5.4}$$

where $\bar{\mathbf{x}}$ is the p-vector of known population x-means in area i, given by

$$\bar{\mathbf{x}}_i = \frac{1}{N_i} \sum_{j \in U_i} \mathbf{x}_j.$$

We denote the *sample*-based estimators of the area-specific y_i-mean and $\mathbf{x}_i$-mean by, respectively,

$$\hat{\bar{\mathbf{x}}}_i = \frac{1}{n_i} \sum_{j \in s_i} \mathbf{x}_j, \quad \text{and} \quad \hat{\bar{y}}_i = \frac{1}{n_i} \sum_{j \in s_i} y_j.$$

In case of non-negligible sampling fractions (i.e., if $N_i \gg n_i$ does not hold), we cannot take the small area mean $\bar{y}_i$ as $\bar{\mathbf{x}}^T \boldsymbol{\beta} + u_i$. However, we can then express $\bar{y}_i$ as

$$\bar{y}_i = f_i \hat{\bar{y}}_i + (1 - f_i) \bar{y}_{U_i \setminus s_i}, \tag{5.5}$$

where $\bar{y}_{U_i \setminus s_i}$ is the y-mean of the *non-sampled* units in area i. Now, (5.5) cannot be exploited since the non-sampled elements $y_{j \in U_i \setminus s_i}$ have not been observed. However, under the population model (5.1), we may estimate the unknown observations by $\bar{\mathbf{x}}_{U_i \setminus s_i}^T \boldsymbol{\beta} + u_i$, where $\bar{\mathbf{x}}_{U_i \setminus s_i}$ denotes the $\mathbf{x}_i$-mean of the non-sampled elements (which is known since all x-observations are supposed known for the whole population); see (Rao, 2003, 142). Sinha and Rao (2009) remark that this approximation may not be adequate in the presence of representative outliers in the sense of Chambers (1986).

The fixed effects parameter $\boldsymbol{\beta}$ and the variance components $\boldsymbol{\theta}$ are unknown in practical applications; therefore, the predicting equation in (5.4) is of limited value. However, for known $\boldsymbol{\theta}$, the best linear unbiased estimator (BLUE), $\tilde{\boldsymbol{\beta}}$, is given by

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\sum_{i=1}^{g} \mathbf{X}_i^T \boldsymbol{\Omega}_i^{-1}(\boldsymbol{\theta}) \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{g} \mathbf{X}_i^T \boldsymbol{\Omega}_i^{-1}(\boldsymbol{\theta}) \mathbf{y}_i \right).$$

Appealing to well-known results of BLUE estimation, we obtain the best linear

unbiased predictor (BLUP) (Rao, 2003, 96–98)

$$\tilde{\mu}_i(\boldsymbol{\theta}) = \bar{\mathbf{x}}_i^T \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \sigma_u^2 \mathbf{1}_i^T \boldsymbol{\Omega}_i^{-1}(\boldsymbol{\theta})\big(\mathbf{y}_i - \mathbf{X}_i \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})\big), \quad i = 1, \ldots, g. \tag{5.6}$$

In practice, $\boldsymbol{\theta}$ has to be estimated. This can been accomplished by several methods; see e.g. Harville (1977); Miller (1977). Once we have computed the estimate $\hat{\boldsymbol{\theta}}$, the empirical BLUP (EBLUP), $\hat{\mu}_i(\hat{\boldsymbol{\theta}})$, is obtained by replacing $\boldsymbol{\theta}$ in (5.6) with the estimate $\hat{\boldsymbol{\theta}}$.

Maximum likelihood estimation

It will prove useful to study the maximum likelihood estimator of the model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Subsequently, we shall derive a robustification of the Fisher-score functions.

Under model (5.2), the non-constant part of the log-likelihood, $l(\boldsymbol{\tau}, \mathbf{X}, \mathbf{y})$, is given by

$$-2l(\boldsymbol{\tau}, \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{g} \log|\boldsymbol{\Omega}_i| + \sum_{i=1}^{g} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \tag{5.7}$$

where $\boldsymbol{\tau} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$. The m.l.e. of $\boldsymbol{\tau}$ is defined as

$$l(\hat{\boldsymbol{\tau}}, \mathbf{X}, \mathbf{y}) = \sup_{\boldsymbol{\tau} \in \boldsymbol{\Theta}} l(\boldsymbol{\tau}, \mathbf{X}, \mathbf{y}),$$

provided that $\hat{\boldsymbol{\tau}}$ is an interior point of $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ denotes the parameter space, defined in Assumption 5.1. Before we study the details of the m.l.e., it will prove useful to express the covariance matrix $\boldsymbol{\Omega}_i$ $(i = 1, \ldots, g)$ as follows (cf. Hartley and Rao, 1967)

$$\boldsymbol{\Omega}_i = \sigma_e^2 \mathbf{I}_i + \sigma_u^2 \mathbf{J}_i = v(\mathbf{I}_i + d\mathbf{J}_i) = v\mathbf{V}_i, \tag{5.8}$$

where

$$v = \sigma_e^2 \qquad d = \frac{\sigma_u^2}{\sigma_e^2} \equiv \frac{a}{v}.$$

The notation in terms of a, d, and v has been chosen for ease of simplicity as it spares us from writing squared terms. The primary advantage of the Hartley–Rao parametrization in (5.8), i.e., parametrization of the covariance matrix in terms of variance components ratios, is that we obtain a separate estimating equation for v. This is easy to see if we rewrite the log-likelihood in (5.7) with the parametrization in (5.8); hence,

$$\begin{aligned} -2l(\boldsymbol{\beta}, v, d; \mathbf{X}, \mathbf{y}) = &\sum_{i=1}^{g} \log|\mathbf{V}_i| + \sum_{i=1}^{g} n_i \log v \\ &+ \frac{1}{v} \sum_{i=1}^{g} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \end{aligned} \tag{5.9}$$

Provided the maximum is not attained on the boundary, the maximum likeli-

hood estimates $\hat{\boldsymbol{\beta}}$, $\hat{v}$, and $\hat{d}$ are a solution to the system of Fisher-score equations, respectively,

$$-2(1/v)\sum_{i=1}^{g}\mathbf{X}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})=\mathbf{0}, \tag{5.10}$$

$$\sum_{i=1}^{g}\frac{n_i}{v}-(1/v^2)\sum_{i=1}^{g}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})=0, \tag{5.11}$$

$$\sum_{i=1}^{g}\mathbf{1}_i^T\mathbf{V}_i^{-1}\mathbf{1}_i-(1/v)\mathbf{1}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}\mathbf{1}_i=0, \tag{5.12}$$

It is evident from (5.11) that the m.l.e. of v is given by

$$\hat{v}=(1/n)\sum_{i=1}^{g}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}), \tag{5.13}$$

where $n=\sum_{i=1}^{g}n_i$. Lindstrom and Bates (1988) – among others – take advantage of (5.13) and propose the variance-profile log-likelihood substituting (5.13) back into (5.9). This leads to an equivalent maximization problem with v eliminated.

At first sight, the effect of parametrizing the variance components in terms of ratios and the implied simplification of the maximization problem seem to be rather limited since only one parameter can be eliminated. From the perspective of computation, though, even such a small reduction simplifies the numerical optimization problem considerably. This parametrization brings along another good property in terms of numerical optimization: Observe from (5.13) that estimates of v are always non-negative due to the quadratic form of the estimator. This property is a great advantage over other methods, which tend to produce negative variance estimates.

On the other hand, Equations (5.10) and (5.12) do not feature a closed-form expression (for unbalanced data) and have to be solved by means of some (iterative) numerical methods.

A criticism of m.l.e.'s of variance is that they are biased downward because they do not take into account the loss in degrees of freedom from the estimation of $\boldsymbol{\beta}$. The restricted (or residual, or reduced) maximum likelihood estimators (REML) correct for this deficiency; see e.g., Harville (1977) for the details. We focus on the m.l.e. and robust M-estimators; see Richardson and Welsh (1995) for the robustification of REML estimators.

5.4. Robust empirical best linear unbiased predictor

Although the classical EBLUP method is useful for estimating the small area means efficiently under the normality assumptions, it can be highly influenced

in the presence of outliers as mixed linear models have – unlike location-scale or regression models – no nice invariance structure. In the presence of contamination, the bias of the m.l.e. can be arbitrarily large which renders it extremely inefficient (Welsh and Richardson, 1997, 349).

A large number of authors proposed methods for robust analysis in mixed linear models, ranging from

- rather algorithmic approaches (Rocke, 1983, 1991; Stahel and Welsh, 1997)
- over robustification of Henderson's mixed-model equations (Fellner, 1986; Stahel and Welsh, 1997; Richardson and Welsh, 1995; Welsh and Richardson, 1997)
- to replacing the Fisher scores by robust Fréchet-differentiable statistical functionals (Bednarski and Zontek, 1996).

Another sensible class of estimating methods are S-estimators. For balanced data (i.e., when n_i is the same for all $i = 1, \ldots, g$), Copt and Victoria-Feser (2006) have suggested an S-estimator under the MLM; they also provide an `R` implementation; cf. supporting website of the book by Heritier, Cantoni, Copt, and Victoria-Feser (2009). Unfortunately, their method does not easily generalized to the case of unbalanced data. Therefore, the methods is of rather limited value for SAE problems (as small areas are virtually never of exactly the same size). Another S-estimator has been suggested by Wellmann (1994, 2000); see Wellmann and Gather (2003). This proposal, however, is restricted to the case of the one-way random effects model.

We restrict attention to M-estimators because of their formidable theoretical and algorithmic properties. In the literature, a distinction is made between the two following M-estimators under a MLM with block-diagonal structure,

(i) direct *robustification of the likelihood*, called RML 1, cf. Richardson and Welsh (1995); Stahel and Welsh (1997),

(ii) M-estimators based on bounded-influence *estimating equations*; method RML 2, cf. Richardson and Welsh (1995); Welsh and Richardson (1997) [the name of RML 2 stems from the fact that this approach is related to Proposal 2 of Huber (1964)]

Both approaches have received considerable attention. We focus on the RML 2 method because it embodies a natural way of restricting the influence of outliers in the response variable and is very closely related to the m.l.e. approach. Furthermore, method RML 2 is *related* to the proposal of Sinha and Rao (2009). In fact, as we will see later, the proposal of Sinha and Rao (2009) is an approximation to method RML 2. For these robust estimators, the potential bias is

bounded, the efficiency is reasonable if the model holds, and the estimators are much more efficient than the m.l.e. if it does not (Welsh and Richardson, 1997).

The following assumptions are crucial to all our robust estimators.

Assumption 5.3.

(i) Outliers occur only in the response variables $\mathbf{y}_i$, $i = 1, \ldots, g$.

(ii) No attempt is made to limit the influence of outliers or influential observations in the design space of the model.

In order to limit the influence of outliers in both the response variable and the design matrix , one has to resort to generalized regression M-estimators (GM) in the context of linear models (e.g., Mallows- or Schweppe-type estimators). Richardson (1997) extended the notion of GM-estimators to include MLM's. Although theoretically convincing, GM-estimators for MLM's lack numerical stability (Richardson, 1995, chap. 6.5).

5.4.1. M-estimators under the MLM

In the presence of contamination, the m.l.e. can be severely biased. Consequently, it is reasonable to replace the system of Fisher-score functions (5.10-5.12) by M-estimators whose influence function (w.r.t. outliers in the response variable) are bounded.

5.4.1.1. M-estimator of parameter β

The Fisher-score function (5.10) can be severely influenced in the presence of contamination; thus, it shall be replaced by an M-estimator estimating equation. M-estimators of location and regression are not scale equivariant, but they can be made equivariant through appropriate scaling. In mixed linear models, however, scaling is more involved than in location or regression problems. What is actually needed is appropriate *decorrelation* not scaling since the $\mathbf{y}_i$'s are correlated within areas. Under the model in (5.2), we have

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, v\mathbf{V}_i), \qquad i = 1, \ldots, g,$$

and this implies a decorrelation matrix of the form

$$\mathbf{B}_i = \frac{1}{\sqrt{v}}\mathbf{V}_i^{-1/2},$$

where $\mathbf{B}_i$, respectively, the (inverse of the) square root, $\mathbf{V}_i^{-1/2}$, must be such that $\mathbf{B}_i^T\mathbf{B}_i = v\mathbf{V}_i$. This then ensures that $\mathbf{B}_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \sim \mathcal{N}(0, 1)$, a property we call

proper decorrelation. A candidate of matrix $\mathbf{V}_i^{-1/2}$ is given by[1]

$$\mathbf{V}_i^{-1/2} = \frac{1}{n_i}\left(\frac{1}{\sqrt{1+dn_i}} - 1\right)\mathbf{1}_i\mathbf{1}_i^T + \mathbf{I}_i, \qquad i = 1, \ldots, g, \tag{5.14}$$

having use the closed-form expressions for the determinant and inverse of matrix $\mathbf{V}_i$ (see e.g., Searle et al., 1992, 79)

$$|\mathbf{V}_i| = |\mathbf{I}_i + d\mathbf{1}_i\mathbf{1}_i^T| = 1 + dn_i \quad \text{and} \quad \mathbf{V}_i^{-1} = \mathbf{I}_i - \frac{d}{1+dn_i}\mathbf{1}_i\mathbf{1}_i^T. \tag{5.15}$$

By straightforward computations, we can check that $\mathbf{V}_i^{-1/2}$ defined in (5.14) indeed implies $\mathbf{B}_i^T\mathbf{B}_i = \mathbf{I}_i$.

Let $\psi_k : \mathbb{R} \to \mathbb{R}$ denote the Huber ψ-function indexed by the tuning constant $k > 0$ (or any other bounded, odd, monotone and almost everywhere differentiable function). Let $\boldsymbol{\psi}_k : \mathbb{R}^n \to \mathbb{R}^n$ be defined as

$$\boldsymbol{\psi}_k(\mathbf{u}) = \big[\psi_k(u_1), \ldots, \psi_k(u_n)\big]^T, \qquad \mathbf{u} \in \mathbb{R}^n, \tag{5.16}$$

then the M-estimator of $\boldsymbol{\beta}$ obtains as the solution to (omitting the dependency of $\mathbf{V}_i$ on the parameter $\boldsymbol{\theta}$)

$$\sum_{i=1}^{g}(1/\sqrt{v})\mathbf{X}_i^T\mathbf{V}_i^{-1/2}\boldsymbol{\psi}_k\big[(1/\sqrt{v})\mathbf{V}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^R)\big] = \mathbf{0}. \tag{5.17}$$

The solution to the (system of) estimating equations in (5.17), given v and $\boldsymbol{\theta}$ (or appropriate estimates, $\hat{v}$ and $\hat{\boldsymbol{\theta}}$), is denoted by $\hat{\boldsymbol{\beta}}^R$. Observe that the argument of the $\boldsymbol{\psi}_k$-function in (5.17) is properly decorrelated. Therefore, we can relate the choice of robustness tuning constant to the standardized Gaussian distribution (e.g., the choice $k = 2$ implies that observations larger in absolute values than 2 standard deviations are downweighted). If decorrelation were not proper, choosing sensible values of k is very difficult as the scale of $\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^R$ depends on the variance parameters, which are not known.

Estimates $\hat{\boldsymbol{\beta}}^R$ shall be obtained by an iteratively re-weighted least square (IRLWS) algorithm, which is the workhorse for computing M-estimates of regression (Maronna et al., 2006, 104–105). The IRWLS approach is numerically much more stable than the Newton–Raphson approach for the problem at hand (Schoch, 2011a). Denote by $\{\boldsymbol{\beta}\}^{(s)}$ the estimate of $\boldsymbol{\beta}$ produced by the algorithm

[1] Formula (5.14) obtains as follows. Denote by $\mathbf{L}_i\mathbf{D}_i\mathbf{L}_i^T$ the eigenvalue decomposition of the $(n_i \times n_i)$ matrix $\mathbf{V}_i$, where $\mathbf{L}_i$ is the $(n_i \times n_i)$ matrix whose columns correspond to the eigenvectors of $\mathbf{V}_i$; $\mathbf{D}_i = \text{diag}(\lambda_1, \ldots, \lambda_{n_i})$ is the $(n_i \times n_i)$ matrix of the eigenvalues λ_j $(j = 1, \ldots, n_i)$. It is not difficult to see that $\mathbf{V}_i$ has only two distinct eigenvalues: the first eigenvalue is $\lambda_1 = 1 + dn_i$ (with multiplicity one) and the remaining $(n_i - 1)$ eigenvalues are one. Now, we define a real-valued function f of the matrix $\mathbf{V}_i$ that corresponds to a function of a scalar as $f(\mathbf{V}_i) = \mathbf{L}_i\text{diag}(f(\lambda_1), \ldots, f(\lambda_{n_i}))\mathbf{L}_i^T$. Some straightforward computations with $f(u) = u^{-1/2}$ imply the result.

on the sth iteration ($s = 1, 2, \ldots$). An updated estimate is obtained from

$$\{\boldsymbol{\beta}\}^{(s+1)} = \left(\sum_{i=1}^{g}(\{\mathbf{W}_i\}^{(s)}\{\mathbf{V}_i^{-1/2}\}^{(s)}\mathbf{X}_i)^T\{\mathbf{W}_i\}^{(s)}\{\mathbf{V}_i^{-1/2}\}^{(s)}\mathbf{X}_i\right)^{-1} \times \times\left(\sum_{i=1}^{g}(\{\mathbf{W}_i\}^{(s)}\{\mathbf{V}_i^{-1/2}\}^{(s)}\mathbf{X}_i)^T\{\mathbf{W}_i\}^{(s)}\{\mathbf{V}_i^{-1/2}\}^{(s)}\mathbf{y}_i\right), \quad (5.18)$$

where $\mathbf{W}_i = \text{diag}(\mathbf{w}_i)$, with

$$\mathbf{w}_i = (w_{i1}, \ldots, w_{in_i})^T, \text{ where } w_{ij} = [\psi_k(r_{ij})/r_{ij}]^{1/2},$$

and

$$\mathbf{r}_i = (r_{i1}, \ldots, r_{in_i})^T = (1/\sqrt{v})\mathbf{V}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

If we pre-multiply $\mathbf{y}_i$ and $\mathbf{X}_i$ conformably,

$$\tilde{\mathbf{X}}_i = (1/\sqrt{v})\mathbf{W}_i\mathbf{V}_i^{-1/2}\mathbf{X}_i \quad \text{and} \quad \tilde{\mathbf{y}}_i = (1/\sqrt{v})\mathbf{W}_i\mathbf{V}_i^{-1/2}\mathbf{y}_i,$$

then the problem takes the form

$$\{\boldsymbol{\beta}\}^{(s+1)} = \left(\sum_{i=1}^{g}\{\tilde{\mathbf{X}}_i^T\}^{(s)}\{\tilde{\mathbf{X}}_i\}^{(s)}\right)^{-1}\left(\sum_{i=1}^{g}\{\tilde{\mathbf{X}}_i^T\}^{(s)}\{\tilde{\mathbf{y}}_i\}^{(s)}\right). \quad (5.19)$$

Now, since (5.19) is a standard least squares problem, we obtain (iteratively) updated estimates of $\boldsymbol{\beta}$ by standard regression technique. First, we note that by Assumption 5.2, $\tilde{\mathbf{X}}_i$ has full rank, given that $\mathbf{w}_i$ is not the null vector. Put $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1^T, \ldots, \tilde{\mathbf{X}}_g^T)^T$ and $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1^T, \ldots, \tilde{\mathbf{y}}_g^T)^T$, which are of size $(n \times p)$ and $(n \times 1)$, respectively, where $n = \sum_{i=1}^{g} n_i$. Hence, we shall decompose $\tilde{\mathbf{X}}$ by means of the "skinny" QR-factorization; see e.g., (Gentle, 2007, 188–189 and 226). Write

$$\tilde{\mathbf{X}} = \mathbf{QR},$$

with $\mathbf{R} = (\mathbf{R}_1^T, \mathbf{0}^T)^T$, where $\mathbf{R}_1$ is an $(p \times p)$ upper triangular matrix. $\mathbf{Q}$ is an $(n \times n)$ orthogonal matrix which can be partitioned as follows: $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$, where $\mathbf{Q}_1$ is an $(n \times p)$ matrix whose columns are orthonormal. This enables us to write $\tilde{\mathbf{X}} = \mathbf{Q}_1\mathbf{R}_1$. Consequently, the overdetermined linear system

$$\tilde{\mathbf{X}}\boldsymbol{\beta} = \tilde{\mathbf{y}},$$

can be expressed as

$$\mathbf{R}_1\boldsymbol{\beta} = \mathbf{Q}_1^T\tilde{\mathbf{y}}.$$

Since $\mathbf{R}_1$ is an $(p \times p)$ triangular matrix, the system is easy to solve, $\boldsymbol{\beta} = \mathbf{R}_1^{-1}\mathbf{Q}_1^T\tilde{\mathbf{y}}$. The IRWLS algorithm now consists of solving

$$\{\boldsymbol{\beta}\}^{(s+1)} = \{\mathbf{R}_1^{-1}\}^{(s)}\{\mathbf{Q}_1^T\}^{(s)}\{\tilde{\mathbf{y}}\}^{(s)}, \quad (5.20)$$

in an iterative manner. The final value is regarded as the estimate $\hat{\boldsymbol{\beta}}^R$.

5.4.1.2. M-estimator of parameter v

The M-estimator of v that replaces the non-robust Fisher score (5.11) is obtained – in the spirit of Huber's Proposal 2 (see Huber, 1964) – as the solution $\hat{v}^R$ to the estimating equation

$$\left[\sum_{i=1}^{g} n_i\right]^{-1} \frac{1}{\delta_k} \sum_{i=1}^{g} \boldsymbol{\psi}_k\left(\frac{\mathbf{V}_i^{-1/2}\mathbf{r}_i}{\sqrt{\hat{v}^R}}\right)^T \boldsymbol{\psi}_k\left(\frac{\mathbf{V}_i^{-1/2}\mathbf{r}_i}{\sqrt{\hat{v}^R}}\right) = 1, \tag{5.21}$$

where function $\boldsymbol{\psi}_k$ is defined in (5.16), and $\delta_k = \mathbb{E}_\Phi[\psi_k^2(u)]$ is a consistency correction term that ensures Fisher consistency of the estimate at the Gaussian core model. Let ϕ and Φ denote, respectively, the p.d.f. and c.d.f. of the standard normal distribution. By symmetry, we have

$$\begin{aligned}\delta_k = \mathbb{E}_\Phi \psi_k^2(u) &= 2k^2 \int_{-\infty}^{k} \phi(u)\mathrm{d}u + \int_{-k}^{k} u^2\phi(u)\mathrm{d}u \\ &= 2k^2\big(1-\Phi(k)\big) + 2\int_0^k u^2\phi(u)\mathrm{d}u\end{aligned}$$

then, using identity $\phi'(u) = -u\phi(u)$,

$$= 2k^2\big(1-\Phi(k)\big) - 2\int_0^k u\phi'(u)\mathrm{d}u,$$

hence, integration by parts gives

$$\begin{aligned}&= 2k^2\big(1-\Phi(k)\big) - 2\big[u\phi(u)\big]_0^k + 2\int_0^k \phi(u)\mathrm{d}u \\ &= 2\big[k^2\big(1-\Phi(k)\big) - k\phi(k) + \Phi(k) - 1/2\big],\end{aligned} \tag{5.22}$$

which is the result given in Maronna et al. (2006, 27). From the perspective of computation, it is worth to consider another representation than equation (5.21). Paralleling the concept of computing M-estimates of scale (cf. Maronna et al., 2006, 40–41), we shall define a weight function

$$w_k(z) = \begin{cases} \psi_k(z)^2/z^2 & \text{if } z \neq 0, \\ \psi_k'(z) & \text{if } z = 0, \end{cases}$$

where $\psi_k'(z)$ denotes the first derivative (which exists for a.e. $z \in \mathbb{R}$, and is set equal to zero if the derivative does not exist) and put for all $i = 1, \ldots, g$

$$\mathbf{W}_i = \mathrm{diag}(w_k(u_{i1}), \ldots, w_k(u_{in_i})) \quad \text{with} \quad \mathbf{u}_i = (1/\sqrt{v})\mathbf{V}_i^{-1/2}\mathbf{r}_i.$$

An updated estimate, $\{v\}^{s+1}$, is then given by

$$\{v\}^{s+1} = \frac{1}{n\delta_k} \sum_{i=1}^{g} \{\mathbf{W}_i\}^s \{\mathbf{r}_i^T\}^s \{\mathbf{V}_i^{-1}\}^s \{\mathbf{r}_i\}^s, \tag{5.23}$$

where $n = \sum_{i=1}^{g} n_i$. The fact that all elements of the diagonal matrix $\mathbf{W}_i$ are non-negative together with the quadratic form in (5.23) implies that the estimates of v are non-negative. This is in sharp contrast compared to the estimates obtained by (the inherently unconstrained) Newton–Raphson approach suggested in Sinha and Rao (2009).

5.4.1.3. M-estimator of parameter d

For the estimator of d, we have to replace the non-robust Fisher-score function (5.12) by an M-estimator. In contrast to v, the estimator of d has no closed-form solution. For $d \in \mathbb{R}^+$, let

$$\mathbf{u}_i(d) = (1/\sqrt{v})\mathbf{V}_i(d)^{-1/2}[\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}], \qquad i = 1, \ldots, g,$$

then a robust estimate of d, say $\hat{d}^R$, is obtained as the solution to

$$\sum_{i=1}^{g} \mathbf{1}_i^T \mathbf{V}_i(\hat{d}^R)^{-1}\mathbf{1}_i = \mathbf{1}_i^T \mathbf{V}_i(\hat{d}^R)^{-1/2}\boldsymbol{\psi}_k\big[\mathbf{u}_i(\hat{d}^R)\big]\boldsymbol{\psi}_k\big[\mathbf{u}_i(\hat{d}^R)\big]^T \mathbf{V}_i(\hat{d}^R)^{-1/2}\mathbf{1}_i, \quad (5.24)$$

where $\boldsymbol{\psi}_k$ is given in (5.16). Note that – for ease of simplicity – we highlighted only the functional dependence on d, i.e., $\mathbf{V}_i(d)^{-1/2}$, instead of reporting all parameters, $\mathbf{V}_i(\boldsymbol{\beta}, v, d)^{-1/2}$.

Among all available methods for finding the root of (5.24) (i.e., a root of a real-valued, continuous function in d), bisection is the most reliable approach, but quite slow. The method of regula falsi has been found to converge at a faster rate than linear. However, it can go quite wrong when the function is not approximately linear over the interval of interest (Small and Wang, 2003, 43–45).[2] The Newton–Raphson (NR) method is well-known to converge with a quadratic convergence rate within some neighborhood of the root, but it has the severe drawback of being very unreliable. In particular, the neighborhood of the root can be very small and (still more important) is not known beforehand. Divergence of the NR algorithm is a severe drawback and happens more frequently than many of the references admit; see also Jiang, Luan, and Wang (2007) or Chaubey and Venkateswarlu (2002). Moreover, NR does not explicitly take into account that the problem at hand is constrained (i.e., that d must be > 0). Obviously, one may deploy a watchdog function which prevents d from becoming negative (through modifying search direction and/or step length). However, this intervention may impede superlinear convergence.

We propose to solve (5.24) by means of Brent's root-finding algorithm (Brent,

[2] Speed of convergence: Suppose the sequence $\{\vartheta\}^{(s)}$ converges to ϑ_0 ($s = 1, 2, \ldots$). In numerical analysis, the speed at which a convergent sequence approaches the limit is determined by the values c and p in $\|\vartheta^{(s+1)} - \vartheta_0\| \leq c\|\vartheta^{(s)} - \vartheta_0\|^p$. For $0 < c \leq 1$ and $p = 1$, we shall say that the algorithm converges linearly. Likewise, we call the convergence superlinear if $p > 1$ (given that a $c > 0$ exists). Note that convergence of order p means that the number of correct decimal places is roughly p times the number of iterations; see e.g., Small and Wang (2003, chap. 3.1).

1973, chap. 4). The search for a root is constrained to the interval $(0, \kappa]$ (where $\kappa > 0$ has to be chosen), and thus ensures non-negativeness of d and $\sigma_u^2 = \sigma_e^2 d$. Brent's method combines the sureness of bisection with the speed of a higher-order method. It keeps track of whether a supposedly superlinear method is actually converging the way it is supposed to, and, if it is not, it intersperses bisection steps so as to guarantee at least linear convergence. Brent's method combines root bracketing, bisection, and inverse quadratic interpolation to converge from the neighborhood of a zero crossing. This kind of super-strategy requires attention to bookkeeping detail, and also careful consideration of how roundoff errors can affect the guiding strategy (Press, Teukolsky, Vetterling, and P.Flannery, 1986, 352–354). We therefore use a modification of Brent's original "zeroin" Fortran 77 code.

5.4.2. Computational issues of the Sinha–Rao proposal

So far, we have worked with the Hartley–Rao parametrization, i.e. the parametrization in terms of ratios of the variances. For ease of comparison with the method of Sinha and Rao (2009), we shall now adhere to the parametrization in terms of (σ_e^2, σ_u^2). The two parametrizations can be converted into each other via the identity $v\mathbf{V}_i = \mathbf{\Omega}_i$; see (5.8).

Recall that our M-estimators use the matrix $(1/\sqrt{v})\mathbf{V}_i^{-1/2}$ defined in (5.14) to decorrelate the residuals. The corresponding decorrelation matrix expressed in terms of (σ_e^2, σ_u^2) is

$$\mathbf{\Omega}_i^{-1/2} = \frac{1}{\sigma_e}\mathbf{V}_i^{-1/2} = \underbrace{\frac{1}{n_i\sqrt{\sigma_e^2 + n_i\sigma_u^2}}\,\mathbf{I}_i}_{\text{diagonal part}} + \underbrace{\frac{1}{n_i}\left(\frac{1}{\sqrt{\sigma_e^2 + n_i\sigma_u^2}} - \frac{1}{\sigma_e}\right)}_{\text{off-diagonal part}}\big(\mathbf{J}_i - \mathbf{I}_i\big), \quad (5.25)$$

where we have separated the diagonal- and off-diagonal parts of the matrix. For ease of referencing, we reproduce the covariance matrix of the $\mathbf{y}_i$'s,

$$\mathbf{\Omega}_i = \underbrace{(\sigma_e^2 + \sigma_u^2)\mathbf{I}_i}_{\text{diagonal part}} + \underbrace{\sigma_u^2(\mathbf{J}_i - \mathbf{I}_i)}_{\text{off-diagonal part}}, \qquad i = 1, \ldots, g. \quad (5.26)$$

Instead of (5.25), Sinha and Rao (2009, 386) use an *approximation* of $\mathbf{\Omega}_i^{-1/2}$ which is defined as

$$\mathbf{U}_i^{-1/2} = \frac{1}{\sqrt{\sigma_e^2 + \sigma_u^2}}\mathbf{I}_i, \qquad i = 1, \ldots, g. \quad (5.27)$$

Note that $\mathbf{U}_i^{-1/2}$ is a diagonal matrix whose elements are equal to the inverse of the square root of the diagonal elements of $\mathbf{\Omega}_i$. Clearly, the proposal of Sinha and Rao (2009) is easier to deal with since it neglects the off-diagonal structure

of the covariance matrix; see (5.26). Let

$$\tilde{\mathbf{r}}_i = \mathbf{U}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \qquad i = 1, \ldots, g, \tag{5.28}$$

to mean the standardized residuals in the proposal of Sinha and Rao (2009). With this and, for $\mathbf{A}_i$ equal to either $\mathbf{I}_i$ or $\mathbf{J}_i$, the M-estimator estimating equations of Sinha and Rao (2009, Eq. 16) for, respectively, σ_e^2 and σ_u^2 are defined as

$$\sum_{i=1}^{g} \boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\mathbf{U}_i^{1/2}\boldsymbol{\Omega}_i^{-1}\mathbf{A}_i\boldsymbol{\Omega}_i^{-1}\mathbf{U}_i^{1/2}\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i) - \operatorname{tr}\{c\boldsymbol{\Omega}_i^{-1}\mathbf{A}_i\} = 0, \tag{5.29}$$

where c is a consistency correction term, determined by (see Sinha and Rao, 2009, 386)

$$c = \mathbb{E}_\Phi[\psi_k(Z)^2] \quad \text{with} \quad Z \sim \mathcal{N}(0,1). \tag{5.30}$$

An expression for c was given in (5.22). Straightforward calculations show that if we let constant $k \to \infty$, the solutions to the estimating equations in (5.29) coincide with the m.l.e.'s.

Now, under the core model (see Definition 5.1) the marginal distribution of the $\mathbf{y}_i$'s is $\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega}_i)$ for all $i = 1, \ldots, g$ (having suppressed the dependency on $\boldsymbol{\theta}$). Consider the normalized residuals of the Sinha–Rao proposal $\tilde{\mathbf{r}}_i$ in (5.28) and observe that under the core model, we have (for all $i = 1, \ldots, g$)

$$\tilde{\mathbf{r}}_i \sim \mathcal{N}\big(\mathbf{0}, \mathbf{U}_i^{-1/2}\boldsymbol{\Omega}_i\mathbf{U}_i^{-1/2}\big),$$

where

$$\mathbf{U}_i^{-1/2}\boldsymbol{\Omega}_i\mathbf{U}_i^{-1/2} = \frac{\sigma_e^2}{\sigma_e^2+\sigma_u^2}\mathbf{I}_i + \frac{\sigma_u^2}{\sigma_e^2+\sigma_u^2}\mathbf{J}_i = \mathbf{I}_i + \underbrace{\frac{\sigma_u^2}{\sigma_e^2+\sigma_u^2}(\mathbf{J}_i - \mathbf{I}_i)}_{\text{off-diagonal part}}. \tag{5.31}$$

From (5.31) it is evident that the resulting covariance matrix of the residuals under the core model is not diagonal (unless $\sigma_u^2 = 0$). This non-diagonal structure is in sharp contrast to our proposal. What is more, our proposal in (5.25) verifies under the core model $\boldsymbol{\Omega}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_i)$ for all $i = 1, \ldots, g$ and therefore obtains *proper decorrelation* of the residuals $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$. Moreover, the proposal of Sinha and Rao (2009) has another issue.

Claim. *The M-estimator of σ_u^2 due to Sinha–Rao is not Fisher consistent.*

Unfortunately, we cannot present a formal proof of the claim as the analytic expression of $\mathbb{E}\big[\psi_k(\tilde{\mathbf{r}}_i)\psi_k(\tilde{\mathbf{r}}_i)^T\big]$ is such a mess that we did not succeed to work it out. Prima vista, one may come to the conclusion that a "sufficiently large collection" of integrals of Gaussian functions, for example, the tables compiled by Owen (1980) may provide enough "help" to solve the problem. Yet, this is not case. Already the first couple of integral evaluations lead to terms that are so "crooked" that the final goal seems out of reach (a sketch of the problem is

given in Appendix D). Therefore, we shall study the problem with the help of numerical methods (and some theoretical reasoning).

Numerical evidence

Instead of a formal proof, we give numerical *evidence* to substantiate the above claim. To this end, observe that the elements of the n_i-vector $\tilde{\mathbf{r}}_i$ defined in (5.28) are identically distributed with unit variance and covariance (in fact, Pearson correlation coefficient) equal to $\rho = \sigma_u^2/(\sigma_e^2+\sigma_u^2)$; see (5.31). Therefore, it suffices w.l.o.g. to restrict attention to the case of a bivariate Gaussian distribution $(X, Y)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ with

$$\mathbf{V} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \qquad \text{where} \quad 1 < \rho < 1. \tag{5.32}$$

Expectation w.r.t. this bivariate law is denoted by $\mathbb{E}_\rho$. The goal is to obtain

$$\kappa(k, \rho) = \mathbb{E}_\rho\big[\psi_k(X)\psi_k(Y)\big]$$

for different values of ρ and k. The values of κ are computed by means of Monte Carlo integration; for each tuple (k, ρ), a series of 10^8 replications have been evaluated. The values of k and ρ are taken to be equidistant in the respective interval, $[0.5,\, 3]$ and $[0, 1)$. The results of κ are symmetric with respect to ρ in the sense that $\kappa(\cdot, \rho) = \kappa(\cdot, -\rho)$; therefore, we restrict attention to $\rho \geq 0$.

The results of the simulation exercise are shown in Figure 5.1. The graph shows κ/c, where κ is normalized by c which has been proposed by Sinha and Rao (2009) and is defined in (5.30). It is evident from Figure 5.1 that κ/c depends on the value of the correlation ρ. For $\rho = 1$, κ and c coincide. On the other hand, κ differs more from c the smaller ρ is. It is also remarkable that for large k, κ and c are virtually identical (note that the curve for $k = 3$ is almost a horizontal line intersecting the ordinate at one). From our point of view, this simulation exercise provides enough evidence to claim that the M-estimator of σ_u^2 due to Sinha and Rao (2009) is not Fisher consistent (unless k is very large or ρ is almost unity).

Theoretical evidence

Although we did not succeed to work out an analytical expression for the consistency correction term, we want to give some theoretical reasoning. First, we consider the M-estimator of σ_e^2 in the Sinha–Rao framework (i.e., when $\mathbf{A}_i = \mathbf{I}_i$). Taking expectations w.r.t. the core model of the ith summand in (5.29), we have [where c is defined in (5.30)]

$$\begin{aligned} &\mathbb{E}\big[\boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\mathbf{U}_i^{1/2}\boldsymbol{\Omega}_i^{-2}\mathbf{U}_i^{1/2}\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i)\big] - \operatorname{tr}\big\{c\boldsymbol{\Omega}_i^{-1}\big\} \\ &\qquad = \operatorname{tr}\big\{\boldsymbol{\Omega}_i^{-2}\mathbf{U}_i\mathbb{E}\big[\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i)\boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\big]\big\} - \operatorname{tr}\big\{c\boldsymbol{\Omega}_i^{-1}\big\}, \end{aligned} \tag{5.33}$$

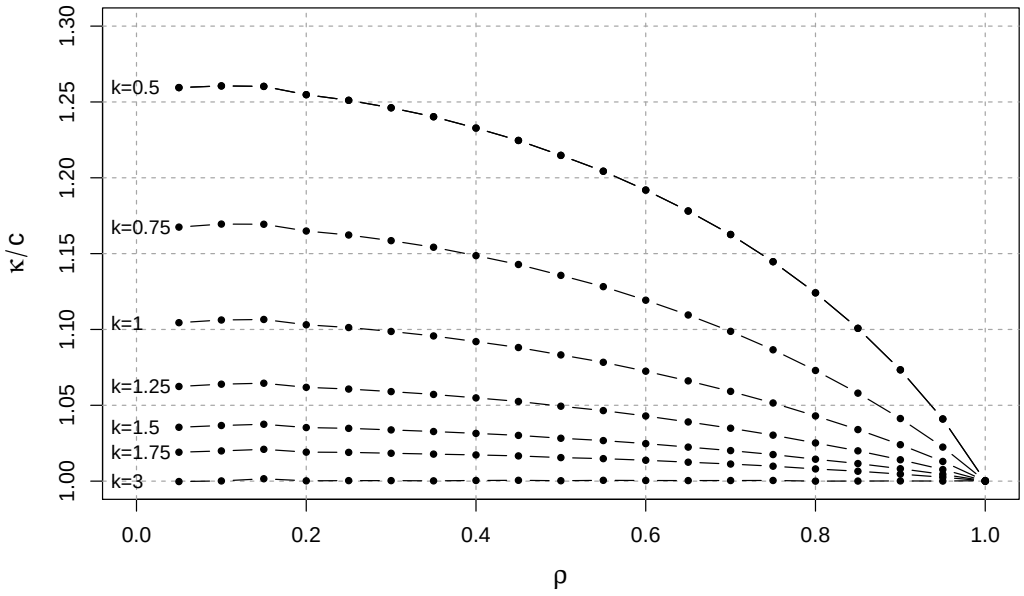


Figure 5.1.: Simulation approximation to the consistency correction term κ, normalized by $c = \mathbb{E}_\Phi \psi_k^2(X)$, for the Sinha–Rao M-estimator of σ_u^2 as a function of the tuning constant k and correlation coefficient ρ under a bivariate Gaussian distribution; for further explanation see text.

where we have used cyclical invariance of the trace operator. We shall approach the solution to this equation by heuristic reasoning. The key to our discussion is the following observation: The trace operator in the first term on the r.h.s. of (5.33) implies that we can restrict attention to the diagonal elements of the matrix under the trace operator. Now, from (5.31) we obtain

$$\operatorname{diag}\big(\mathbb{E}[\boldsymbol{\psi}(\tilde{\mathbf{r}}_i)\boldsymbol{\psi}(\tilde{\mathbf{r}}_i)^T]\big) = c\mathbf{I}_i,$$

and likewise, we have $\operatorname{diag}(\boldsymbol{\Omega}_i^{-2}\mathbf{U}_i) = \operatorname{diag}(\boldsymbol{\Omega}_i^{-1})$. From this it is easy to see that, for c defined in (5.30), (5.33) equals zero (for all $i = 1, \ldots, g$). Therefore, the choice of c in (5.30) is sufficient to make the M-estimator of σ_e^2 Fisher consistent at the Gaussian core model.

What about the M-estimator of σ_u^2 in the Sinha–Rao setup? We shall consider only the non-trivial case when $k \in (0, \infty)$ and $\sigma_u^2 > 0$. The estimating equation for σ_u^2 obtains from (5.29) with $\mathbf{A}_i = \mathbf{1}_i\mathbf{1}_i^T$ ($\equiv \mathbf{J}_i$), and its ith summand writes

$$\begin{aligned}
&\mathbb{E}\big[\boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\mathbf{U}_i^{1/2}\boldsymbol{\Omega}_i^{-1}\mathbf{1}_i\mathbf{1}_i^T\boldsymbol{\Omega}_i^{-1}\mathbf{U}_i^{1/2}\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i)\big] - \operatorname{tr}\big\{c\boldsymbol{\Omega}_i^{-1}\mathbf{1}_i\mathbf{1}_i^T\big\} \\
&\quad = \operatorname{tr}\big\{\mathbf{1}_i^T\boldsymbol{\Omega}_i^{-1}\mathbf{U}_i^{1/2}\mathbb{E}\big[\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i)\boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\big]\mathbf{U}_i^{1/2}\boldsymbol{\Omega}_i^{-1}\mathbf{1}_i\big\} - \operatorname{tr}\big\{c\boldsymbol{\Omega}_i^{-1}\mathbf{1}_i\mathbf{1}_i^T\big\} \\
&\quad = \mathbf{1}_i^T\boldsymbol{\Omega}_i^{-1}\mathbf{U}_i^{1/2}\mathbb{E}\big[\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i)\boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\big]\mathbf{U}_i^{1/2}\boldsymbol{\Omega}_i^{-1}\mathbf{1}_i - c\mathbf{1}_i^T\boldsymbol{\Omega}_i^{-1}\mathbf{1}_i, \qquad (5.34)
\end{aligned}$$

where, to it make clear, ($[\boldsymbol{\Omega}_i^{-1}]_{jk}$ denotes element (j,k) of $\boldsymbol{\Omega}_i^{-1}$)

$$c\,\mathbf{1}_i^T\boldsymbol{\Omega}_i^{-1}\mathbf{1}_i = c\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\left[\boldsymbol{\Omega}_i^{-1}\right]_{jk}.$$

Already at this stage of reasoning it is apparent that our above argument of using the trace operator (and the implied restriction to consider only diagonal elements) does not apply to (5.34). Here, we cannot disregard the off-diagonal structure of $\boldsymbol{\Omega}_i^{-1}$. Furthermore, it is evident from (5.34) that the expression in (5.34) equates to zero (which is a necessary condition for Fisher consistency at the core model) if $\mathbb{E}\big[\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i)\boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\big]$ is equal to $c\,\mathbf{U}_i^{-1/2}\boldsymbol{\Omega}_i\mathbf{U}_i{}^-1/2$, where c is determined – according to the Sinha–Rao proposal – under the standard normal model; see (5.30). Now, it is very "unlikely" that the consistency correction term related to the off-diagonal elements of $\mathbb{E}\big[\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i)\boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\big]$, which shall be denoted by $\tilde{c}$, effectively behaves exactly like c. More precisely, the diagonal elements $\operatorname{diag}(\mathbb{E}\big[\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i)\boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\big])$ indeed equate to $c\,\mathbf{I}_i$; this can be seen using an argument like the one used in case of σ_e^2. However, in case of the off-diagonal elements, there is sufficient evidence that c is not appropriate (even worse, a correction term will be needed that depends on $\mathbf{U}_i^{-1/2}\boldsymbol{\Omega}_i\mathbf{U}_i^{-1/2}$). Eventually, it is seen from (5.34) that instead of $c\,\mathbf{1}_i^T\boldsymbol{\Omega}_i^{-1}\mathbf{1}_i$, a linear combination of c and $\tilde{c}$ is needed for (5.34) to be equal to zero.

Implication

What does the lack of Fisher consistency imply? This essentially means that the M-estimator of σ_u^2 is biased at the Gaussian core model. How severe is the bias? Frankly, we don't know. The best way to quantify this bias is (in our opinion) by simulation. Unfortunately, we do not own a reliable software implementation of the Sinha–Rao method; hence, this simulation study could not have been undertaken in the thesis and remains subject to future research.

Our numerical approximation of the correct consistency correction term (via simulation) lets us conjecture that the bias is, for k sufficiently large, negligible. For small values of k (i.e., higher robustness, e.g. $1 < k < 1.3$), the incurred bias can be substantial, e.g., 5-10%. Furthermore, any bias in estimating σ_u^2 translates through the model and, as a rule, also affects the estimators of σ_e^2 and $\boldsymbol{\beta}$. The lack of Fisher consistency at the core model is one issue of the Sinha–Rao method. Another issue is that the improper decorrelation of the residuals makes it really hard (if not impossible) to choose an appropriate tuning constant k in order to obtain a certain degree of robustness since we cannot refer to the behavior of the ψ-function at the standard normal distribution. This is actually a real problem (since the choice of a fixed k has a different effect, depending on the scale of the observations; viz. $\sigma = 2$ vs. $\sigma = 15$).

One may go through the effort studying the numerical accuracy of the approximation in detail. We renounced to do so for essentially one single reason.

It is not worth the effort. The very same argument applies to working out an analytic expression for κ (see Appendix D). The major issue is that both, an even more sophisticated approximation and the analytic solution to κ, *depend on* $\rho = \sigma_u^2/(\sigma_e^2 + \sigma_u^2)$. The real tragedy is that ρ is required prior to computation *but depends itself on the estimand* σ_u^2 (snake biting in its own tail)! The only way to escape this trap is (presumably) to add the expression for κ to the estimation problem and update, say, $\kappa(\hat{\rho}, k)$ and the estimate of σ_u^2 jointly in some iterative manner (assuming that such a method is feasible). However, this approach seems to be ridiculously complicated; instead, it is considerably simpler to use the method we proposed (or live with the bias incurred by the Sinha–Rao proposal). Again, we point out that further simulation evidence is required in order to quantify the bias in real-world applications.

5.4.3. Measures of robustness

In the preceding sections, we have discussed robust M-estimators of the parameters at the core model. Now, we study the methods' robustness with regard to outlying observations in more detail using the principal tools of robust statistics, influence function and breakdown point. The influence function (IF; see Hampel et al., 1986, chap. 2.1), given in Definition 2.5 of Section 2.3.1, is a measure of robustness which indicates the extent to which an estimator is influenced by an infinitesimal amount of contamination.

5.4.3.1. Influence curve

First, we shall study the IF of the m.l.e. Consider the working model in (5.2) and recall that the parameter of interest is $\boldsymbol{\tau} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$, where $\boldsymbol{\theta}$ denotes the vector of variance components (σ_e^2, σ_u^2). For ease of discussion, we utilize the "classical" parametrization of the covariance matrix, given by $\boldsymbol{\Omega}_i(\boldsymbol{\theta}) = \sigma_e^2 \mathbf{I}_i + \sigma_u^2 \mathbf{J}_i$, $i = 1, \ldots, g$, (see Eq. 5.8) instead of the Hartley–Rao parametrization.[3] Let

$$\mathbf{u}_i = \boldsymbol{\Omega}_i^{-1/2}\big(\hat{\boldsymbol{\theta}}_{ML}(F)\big)\big(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{ML}(F)\big), \quad i = 1, \ldots, g,$$

where the m.l.e.'s $\hat{\boldsymbol{\beta}}_{ML}(F)$ and $\hat{\boldsymbol{\theta}}_{ML}(F)$ are expressed as functionals of the c.d.f. F. The influence of an infinitesimal contamination in the ith area on the joint m.l.e. $\hat{\boldsymbol{\tau}}_{ML}(F) = \big(\hat{\boldsymbol{\beta}}_{ML}^T(F), \hat{\boldsymbol{\theta}}_{ML}^T(F)\big)^T$ is given by (see Welsh and Richardson, 1997, 352)

$$\mathrm{IF}_{\boldsymbol{\beta}}\big(\mathbf{u}_i, \mathbf{X}_i, \hat{\boldsymbol{\beta}}_{ML}(F)\big) = \mathbf{G}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\mathbf{X}_i^T\mathbf{V}_i^{-1/2}\mathbf{u}_i \tag{5.35}$$

$$\mathrm{IF}_{\boldsymbol{\theta}}\big(\mathbf{u}_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}_{ML}(F)\big) = \mathbf{G}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\begin{bmatrix} \mathbf{u}_i^T\boldsymbol{\Omega}_i^{-1}\mathbf{u}_i - \mathrm{tr}\{\boldsymbol{\Omega}_i^{-1}\} \\ \mathbf{u}_i^T\boldsymbol{\Omega}^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\mathbf{u}_i - \mathrm{tr}\{\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\} \end{bmatrix}, \tag{5.36}$$

[3] Since the two forms of parametrizations are related by a continuous bijective function, the invariance property for m.l.e.'s ensures that we can obtain the m.l.e. of one parametrization from the other parametrization.

where

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{\boldsymbol{\beta\beta}} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\boldsymbol{\theta\theta}} \end{bmatrix} \tag{5.37}$$

with

$$\mathbf{G}_{\boldsymbol{\beta\beta}} = \frac{1}{n} \sum_{i=1}^{g} \mathbf{X}_i^T \boldsymbol{\Omega}_i^{-1} \mathbf{X}_i$$

and

$$\mathbf{G}_{\boldsymbol{\theta\theta}} = \frac{1}{2n} \sum_{i=1}^{g} \begin{bmatrix} \operatorname{tr}\{\boldsymbol{\Omega}_i^{-2}\} & \operatorname{tr}\{\boldsymbol{\Omega}_i^{-2}\mathbf{J}_i\} \\ \operatorname{tr}\{\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1}\} & \operatorname{tr}\{\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\} \end{bmatrix}.$$

Remarks. (i) $\mathrm{IF}_{\boldsymbol{\beta}}(\cdot)$ and $\mathrm{IF}_{\boldsymbol{\theta}}(\cdot)$ denote, respectively, the influence function w.r.t. to the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The influence function of the joint parameter $(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ is denoted by $\mathrm{IF}(\cdot)$, and is composed of $\mathrm{IF}_{\boldsymbol{\beta}}(\cdot)$ and $\mathrm{IF}_{\boldsymbol{\theta}}(\cdot)$.

(ii) The matrix $\boldsymbol{\Omega}_i$ depends on the variance components $\boldsymbol{\theta}$, more specifically, $\boldsymbol{\theta}(F)$; for ease of reading, we have suppressed this dependency.

(iii) The block-diagonal structure of $\mathbf{G}$ in (5.37) is implied by Definition 5.1.

It is evident from (5.35) and (5.36) that the IF of the m.l.e. is unbounded in $\mathbf{u}_i$ and $\mathbf{X}_i$. By Assumption 5.3, however, we have ruled out influential leverage observations/ outliers in the model's design space. Therefore, we shall restrict attention to the influence w.r.t. $\mathbf{u}_i$, $i = 1, \ldots, g$. Since the IF of the joint m.l.e. is an unbounded function in $\mathbf{u}_i$, deviations from the core model can cause an arbitrarily large bias of the estimators $\hat{\boldsymbol{\beta}}_{ML}$ and $\hat{\boldsymbol{\theta}}_{ML}$.

Next, we address the influence function of the robust estimator. Here too, we make use of the parametrization in terms of $\boldsymbol{\Omega}_i$, not the Hartley–Rao representation. Therefore, the robust estimator of the model parameters is given by $\hat{\boldsymbol{\tau}}^R = ([\hat{\boldsymbol{\beta}}^R]^T, [\hat{\boldsymbol{\theta}}^R]^T)^T$, where $\hat{\boldsymbol{\beta}}^R$ is defined as the solution to (5.17) and $\hat{\boldsymbol{\theta}}^R$ equals $(\hat{v}^R, \hat{v}^R\hat{d}^R)$, where $\hat{v}^R$ and $\hat{d}^R$ are the solutions to, respectively, (5.21) and (5.24). For ease of discussion, we work with the estimator $\hat{\boldsymbol{\theta}}^R$ of (σ_e^2, σ_u^2) given by the solution to

$$\sum_{i=1}^{g} \begin{bmatrix} \boldsymbol{\psi}(\mathbf{u}_i)^T \boldsymbol{\Omega}_i(\boldsymbol{\theta})^{-1} \boldsymbol{\psi}\{\mathbf{u}_i(\boldsymbol{\theta})\} - \operatorname{tr}\{K_2 \boldsymbol{\Omega}_i(\boldsymbol{\theta})^{-1}\} \\ \boldsymbol{\psi}\{\mathbf{u}_i(\boldsymbol{\theta})\}^T \boldsymbol{\Omega}_i(\boldsymbol{\theta})^{-1/2} \mathbf{J}_i \boldsymbol{\Omega}_i(\boldsymbol{\theta})^{-1/2} \boldsymbol{\psi}\{\mathbf{u}_i(\boldsymbol{\theta})\} - \operatorname{tr}\{K_2 \boldsymbol{\Omega}_i(\boldsymbol{\theta})^{-1} \mathbf{J}_i\} \end{bmatrix} = \mathbf{0}, \tag{5.38}$$

instead of the equivalent representation $(\hat{v}^R, \hat{v}^R\hat{d}^R)$. The reason for doing so is that the "classical" parametrization enables us to utilize the work of Welsh and Richardson (1997). The term K_2 in (5.38) denotes a consistency correction term; see Welsh and Richardson (1997, 362). The IF of $\hat{\boldsymbol{\tau}}^R$ is given by

$$\mathrm{IF}\big(\mathbf{u}_i, \mathbf{X}_i, \hat{\boldsymbol{\tau}}^R\big) =$$

$$(1/2)\big(\mathbf{G}^{\psi}\big)^{-1}\begin{bmatrix} 2\mathbf{X}_i^T\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}(\mathbf{u}_i) \\ \boldsymbol{\psi}(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{\psi}(\mathbf{u}_i) - \mathrm{tr}\{K\boldsymbol{\Omega}_i^{-1}\} \\ \boldsymbol{\psi}(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}(\mathbf{u}_i) - \mathrm{tr}\{K\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\} \end{bmatrix}, \tag{5.39}$$

where (see Welsh and Richardson, 1997, 363)[4]

$$\mathbf{G}^{\psi} = \begin{bmatrix} \mathbf{G}^{\psi}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{G}^{\psi}_{\boldsymbol{\beta}\sigma_e^2} & \mathbf{G}^{\psi}_{\boldsymbol{\beta}\sigma_u^2} \\ \big(\mathbf{G}^{\psi}_{\sigma_e^2\boldsymbol{\beta}}\big)^T & G^{\psi}_{\sigma_e^2\sigma_e^2} & G^{\psi}_{\sigma_e^2\sigma_u^2} \\ \big(\mathbf{G}^{\psi}_{\sigma_u^2\boldsymbol{\beta}}\big)^T & G^{\psi}_{\sigma_u^2\sigma_e^2} & G^{\psi}_{\sigma_u^2\sigma_u^2} \end{bmatrix},$$

and

$$\begin{aligned}
\mathbf{G}^{\psi}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= \frac{1}{n}\sum_{i=1}^{g}\mathbb{E}_F\big\{\mathbf{X}_i^T\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}'(\mathbf{u}_i)\boldsymbol{\Omega}_i^{-1/2}\mathbf{X}_i\big\} \\
\mathbf{G}^{\psi}_{\boldsymbol{\beta}\sigma_e^2} &= \frac{1}{2n}\sum_{i=1}^{g}\mathbb{E}_F\big\{\mathbf{X}_i^T\boldsymbol{\Omega}_i^{-3/2}\boldsymbol{\psi}(\mathbf{u}_i) + \mathbf{X}_i^T\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}'(\mathbf{u}_i)\boldsymbol{\Omega}_i^{-1}\mathbf{u}_i\big\} \\
\mathbf{G}^{\psi}_{\boldsymbol{\beta}\sigma_u^2} &= \frac{1}{2n}\sum_{i=1}^{g}\mathbb{E}_F\big\{\mathbf{X}_i^T\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}(\mathbf{u}_i) + \mathbf{X}_i^T\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}'(\mathbf{u}_i)\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\mathbf{u}_i\big\} \\
G^{\psi}_{\boldsymbol{\beta}\sigma_e^2} &= \frac{1}{2n}\sum_{i=1}^{g}\mathbb{E}_F\big\{\mathbf{X}_i^T\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}'(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{\psi}(\mathbf{u}_i) + \boldsymbol{\psi}(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{\psi}'(\mathbf{u}_i)\boldsymbol{\Omega}_i^{-1/2}\mathbf{X}_i\big\} \\
G^{\psi}_{\boldsymbol{\beta}\sigma_u^2} &= \frac{1}{2n}\sum_{i=1}^{g}\mathbb{E}_F\big\{\mathbf{X}_i^T\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}'(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}(\mathbf{u}_i) \\
&\qquad + \boldsymbol{\psi}(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}'(\mathbf{u}_i)\boldsymbol{\Omega}_i^{-1/2}\mathbf{X}_i\big\} \\
G^{\psi}_{\sigma_e^2\sigma_e^2} &= \frac{1}{4n}\sum_{i=1}^{g}\big\{2\boldsymbol{\psi}(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-2}\boldsymbol{\psi}(\mathbf{u}_i) + 2\mathbf{u}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{\psi}'(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{\psi}(\mathbf{u}_i) - \mathrm{tr}\{K_2\boldsymbol{\Omega}_i^{-2}\}\big\} \\
G^{\psi}_{\sigma_e^2\sigma_u^2} &= \frac{1}{4n}\sum_{i=1}^{g}\big\{2\boldsymbol{\psi}(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-3/2}\boldsymbol{\psi}(\mathbf{u}_i) + \mathbf{u}_i^T\boldsymbol{\Omega}_i^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}'(\mathbf{u}_i)\boldsymbol{\Omega}_i^{-1}\boldsymbol{\psi}(\mathbf{u}_i) \\
&\qquad + \boldsymbol{\psi}(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{\psi}'(\mathbf{u}_i)\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\mathbf{u}_i + \mathrm{tr}\{K_2\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1}\}\big\} \\
G^{\psi}_{\sigma_u^2\sigma_e^2} &= \frac{1}{4n}\sum_{i=1}^{g}\big\{2\boldsymbol{\psi}(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-3/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}(\mathbf{u}_i) + \mathbf{u}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{\psi}'(\mathbf{u}_i)\boldsymbol{\Omega}_i^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}(\mathbf{u}_i) \\
&\qquad + \boldsymbol{\psi}(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}'(\mathbf{u}_i)\boldsymbol{\Omega}_i^{-1}\mathbf{u}_i + \mathrm{tr}\{K_2\boldsymbol{\Omega}_i^{-2}\mathbf{J}_i\}\big\} \\
G^{\psi}_{\sigma_u^2\sigma_u^2} &= \frac{1}{4n}\sum_{i=1}^{g}\big\{2\boldsymbol{\psi}(\mathbf{u}_i)^T\boldsymbol{\Omega}_i^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}(\mathbf{u}_i) \\
&\quad + \quad 2\mathbf{u}_i^T\boldsymbol{\Omega}_i^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}'(\mathbf{u}_i)\boldsymbol{\Omega}_i^{-1/2}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\psi}(\mathbf{u}_i) + \mathrm{tr}\{K_2\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\}\big\},
\end{aligned}$$

where $\mathbb{E}_F$ denotes expectation w.r.t. the c.d.f. F and $\boldsymbol{\psi}'(\mathbf{z})$ is the partial derivative of $\boldsymbol{\psi}(\mathbf{z})$ w.r.t. $\mathbf{z}$ [a.e. F] (and is w.l.o.g. set equal to zero on the set of measure

[4] In the framework of Welsh and Richardson (1997), we set $\psi_0 = \psi_1 = \psi_2$ (i.e., we do not consider different ψ-functions), $\mathbf{U}_{oj} = \mathbf{U}_{1j} = \mathbf{I}_j$, $\mathbf{W}_{1j} = \mathbf{W}_{oj} = \mathbf{I}_j$, where $\mathbf{I}_j$ denotes the $(n_j \times n_j)$ identity matrix of subgroup/ area j; the matrices $\mathbf{W}$ and $\mathbf{U}$ (subscripts are suppressed) refer to GM-estimators of the type Mallows, Schweppe, Andrews or Hill–Ryan.

zero); K_2 is a consistency correction term.

Remarks. (i) It is important to point out that when there is no single criterion function (e.g., likelihood function) from which the estimating equations are derived, then matrix $\mathbf{G}^{\psi}$ need not be symmetric (Welsh and Richardson, 1997, 363).

(ii) Provided ψ is a bounded function (and given some further regularity conditions), the influence of an infinitesimal amount of contamination in the ith area on the estimator $\hat{\boldsymbol{\tau}}^T$ is bounded; see (5.39). Thus, the potential bias of the estimator caused by contamination is bounded. [The reader shall be reminded that the proposed robust method limits the influence an outlying area exerts on the estimators.]

(iii) As pointed out, the IF in (5.39) refers to estimator $\hat{\boldsymbol{\tau}}^R$ with subvector $\hat{\boldsymbol{\theta}}^R$ as the solution to (5.38). The IF for the parametrization in terms of the ratio of variances, $(\hat{v}^R, \hat{d}^R)$, obtains by taking into account the transformation $(\sigma_e^2, \sigma_u^2) \mapsto (\sigma_e^2, \sigma_u^2/\sigma_e^2)$ and the Jacobian matrix of the transformation; hence, the IF obtains in straightforward manner; see Hampel et al. (1986, 229 and chap. 4.5).

5.4.3.2. Sensitivity curve

The influence function in (5.39) illustrates the qualitative robustness of the proposed method, but the analytic expression is laborious and hard to work with. Therefore, we introduce the sensitivity curve (SC; see also Section 2.3), which is – loosely speaking – a measure of the (standardized) effect of an estimator $T(\mathcal{X})$ based on sample $\mathcal{X} = \{x_1, \ldots, x_n\}$ when one observation is replaced by, say, z_0. The finite-sample nature of the SC eases the study of robustness.

In the problem at hand, the above definition of the sensitivity curve must be slightly altered to be meaningful (since the core model imposes more structure than is present in the simple location model). It is straightforward to define the following two types of SC,

(i) area-level sensitivity curve,

(ii) unit-level sensitivity curve.

The former refers to the effect that obtains if all elements of a particular area are altered (i.e., set equal to z_0); the latter is conceptually equivalent to the standard SC, where only one observation is affected.

Under the basic unit-level model, three parameters have to be estimated ($\boldsymbol{\beta}$, σ_e^2, and σ_u^2). We consider only the variance parameters – which are, as a rule, more sensitive – and their respective SC (i.e. area-level SC for σ_u^2 and unit-level

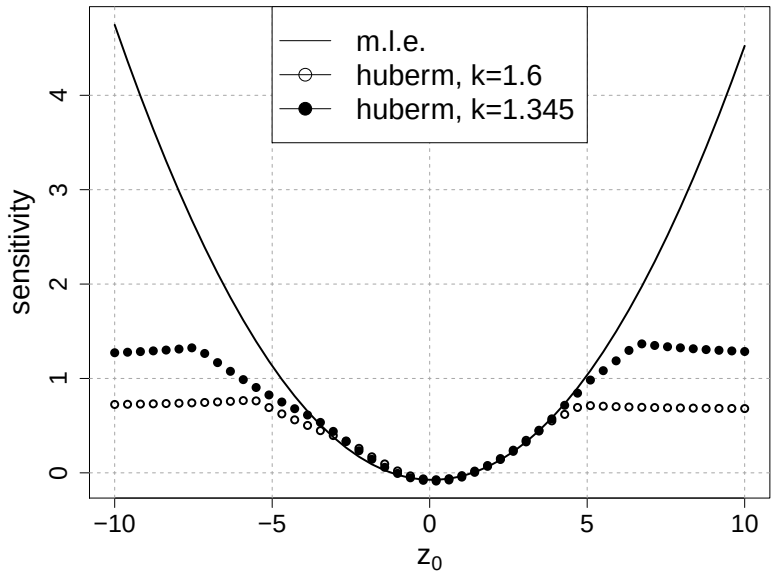


Figure 5.2.: Area-level sensitivity plot for three estimators of σ_u^2 (model: balanced data, $n_i = 4$ units in each of the $g = 20$ areas, $\sigma_u^2 = \sigma_e^2 = \beta = \alpha = 1$; the sensitivity refers to the effect resulting when all observations in a single area are replaced by z_0).

SC in case of σ_e^2). For the numerical analysis, the following model specification is adopted,

$$y_{ij} = \alpha + x_{ij}\beta + u_i + e_i, \qquad j = 1, \ldots, n_i, \quad i = 1, \ldots, g,$$

with

$$e_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_e^2), \qquad u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2),$$

where $\alpha = \beta = 1$, $\sigma_e^2 = \sigma_u^2 = 1$, and $x_{ij} \sim \mathcal{N}(0, 1)$, $j = 1, \ldots, n_i$, $i = 1, \ldots, g$; $g = 20$ areas, each of which includes $n_i = 4$ units (balanced data). In the numerical analysis, the point measure z_0 (see definition of the SC, above) is taken on 50 equidistant points in the interval $[-10, 10]$.[5]

The results of the area-level SC are shown in Figure 5.2 for three different estimators of σ_u^2 (i.e., m.l.e. and the proposed method, denoted by `huberm`, with two different robustness tuning constants k). It is apparent from the visual display in Figure 5.2 that the area-level SC for both robust estimators are bounded. Moreover, the shape of the curves clearly illustrate that the M-estimator of σ_u^2 derives from the "Proposal 2" estimator of Huber (1964); i.e., the ψ-function of the M-estimator of scale is proportional to ψ^2. On the other hand, the sensitivity

[5] The empirical implementation of the SC utilizes the function `fitsaemodel` in the `R`-package `rsae` with argument `method` equal to `huberm`. Area-level SC: For each z_0, all units in one particular area are replaced by z_0, then `fitsaemodel` is computed and the estimated parameters were stored and used to compute the SC.

curve for the m.l.e. is unbounded (by design).

Since the SC's for the estimators of σ_e^2 show an analogous qualitative behavior to the ones depicted in Figure 5.2, we do not show them. Furthermore, the plotted SC can also be seen as evidence that the methods in package `rsae` work properly (in the sense of a proof of concept).

5.4.3.3. Breakdown point

Although M-estimators in MLM are qualitatively robust (i.e., have a bounded influence function), the breakdown point (BP) can be low; see e.g. Richardson (1995), Welsh and Richardson (1997), Wellmann (1994), Wellmann (2000), Müller and Uhlig (2001) or Wellmann and Gather (2003).

The notion of breakdown point (see Section 2.3.1) provides a crude *quantification* of the robustness of an estimator (beyond infinitesimal contamination). For ease of discussion, we restrict attention to the basic unit-level model with *balanced data*. The insights gained from this simplified setup also apply to estimators under the general model.

Simplified model

To motivate the problem, we consider the balanced one-way classification model,

$$y_{ij} = \mu + u_i + e_{ij}, \qquad j = 1, \dots, n, \qquad j = 1, \dots, g,$$

where $u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2)$, $e_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_e^2)$, and $\mu \in \mathbb{R}$. Instead of the parametrization in terms of (σ_u^2, σ_e^2), we put $\vartheta = \sigma_u^2 + \sigma_e^2/n$ (see Searle et al., 1992, chap. 3.7a). The m.l.e. is

$$\left.\begin{aligned} \tilde{\mu} &= \bar{\bar{y}} \\ \tilde{\sigma}_e^2 &= \frac{1}{g(n-1)} \sum_{i=1}^{g} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 \\ \tilde{\vartheta} &= \frac{1}{g} \sum_{i=1}^{g} (\bar{y}_i - \bar{\bar{y}})^2 \end{aligned}\right\},$$

where

$$\bar{\bar{y}} = \frac{1}{g} \sum_{i=1}^{g} \bar{y}_i, \qquad \text{and} \qquad \bar{y}_i = \frac{1}{n} \sum_{j=1}^{n} y_{ij}, \qquad i = 1, \dots, n.$$

Under the balanced one-way classification model, the most B-robust estimator is (Stahel and Welsh, 1997, 303)

$$\left.\begin{aligned} \hat{\mu} &= \operatorname{median}\big\{\bar{y}_i \,:\, i = 1, \dots, g\big\} \\ \hat{\vartheta} &= 1.4826^2 \cdot \operatorname{median}\big\{|\bar{y}_i - \hat{\mu}|^2 \,:\, i = 1, \dots, g\big\} \\ \hat{\sigma}_e^2 &= \operatorname{median}\Big\{ \sum_{j \le n} (y_{ij} - \bar{y}_i)^2 \,:\, i = 1, \dots, g \Big\} / \operatorname{median}\big(\chi^2_{n-1}\big) \end{aligned}\right\}, \tag{5.40}$$

A somewhat surprising feature is the appearance of the group/ area means in

the formula of $\hat{\vartheta}$, a fact that has also been discussed by Stahel and Welsh (1997, 303). They point out that this property is a consequence of treating the vector of all y_{ij}'s in a group i (area) as one observation and of using the Gaussian model as the "central" model.

No affine equivariant scale estimator is going to do any better in terms of robustness than the median. For sample data without ties, the median (respectively, MAD from zero) has an ϵ-replacement FBP [see Definition 2.8 and the discussion in Section 2.3] $\varepsilon^*(MAD, \mathbf{y}) = \lfloor n/2 \rfloor / n$ which is no larger than $1/2$ (Donoho and Huber, 1983, 161). However, the appearance of the $\bar{y}_i$'s in the formulas of the most B-robust estimator imply that the FBP can be rather low although the estimators are based on the median. The FBP of estimator $\hat{\sigma}_e^2$ is given by $\varepsilon^*(\hat{\sigma}_e^2, \mathbf{y}) = \lfloor g/2 \rfloor / (gn) \approx 1/(2n)$ (for sufficiently large g). From this it is evident that the FBP can be dangerously low when n (i.e. within-area sample size) is large.

Un-balanced sample

Under the basic unit-level model with un-balanced data, we also encounter a relatively low FBP for the M-estimator. Note that the square root of the inverse covariance matrix [using the "canonical" parametrization, not the Hartley–Rao parametrization, see Formula (5.25)] can be written as

$$\boldsymbol{\Omega}_i^{-1/2} = \underbrace{\frac{1}{\sqrt{\sigma_e^2}} \left(\frac{1}{\sqrt{1 + n_i(\sigma_u^2/\sigma_e^2)}} - 1 \right)}_{c_1} \frac{1}{n_i} \mathbf{1}_i \mathbf{1}_i^T + \underbrace{\frac{1}{\sqrt{\sigma_e^2}}}_{c_2} \mathbf{I}_i, \qquad i = 1, \ldots, g.$$

Now, the joint M-estimator of $(\boldsymbol{\beta}, \sigma_e^2, \sigma_u^2)$ discussed in Section 5.4.1 depends on the standardized residuals, $\boldsymbol{\Omega}_i^{-1/2}\mathbf{r}_i$, where $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$, which can be expressed as

$$\boldsymbol{\Omega}_i^{-1/2}\mathbf{r}_i = c_1 \bar{y}_i \mathbf{1}_i + c_2 \mathbf{y}_i - \boldsymbol{\Omega}_i^{-1/2}\mathbf{X}_i\boldsymbol{\beta}, \qquad i = 1, \ldots, g,$$

where $\bar{y}_i = (1/n_i) \sum_{j \leq n_i} y_{ij}$. Observe that the appearance of the area-level means $\bar{y}_i$ are to blame for the relatively low BP of the M-estimator. A simple but effective remedy is to replace the $\bar{y}_i$'s by the area-specific median.

5.4.4. Robust prediction

So far, we have studied robust estimating methods of the model parameters. Now, we consider robust prediction of the random effects (and subsequently the small-area means). We assume that $N_i \gg n_i$ ($\forall i = 1, \ldots, g$). On rewriting the BLUP equation in (5.6) using the Hartely–Rao variance parametrization, the area-level predicting equations are given by

$$\hat{\mu}_i = \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}} + \hat{u}_i, \qquad i = 1, \ldots, g,$$

with

$$\hat{u}_i = \hat{a}\mathbf{1}_i^T\boldsymbol{\Omega}_i(\hat{v},\hat{a})^{-1}\big[\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\big], \tag{5.41}$$

where

$$\boldsymbol{\Omega}_i(\hat{v},\hat{a})^{-1} \equiv \frac{1}{\hat{v}}\mathbf{V}_i(\hat{d})^{-1}.$$

If $N_i \gg n_i$ does not hold we may proceed as in Section 5.3. From (5.41) it is apparent that only replacing $\hat{\boldsymbol{\beta}}$, $\hat{v}$, and $\hat{a}$ by robust estimates is not sufficient in order to get robust predictions of μ_i [since Formula (5.41) does not limit the influence of outliers in $\mathbf{y}_i$]. As Sinha and Rao (2009) indicate, $\hat{u}_i$ has to be replaced by a robustly predicted random effect, $\hat{u}_i^R$, as well. They therefore propose to solve Fellner's robust mixed-model equation (Fellner, 1986)

$$\mathbf{1}_i^T\frac{1}{\sqrt{v}}\boldsymbol{\psi}_k\left(\frac{1}{\sqrt{v}}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{1}_i\hat{u}_i^R)\right) - \frac{1}{\sqrt{a}}\boldsymbol{\psi}_k\left(\frac{1}{\sqrt{a}}\hat{u}_i^R\right) = 0, \tag{5.42}$$

for u_i. In order to solve (5.42), Sinha and Rao (2009) suggest a Newton–Raphson algorithm applied to a first-order Taylor series expansion of (5.42). Consequently, computation is very involved. However, one can obtain robust predictions far more easily (cf. Copt and Victoria-Feser, 2009). If we put $\boldsymbol{\psi}_c(\mathbf{u}_i) = \big(\psi_c(u_{i1}),\ldots,\psi_c(u_{in_i})\big)^T$, where $\psi_c(\cdot)$ is the Huber ψ-function indexed by the robustness tuning constant c, then we may write

$$\hat{u}_i^R = \kappa\frac{\hat{a}^R}{\sqrt{\hat{v}^R}}\mathbf{1}_i^T\mathbf{V}_i^{-1/2}(\hat{d}^R)\boldsymbol{\psi}_c\left[\frac{1}{\sqrt{\hat{v}^R}}\mathbf{V}_i^{-1/2}(\hat{d}^R)[\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}]\right], \tag{5.43}$$

where

$$\kappa = [-2c\phi(c) + 2\Phi(c) - 1 + 2c^2(1-\Phi(c))]^{-1/2}$$

with ϕ and Φ, respectively, the p.d.f. and c.d.f. of the standard normal distribution. Note that κ is kind of a consistency correction term which has been chosen such that $\hat{u}_i^R$ behaves like $\tilde{u}_i$ at the core model. In essence, we follow Heritier et al. (2009) and impose the (implicit) moment conditions that $\mathbb{E}[\hat{u}_i^R] = 0$ and $\mathbb{V}[\hat{u}_i^R] = \mathbb{V}[\tilde{u}_i]$ (Heritier et al., 2009, 113–114). Thus, the robust predictor of the area mean $\hat{\mu}_i^R$ – referred to as robust EBLUP (REBLUP) of μ_i – is given by

$$\hat{\mu}_i^R = \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}}^R + \hat{u}_i^R, \qquad i = 1,\ldots,g. \tag{5.44}$$

This robust-projective alternative to the EBLUP estimates the y-mean of the ith area. A minor modification to (5.44) is the following predictor (see Chambers et al., 2014)

$$\hat{\mu}_i^{R*} = N_i^{-1}\big\{n_i\bar{y}_{is} + (N_i - n_i)(\bar{\mathbf{x}}_{ir}^T\hat{\boldsymbol{\beta}}^R + \hat{u}_i^R)\big\}, \qquad i = 1,\ldots,g, \tag{5.45}$$

which restricts the model dependency to the non-sampled units; here, $\bar{y}_{is}$ denotes the sample y-mean in area i and $\bar{\mathbf{x}}_{ir}$ is the x-mean of the non-sampled elements in the ith area.

Remarks. (i) A problem with the robust-projective approach (in Eq. 5.44 or 5.45) is that it assumes that all non-sampled elements follow the underlying model. This implies, as has been pointed out by Chambers et al. (2014, 51), that any deviations from the model are considered as noise and therefore cancel out "on average". In particular, under the basic unit-level model and provided that the individual errors e_{ij} are symmetrically distributed about zero, REBLUP will perform well because the averages of these errors over the non-sampled units in the ith area cancel out. However, "[t]his does not mean that these non-sample units are not outliers. It is just that our best prediction of the corresponding small area average value of their model errors is 0" (Chambers et al., 2014, 51). In order to limit (or control) the *potential* prediction bias of the REBLUP method incurred by representative outliers, Chambers et al. (2014) propose a bias-corrected robust predictive estimator of the small area means – relating to the ideas of Welsh and Ronchetti (1998) and Chambers (1986) – which is defined as

$$\hat{\mu}_i^{R^*-BC} = \hat{\mu}_i^{R^*} + (n_i^{-1} - N_i^{-1}) \sum_{j \in s_i} \hat{s}_{ij} \psi_{c_2}\big\{(y_j - \mathbf{x}_j^T \hat{\boldsymbol{\beta}}^R - \hat{u}_j^R)/\hat{s}_{ij}\big\}, \qquad (5.46)$$

where the $\hat{s}_{ij}$'s denote robust estimates of the scale of the residuals $y_j - \mathbf{x}_j^T \hat{\boldsymbol{\beta}}^R - \hat{u}_j^R$ in area i; c_2 is a tuning constant that indexes function ψ_{c_2}. For this proposal to work properly, the analyst has to specify three robustness tuning constants (k for the robust estimation of the model parameters, c for the prediction of the random effects $\hat{u}_i^R$, and c_2 for the bias correction). Jiongo et al. (2013) take the level of sophistication even one step further. They point out that the bias correction term in (5.46) takes only within-area units into account (i.e., elements $j \in s_i$) and neglects the information associated with units which are not in the ith area, and which may still influence the estimators. Therefore, Jiongo et al. (2013, 848–49) come up with what they call the fully bias-corrected method by adding an additional correction term to (5.46) that accounts for deviations in the residuals for sampled units in all but the area under consideration (i.e., $j \in s \setminus s_i$). As a consequence, their method requires to specify four tuning constants instead of three.

(ii) Note that pre-multiplying the area-specific vector of residuals, $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$, in (5.43) by $\mathbf{V}_i^{-1/2}$ from (5.14) will transmit the effect of even one single outlying residual, $r_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}$, say, to the vector of all other within-area residuals. That is, on pre-multiplying $\mathbf{r}_i$ by $\mathbf{V}_i^{-1/2}$, the first term of (5.14) yields the mean of $\mathbf{r}_i$ (times a constant), which is non-robust per se. Thus, from the perspective of robustness (and with regard to breakdown point considerations), the term $\mathbf{V}_i^{-1/2}\mathbf{r}_i$ with $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$ in (5.43) should

be replaced by

$$\left(\frac{1}{\sqrt{1+dn_i}}-1\right)\mathbf{1}_i\bar{r}_i^{med}+\mathbf{r}_i, \tag{5.47}$$

$\bar{r}_i^{med}$ denoting the median of $\mathbf{r}_i$. Alternatively, and under Assumption 5.3 it is sufficient to use $\bar{y}_i^{med}-(1/n_i)\mathbf{1}_i^T\mathbf{X}_i\hat{\boldsymbol{\beta}}$, where $\bar{y}_i^{med}$ is the median of $\mathbf{y}_i$, instead of $\bar{r}_i^{med}$ in (5.47).

(iii) The function $\psi_c(\cdot)$ in (5.43) can be replaced by any other non-redescending, odd, bounded function. The restriction on non-redescending functions is crucial since redescending ψ-functions lead, for sufficiently large residuals in (5.43), to realizations of u_i equal to zero (i.e., mimicking a synthetic estimator) which is unappealing under model (5.2).

(iv) In what follows, we focus on predictor $\hat{\mu}_i^R$ defined in (5.44).

5.5. Mean squared error estimation

While EBLUP is fairly easy to obtain, the estimation of a reasonable measure of uncertainty for the predicted area-level means is a challenging problem. In their seminal paper, Prasad and Rao (1990) studied a second-order approximation to the mean square prediction error (MSPE) of the EBLUP. Datta and Lahiri (2000) extended the Prasad–Rao setting to a wider range of variance estimators, including the ML estimator. Given the complex nature of the REBLUP predictor and the lack of knowledge of the underlying distribution of the u_i and e_{ij}, Sinha and Rao (2009) noted that it is not possible to adopt the existing methods for MSPE estimation. Instead, they proposed a parametric bootstrap procedure [see Lahiri (2003) and Hall and Maiti (2006) for more details on bootstrap estimates]. We adopt the parametric bootstrap method of Sinha and Rao (2009) based on the robust quantities $\hat{\boldsymbol{\beta}}^R$ and $\hat{\boldsymbol{\theta}}^R$ to estimate

$$\text{MSPE}(\hat{\mu}_i^R)=\mathbb{E}\{\hat{\mu}_i^R-\mu_i\}^2.$$

The method works as follows.

1. For given $\hat{\boldsymbol{\beta}}^R$ and $\hat{\boldsymbol{\theta}}^R=(\hat{v}^R,\hat{a}^R)^T$, generate area-specific random effects u_i^* and random errors e_{ij}^* from $\mathcal{N}(0,\hat{a}^R)$ and $\mathcal{N}(0,\hat{v}^R)$, respectively. Then we create a bootstrap sample from the model

$$y_{ij}^*=\mathbf{X}_i\hat{\boldsymbol{\beta}}^R+u_i^*+e_{ij}^*,\quad j=1,\ldots,n_i;\ i=1,\ldots,g, \tag{5.48}$$

2. Generate $b=1,\ldots,B$ bootstrap samples $\{\mathbf{y}^{*[1]},\mathbf{y}^{*[2]},\ldots,\mathbf{y}^{*[B]}\}$ from the bootstrap population model (5.48). For each bootstrap sample $\mathbf{y}^{*[b]}$, compute robust bootstrap estimates $\hat{\boldsymbol{\beta}}^{R[b]}$, $\hat{\boldsymbol{\theta}}^{R[b]}$, and $\hat{u}^{R[b]}$ and robustly predict $\hat{\mu}_i^{R[b]}=\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}}^{R[b]}+\hat{u}_i^{R[b]}$.

3. For each area $i = 1, \ldots, g$, compute a bootstrap estimate of MSPE$\{\hat{\mu}_i^R\}$ as

$$\text{MSPE}_B\{\hat{\mu}_i^R)\} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\mu}_i^R(\hat{\boldsymbol{\beta}}^{[R[b]}, \hat{\boldsymbol{\theta}}^{R[b]}, \hat{u}_i^{R[b]}) - \mu_i(\hat{\boldsymbol{\beta}}^R, \hat{\boldsymbol{\theta}}^R, \hat{u}_i^{*[b]})\right)^2.$$

Remarks. (i) The parametric bootstrap method tends to slightly underestimate the true MSPE. The underestimation results mainly because the uncertainty of estimating $\boldsymbol{\beta}$ has not been taken into account.

(ii) A major objection against bootstrap procedures has been the time-consuming computational effort required to compute the estimates. Therefore, a lot of research effort was dedicated to developing analytical estimators; see e.g. Prasad and Rao (1990). For robust approaches, Chambers et al. (2014) proposed two (approximate) methods for MSPE estimation under the assumptions that the model conditions (for inference) on the realized values of the area-level effects (hence, the proposed MSPE estimators are conditional estimators). However, when the area-specific sample sizes are very small, the method leads to MSPE estimates with high variability.

(iii) We stick with the parametric bootstrap method because computing time is not really an issue for two reasons: (a) the software implementation of the proposed method is highly optimized (see `R`-package `rsae`, Schoch, 2011c), and (b) the bootstrap procedure can be easily parallelized. By way of illustration: the computation of 500 bootstrap replicates for twelve area totals [using the "landsat" data of Battese et al. (1988)] takes less than 3 seconds (without parallelization, i.e. single-core computation; see vignette of the `rsae` package). The computational aspects will be discussed in more detail in the following section.

5.6. Algorithm

5.6.1. Estimation bounds

Given some initial values, $\boldsymbol{\beta}_0$, v_0, and d_0, we may consider updating these estimates by solving equations (5.20), (5.23), and (5.24) in an alternating order. From a theoretical point of view, there is no objective against doing so. From the perspective of numerical computation, it will prove useful to introduce two (pre-) estimation bounds for the variance components. As d is concerned, we have to consider two limiting situations: $d = 0$ and $d \to \infty$. Accordingly, we obtain v_{zero} (if $d = 0$) and v_∞ (if $d \to \infty$), respectively. These two cases depict a lower and an upper bound of estimates of v. It is easy to prove that the following relations hold

$$v_\infty \leq v_{ML} \leq v_{zero},$$

where v_{ML} denotes the ML estimator; see e.g., (Demidenko, 2004, 78–79). From the perspective of numerical optimization, these bounds are extremely useful since they determine the range of plausible values which may guide the choice of initial values and identify potential run-away values. Subsequently, we shall study robust estimators of v_{zero} and v_{∞}.

Case I: Robust estimate of v_{zero}

In the first case, we have $d = 0$ which implies that $\mathbf{\Omega}_i(v, d) = v(\mathbf{I}_i + d\mathbf{J}_i)$ reduces to $v_{zero}\mathbf{I}_i$ $(i = 1, \ldots, g)$. As a consequence, the estimator of $\boldsymbol{\beta}$ collapses to the ordinary least squares (OLS) estimator, $\hat{\boldsymbol{\beta}}_0$, and the corresponding estimator $\hat{v}_{zero}$ of the residual variance. A robust estimate $\hat{v}^R_{zero}$ obtains as a byproduct from e.g. the least trimmed squares (LTS) regression (see Rousseeuw, 1984) or an M- or S-estimator of regression (see e.g., Maronna et al., 2006, chap. 5). This robust regression exercise not only yields a robust estimate of v_{zero}, but also provides us with a starting value for $\boldsymbol{\beta}$ in order to initialize the iterative algorithm (see below). From the perspective of computation, the fast LTS method of (Rousseeuw and Van Driessen, 2006) offers a good trade-off between robustness and computation time for sample sizes up to about 20,000 (this limit depends heavily on the number of auxiliary variables). For larger data sets, a regression S-estimator is considerably faster.

Case II: Robust estimate of v_{∞}

In the second case, we consider letting $\lim_{d\to\infty} \mathbf{V}_i^{-1}(v, d)$ (for all $i = 1, \ldots, g$ and holding v fixed) and obtain

$$\lim_{d\to\infty} [\mathbf{I}_{n_i} - d/(1 + dn_i)\mathbf{J}_{n_i}] = [\mathbf{I}_{n_i} - (1/n_i)\mathbf{J}_{n_i}]$$

Note that letting the random-effect variance d approach infinity corresponds to treating the random effects u_i, $i = 1, \ldots, g$, in the model as if they were fixed effects. Indeed, we shall consider the fixed-effects model as an alternative to the mixed linear model. Now, let the model matrix be partitioned as $[\mathbf{1}_n, \mathbf{X}]$; the corresponding parameter vector shall be partitioned accordingly, $\boldsymbol{\beta} = (\alpha, \boldsymbol{\gamma}^T)^T$. The fixed-effects model writes

$$\mathbf{y}_i = \alpha\mathbf{1}_{n_i} + \mathbf{X}_i\boldsymbol{\gamma} + \mathbf{1}_{n_i}u_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, g, \tag{5.49}$$

where the u_i's are unknown, but fixed parameters (not realizations of the area-level random effects). Model (5.49) is traditionally called the one-way classification model for the analysis of covariance (Searle, 1987, chap. 11.2). If we put $\mathbf{K} = [\mathbf{1}_n, \mathbf{Z}]$, where $\mathbf{Z} = \text{blockdiag}(\mathbf{1}_1, \ldots, \mathbf{1}_g)$, then model (5.49) can be written as a general linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{K}\mathbf{d} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{d} = (\alpha, \mathbf{u}^T)^T \qquad \text{with} \qquad \mathbf{u} = (u_1, \ldots, u_g)^T,$$

and $\varepsilon = \text{blockdiag}(\varepsilon_1, \ldots, \varepsilon_g)$. The normal equations of the general linear model are

$$\begin{bmatrix} \mathbf{K}^T\mathbf{K} & \mathbf{K}^T\mathbf{X} \\ \mathbf{X}^T\mathbf{K} & \mathbf{X}^T\mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{d} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{K}^T\mathbf{y} \\ \mathbf{X}^T\mathbf{y} \end{bmatrix}. \tag{5.50}$$

The generalized matrix inverse shall be denoted by superscript "-". The first normal equation can be written as

$$\mathbf{d} = (\alpha, \mathbf{u}^T)^T = (\mathbf{K}^T\mathbf{K})^-\mathbf{K}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}), \tag{5.51}$$

which can be substituted into (5.50) to yield

$$\mathbf{X}^T\mathbf{P}^*\mathbf{X}\boldsymbol{\gamma} = \mathbf{X}^T\mathbf{P}^*\mathbf{y}, \quad \text{with } \mathbf{P}^* = \mathbf{I} - \mathbf{K}(\mathbf{K}^T\mathbf{K})^-\mathbf{K}^T, \tag{5.52}$$

where $\mathbf{P}^*$ is a symmetric and idempotent matrix (Searle, 1971, 341–342). Note that, although $(\mathbf{K}^T\mathbf{K})^-$ is not unique (since $\mathbf{K}^T\mathbf{K}$ has not full rank), it enters only in the form $\mathbf{K}(\mathbf{K}^T\mathbf{K})^-\mathbf{K}^T$, which is invariant to the choice of the generalized matrix inverse. Thus, the non-full rank property does not itself lead to manifold solutions of $\boldsymbol{\gamma}$. However, for reasons of numerical stability, we will avoid computing a brute-force generalized matrix inverse. Instead, we derive a very simple variant (approximation) of the generalized matrix inverse, say, $\mathbf{G}$.

To this end, note that (5.51) is not invariant to the particular choice of $\mathbf{G} = (\mathbf{K}^T\mathbf{K})^-$. However, since any linear combination of $\mathbf{d}$ in (5.51), say, $\boldsymbol{\lambda}^T\mathbf{d}$ is estimable when $\boldsymbol{\lambda}^T = \mathbf{t}^T\mathbf{K}$ for some $\mathbf{t} \in \mathbb{R}^n$, we deliberately put one element of $\mathbf{d}$ equal to zero, and cross out the corresponding element in the normal equations (cf. Searle, 1971, 232–233). The obvious element to equate to zero in (5.51) is α. Thus, our generalized inverse shall be given by

$$\mathbf{G} = \begin{bmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & (\mathbf{Z}^T\mathbf{Z})^{-1} \end{bmatrix}, \quad \text{where } (\mathbf{Z}^T\mathbf{Z})^{-1} = \text{diag}(1/n_1, \ldots, 1/n_g).$$

Substituting $\mathbf{G}$ into (5.51) yields $\mathbf{d} = (\alpha, \mathbf{u}^T)^T$ with $\alpha = 0$ (by assumption) and $\mathbf{u} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. In line with this, we replace $\mathbf{P}^*$ in (5.52) by

$$\mathbf{P} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$$

because it is computationally much simpler than $\mathbf{P}^*$. Note that pre-multiplying a matrix by $\mathbf{P}$ corresponds to centering the matrix by its column-wise arithmetic means. In the present context, column-wise centering corresponds to centering by the area-specific means. By symmetry and idempotency of $\mathbf{P}$ we shall use

$$\hat{\boldsymbol{\gamma}} = \left[(\mathbf{PX})^T\mathbf{PX}\right]^{-1}(\mathbf{PX})^T\mathbf{Py}, \tag{5.53}$$

instead of (5.52), where $\boldsymbol{\gamma}$ is based on the centered data, $\mathbf{PX}$ and $\mathbf{Py}$. It is

evident from (5.53) that the influence of outliers in $\mathbf{y}$ on the estimate $\hat{\boldsymbol{\gamma}}$ is unbounded. Thus, we obtain robust estimates of $\boldsymbol{\gamma}$, say $\hat{\boldsymbol{\gamma}}^R$, by means of M-estimation of regression which is defined by the estimating equations

$$(\mathbf{PX})^T \boldsymbol{\psi}_k\big[(\mathbf{Py} - \mathbf{PX}\hat{\boldsymbol{\gamma}}^R)/\hat{S}\big] = \mathbf{0}, \tag{5.54}$$

where $\boldsymbol{\psi}_k$ is the Huber ψ-function, and $\hat{S}$ is a robust estimate of scale of the non-zero residuals from (5.54), e.g. the normalized median absolute deviation (MAD) about zero. The restriction of taking only non-zero residuals prevents underestimating the scale, a problem which becomes an issue if the number of auxiliary variables is relatively large (cf. Maronna et al., 2006, 100).

It is fruitful to note that $\mathbf{Py}$ can be expressed as $\mathbf{y} - \mathbf{Z}\boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is the g-vector of area-specific means, $(\hat{\bar{y}}_1, \ldots, \hat{\bar{y}}_g)^T$ with $\hat{\bar{y}}_i = (1/n_i)\sum_{j\in s_i} y_{ij}$, $i = 1, \ldots, g$. This representation indicates that the breakdown point of (5.54) may be much lower than the one of a regression M-estimator. This is a consequence of the centering procedure, which centers the y_{ij} in area i by the area-specific arithmetic mean. The procedure may thus turn $(n_i - 1)$ ordinary observations into outliers (typically with a reversed sign) if the area contains one single heavy outlier. From the perspective of breakdown point, a simple remedy is to center the response variable by the area-specific median instead. This corresponds to replacing $\mathbf{Py}$ in (5.54) by $\bar{\mathbf{y}}^{med} = \mathbf{y} - \mathbf{Z}\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is the g-vector of area-specific medians, $(\text{med}_{i=1}(y_{1j}), \ldots, \text{med}_{i=g}(y_{gj}))^T$. This approach resembles the "median polish" strategy, which has been proposed for the two-way analysis of variance (Tukey, 1977).

Now, in order to obtain a robust estimate of the variance pre-estimation bound v_∞, we have to solve (5.54) with $\mathbf{Py}$ replaced by $\bar{\mathbf{y}}^{med}$ and obtain $\hat{v}_\infty^R$ from

$$\hat{v}_\infty^R = \hat{S}^2. \tag{5.55}$$

An alternative approach has been proposed by Birch and Myers (1982). They obtain M-estimates of $\boldsymbol{\gamma}$ and all u_i's, $i = 1, \ldots, g$, as the solutions to the following system of estimating equations,

$$\begin{aligned} \sum_{i=1}^{g}\sum_{j=1}^{n_i} \psi_k\left(\frac{r_{ij}}{\sigma}\right)\mathbf{x}_{ij} &= \mathbf{0}, \\ \sum_{j=1}^{n_i} \psi_k\left(\frac{r_{ij}}{\sigma}\right) &= 0, \quad i = 1, \ldots, g, \end{aligned}$$

where $r_{ij} = y_{ij} - u_i - \mathbf{x}_{ij}^T\boldsymbol{\gamma}$, and σ denotes the normalized MAD of the residuals. In essence, this strategy consists of computing a relatively large number of M-estimates which is rather time consuming and therefore not the optimal strategy to compute pre-estimation bounds.

5.6.2. Computational details

We arranged all vector and matrix operations to make them rich in level-1 procedures (i.e., procedures which operate on vectors of size n and involve $\mathcal{O}(n)$ floating-point operations, see Golub and van Loan, 1996) and paid attention to potential floating-point arithmetic issues. With respect to elementary operations, we rely on the procedures in BLAS (Blackford et al., 2002) and LAPACK (Anderson et al., 2000). Furthermore, we have avoided to compute any "brute-force" matrix inverse and the likes.

5.6.3. Algorithm

The choice of starting values is crucial in terms of speed and numerical stability of the algorithm. Extensive simulation showed that the method is best initialized by the choice

$$\boldsymbol{\beta}_0^T \leftarrow \big(0, \hat{\boldsymbol{\gamma}}^R\big), \qquad v_0 \leftarrow \hat{v}_\infty^R, \qquad d_0 \leftarrow 200,$$

where $\hat{\boldsymbol{\gamma}}^R$ and $\hat{v}_\infty^R$ are the pre-estimation estimates defined in (5.54) and (5.55), respectively. A reasonable initial value for d is $d_0 = 200$ (or any other large number).

The display Algorithm 5.1 shows the most important aspects of the algorithm. The numerical tests whether an updated value behaves well (e.g., lies within the pre-estimation bounds) have been omitted in the display. The final estimates are given by: $\hat{\boldsymbol{\beta}}^R \leftarrow \boldsymbol{\beta}_{i+1}$, $[\hat{\sigma}_e^2]^R = \hat{v}^R \leftarrow v_{i+1}$, and $\hat{d}^R \leftarrow d_{i+1}$. By means of identity (5.8), we obtain $[\hat{\sigma}_a^2]^R = [\hat{\sigma}_e^2]^R \hat{d}^R$.

Algorithm 5.1. Let $\boldsymbol{\beta}_0$, v_0, and d_0 denote starting values. Define the termination-rule constants δ, $\delta_{\boldsymbol{\beta}}$, and δ_v such that $\epsilon^{1/2} \leq \delta < 0.001$, $\epsilon^{1/2} \leq \delta_{\boldsymbol{\beta}} < 0.001$, and $\epsilon^{1/2} \leq \delta_v < 0.001$, where $\epsilon = 2.2 \times 10^{-16}$ is the machine epsilon.

```
MAIN
  WHILE  ‖τ_{i+1} − τ_i‖_{p+2} ≤ δ    (INIT τ_0 ← (β_0, v_0, d_0) , INDEX i = 0, 1, ...)
    WHILE ‖β_i^{j+1} − β_i^j‖_p ≤ δ_β    (INIT β_i^0 ← β_i, INDEX j = 0, 1, ...)
      update β_i^{j+1} ← f(β_i^j; ·), where f denotes (5.20)
    END WHILE
    WHILE |v_i^{j+1} − v_i^j| ≤ δ_v    (INIT v_i^0 ← v_i, INDEX j = 0, 1, ...)
      update v_i^{j+1} ← f(v_i^j; ·), where f denotes (5.23).
    END WHILE
    solve d_{i+1} ← f(d_i; ·), where f denotes (5.24); Brent's algorithm
  END WHILE
END
```

5.7. Simulations

In order to study the behavior of the proposed method in more detail, we consider a small model-based simulation study. The simulation data are generated using the following model,

$$y_{ij} = (1, x_{ij})^T(\alpha, \beta) + u_i + e_{ij}, \quad j = 1, \ldots, n;\ i = 1, \ldots, g,$$

where $x_{ij} \sim \mathcal{N}(0, 1)$ and $\alpha = \beta = 1$. In each Monte Carlo sample, a configuration of $g = 20$ areas and $n = 5$ within-area units (overall, $N = 80$ observations; balanced data) are generated. In line with Stahel and Welsh (1997), we allow for contamination in terms of normal mixture distributions,

$$(1 - \varepsilon) \cdot \mathcal{N}(0, 1) + \varepsilon \cdot \mathcal{N}(0, \zeta),$$

where $\epsilon \in [0, 1)$ and $\zeta \in \mathbb{R}_+$ can be chosen in either or both of the random effect distributions, giving rise to the following (relevant) combinations:

(0,0) no contamination; $e_{ij} \sim \mathcal{N}(0, 1)$ and $u_i \sim \mathcal{N}(0, 1)$,

(e,0) $e_{ij} \sim (1 - \varepsilon) \cdot \mathcal{N}(0, 1) + \varepsilon \cdot \mathcal{N}(0, 41)$ and $u_i \sim \mathcal{N}(0, 1)$,

(0,u) $e_{ij} \sim \mathcal{N}(0, 1)$ and $u_i \sim (1 - \varepsilon) \cdot \mathcal{N}(0, 1) + \varepsilon \cdot \mathcal{N}(0, 41)$,

(e,u) $e_{ij} \sim (1 - \varepsilon) \cdot \mathcal{N}(0, 1) + \varepsilon \cdot \mathcal{N}(0, 41)$ and $u_i \sim (1 - \varepsilon) \cdot \mathcal{N}(0, 1) + \varepsilon \cdot \mathcal{N}(0, 41)$.

For the evaluation of the simulation results, we study robust and non-robust numerical criteria; namely,

- (average) bias of the point estimates ($bias$) and median error of point estimates ($mede$),
- mean squared error of the point estimates (mse) and median absolute error of the point estimates ($medae$),

see Appendix B for a rigorous definition of the criteria.

Robust parameter estimation

The findings of the small simulation exercise can be summarized as follows.

- The "huberm" *method converged in all 1000 Monte Carlo trials* for each simulation configuration (i.e., contamination scenario and choice of k). This is in sharp contrast to the results reported by Richardson (1995) and Chaubey and Venkateswarlu (2002) (among others). The proposed algorithm may, on the other hand, fail to converge when the amount of contamination, ε, is larger than the breakdown point (which can be rather low in the case of unbalanced data).

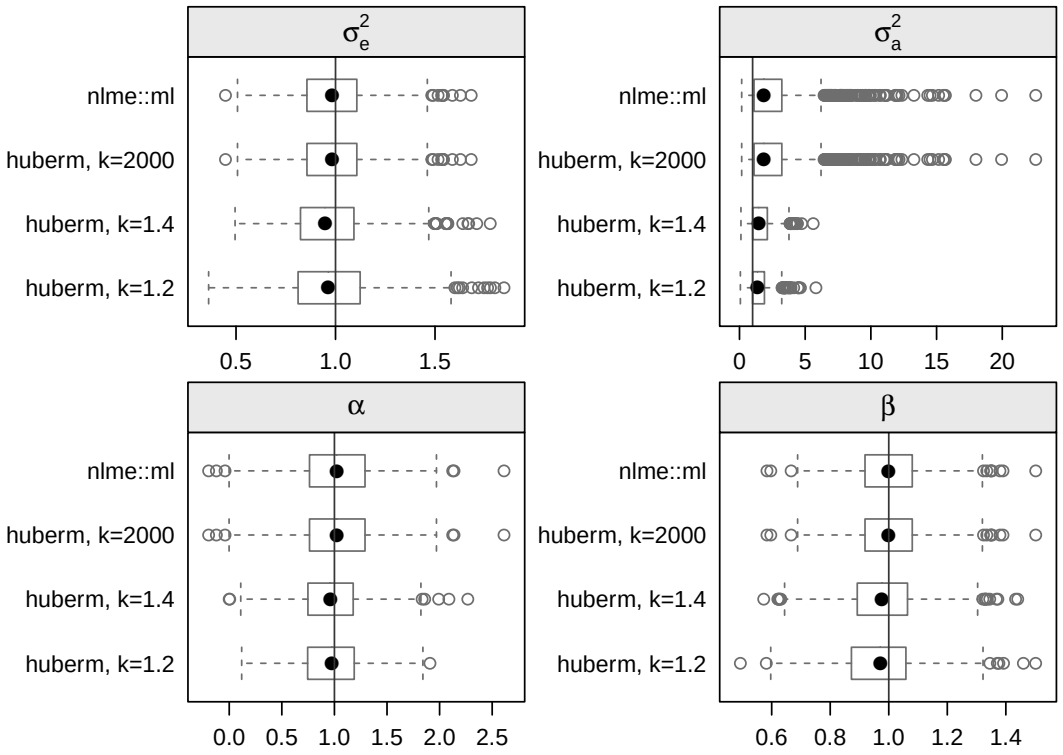


Figure 5.3.: Impact of area-level contamination on the estimators of α, β, σ_e^2, and σ_u^2 for four different estimation methods (see text for further details).

- The results of the "huberm" method mimicking the m.l.e. (i.e., when $k = 2000$) are equal (up to the 6th or 7th decimal place) with those of the gold-standard method "lme" in the `R` package "nlme".[6]

- In the presence of contamination, the M-estimator has a smaller bias than the corresponding m.l.e. The mean squared error is also considerably smaller. In contrast, the loss of efficiency of the M-estimator in the absence of contamination is almost negligible. These findings remain valid if one considers the robust criteria ($mede$ and $medae$ in Table 5.2). In the presence of contamination, $medae$ tends to be smaller than mse which indicates that the simulated distribution is skewed; see also Figures 5.3 and 5.4.

- Contamination of the model error, e_{ij}, affects the robust estimates of σ_u^2 very little since the contamination affects the diagonal elements of the variance of y_{ij} but not the off-diagonal elements (see Figure 5.4). Contamination of the area-specific random effects, u_i, affects both diagonal and off-diagonal elements of the variance (see also Welsh and Richardson,

[6] See Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2017). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131.

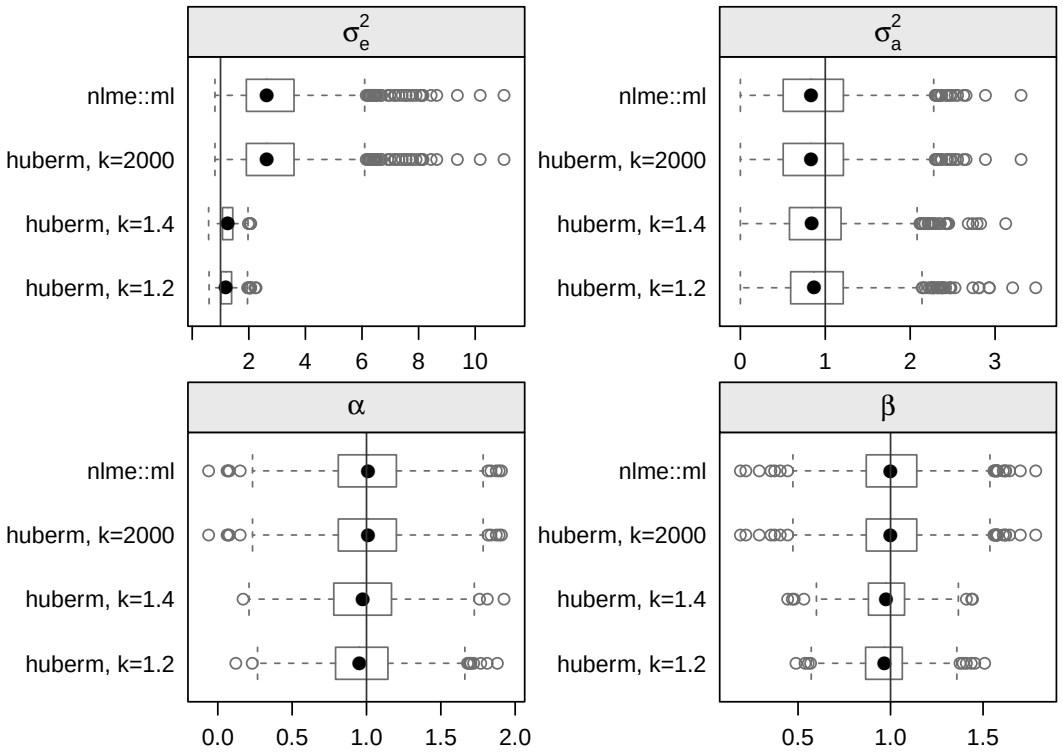


Figure 5.4.: Impact of unit-level contamination on the estimators of α, β, σ_e^2, and σ_u^2 for four different estimation methods (see text for further details).

1997, 348). When both components are contaminated (see Figure 5.5), the effects on the estimates are the combination of the effects of contaminating the components one at a time (see also Stahel and Welsh, 1997, 315).

- In the simulation exercise, we had focused on two choices of the tuning constant k. It goes without saying that one may obtain better estimates trying different choices of k. Our experience supports the finding of Stahel and Welsh (1997) that fine tuning pays more in estimating these models than it does with simpler models (p. 315).

- Computing the robust estimates based on data consisting of $g = 500$ areas, each of which has $n = 20$ units (i.e., $500 \times 20 = 10,000$ observations), takes on average 1.3 seconds on an ordinary desktop computer.

Robust prediction

So far, we have discussed the behavior of robust parameter estimates in the presence of contamination. As far as small area estimation is concerned, computing robust parameter estimates is only an intermediate step on the way to robust predictions of area-specific means. However, robust prediction is an easy

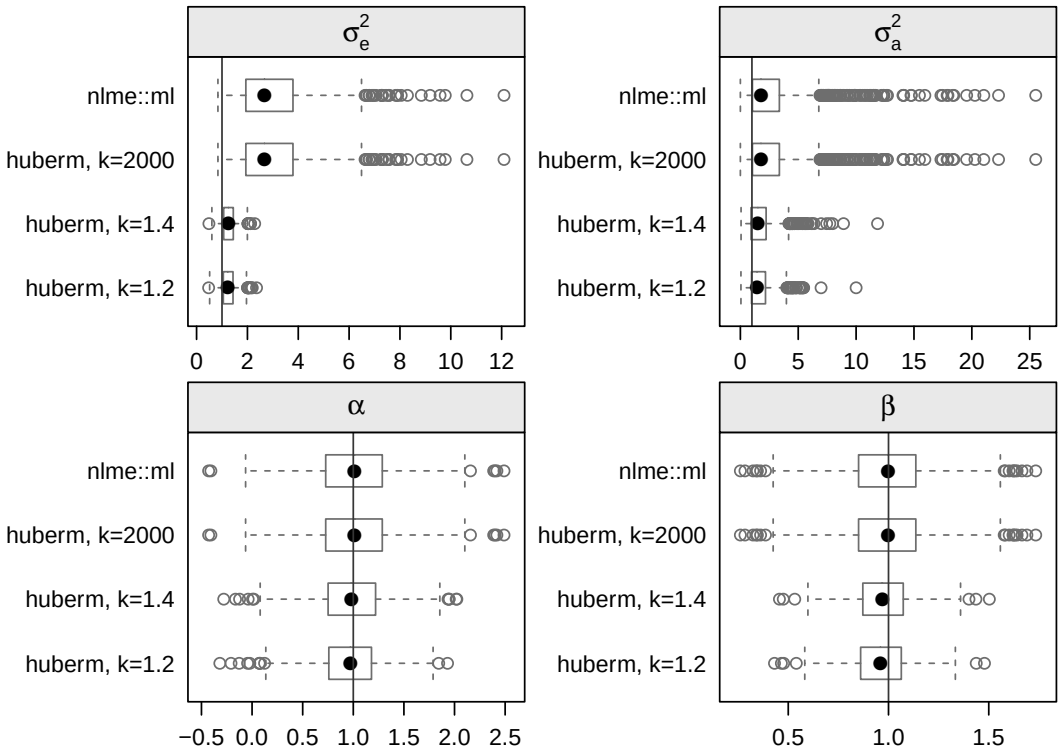


Figure 5.5.: Impact of joint area- and unit-level contamination on the estimators of α, β, σ_e^2, and σ_u^2 for four different estimation methods (see text for further details).

task once the parameters have been estimated; see Section 5.4.4. Moreover, it is evident that biased parameter estimates lead to biased and (in most cases) inefficient predictions; cf. the simulation study in Sinha and Rao (2009). In this regard, carrying out another simulation study would reveal only few new insights (if any). Therefore, we do not show the results. Though, the influence that outliers or excessive variability in the data can exert on the predicted values and the associated MSPE estimates can be substantial. To really understand when robust methods are advantageous, we shall study robust prediction and MSPE estimation in the context of real sample data; see the case study, below.

Table 5.1.: Bias and MSE estimates of the variance components for the four contaminations scenarios

(e,a)%	method	k	$bias(\hat{\sigma}_a^2)$	$mse(\hat{\sigma}_a^2)$	$bias(\hat{\sigma}_e^2)$	$mse(\hat{\sigma}_e^2)$
(0,0)	lme(ml)	–	-0.0801	0.1437	-0.0018	0.0348
	huberm	2000	-0.0801	0.1437	-0.0018	0.0348
	huberm	1.4	-0.0463	0.1655	-0.0235	0.0453
	huberm	1.2	-0.0701	0.1793	-0.0235	0.0462
(5,0)	huberm	2000	-0.1027	0.2834	1.9474	2.0677
	huberm	1.4	-0.0764	0.2301	0.2587	0.0700
	huberm	1.2	-0.0450	0.2579	0.2164	0.0834
(0, 5)	huberm	2000	1.7364	7.1491	-0.0097	0.0352
	huberm	1.4	0.6322	0.6821	-0.0321	0.0428
	huberm	1.2	0.4996	0.5373	-0.0204	0.0545
(5, 5)	huberm	2000	1.8484	9.7617	2.0203	2.1784
	huberm	1.4	0.7560	1.4080	0.2628	0.0734
	huberm	1.2	0.6732	1.0512	0.2521	0.0838

Notes: Each criterion is computed based on 1000 Monte Carlo replications; $(e, a)\%$ denotes the contamination scheme, where a and e denote the percentage of contamination; k is the robustness tuning constant of the Huber-type M-estimator; `R` packages: `nlme` (vers. 3.1-128) and `rsae` (vers. 0.1-5)

Table 5.2.: Robust evaluation criteria ($mede$ and $medae$ instead of $bias$ and mse) of the variance components for the four contamination scenarios

(e,a)%	method	k	$mede(\hat{\sigma}_a^2)$	$medae(\hat{\sigma}_a^2)$	$mede(\hat{\sigma}_e^2)$	$medae(\hat{\sigma}_e^2)$
(0,0)	lme(ml)	–	-0.1064	0.2707	-0.0134	0.1279
	huberm	2000	-0.1064	0.2707	-0.0134	0.1279
	huberm	1.4	-0.0920	0.2824	-0.0329	0.1490
	huberm	1.2	-0.1460	0.3064	-0.0318	0.1509
(5,0)	huberm	2000	-0.1675	0.3790	1.6308	1.6308
	huberm	1.4	-0.1579	0.3401	0.2525	0.2630
	huberm	1.2	-0.1330	0.3427	0.1902	0.2230
(0, 5)	huberm	2000	0.8582	0.8582	-0.0177	0.1277
	huberm	1.4	0.4690	0.5333	-0.0526	0.1448
	huberm	1.2	0.3645	0.4583	-0.0365	0.1632
(5, 5)	huberm	2000	0.7833	0.8540	1.6719	1.6719
	huberm	1.4	0.5175	0.5856	0.2540	0.2642
	huberm	1.2	0.4392	0.5256	0.2300	0.2451

Notes: Each criterion is computed based on 1000 Monte Carlo replications; $(e, a)\%$ denotes the contamination scheme, where a and e denote the percentage of contamination; k is the robustness tuning constant of the Huber-type M-estimator; `R` packages: `nlme` (vers. 3.1-128) and `rsae` (vers. 0.1-5)

5.8. Case study: Estimates of average above-ground forest biomass in Norwegian municipalities

The Norwegian National Forest Inventory provides data on forest characteristics at the national and regional level. Forest attribute information for small domains such as municipalities or protected areas is only available at a limited number of sampled points. The established domain-specific samples are often not large enough to compute estimates with acceptable precision. Breidenbach and Astrup (2012) therefore proposed the application of SAE techniques to obtain reliable estimates of average above-ground forest biomass for 14 Norwegian municipalities. The municipalities cover an overall area of 2184 km^2; the territory of a single municipality amounts to between 31 and 527 km^2 and contains 1–35 sample measurements on the variable biomass (measured in $10^6 g/ha$, where a hectare (ha) equals $10^4 m^2$). Breidenbach and Astrup (2012) relate the variable biomass to the auxiliary variable mean canopy height, which is obtained from a photogrammetric canopy height model. The data on biomass come from a sample survey, the auxiliary data are known for the whole forest population. Breidenbach and Astrup (2012) made both data sets available to the public via the `R` package `JoSAE`; see Breidenbach (2015). The authors considered the following unit-level SAE model

$$biomass_{ij} = \beta_0 + \beta_1 canopy_{ij} + u_i + e_{ij}, \qquad i = 1, \ldots, 14,$$

where $canopy_{ij}$ is a shorthand notation for the variable mean canopy height; the index j runs over the set of municipality-specific sample sizes, $j = 1, \ldots, n_i$. The β-parameters are unknown, so are the underlying parameters of the stochastic components u_i and e_{ij}. Figure 5.6 shows a scatter plot of the variables `biomass` vs. `canopy` and supports the linearity assumption of the model pretty good. An analogous graph is shown in Figure 5.7 for each of the 14 municipalities.

Table 5.3.: Parameter estimates (with t-values in parentheses)

	REML	ML	RML ($k = 2$)
$\hat{\beta}_0$	6.695	7.140	13.837
	(0.80)	*(0.88)*	*(1.93)*
$\hat{\beta}_1$	1.376	1.373	1.255
	(17.74)	*(17.82)*	*(17.19)*
$\hat{\sigma}_u^2$	10.304	8.619	2.979
$\hat{\sigma}_e^2$	49.858	49.694	47.567

Note: REML is due to Breidenbach and Astrup (2012), all other estimates are computed with package `rsae`.

Under the assumption that the r.v.'s u_i and e_{ij} are Gaussian with zero mean and variances, respectively, σ_u^2 and σ_e^2, Breidenbach and Astrup (2012) obtained

estimates of the unknown parameters by means of reduced m.l.e. (REML, see Table 5.3) and computed EBLUP estimates for the municipality-specific average above-ground forest biomass. The m.l.e. estimates (ML) are also shown in Table 5.3 [the M-estimator RML will be discussed later]. As expected, the m.l.e. of variance tends to underestimate, compared with REML. However, the amount of underestimation is modest because the loss in degrees of freedom not accounted for is small with only one auxiliary variable.

It is questionable whether the model intercept should be kept in the model specification as it is not statistically different from zero (p-value of 0.423 for method reduced m.l.e.). For reasons of comparability with the results of Breidenbach and Astrup (2012), we stick with the current model specification. More important, it is evident from Figure 5.6 that the variable `biomass` shows a rather high variability (and some tendency of heteroscedasticity towards the right of the plot). The high variability is reason enough to consider robust parameter estimates. Therefore, we computed M-estimates (RML) under the basic unit-level model (see Table 5.3). It is absolutely remarkable by how much the estimates of variances $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ differ for the methods reduced m.l.e. and M-estimator. The M-estimates of variance are much smaller, also compared with the m.l.e. Moreover, the M-estimate of the regression intercept is significantly different from zero at the 5% significance level (see Table 5.3).

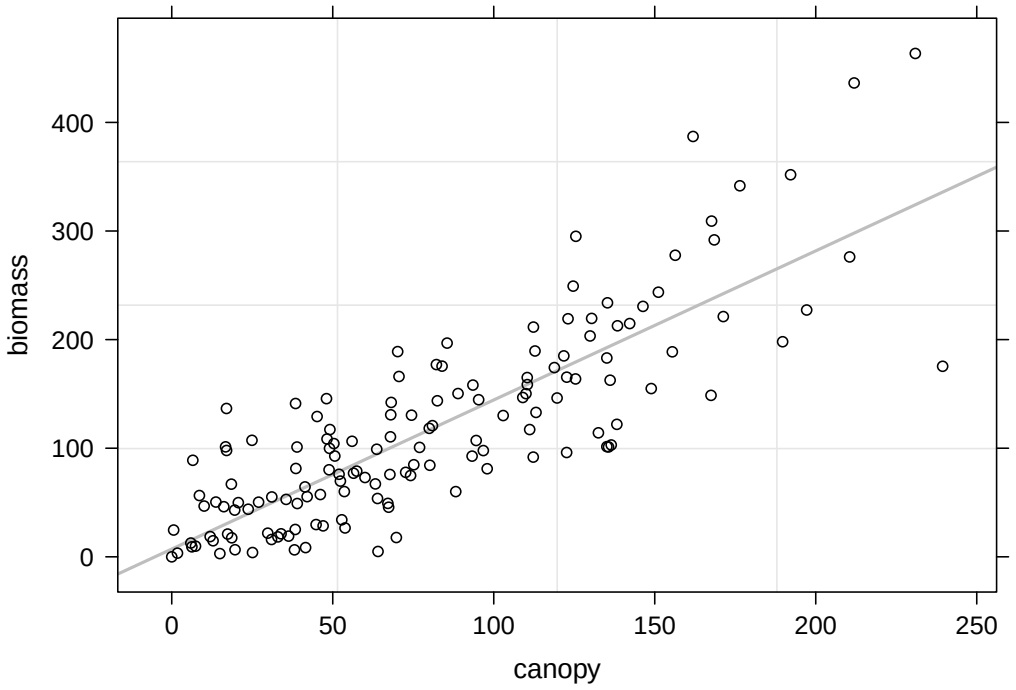


Figure 5.6.: Scatter plot of the variables `biomass` vs. `canopy`; the straight line refers to the reduced m.l.e. estimate [see text for further details].

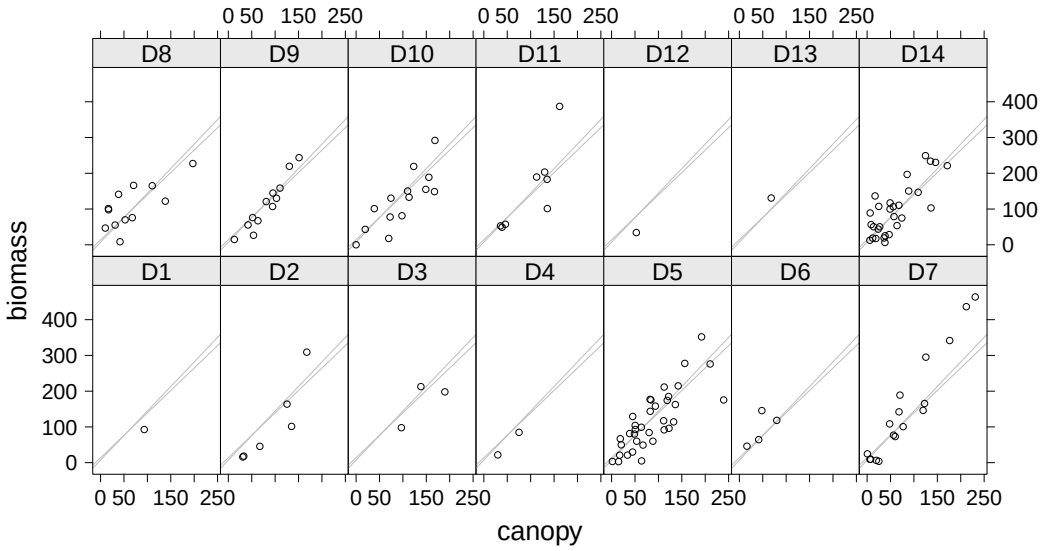


Figure 5.7.: Scatter plot of the variables `biomass` vs. `canopy` by municipalities (D1, ..., D14); the straight lines refer to the reduced m.l.e. estimate and the M-estimate, which are hardly distinguishable [see text for further details].

Robust estimation with `R` package `rsae`

In this paragraph, we briefly sketch the computation of the M-estimates reported in Table 5.3 using the functions in `R` package `rsae`. The discussion of the data preparation prior to estimation is deferred to the appendix, see Section C.1. First, we have to specify a `saemodel`-object.

```
> model <- saemodel(biomass ~ canopy, area = ~domainID,
+                   data = sample)
```

The M-estimator (method RML) with tuning constant $k = 2$ can then be obtained as follows.

```
> mest <- fitsaemodel("huberm", model, k = 2)

ESTIMATES OF SAE-MODEL (model type B)
Method:  Maximum likelihood estimation
---
Fixed effects
Model: biomass ~ (Intercept) + canopy
  Coefficients:
(Intercept)       canopy
    7.14007      1.37290
---
Random effects
  Model: ~1| domainID
           (Intercept)  Residual
Std. Dev.    8.61875     49.69384
---
```

```
Number of Observations:  145
Number of Areas:  14
```

Prediction of area-level mean

Given the M-estimate `mest`, we obtain predictions of the area-level means and an estimate of the mean squared prediction error (MSPE) by parametric bootstrap with `reps=5000` replications.

```
> mestpred <- robpredict(mest, population, k = 2, reps = 5000)
```

The object `population` is a `data.frame` that contains the population means of the variable `canopy` for each municipality; see Section C.1. The tuning constant `k` specified in function `robpredict` controls the degree of robustness to be achieved for the prediction. The chosen number of bootstrap replications, `reps=5000`, is rather large, but this choices ensures precise estimates. We compared two successive batches of 5000 bootstrap replicates and found that the MSPE estimates of the two batches differed on average by less than 1.2%. For `reps=500`, the batch-by-batch comparison of the MSPE estimates showed large differences. It is therefore recommended to choose rather larger numbers of bootstrap replicates.

Table 5.4.: Predicted area-level means and root mean squared prediction error

	EBLUP		REBLUP	root MSPE		
Domain	REML	ML	RML	REML	ML[a)]	RML[a)]
1	153.76	154.39	149.37	11.93	10.34	5.59
2	107.82	109.34	110.33	12.62	9.78	5.58
3	132.74	133.90	131.81	12.26	9.99	5.63
4	123.88	124.62	122.31	11.59	10.15	5.39
5	118.49	119.31	119.85	7.74	7.11	5.11
6	116.91	116.21	111.63	11.77	9.79	5.53
7	117.73	115.40	105.02	15.63	8.67	5.70
8	99.86	99.39	96.50	9.83	9.21	5.76
9	116.84	117.47	116.10	9.48	9.16	5.72
10	110.76	112.19	113.60	11.11	8.81	5.66
11	135.89	135.47	129.63	10.29	9.60	5.79
12	118.19	118.92	117.00	11.80	10.01	5.28
13	95.01	94.91	93.37	11.65	10.13	5.41
14	102.46	101.81	97.66	8.28	7.59	5.32

Notes: EBLUP(REML) and MSPE(REML) are due to Breidenbach and Astrup (2012), all other estimates are computed with `rsae`.
Superscript a) indicates parametric bootstrap estimates.

In terms of predicted area-level means, the three methods produce very similar results (see columns EBLUP and REBLUP in Table 5.4). The robust REBLUP(RML) estimates differ on average by 2.2% from the ones of EBLUP with

REML estimates; the largest difference is encountered in domain no. 7 [REBLUP differs by 10.8% from EBLUP(REML)]. Table 5.4 also shows estimates of (root) mean squared prediction error. The root MSPE estimates for the EBLUP method with REML estimates are due to Breidenbach and Astrup (2012), and are obtained via analytic approximation to the root MSPE (see e.g. Rao, 2003, chap. 7.2). The root MSPE estimates for the methods ML and RML are obtained by parametric bootstrap (see above); hence, they are comparable with each other. It is apparent from the tabulated numbers that the root MSPE estimates for RML are substantially better than the estimates for the ML methods. In fact, the differences are surprisingly large. The prime reason for these tremendous differences lies in the magnitude of the estimates for σ_u^2 (variance of the area-level random effect). The robust estimate of σ_u^2 is much smaller and thus leads to less prediction uncertainty.

Findings from the case study

Outside robust statistics, people quite often regard robust methods as a mean to limit the influence of a couple of massive outliers in the data. However, our case study shows that robust methods can indeed play off much of their advantage over standard methods when the data feature a “a bit” more variability “than expected”, but are still considered to be “well-behaved”. In our example on estimation of above-ground forest biomass domain means, the gains in efficiency obtainable through M-estimation are surprisingly big.

5.9. Summary and discussion

Unlike location-scale or regression models, mixed linear models (MLM) with block-diagonal covariance matrix have no nice invariance structure. As a result, the model parameters cannot be estimated consistently in the presence of contamination (Welsh and Richardson, 1997, 349). The potential bias of the maximum likelihood estimator can be arbitrarily large.

In view of the robustness issues encountered with estimation and prediction under the basic unit-level model, Sinha and Rao (2009) proposed an M-estimator and established a robustification of the EBLUP method. Their M-estimator is an approximation of the RML 2 method which was suggested by Richardson and Welsh (1995). The approximation due to Sinha and Rao (2009) together with their Newton–Raphson algorithm does not solve the numerical issues experienced with fitting robust MLM's. In general, the existing algorithms for fitting robust MLM's are notoriously instable (see also Richardson, 1995, or Chaubey and Venkateswarlu, 2002, who report a high susceptibility to failure of convergence).

Parametrization is key

Conceptionally, the joint M-estimator we suggested under the basic unit-level model is equivalent to method RML 2. The major difference between RML 2 and our approach lies in the parametrization of the model's covariance matrix. Instead of the "canonical" covariance matrix specification, $\mathbf{\Omega}_i = \sigma_e^2\mathbf{I}_i + \sigma_u^2\mathbf{J}_i$, we use a parametrization in terms of variance ratios, which is due to Hartley and Rao (1967), and given by

$$v(\mathbf{I}_i + d\mathbf{J}_i) \quad \text{for all areas} \quad i = 1, \ldots, g,$$

where $\mathbf{I}_i$ and $\mathbf{J}_i$ are, respectively, the $(n_i \times n_i)$ identity matrix and the matrix of ones; $v = \sigma_e^2$ and $d = \sigma_u^2/\sigma_e^2$. The main advantage of the Hartley–Rao parametrization is that the m.l.e. of v $(\equiv \sigma_e^2)$ has a closed-form solution which is a quadratic form. Therefore, m.l. estimates of v are non-negative by design; a property which is not guaranteed in case of the canonical parametrization. These advantageous properties (in terms of numerical computations) are "passed on" to the M-estimator that obtains as a robustification of the m.l.e. of v. The variance ratio parameter d does not feature such nice properties. However, its estimating equation (m.l.e. or M-estimator) is a root-finding problem in only one dimension; hence, it is easy to solve numerically. Finally, estimates of the regression coefficients (fixed effects) are obtained with an M-estimator.

The M-estimators of the variance parameters are Fisher consistent. In addition, we showed that their (empirical) sensitivity curves behave as was to be expected from the point of M-estimator theory. The M-estimator of the area-level random effect variance (σ_u^2) proposed by Sinha and Rao (2009), on the

other hand, is not Fisher consistent. The resulting bias is presumably rather small, but a detailed evaluation of the bias was beyond the scope of this thesis since we do not own a reliable software implementation of the Sinha–Rao proposal. Only a broad simulation study can shed light on the magnitude of bias. Another remark is in order. Our discussion also showed that their M-estimator of σ_u^2 suffers from improper decorrelation. When the residuals enter the ψ-function of the M-estimator of σ_u^2 not properly decorrelated, the scale of the residuals is not normalized. This implies that it can be quite difficult to achieve a per-determined level of robustness via the choice of robustness tuning constant.

Robust prediction of the random effects

Sinha and Rao (2009) have proposed a method to robustly predict the area-level random effects which relies on the solution to Fellner's robust mixed-model equation (Fellner, 1986). Since Fellner's equations are non-linear, Sinha and Rao (2009) suggested a Newton–Raphson algorithm to solve the equations. We have shown (see Schoch, 2012) that robust predictions can be obtained far more easily on the basis of an approach suggested by Copt and Victoria-Feser (2009). Our method is theoretically convincing, computationally much less demanding, and does – unlike the Newton–Raphson approach – not suffer from failure of convergence. In the second edition of "the" SAE book (Rao and Molina, 2015), the authors came up with a robust prediction method that is very similar to ours [see p. 196].

Achieving high numerical stability

Numerical stability and reliability were the core criteria we had in mind when developing robust methods for the basic unit-level model. This requires paying attention to bookkeeping details, and also careful consideration of how roundoff errors can affect the overall computation. Important contributions to numerical stability are implied via our choice of parametrization, the choice of starting values, and the so-called pre-estimation bounds. The pre-estimation bounds are parameter estimates (obtained before the main algorithm is started) that provide upper and lower bounds within which the final estimates are supposed to lie. By means of these bounds, the algorithm keeps track whether the iteratively updated values are converging the way they are supposed to, and, if they are not, the method intersperses some computation steps to let potential run-away values settle down.

Another crucial ingredient that contributes to numerical stability is the choice of initial estimates of the regression coefficients. Unlike Sinha and Rao (2009) and Rao and Molina (2015, 196), who start their iterative updating from the generalized least squares estimate, we use robust estimates (either the least trimmed squares (LTS) estimator or an S-estimator of regression).

A note on a competing algorithm

Recently, Rao and Molina (2015) have introduced another method to compute M-estimates under the basic unit-level model. Their starting point is the system of estimating equations of the variance components $\boldsymbol{\theta} = (\sigma_e^2, \sigma_u^2)^T$, which has been studied in Sinha and Rao (2009),

$$\left.\begin{aligned} \sum_{i=1}^{g} \boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\mathbf{U}_i^{1/2}\boldsymbol{\Omega}_i^{-2}\mathbf{U}_i^{1/2}\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i) - \operatorname{tr}\{c\boldsymbol{\Omega}_i^{-1}\} = 0, \\ \sum_{i=1}^{g} \boldsymbol{\psi}_k^T(\tilde{\mathbf{r}}_i)\mathbf{U}_i^{1/2}\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\boldsymbol{\Omega}_i^{-1}\mathbf{U}_i^{1/2}\boldsymbol{\psi}_i(\tilde{\mathbf{r}}_i) - \operatorname{tr}\{c\boldsymbol{\Omega}_i^{-1}\mathbf{J}_i\} = 0, \end{aligned}\right\} \Leftrightarrow: \boldsymbol{\Psi}(\boldsymbol{\theta}) = \mathbf{0}, \quad (5.56)$$

where $\tilde{\mathbf{r}}_i = \mathbf{U}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$, and $\mathbf{U}_i^{1/2}$ is the diagonal matrix whose elements are equal to the inverse of the square root of the diagonal elements of $\boldsymbol{\Omega}_i$; $\mathbf{J}_i$ is the matrix of ones. Chatrchi (2012), a student of J.N.K. Rao and S.K. Sinha, noticed[7] that we may write (see Rao and Molina, 2015, 196)

$$\begin{aligned} \operatorname{tr}\left\{\boldsymbol{\Omega}_i^{-1}\right\} &= \operatorname{tr}\left\{\boldsymbol{\Omega}_i^{-1}\boldsymbol{\Omega}_i^{-1}\boldsymbol{\Omega}_i\right\} = \operatorname{tr}\left\{\boldsymbol{\Omega}_i^{-2}(\sigma_e^2\mathbf{I}_i + \sigma_u^2\mathbf{J}_i)\right\} \\ &= \sigma_e^2 \operatorname{tr}\left\{\boldsymbol{\Omega}_i^{-2}\right\} + \sigma_u^2\left\{\boldsymbol{\Omega}_i^{-2}\mathbf{J}_i\right\} \end{aligned} \quad (5.57)$$

which enables us to "extract" σ_e^2 in the first equation of the system in (5.56). We then proceed in the same manner for the second equation and σ_u^2. Eventually, this enables us to formulate the fixed-point equations

$$\begin{bmatrix} \{\sigma_e^2\}^{t+1} \\ \{\sigma_u^2\}^{t+1} \end{bmatrix} = \boldsymbol{f}\left(\begin{bmatrix} \{\sigma_e^2\}^{t} \\ \{\sigma_u^2\}^{t} \end{bmatrix}\right), \qquad t = 0, 1, 2, \ldots \quad (5.58)$$

which can be solved iteratively (together with a method to compute $\boldsymbol{\beta}$); see Rao and Molina (2015, 196) for a specification of map $\boldsymbol{f}$. Key to this approach is identity (5.57) which expands the within-trace term by $\boldsymbol{\Omega}_i^{-1}\boldsymbol{\Omega}_i$ and then kind of "factors out" $\boldsymbol{\Omega}_i$. However, there is a price to pay since we have modified the original problem by the factor $\boldsymbol{\Omega}_i^{-1}$ (and an additive term). To see this, consider the fixed-point equation in (5.58). If we replace map $\boldsymbol{f}$ by the map $\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{\theta} - [\boldsymbol{\Psi}'(\boldsymbol{\theta})]^{-1}\boldsymbol{\Psi}(\boldsymbol{\theta})$, where $\boldsymbol{\Psi}'(\boldsymbol{\theta})$ denotes the Jacobian matrix of $\boldsymbol{\Psi}$, the resulting fixed-point equation is the Newton–Raphson (NR) method. Within a neighborhood of the fixed point, the NR method has quadratic convergence and is thus optimal (since the core model is Gaussian), but the methods tends to be somewhat unreliable outside the neighborhood. Now, the point we want to make is that a fixed-point equation, say, $\boldsymbol{h}$ proves to be effective if it is numerically "more reliable" than the NR method and if it minimizes $\|\mathbf{g} - \mathbf{h}\|_M$ for some suitable metric $\|\cdot\|_M$. Then, $\boldsymbol{h}$ preserves much of the optimality of the NR method while being "more reliable".

[7] A similar idea has been proposed earlier by R. Fried, I. Molina, B. Perez, and A. Thieler (2011): Robustness analysis of unbalanced linear mixed modeling. Presentation at the ERCIM conference, London.

If we manage to proof that our M-estimators expressed as weighted estimators are closer to the NR method in terms of a suitable metric, compared with f, we could prove the superiority of our approach. But we did not try to do that. Instead, we shall only give some informal arguments. First, our weighted estimators are very closely related to the NR method. In fact, they are approximations to the NR method insofar that the major source causing numerical instabilities, the Jacobian matrix, is approximated by a matrix that is rather "inert" or less erratic w.r.t. to changes in the parameters. Hence, we conjecture that the weighted estimators are "closer" to the NR fixed-point equations than the method due to Chatrchi (2012). In particular, the weighted estimators do not fundamentally change the scaling of the fixed-point equations the way it happens with the proposal of Chatrchi (2012).

Second, experience with simulations shows that the fixed-point method of Chatrchi (2012) requires an enormous number of iterations until the termination criterion is satisfied, compared with our method. We come to the same conclusion when we study the numerical example discussed in Chatrchi (2012, chap. 2.2.3). The convergence profiles he shows (i.e. the estimates plotted vs. the number of iteration steps) decrease rapidly for the first couple of iterations, but then the profiles flatten out remarkably. There is evidence that the step length in the search for the fixed point tends to become too small after a couple of iterations. As a result, the number of updating steps growth rapidly until the method eventually converges. Such an inferior behavior is not encountered with our method. On the contrary, our method converges rather quickly. There is some evidence that the re-scaling of the estimating equations that results via the application of identity (5.57) is to blame for the bad convergence. Further evidence comes from a related application: Warnholz (2016b, chap. 3.3.3) suggests a Chatrchi-type fixed-point equation to compute the M-estimator under the Fay–Herriot model; see also method `rfh` in the `R` package `saeRobust` (Warnholz, 2016a). His method is discussed in Chapter 6. There, we point out that the estimators due to S. Warnholz are computationally extremely inefficient because they require on average 16.8 iterations until converge whereas our method is done after only 2.4 iterations; see Section 6.5. This is further evidence that the Chatrchi-type fixed-point equations are suboptimal to compute the parameter estimates.

How well does the method work in practice?

In order to study the behavior of the method in a practical context, we computed estimates of average above-ground forest biomass for a sample of Norwegian municipalities. The findings show that robust methods can indeed play off much of their advantage over standard methods when the data feature a "a bit" more variability "than expected". In our example on estimation of above-ground forest biomass means, the gains in efficiency obtainable through M-estimation are

surprisingly big.

6. Robust estimation under the Fay–Herriot model

6.1. Introduction

Consider a finite survey population U. The population U is assumed to be partitioned into n small areas $U_1, \ldots, U_n$ (see domain structure in Definition 2.1). A sample s_i of $n_i > 0$ units has been drawn from U_i ($i = 1, \ldots, n$) and the values of the study variable y_{ij} and a q-vector of auxiliary variables $\mathbf{x}_{ij}$ have been recorded for all elements $j \in s_i$. The measurements for both variables are supposed to be error-free. The vector of *area-level means* (direct estimators of the arithmetic mean) is denoted by $(y_i, \mathbf{x}_i) = (1/n_i) \sum_{j \in s_i} (y_{ij}, \mathbf{x}_{ij})$, $i = 1, \ldots, n$. Associated with the *mean*, y_i, is a measure of uncertainty of the area-specific estimator of the y-mean (e.g., variance of the estimator of the mean) denoted by D_i (such that $D_i > 0$ for all $i = 1, \ldots, n$). The unit-level elements y_{ij} and $\mathbf{x}_{ij}$ are known to the sampler but are *not* released to the analyst (who is different from the sampler). We take the view of the analyst. Let the data available to the analyst be

$$\{y_1, \ldots, y_n\}, \qquad \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \qquad \text{and} \qquad \{D_1, \ldots, D_n\}.$$

The above data structure is the starting point of the Fay–Herriot (FH) model, which was proposed by Fay and Herriot (1979) as an application of the celebrated *James–Stein* estimator (see James and Stein, 1961) for estimating small-area means on the basis of census income data. R.E. Fay and R.A. Herriot showed that the model-based estimators tend to be superior in terms of efficiency compared with the direct estimators y_i. Let Y_i, $i = 1, \ldots, n$, be real-valued random variables; the realizations (i.e. the area-specific means) of which are denoted by y_i. The basic area-level model is defined as follows (Rao, 2003, 115–116).

Definition 6.1 (Basic area-level model / Fay–Herriot model)**.**

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + b_i V_i + E_i, \quad i = 1, \ldots, n, \tag{6.1}$$

where

(i) $\mathbf{x}_i$ *is a* q*-vector* $(q < n)$ *of known area-level means, regarded as fixed regression carriers; for convenience, it is assumed that the matrix* $(\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ *has rank* q*,*

(ii) $\boldsymbol{\beta} \in \mathbb{R}^n$ *is an unknown parameter vector,*

(iii) the b_i*'s are known constants,*

(iv) the V_i*'s and* E_i*'s are independent random variables (to be specified separately).*

The Fay–Herriot model obtains from the basic area-level model if we take the constants b_i equal to one. For the moment, we shall assume that the r.v.'s E_i and V_i satisfy the laws

$$V_i \overset{\text{i.i.d.}}{\sim} P_v(0, A) \qquad \text{and} \qquad E_i \overset{\text{ind.}}{\sim} P_e(0, D_i), \quad i = 1, \ldots, n, \tag{6.2}$$

where $A \in \mathbb{R}^+$ is a variance parameter that in assumed unknown (in most cases). The D_i's are strictly positive and are assumed to be known (throughout the discussion) for all $i = 1, \ldots, n$; P_e and P_v denote absolutely continuous distribution functions.

The model defined by (6.1–6.2) is not completely specified. The definition yet lacks a specification for the d.f.'s P_e and P_v. For the canonical FH model, we have

$$P_e \equiv \mathcal{N}(0, D_i),\ i = 1, \ldots, n, \quad \text{and} \quad P_v \equiv \mathcal{N}(0, A), \tag{6.3}$$

where $\mathcal{N}$ denotes the c.d.f. of the Gaussian distribution.

The Gaussian assumption on the r.v.'s E_i and V_i is attractive since it simplifies the estimators of the unknown parameters considerably. Clearly, we can consider a setup that is more general than assuming P_e and P_v to be Gaussian distributions. Following Lahiri and Rao (1995) one may assume that the V_i's are i.i.d. r.v. which are defined only in terms of the first two moments, i.e. $\mathbb{E}[V_i] = 0$ and $\mathbb{V}[V_i] = A$; see also Rao (2003, 76). In line with this modification, we may also drop the normal law for the r.v.'s E_i and assume instead that the E_i's are independently distributed with conditional expectation and variance, given by

$$\mathbb{E}[E_i \mid Y_i = y_i] = 0 \quad \text{and} \quad \mathbb{V}[E_i \mid Y_i = y_i] = D_i, \quad i = 1, \ldots, n.$$

The specification of the d.f.'s P_e and P_v in terms of only mean and variance, is said to yield "robust" inference procedures (Rao, 2003, 77 and 130). This result is due to Lahiri and Rao (1995) who show that the usual estimator of the mean square prediction error (MSPE) of the EBLUP-estimator (with a moment-type estimator for A) remains valid under non Gaussian distributions of the r.v.'s V_i (given that $\mathbb{E}|V_i|^{8+\epsilon} < \infty$ with $0 < \epsilon < 1$). In this context, robustness means that the formula for obtaining estimates of the MSPE still applies despite the deviation from the normality assumption on the r.v.'s V_i.

Motivation and goal

Replacing the Gaussian laws by a distributional assumption that is only spec-

ified in terms of the first two moments leads to "robust" inference procedures (a view that is widely shared in the SAE community). However, the qualifier "robust" may create a false sense of security that the procedure is indeed robust w.r.t. outlying observations and other deviations from the core model, which is not the case: the "standard" estimators under the FH model (either m.l.e. or the moment-based estimators) are not robust in the sense of Principles 1 and 2.

The lack of robustness has been known for quite some time; see e.g. Datta and Lahiri (1995). These authors took a formal robust Bayes point of view, referring to the concepts in West (1984) and Berger (1994), and developed outlier robust Bayes-type predictors for the area-level model. Another Bayesian robustification has been suggested by Ghosh, Maiti, and Roy (2008). Under the frequentist inference paradigm, an *explicit* robustification has not been published. However, there are (at least) two obvious ways to attack the robustification problem, namely,

(i) robust estimation and prediction under a special (in fact, degenerate) case of the mixed linear model with block-diagonal covariance matrix (MLM; see Section 2.5);

(ii) robust etimation and prediction under the linear regression model that includes the variances D_i (i.e. variances of the direct estimator).

The first approach takes advantage that the area-level model can indeed be regarded as an MLM with block-diagonal covariance matrix (thus, EBLUP theory can be utilized), pretending certain variance parameters are known. Hence, the robustification can be tackled using the M-estimator theory for MLM's (methods RML 1 or 2) developed by Richardson and Welsh (1995) and Welsh and Richardson (1997); see Chapter 5. In the context of SAE, the method suggested by Sinha and Rao (2009) [which is an approximation of the RML 2 method] attracted a lot of attention. So, the robustification under the Fay–Herriot model (and more general area-level models) is an immediate consequence of applying the method of Sinha and Rao (2009) to the FH model. The details of this approach have been written up in the Ph.D. thesis of Warnholz (2016b, chap. 3.3).

The second approach can be seen as a "bottom-up" approach. In contrast to the "top-down" approach in (i), which starts with the general class of MLM's and boils it down to some special case, we may start in (ii) from the regression model and include the known variances D_i ($i = 1, \ldots, n$) in an appropriate way. Two remarks are in order. First, the D_i can be "included" in a formal way by considering the heteroscedastic model $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + E_i$, where the E_i's have mean zero an variances equal to $A + D_i$, $i = 1, \ldots, n$. Second, the first approach is theoretically much less convincing since the boiled down MLM has lost its essential characteristic of being a mixed linear model—it does not include random effects.

In fact, neither of the two approaches is truly satisfactory. Consequentially, we have chosen yet another approach for obtaining robust predictors under the FH model. To this end, we go back to the roots of the FH model, that is James–Stein estimation. In their seminal paper, Fay and Herriot (1979) motivate their SAE method as an extension of the James–Stein estimator that obtains via an empirical Bayes argument, which has been popularized by Efron and Morris (1972, 1973b). We pick up this line of argument, adopt a robust Bayes view and derive from there a robust empirical Bayes method that creates a direct link to the seminal work of Huber (1964), but under a slightly more general model. From this, we derive M-estimators under the FH model. [Furthermore, instead of merely reproducing the 1970's Bayes and empirical Bayes theory developed by B. Efron and C. Morris, we state some new results and/or give our own (simplified) proofs to some of their results.]

Technically, the resulting M-estimator under the FH model coincides largely with the Sinha–Rao proposal applied to the FH model. Nevertheless, our major contribution is to put the derivation of M-estimators under the area-level model onto a formal, theoretical foundation, instead of, what Stahel and Welsh (1997) called, merely "huberizing suitable quantities". Via this theoretical motivation, we can view the existing estimators and their behavior from a new perspective. In addition to M-estimators, we introduce the class of generalized regression M-estimator (GM) to the field of fitting area-level models. GM-estimators are an indispensable tool in the presence of influential observations in the design space of the model.

Outline of the chapter

The chapter is organized into six sections. To start with, we study a simplified hierarchical Gaussian model and discuss the Bayes rule. Then, we introduce the Hodges–Lehmann theory and study compromise rules that dominate the m.l.e. in terms of compound risk (this will be made precise later), but do not exhibit extreme component-wise risk as it is experienced with the Bayes rule. It will be seen that the limited translation rule of Efron and Morris (1971) is an approximation to this theory. In Section 6.3, we consider the empirical Bayes situation and an estimate of the limited translation rule. Then, in Section 6.4, we discuss robust estimators and predictors. Finally, the robust procedures are studied in a small-scale model-based simulation study (see Section 6.5) and a three case studies in Section 6.6. Finally, section 6.7 draws together the major findings.

Note on the notation

The notation used in this chapter will differ slightly from the rest of the thesis. We have chosen to follow a style of notation that is commonly adopted in the literature on Bayesian statistics. In addition, we use the opportunity to remind

the reader of some important definition.

(i) Unless otherwise stated, the notions of "measurability", "almost everywhere", and "integrable" are understood w.r.t. Lebesgue measure; for further details, see Chapter A.1 in the appendix.

(ii) We write $\mathbb{E}_{\square}$ to mean *conditional expectation* given $\square$. Depending on the context, the meaning of $\square$ will be clear. For instance, $\mathbb{E}_{\pi_0}$ will denote expectation versus the p.d.f. π_0 of a prior distribution.

(iii) $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ denotes the extended real line.

(iv) $\mathscr{P}(\mathbb{R})$ denotes the class of probability measures supported on $\mathbb{R}$.

(v) π_0 denotes the p.d.f. (provided it exists) of a probability measure $P_0 \in \mathscr{P}(\mathbb{R})$. Whenever the symbol π is used to denote the area of the unit circle, we draw explicitly attention to this in the text.

(vi) We use the notations f' and ∇f interchangeably to mean the first derivative of f.

(vii) A function $f : [a, b] \to \mathbb{R}$ is said to be absolutely continuous (a.c) if it satisfies Definition A.2. On $\mathbb{R}$, we write $f \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$ to mean a function $f : \mathbb{R} \to \mathbb{R}$ that is a.c. on bounded intervals; a formal definition is given in (A.5).

(viii) Φ denotes the c.d.f. of the standardized normal / Gaussian distribution. We write Φ_{μ,σ^2} to mean the c.d.f. of the Gaussian distribution with mean μ and variance σ^2. In view of this, we have $\Phi \equiv \Phi_{0,1}$.

6.2. Hierarchical normal model: A Bayesian view

In what follows, we shall be concerned with the definition of the model given in (6.1), (6.2) and the normality assumption on both error terms in (6.3). For ease of exposition, we simplify the model according to the subsequent assumptions.

(i) The linear predictor $\mathbf{x}_i^T\boldsymbol{\beta}$ $(i = 1\ldots, n)$ is replaced by the grand mean μ, which is supposed to be zero for ease of discussion.

(ii) Let $D_i = D$, $\forall i = 1, \ldots, n$, in (6.2); D is supposed known.

These simplifications are not really consequential and will be brought back into action in the further course of the discussion. Under the simplifying assumptions, the FH model can be expressed as a *hierarchical* Gaussian model, defined as

$$\text{(sampling model)} \qquad (Y_i \mid \Theta_i = \theta_i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_i,\, D), \tag{6.4}$$

$$\text{(prior distr. } \pi_0\text{)} \qquad (\Theta_i \mid A) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\, A), \quad \text{for all } i = 1, \ldots, n. \tag{6.5}$$

The law in (6.5) is a Bayesian *prior (distribution)* that represents our a priori beliefs in the distribution of the Θ_i's. Therefore, the Θ_i's (with mean $\mu = 0$) are regarded as r.v.; the realizations of which, θ_i, are called parameters; A and D, on the other hand, are called hyperparameters. We use the symbol π_0 in the annotation of (6.5) to mean the prior p.d.f., the c.d.f. of which is denoted by P_0. The prior d.f. in (6.5) expresses our belief that the Θ_i's share a common parent distribution. This situation is also called *exchangeable* prior distribution; see e.g. Lindley and Smith (1972, 2-3). Since D is a known quantity, we shall from here on assume w.l.o.g. that $D = 1$ (in cases where D does not equal one, we may transform the data such that we get $D = 1$, then apply the methods to the transformed data and finally back transform to the original coordinates). In terms of notation, we shall write $\mathbf{y}$ and $\boldsymbol{\theta}$ to mean, respectively, $(y_1, \ldots, y_n)^T$ and $(\theta_1, \ldots, \theta_n)^T$.

Before we continue with our Bayesian argument, we remind the reader that our goal is to estimate $\boldsymbol{\theta}$ by an estimation rule $\boldsymbol{\delta} : \mathbb{R}^n \to \mathbb{R}^n$ using average squared error loss,

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{n}\big\|\boldsymbol{\delta}(\mathbf{y}) - \boldsymbol{\theta}\big\|^2,$$

to assess the performance of the rule, where $\|\cdot\|$ denotes the Euclidean norm (for ease of reading, we drop the subscript 2 in our standard notation $\|\cdot\|_2$). Note that $L(\boldsymbol{\theta}, \cdot)$ is average squared error loss; some authors use sum of squared error loss instead.[1] Also, if we are not willing to make a truly Bayesian prior assumption,

[1] B. Efron and C. Morris also use both notions; in their 1971 paper on limiting the risk of Bayes estimators, they use the *sum* of squared error loss whereas in their 1973b paper they work with the average.

the optimal rule for estimating $\boldsymbol{\theta}$ is the m.l.e., given by

$$\boldsymbol{\delta}^0(\mathbf{y}) = \mathbf{y}.$$

Under model (6.4), the m.l.e. has *risk* (function) or expected loss

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^0) = \mathbb{E}_{\boldsymbol{\theta}}\big(L(\boldsymbol{\theta}, \boldsymbol{\delta}^0)\big) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\theta}}\Big(\big\|\boldsymbol{\delta}^0(\mathbf{y}) - \boldsymbol{\theta}\big\|^2\Big) = 1$$

for every value of $\boldsymbol{\theta} \in \mathbb{R}^n$, where $\mathbb{E}_{\boldsymbol{\theta}}$ denotes expectation conditional on $\boldsymbol{\theta}$. For ease of display, we will sometimes use the shorthand notation $\boldsymbol{\delta}^0$ instead of $\boldsymbol{\delta}^0(\mathbf{y})$.

Now, let us come back to our Bayesian argument. The conditional distribution of Θ_i given $Y_i = y_i$ is called the *posterior* (distribution) of the model defined in (6.4–6.5). The posterior expectation is denoted by $\mathbb{E}_Y$.

A prior distribution is said to be *conjugate* if the resulting posterior distribution is in the same family of distributions as the prior d.f. (Lehmann and Casella, 1998, 236). Conjugate priors are typically chosen for algebraic convenience because they yield closed-form expressions for the posterior; otherwise numerical integration may be necessary. Since the Gaussian prior over the mean is self-conjugate w.r.t. the Gaussian sampling model's likelihood function, the posterior distribution is also Gaussian; see e.g. Lindley and Smith (1972, Lem. 1). For the hierarchical Gaussian model postulated in (6.4) and (6.5), the posterior distribution is given by (see e.g. Efron and Morris, 1973b, 117)

$$(\Theta_i \mid Y_i = y_i) \overset{\text{ind.}}{\sim} \mathcal{N}\big((1-B)y_i,\, 1-B\big), \tag{6.6}$$

for all $i = 1, \ldots, n$, where

$$B = \frac{1}{A+1}. \tag{6.7}$$

The posterior mean is a smooth blend of $(1-B)$ times the m.l.e., y_i, and B times the a priori mean (which equals zero here). In a formal Bayesian analysis, the hyperparameter A is a known quantity. We will consider both cases: A is either assumed known or not known to the analyst. In latter case, the analyst can attempt to estimate A from the data (empirical Bayes case; see below).

Remark. Identity (6.7) is of crucial importance for the subsequent discussion since the parametrization in terms of B is considerably simpler than the one using A directly.

When A is a known quantity, a natural estimator of θ_i under the Gaussian model in (6.4–6.5) is given by

$$\delta_i^*(y_i) = (1-B)y_i, \quad i = 1, \ldots, n, \tag{6.8}$$

where B is defined in (6.7). We shall denote the rule for estimating $\boldsymbol{\theta}$ by $\boldsymbol{\delta}^* = (\delta_1^*, \ldots, \delta_n^*)^T$, having suppressed the dependency on $\mathbf{y}$. The estimator δ_i^* of θ_i is

the *Bayes rule* (see Definition 2.12) under average squared error loss function since it minimizes the *Bayes risk* (see below and Definition 2.11). This follows from the fact that δ_i^* is the posterior mean under the hierarchical model; see (6.6). It is also remarkable that the choice of squared error loss function is not essential to the optimality of $\boldsymbol{\delta}^*$. Indeed, it can be shown that $\boldsymbol{\delta}^*$ will minimize the posterior risk for any loss function $l(\boldsymbol{\theta}, \boldsymbol{\delta})$ which is an increasing function of $\|\boldsymbol{\delta} - \boldsymbol{\theta}\|$ (Efron and Morris, 1971, 807). Actually, existence and uniqueness of the Bayes estimator are guaranteed under more general conditions (e.g., when the non-negative loss function $l(\theta, a)$ is strictly convex in a); see Thm. 1.1.1 and Cor. 1.1.4 in Lehmann and Casella (1998). We stick with average squared error loss $L(\boldsymbol{\theta}, \cdot)$ because it leads to estimators that can be obtained explicitly.

The Bayes risk of $\boldsymbol{\delta}^*$ under $L(\boldsymbol{\theta}, \cdot)$ versus the Gaussian prior p.d.f. π_0 defined in (6.5) is given by (Efron and Morris, 1973b, 117)

$$r(\pi_0, \boldsymbol{\delta}^*) = \mathbb{E}_{\pi_0}\big[R(\boldsymbol{\theta}, \boldsymbol{\delta}^*)\big] = 1 - B, \tag{6.9}$$

where $\mathbb{E}_{\pi_0}$ denotes expectation w.r.t. prior π_0. The Bayes risk is also called *ensemble risk* (or compound risk) since it is averaged over all $i = 1, \ldots, n$ coordinate-specific estimation problems [Morris (1983, 48) uses yet another term, namely, empirical Bayes MSE].

Under the Bayes prior assumption, the Bayes estimator $\boldsymbol{\delta}^*$ provides a reduction in Bayes risk over the m.l.e. In line with Efron and Morris (1971, 809), we call the reduction of Bayes risk below that of the m.l.e. *savings* versus prior π_0, and define

$$\mathrm{S}(\pi_0) = r(\pi_0, \boldsymbol{\delta}^0) - r(\pi_0, \boldsymbol{\delta}^*).$$

The maximal possible savings of $\boldsymbol{\delta}^*$ versus the prior in (6.5) equals $S(\pi_0) = B$. In order to facilitate risk comparison among different estimating rules, it is convenient to work with normalized risk measures. The first measure is the *relative savings loss* of Efron and Morris (1971, 809) which gives the proportional reduction in savings that we sacrifice if we use the (generic) rule $\boldsymbol{\delta}$ instead of the Bayes rule. The second measure is used to evaluate the robustness of a rule $\boldsymbol{\delta}$ by a scaled version of $\sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \boldsymbol{\delta})$, i.e., the maximum harm that could be encountered; this measure has been studied by Berger (1982, 361).

Definition 6.2 (Relative savings loss, Efron and Morris, 1971)**.** *The relative savings loss for an arbitrary rule $\boldsymbol{\delta}$ versus prior π_0 is defined as*

$$\mathrm{RSL}(\pi_0, \boldsymbol{\delta}) = \frac{r(\pi_0, \boldsymbol{\delta}) - r(\pi_0, \boldsymbol{\delta}^*)}{\mathrm{S}(\pi_0)}.$$

Definition 6.3 (RRR measure of Berger, 1982). *Let δ denote an arbitrary estimation rule versus prior π_0; we define*

$$\mathrm{RRR}(\pi_0, \boldsymbol{\delta}) = \frac{\left[\sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \boldsymbol{\delta})\right] - \left[\sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \boldsymbol{\delta}^0)\right]}{\mathrm{S}(\pi_0)}.$$

Remarks. (i) RSL near zero is optimal. The Bayes estimator has RSL equal to zero, while RSL near one indicates performance as bad as that of the m.l.e. The main purpose in using the normalized versions of risk is that RSL is often independent of the model parameters.

(ii) The measure RRR gives the maximum possible *harm* that could be encountered using $\boldsymbol{\delta}$ instead of the minimax (i.e., most robust) estimator $\boldsymbol{\delta}^0$, relative to the potential Bayes risk improvement over $\boldsymbol{\delta}^0$.

Under the assumption of (sum or average) squared error loss, it proves useful to work with *Stein's lemma* (a.k.a. Stein's identity), which provides a useful device for handling covariance terms under the n-variate Gaussian model with zero mean and identity covariance matrix, $\mathbf{I}_n$. Before giving the lemma, we introduce some notation. Consider the functions $h : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$, where the component functions of $\mathbf{h}$ are denoted by h_i; we define

$$\nabla_i = \frac{\partial}{\partial x_i}, \quad i = 1, \ldots, n, \qquad \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i, \qquad \nabla \cdot \mathbf{h} = \sum_{i=1}^{n} \nabla_i h_i.$$

Definition 6.4 (Almost differentiable function, Stein, 1981). *A function $h : \mathbb{R}^n \to \mathbb{R}$ is called almost differentiable if there exists a function $\nabla h : \mathbb{R}^n \to \mathbb{R}^n$ such that, for all $\mathbf{z} \in \mathbb{R}^n$,*

$$h(\mathbf{y} + \mathbf{z}) - h(\mathbf{y}) = \int_0^1 \mathbf{z} \cdot \nabla h(\mathbf{y} + t\mathbf{z}) \mathrm{d}t \tag{6.10}$$

for almost all $\mathbf{y} \in \mathbb{R}^n$.

A vector-valued function $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$ is called almost differentiable if all its coordinate functions are. Stein's notion of an almost differentiable function is rather uncommon. Observe that Formula (6.10) has similarities with the fundamental theorem of calculus (for Lebesgue integrable functions); see Theorem A.1 and also Theorem A.2 in the appendix. In fact, functions that satisfy (6.10) can be seen to satisfy the mean value theorem. A useful and sufficiently broad class of functions that satisfy (6.10) is the family of Lipschitz continuous functions $h : U \to \mathbb{R}$ (see Definition A.3), where U is an open and connected subset of $\mathbb{R}^n$ (such that the line segment $\mathbf{y} + t\mathbf{z}$ with $0 \leq t \leq 1$ is contained in U); this class can easily be extended to vector-valued functions $\mathbf{h} : U \to \mathbb{R}^n$.

Lemma 6.1 (Stein, 1981). *Consider $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\theta}, \mathbf{I}_n)$ for some $\boldsymbol{\theta} \in \mathbb{R}^n$. Let $h : \mathbb{R}^n \to$*

$\mathbb{R}$ *be an almost differentiable function with* $\mathbb{E}_{\boldsymbol{\theta}}\|\nabla h(\mathbf{Y})\| < \infty$*, then*

$$\mathbb{E}_{\boldsymbol{\theta}}\big[(\mathbf{Y}-\boldsymbol{\theta})h(\mathbf{Y})\big] = \mathbb{E}_{\boldsymbol{\theta}}\big[\nabla h(\mathbf{Y})\big].$$

Proof. See Lem. 2 in Stein (1981). We shall sketch the proof under the simplifying assumption that $h : \mathbb{R} \to \mathbb{R}$ is an almost everywhere differentiable [w.r.t. Lebesque measure] function of only one variable. Hence,

$$\int_{\overline{\mathbb{R}}} h'(x)\Phi(\mathrm{d}x) = \int_0^\infty h'(x)\Big[\int_x^\infty t\Phi(\mathrm{d}t)\Big]\mathrm{d}x - \int_{-\infty}^0 h'(x)\Big[\int_{-\infty}^x t\Phi(\mathrm{d}t)\Big]\mathrm{d}x$$

then, by identity $\phi'(u) = -u\phi(u)$,

$$= \int_0^\infty t\phi(t)\Big[\int_0^t h'(x)\mathrm{d}x\Big]\mathrm{d}t - \int_{-\infty}^0 t\phi(t)\Big[\int_t^0 h'(x)\mathrm{d}x\Big]\mathrm{d}t$$

and using Fubini's theorem (see Thm. A.4), we have

$$= \int_0^\infty t\phi(t)\big[h(t)-h(0)\big]\mathrm{d}t - \int_{-\infty}^0 t\phi(t)\big[h(0)-h(t)\big]\mathrm{d}t = \int_{\overline{\mathbb{R}}} th(t)\Phi(\mathrm{d}t).$$

Stein's lemma for a n-variate function $\mathbf{h}$ can be proved by analogous arguments. ■

Remark. Early papers, see e.g. James and Stein (1961), Baranchik (1964), or Efron and Morris (1971, 1972), used a (rather tedious) Poisson representation of the non-central chi-squared distribution for risk comparisons; see Sect. A.2 and in particular Formula (A.39). Stein's lemma simplifies proofs substantially. Stein's lemma and theorem were known to the statistical community since the mid to late 1970s. The earliest appearance seems to be Stein's announcement of the results in the discussion to the paper Efron and Morris (1973a). Rumours go that Stein wasn't planning to publish the paper until colleagues at Stanford convinced him to do so in 1981.

Stein's lemma is useful to compute expected loss for – what he calls – a nearly arbitrary estimating rule.

Definition 6.5 (Nearly arbitrary estimating rule, Stein, 1981)**.** *Consider* $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I}_n)$ *for some* $\boldsymbol{\theta} \in \mathbb{R}^n$*. Let* $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$ *be a known almost differentiable function. The nearly arbitrary rule for estimating* $\boldsymbol{\theta}$ *is defined as*

$$\boldsymbol{\delta}(\mathbf{Y}) = \mathbf{Y} + \mathbf{h}(\mathbf{Y}).$$

This class of rules is general enough to be considered interesting and will play a crucial role in the further course of discussion. Observe that the Bayes rule $\boldsymbol{\delta}^*$ defined in (6.8) is a member of this class of rules with $\mathbf{h}(\mathbf{Y}) = -B\mathbf{Y}$. The following theorem provides an expression for the expected loss of the family of

nearly arbitrary rules.

Theorem 6.1 (Stein, 1981). *Let* $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I}_n)$ *for some* $\boldsymbol{\theta} \in \mathbb{R}^n$. *Let* $\boldsymbol{\delta}$ *satisfy Definition 6.5 with function* $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$, *such that*

$$\mathbb{E}_{\boldsymbol{\theta}}\Big[\sum_{i=1}^{n} \big|\nabla_i h_i(\mathbf{Y})\big|\Big] < \infty,$$

where h_i *denotes the* i*th coordinate-specific function of* $\mathbf{h}$. *Then, for each* $i = 1, \ldots, n$,

$$R(\theta_i, \delta_i) = \mathbb{E}_{\boldsymbol{\theta}}\Big[\big(Y_i + h_i(\mathbf{Y}) - \theta_i\big)^2\Big] = 1 + \mathbb{E}_{\boldsymbol{\theta}}\big[h_i^2(\mathbf{Y}) + 2\nabla_i h_i(\mathbf{Y})\big],$$

and consequently

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) = \mathbb{E}_{\boldsymbol{\theta}}\big\|\mathbf{Y} + \mathbf{h}(\mathbf{Y}) - \boldsymbol{\theta}\big\|^2 = 1 + \frac{1}{n}\mathbb{E}_{\boldsymbol{\theta}}\Big[\big\|\mathbf{h}(\mathbf{Y})\big\|^2 + 2\nabla \cdot \mathbf{h}(\mathbf{Y})\Big].$$

Proof. See Thm. 1 in Stein (1981); the proof is straightforward and obtains by application of Stein's lemma. ■

Remark. Let $\mathbf{y}$ denote a realization of the r.v. $\mathbf{Y}$ under the assumptions of Thm. 6.1. Stein's theorem provides

$$\hat{R}(\boldsymbol{\theta}, \boldsymbol{\delta}) = 1 + \frac{1}{n}\big\|\mathbf{h}(\mathbf{y})\big\|^2 + \frac{2}{n}\nabla \cdot \mathbf{h}(\mathbf{y}),$$

which is an *unbiased estimator* of the risk $R(\boldsymbol{\theta}, \cdot)$ for rules of the form $\mathbf{Y} + \mathbf{h}(\mathbf{Y})$. Yet, this device proves also useful to compute the Bayes risk. That is, instead of directly computing the Bayes risk it is usually easier to obtain an unbiased estimate of the risk $R(\boldsymbol{\theta}, \cdot)$ first and then integrate over $\boldsymbol{\theta}$ versus prior π_0. Since $R(\boldsymbol{\theta}, \cdot)$ is constant in $\boldsymbol{\theta}$ for a large number of rules, integration versus π_0 is easy.

Next, we give formulas of RRR and RSL for the estimating rules discussed so far. Under model (6.4) and prior (6.5) with a priori variance $A = 1/B - 1$, we obtain the following formulae for RSL and RRR,

$$\mathrm{RSL}(\pi_0, \boldsymbol{\delta}) = 1 + \frac{1}{B}\big(r(\pi_0, \boldsymbol{\delta}) - 1\big) \qquad \text{and} \qquad \mathrm{RRR}(\pi_0, \boldsymbol{\delta}) = \frac{1}{B}\Big(\sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \boldsymbol{\delta}) - 1\Big).$$

Hence, we have for the m.l.e. and the Bayes estimator, respectively,

$$\begin{aligned} &\mathrm{RSL}(\pi_0, \boldsymbol{\delta}^0) = 1, \qquad &&\text{and} \qquad &&\mathrm{RRR}(\pi_0, \boldsymbol{\delta}^0) = 0, \\ &\mathrm{RSL}(\pi_0, \boldsymbol{\delta}^*) = 0, \qquad &&\text{and} \qquad &&\mathrm{RRR}(\pi_0, \boldsymbol{\delta}^*) = B\bigg(1 + \sup_{\boldsymbol{\theta}} \frac{\|\boldsymbol{\theta}\|^2}{p}\bigg) - 2, \end{aligned}$$

where, for the last claim, we have used Theorem 6.1 with $\mathbf{h}(\mathbf{Y}) = -B\mathbf{Y}$. From the above formulae, it is clear that $\mathrm{RRR}(\pi_0, \boldsymbol{\delta}^*)$ will be problematic for large $\|\boldsymbol{\theta}\|$. The m.l.e., on the other hand, has no savings versus prior π_0 but minimax risk $R(\boldsymbol{\theta}, \boldsymbol{\delta}^0) = 1$ for every value of $\boldsymbol{\theta} \in \mathbb{R}^n$. In other words, the Bayes rule and

the m.l.e. behave opposite to each other and are both rather "extreme" (either overly "optimistic" or "pessimistic" when dealing with prior knowledge).

It is thus natural to ask whether a compromise exists. The central idea is that we might be willing to be worse than $\boldsymbol{\delta}^0$ in terms of RRR only in proportion to the potential gain obtainable in using prior information. Ideally, we hope to find estimators with small values for both RSL and RRR.

6.2.1. Hodges–Lehmann theory

If *complete* knowledge of the prior distribution π_0 is available, the Bayes rule $\boldsymbol{\delta}^*$ versus the elicited prior π_0 is a sensible choice (or, in the eyes of a formal Bayesian, the only relevant choice since $\boldsymbol{\delta}^*$ minimizes the Bayes risk). When *no* prior knowledge or beliefs regarding $\boldsymbol{\theta}$ are available (or trusted), then choosing the *minimax* rule may seem appropriate. However, choosing the minimax rule forces us to act as if $\boldsymbol{\theta}$ follows the probability distribution least favorable to us even though we feel pretty sure that $\boldsymbol{\theta}$ actually follows another distribution (Hodges and Lehmann, 1952, 396). Therefore, the minimax approach tends to be overly pessimistic. On the other hand, the only weak point in the fundamentally Bayesian argument is, of course, the assumption that π_0 is completely known. All (serious) Bayesian will acknowledge that, in a finite amount of time, only subjective *approximations* to π_0 can typically be constructed (through a prior elicitation process and initial beliefs); see Berger (1982) and Berger (1994). It is thus natural to desire an analysis that is robust with respect to possible misspecification of the prior π_0. For the situation that we do have *rather strong beliefs* in the prior but are *not willing to entirely rely* on such information, Hodges and Lehmann (1952) suggested a formal compromise of the two principles. The compromise is expressed through restricting either the Bayes or the minimax principle.

The two principles play a fundamental role in the further course of discussion, therefore we should stop our argument for a moment to make important assumptions and give some definitions.

(i) For ease of discussion, we restrict attention to the univariate location model

$$Y = \Theta + E, \tag{6.11}$$

where E is a r.v. with law $E \sim \mathcal{N}(0,1)$ and Θ denotes a r.v. with prior p.d.f. π_0, which is supposed to be absolutely continuous (a.c.) on $\mathbb{R}$, formally $\pi_0 \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$. The realization of Θ is denoted by θ.

(ii) Although we assume that our prior beliefs can be formalized in terms of a p.d.f. π_0, it will prove useful to work under a more general setup. Let $P_0 \in \mathscr{P}(\mathbb{R})$ denote any prior probability measure whose density is π_0. In

line with this, we shall express the Bayes risk of δ in terms of P_0,

$$r(P_0, \delta) := \int_{\overline{\mathbb{R}}} R(\theta, \delta) P_0(\mathrm{d}\theta),$$

that is, the Bayes risk is regarded as a functional on $\mathscr{P}(\mathbb{R})$.

(iii) The marginal density of r.v. Y under model (6.11) versus prior d.f. P_0 is

$$(\Phi * P_0)(y) := \int_{\overline{\mathbb{R}}} \Phi'(y - \theta) P_0(\mathrm{d}\theta), \tag{6.12}$$

where $*$ denotes *convolution*; Φ is the c.d.f. of the standard Gaussian. The marginal p.d.f. of Y is well-defined for a.e. $y \in \mathbb{R}$ (this follows from Theorem A.6).

(iv) Let $F \in \mathscr{P}(\mathbb{R})$ with p.d.f. f. If $f \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$, then the *Fisher information* for location is given by

$$\mathcal{I}(f) := \int_{\overline{\mathbb{R}}} \frac{\{f'(y)\}^2}{f(y)} \mathrm{d}y.$$

Under a slightly extended definition of Fisher information (which is not consequential for our purposes), Thm. 3 in Huber (1964) shows that $\mathcal{I}(f) < \infty$ provided f is absolutely continuous on $\mathbb{R}$; a much more elegant proof of this result was later given in Huber (1981, see Thm 4.4.2).

We are now in the position to present the principles that originate in the work of Hodges and Lehmann (1952). To this end, suppose that $\mathscr{D}$ is a family of estimating rules (see Section 2.4). For the moment, we shall only assume that the set $\mathscr{D}$ is nonempty and that it contains relevant candidate estimators; later, we will give a rigorous definition of this family. Also, let ϵ be a fixed number in $[0, 1]$ and let $M_0 > 0$ be a given number such that $M_0 > R(\theta, \delta^0)$ (otherwise principle PI, see below, does not have solutions). The two principles, referred to as PI and PII, are as follows.

PI: restricted Bayes principle

Find δ_ϵ such that

$$r(P_0, \delta_\epsilon) = \inf_{\delta \in \mathscr{D}} r(P_0, \delta)$$

subject to

$$R(\theta, \delta_\epsilon) \leq M_0 \quad \text{for all } \theta \in \mathbb{R}.$$

PII: restricted minimax principle

Find δ_ϵ such that

$$\sup_{P \in \mathcal{P}_\epsilon} r(P, \delta_\epsilon) = \inf_{\delta \in \mathscr{D}} \sup_{P \in \mathcal{P}_\epsilon} r(P, \delta)$$

where, for arbitrary d.f. H supported on $\mathbb{R}$ and $\epsilon \in (0,1)$,

$$\mathcal{P}_\epsilon = \big\{P \,:\, P(y) = (1-\epsilon)P_0(y) + \epsilon H(y)\big\}.$$

Remarks. (i) In their Theorems 1 and 2, Hodges and Lehmann (1952) gave sufficient conditions for the two principles to be equivalent in the following sense: If δ_ϵ is restricted minimax, then δ_ϵ is a restricted Bayes solution with expected loss bounded by $\sup R(\theta, \delta_\epsilon)$ et vice versa.

(ii) From a Bayesian perspective, the "natural" principle is the restricted Bayes principle, where we obtain the minimizer of the Bayes risk, δ_ϵ, subject to a predetermined bound on the expected loss.

(iii) The rule δ_ϵ is a compromise between the formal Bayes rule versus the ideal, elicited base prior P_0 and the minimax rule. In what follows, we shall typically assume that the elicited prior is taken to be $P_0 \equiv \Phi$, i.e. the c.d.f. of the standard Gaussian distribution.

Family of estimating rules

So far, we have worked with a rather unspecific class of estimating rules $\mathscr{D}$. We shall now be more precise in this regard. Let $P_0 \in \mathscr{P}(\mathbb{R})$ and denote expectation w.r.t. to this measure by $\mathbb{E}_{P_0}$. The formal Bayes rule under model (6.11) and prior P_0 minimizes the Bayes risk (see e.g. Stein, 1981, 1139–40)

$$\begin{aligned} r(P_0, \delta) &= \mathbb{E}_{P_0}\mathbb{E}_\theta\big[\big(\theta - \delta(Y)\big)^2 \mid Y = y\big] \\ &= \mathbb{E}_{P_0}\left[\frac{1}{(\Phi * P_0)(y)} \int \big(\theta - \delta(y)\big)^2 \phi(y-\theta) P_0(\mathrm{d}\theta)\right] \end{aligned}$$

and is given by the formal posterior expectation,

$$\delta^*(y) = \mathbb{E}_\theta[\theta \mid Y = y] = y + \mathbb{E}_\theta[\theta - Y \mid Y = y] = y + \nabla \log(\Phi * P_0). \tag{6.13}$$

In order to see the "mechanics" behind (6.13), we shall for the moment suppose that $P_0 \equiv \Phi_{\mu,A}$, i.e. the Gaussian prior distribution with mean μ and variance A. Then δ^* in (6.13) coincides – as expected – with the Bayes rule in (6.8). The key to see this is the observation that the family of normal densities is closed under convolutions; in particular, we have $(\Phi_{\mu_1,\sigma_1^2} * \Phi_{\mu_2,\sigma_2^2})(y) = \Phi_{\mu_1+\mu_2,\sigma_1^2+\sigma_2^2}(y)$, for any $y \in \mathbb{R}$ (see e.g. Feller, 1971, 46). Hence, in the case of a Gaussian prior with mean μ and variance A, we have

$$\nabla \log(\Phi_{0,1} * \Phi_{\mu,A})(y) = \nabla \log \Phi_{\mu,A+1}(y) = -\frac{y-\mu}{A+1},$$

which then implies the Bayes rule via (6.13); note that $\Phi_{0,1} \equiv \Phi$.

Stein (1981, 1140) goes even a step further in his discussion of (6.13) and defines the Bayes rule δ^* when P_0 is possibly an infinite measure. We stick with the assumption that P_0 is a well-defined probability measure. More importantly,

in view of the form of the rule δ^* in (6.13), we shall define the following class of estimating rules

$$\mathscr{D} = \big\{\delta \,:\, \delta(y) = y + \psi(y), \quad \psi \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})\big\}. \tag{6.14}$$

The next result shows that the formal Bayes rule defined in (6.13) is an element of the family $\mathscr{D}$.

Proposition 6.1. *The formal Bayes rule defined in (6.13) is an element of the set $\mathscr{D}$ with*

$$\psi(y) = \nabla \log(\Phi * P_0)(y)$$

provided that P_0 is such that $\mathrm{d}P_0 = \pi_0$ with $\pi_0 \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$.

Proof. Define $f(y) = (\Phi * P_0)(y)$ and observe that $\psi(y) \equiv f'(y)/f(y)$. By Proposition A.1, we have $f \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$. Next, we shall show that $f' \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$. To this end, we note first that $\phi \in C^\infty$, i.e. an infinitely continuously differentiable function since ϕ is a real analytic function. Hence, we can apply Theorem A.1 to get

$$f'(y) = \int_{\overline{\mathbb{R}}} \phi'(y-\theta)\pi_0(\theta)\mathrm{d}\theta. \tag{A}$$

The function ϕ' is an element of $L^1(\mathbb{R})$ since all derivatives of ϕ are absolutely integrable; this implies that (A) is well-defined for a.e. $y \in \mathbb{R}$; it now follows that $\phi'(y) = -y\phi(y)$ is also an element of $\mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$. Therefore, Proposition A.1 applied to (A) implies that $f' \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$. It remains to check that $f'/f \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$, but this follows from Lemma A.1 for any interval $[a, b]$ and thus locally in $\mathbb{R}$. ■

Our next result is of principal importance and has been independently discovered in 1980 by Bickel (1980, 1983), Marazzi (1980, 1985) and Levit (1980)[2], respectively, Berkhin and Levit (1980). The result shows that the Bayes risk for rules $\delta \in \mathscr{D}$ can be expressed in terms of the Fisher information for location.

Proposition 6.2 (Marazzi, 1980)**.** *Suppose $Y \sim \mathcal{N}(\theta, 1)$ for some $\theta \in \mathbb{R}$. Let $\delta \in \mathscr{D}$ and consider estimation of θ versus prior d.f. $P \in \mathscr{P}(\mathbb{R})$ under squared error loss. The rule*

$$\delta(y) = y + \psi(y), \qquad \textit{with} \quad \psi = \nabla \log(\Phi * P)$$

is Bayes with risk (in terms of the Fisher information for location)

$$r(P, \delta) = 1 - \mathcal{I}(\Phi * P).$$

Proof. We follow Marazzi (1980, 7) and consider rules of the form $\delta_a(y) = y + a\psi(y)$, where $a \in \mathbb{R}$ (these rules form a slightly more general class than $\mathscr{D}$). Introducing the parameter a spares us from using variational methods. Put $f \equiv \Phi * P$. By application of Stein's Theorem and integrating versus prior d.f.

[2] Some of the results have already been announced in 1979; see Levit (1979).

P, we have

$$r(P, \delta_a) = 1 + a^2 \mathbb{E}\psi^2 + 2a\mathbb{E}\psi',$$

where $\mathbb{E}$ stands for the compound expectation $\mathbb{E}_P\mathbb{E}_\theta$. First, we consider minimization on a, and obtain the sufficient condition

$$a_0 = -\frac{\mathbb{E}\psi'}{\mathbb{E}\psi^2};$$

hence, the respective rule writes

$$\delta_{a_0}(y) = y - \frac{\mathbb{E}\psi'}{\mathbb{E}\psi^2}\psi$$

with Bayes risk

$$r(P, \delta_{a_0}) = 1 - \frac{\big(\mathbb{E}\psi'\big)^2}{\mathbb{E}\psi^2}. \tag{A}$$

Before we consider minimizing (A) on ψ, we shall make the following important observation. For $f, \psi \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$, we have by integration by parts (see Lem. A.2)

$$\mathbb{E}\psi' = \int_{\overline{\mathbb{R}}} \psi' f \mathrm{d}t = -\int_{\overline{\mathbb{R}}} f'\psi \mathrm{d}t$$

since $\psi(t)f(t) \to 0$ as $t \to \pm\infty$. Hence,

$$\big(\mathbb{E}\psi'\big)^2 = \Big|\int_{\overline{\mathbb{R}}} \psi' f \mathrm{d}t\Big|^2 = \Big|\int_{\overline{\mathbb{R}}} \psi f' \mathrm{d}t\Big|^2 = \Big|\int_{\overline{\mathbb{R}}} \psi f' \frac{\sqrt{f}}{\sqrt{f}} \mathrm{d}t\Big|^2$$

and by the Schwarz inequality (see e.g. Huber, 1981, 78)

$$\leq \int_{\overline{\mathbb{R}}} \psi^2 f \mathrm{d}t \int_{\overline{\mathbb{R}}} \left(\frac{f'}{f}\right)^2 f \mathrm{d}t.$$

Therefore, we get

$$\frac{\big(\mathbb{E}\psi'\big)^2}{\mathbb{E}\psi^2} \leq \int_{\overline{\mathbb{R}}} \left(\frac{f'}{f}\right)^2 f \mathrm{d}t = \mathcal{I}(f),$$

with equality if and only if $\psi = f'/f$. This implies that the minimum for ψ in (A) obtains for the choice $\psi = f'/f$ and the minimum risk $r(P, \cdot)$ equals $1 - \mathcal{I}(f)$. ■

Remark. Suppose the sampling model $(Y \,|\, \Theta = \theta) \sim \mathcal{N}(\theta, D)$, where $D > 0$. Proposition 6.2 treats only the special case when $D = 1$. Under the more general sampling model with arbitrary $D > 0$, we obtain for rules of the form $\delta(y) = y + D\psi(y)$, the risk $R(\theta, \delta) = D + D^2\mathbb{E}_\theta\big[\psi^2(y) + 2\psi'(y)\big]$ and the Bayes rule is

$$\delta^*(y) = y + D\psi(y), \qquad \text{where} \quad \psi(y) = \nabla \log(\Phi_{0,D} * P)(y)$$

with (minimum) Bayes risk $r(P, \delta^*) = D - D^2\mathcal{I}(\Phi_{0,D} * P)$.

Finding the least favorable distribution

In view of Proposition 6.2, one can obtain for a given $\epsilon \in (0,1)$ the least favorable prior distribution, P_ϵ, in the set (see PII, above)

$$\mathcal{P}_\epsilon = \big\{P \,:\, P(y) = (1-\epsilon)P_0(y) + \epsilon H(y)\big\}$$

via minimization of $\mathcal{I}(g)$ on the set

$$\begin{aligned}\mathcal{G}_\epsilon &= \big\{g \,:\, g(y) = (\Phi * P)(y), \quad P \in \mathcal{P}_\epsilon\big\} \\ &= \big\{g \,:\, g(y) = (1-\epsilon)(\Phi * P_0)(y) + \epsilon(\Phi * H)(y),\ H \text{ arbitrary}\big\}. \qquad (6.15)\end{aligned}$$

The result to this minimization is the least favorable p.d.f g_ϵ and the corresponding estimating rule is

$$\delta_\epsilon(y) = y + \frac{g'_\epsilon(y)}{g_\epsilon(y)}.$$

Unfortunately, this minimization exercise is extremely difficult to solve; the major obstacle is the convolution $\Phi * H$ in the definition of $\mathcal{G}_\epsilon$.

What is known about g_ϵ and the corresponding rule δ_ϵ? Let us state the main finding first: very little is known how the rule δ_ϵ exactly behaves. In fact, no one has yet succeeded to derive an exact representation of the least favorable distribution g_ϵ and the associated rule δ_ϵ. Among the earliest researchers who worked on the problem are Efron and Morris (1971); they come to the conclusion through (clever) guesswork and some computer exercises that the rule oscillates. Bickel and Collins (1983, see their Case iib and Thm. 3) prove that g_ϵ is symmetric and that it concentrates (granted some regularity conditions) its mass on a countable set of isolated points (i.e. discrete distribution), possibly including $\{\pm\infty\}$. Earlier (but not widely known), Marazzi (1980) has shown the discreteness of the distribution. Moreover, Bickel and Collins (1983, Remark 3.5) show that g_ϵ obtains if we take $l \to \infty$ in

$$\bar{g}_\epsilon(y) = (1-\epsilon)\phi(y) + \frac{\epsilon}{2}\sum_{j=1}^{l} p_j\big[\phi(y-\gamma_j) - \phi(y+\gamma_j)\big], \qquad (6.16)$$

where $0 < \gamma_1 < \gamma_2 < \cdots < \gamma_l < \infty$, $0 < p_j < 1$ for all $j = 1, \ldots, l$ and $\sum_{j \le l} p_j = 1$. Furthermore, the tuples $\{(p_j, \gamma_j) : j = 1, 2, \ldots\}$ satisfy certain side conditions, which can be seen in Bickel and Collins (1983, 17); we are not going into this any further. More importantly, Mallows (1978) *conjectured* that

(i) the $\gamma'_j s$ obey $\gamma_j = hj$ for all $j = 1, 2, \ldots$ and for some real-valued $h > 0$, and

(ii) that the p_j's have a geometric distribution, i.e. that we have $p_j = \lambda(1-\lambda)^{j-1}$ for all $j = 1, 2, \ldots$ and some $\lambda > 0$.

D. Donoho (see Bickel and Collins, 1983; Marazzi, 1985) gave a slight modification of this conjecture in which the γ_j's are translated by some real-valued

$a > 0$. Neither of the two conjectures have – as far as we know – been rigorously proved or rejected, but Marazzi (1985) provides numerical evidence that Mallows' conjecture appears to be wrong. Beyond this point the problem remains open.[3]

Marazzi (1985) managed to approximate g_ϵ numerically using $\bar{g}_\epsilon$ in (6.16). In an elaborate numerical simulation, he minimizes Fisher information over the set of parameters $\{\gamma_j, p_j, j = 1, \ldots, l\}$ for some fixed l, which determine $\bar{g}_\epsilon$. Somewhat surprisingly, he has chosen rather small values for l, e.g., $l = 3$ or $l = 5$. An approximation of the rule using $l = 3$ is shown in Figure 6.1. The oscillating behavior of the rule (and the corresponding risk function) is evident from the display. The graph in Figure 6.1 is a taken from Marazzi (1985) because we have been unable (or did not try hard enough) to reproduce the results of A. Marazzi. The other rule depicted in the figure, denoted by $-\hat{g}'_\epsilon/\hat{g}_\epsilon$, will be discussed later.

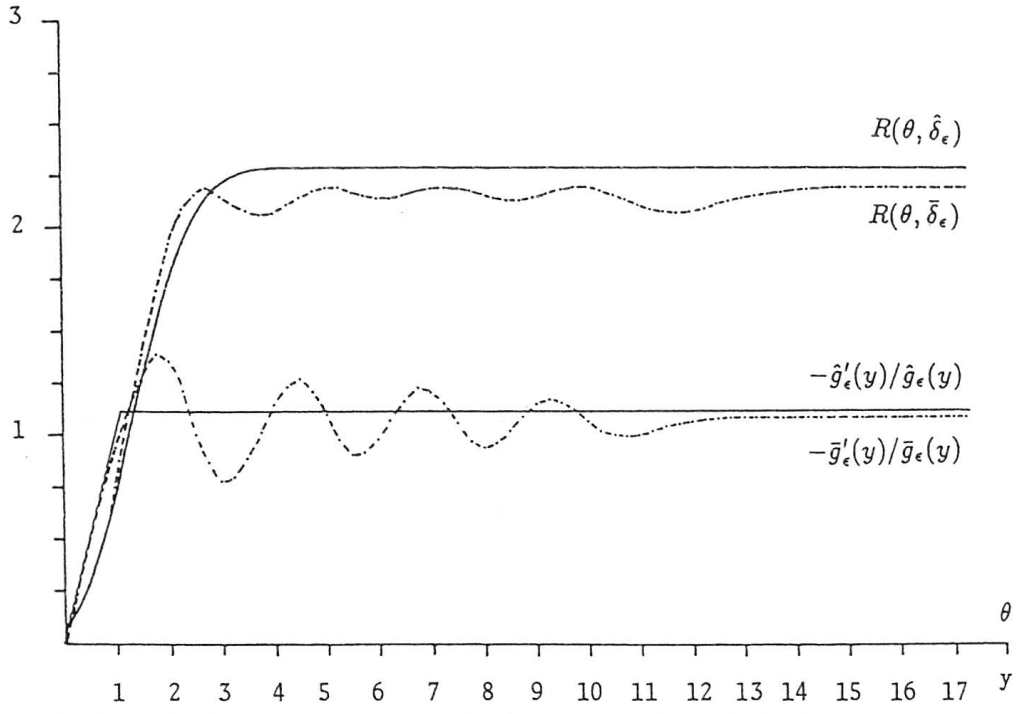


Figure 6.1.: The display shows an approximation to the restricted Bayes rule, $-\bar{g}'_\epsilon/\bar{g}_\epsilon$ (with the oscillating behavior), and the limited translation rule, $-\hat{g}'_\epsilon/\hat{g}_\epsilon$. Above the rules, the corresponding risk functions are displayed. Here, $\epsilon = 0.1$; the figure is taken from Marazzi (1985, 287).

Remark. We have restricted attention to the univariate Gaussian location model. The papers we referred to, namely, Marazzi (1980, 1985, 1990), Bickel (1981, 1983, 1984) and Berger (1982), study generalizations to our model, e.g., n-

[3] There are attempts to study the asymptotic behavior of g_ϵ; see Bickel (1981, 1984). Although this provide further insight, it did not resolve the major obstacles.

variate models, sampling model with a distribution in the exponential family, etc.

6.2.2. Limited translation rule: An approximate compromise

The oscillating nature of δ_ϵ (see Figure 6.1) is a rather unpleasant property. It it thus natural to ask, whether δ_ϵ could be approximated in a way that is consistent with the underlying theory. This is equivalent to the question, whether there exist approximations to the Principles PI and PII. The answer to both questions is yes, which will be seen immediately.

In line with Marazzi (1980), we define the set of functions

$$\mathcal{G}_\epsilon^* = \big\{g \,:\, g(y) = (1-\epsilon)(\Phi * P_0)(y) + \epsilon H(y),\ H \text{ arbitrary}\big\},$$

which is an "approximation" to $\mathcal{G}_\epsilon$ in (6.15) insofar that the convolution $\Phi * H$ is replaced by H. Consider the class of estimating rules $\mathscr{D}$ defined in (6.14). The constitutive element of these rules is the function ψ, which completely determines the rule's behavior. In view of this, we shall express the Bayes risk of such rules in terms of ψ, that is

$$\bar{r}(g, \psi) = 1 + \mathbb{E}\psi^2 + 2\mathbb{E}\psi',$$

where $\mathbb{E}$ is a shorthand notation for the expectation w.r.t. to $g \in \mathcal{G}_\epsilon^*$. We also remark that $\bar{r}(g, \psi)$ coincides with $r(P, \delta)$ for $\delta \in \mathscr{D}$ and $g \equiv \Phi * P$ with $P \in \mathcal{P}_\epsilon$. The *approximation* to principle PII reads as follows.

PII*: Approximate restricted minimax principle

Find the defining ψ_ϵ of the rule $\delta \in \mathscr{D}$ such that

$$\sup_{g \in \mathcal{G}_\epsilon^*} \bar{r}(g, \psi_\epsilon) = \inf_{\psi | \delta \in \mathscr{D}} \sup_{g \in \mathcal{G}_\epsilon^*} \bar{r}(g, \psi).$$

Remark. The formulation of Principle PII* is due to Marazzi (1980); see also Marazzi (1985) and Bickel (1980, 1983). An approximate restricted Bayes principle has been introduced by Berger (1982).

By standard arguments, Principle PII* leads to minimization of $\mathcal{I}(g)$ over the set $\mathcal{G}_\epsilon^*$. Let $g_0 \equiv \Phi * P_0$. If $-\log g_0$ is convex, we can directly apply the results of Huber (1964); hence, the *approximate* least favorable distribution is

$$\bar{g}_\epsilon^0(y) = \begin{cases} (1-\epsilon) g_0(y_0) \exp\big(k(y - y_0)\big) & \text{if } y \le y_0, \\ (1-\epsilon) g_0(y) & \text{if } y_0 < y < y_1, \\ (1-\epsilon) g_0(y_1) \exp\big(-k(y - y_0)\big) & \text{if } y_1 \le y, \end{cases}$$

where k, y_0, and y_1 are numbers that depend on ϵ and have to be compute; see

Marazzi (1980, 15). Of particular importance is the case when P_0 is a Gaussian prior d.f. with zero mean and variance A. Then, the approximate least favorable distribution implies $\psi \equiv \hat{g}'_\epsilon/\hat{g}_\epsilon$ where

$$\psi(y) = \begin{cases} -k & \text{if } y < y_0, \\ By & \text{if } y \in [y_0, y_1], \\ k & \text{if } y_1 < y, \end{cases}$$

where $-y_0 = y_1 = k(A+1) = k/B$ and k is a "tuning constant" that depends on ϵ. The corresponding estimating rule coincides with the *limited translation rule* of Efron and Morris (1971, 809), for the n-variate problem, $i = 1, \ldots, n$,

$$\delta_i^{EM,k}(y_i) = \begin{cases} y_i + k & \text{if} \quad y_i < -k/B, \\ (1-B)y_i & \text{if} \quad |y_i| \le k/B, \\ y_i - k & \text{if} \quad y_i > k/B. \end{cases} \tag{6.17}$$

Efron and Morris (1971) suggested this rule on grounds of ad hoc arguments; it was A. Marazzi and P. Bickel (and B. Levit and J. Berger) who introduced the rigorous mathematical argument behind this estimating rule.

For estimating $\boldsymbol{\theta} \in \mathbb{R}^n$, let the corresponding rule be $\boldsymbol{\delta}^{EM,k} = (\delta_1^{EM,k}, \ldots, \delta_n^{EM,k})^T$, having suppressed the dependency on $\mathbf{y}$. The limited translation rule and its ψ-function are shown in Figure 6.2. The limited translation rule behaves at the center of the data like the Bayes rule but limits the maximum translation from the m.l.e. by k for observations that are far from the prior mean (i.e. inhibits shrinkage). The graph of the function $\psi \equiv \hat{g}'_\epsilon/\hat{g}_\epsilon$ (for the ith coordinate and only positive y) is shown in Figure 6.1. From the visual display it is evident that the limited translation rule "approximates" the oscillating behavior of the optimal rule.

How does $\boldsymbol{\delta}^{EM,k}$ perform in terms of RSL and RRR? Let $\kappa = k/\sqrt{B}$, then by Thm. 3.1 of Efron and Morris (1971), the relative savings loss of the ith component of $\boldsymbol{\delta}^{EM,k}$ is given by

$$\mathrm{RSL}\big(\pi_0, \delta_i^{EM,k}\big) = 2\big(\kappa^2+1\big)\big(1-\Phi(\kappa)\big) - 2\kappa\phi(\kappa), \qquad i = 1, \ldots, n, \tag{6.18}$$

where ϕ and Φ are, respectively, the p.d.f. and c.d.f of the standard Gaussian distribution. Furthermore, by Thm. 3.2 of Efron and Morris (1971), we have

$$\sup_{\theta_i} R\big(\theta_i, \delta_i^{EM,k}\big) = 1 + k^2, \qquad i = 1, \ldots, n,$$

hence,

$$\mathrm{RRR}\big(\pi_0, \delta_i^{EM,k}\big) = \frac{k^2}{B}.$$

It is remarkable that the expressions for both risk measures are so simple. Also observe that the formula for RSL is very similar to the consistency correction

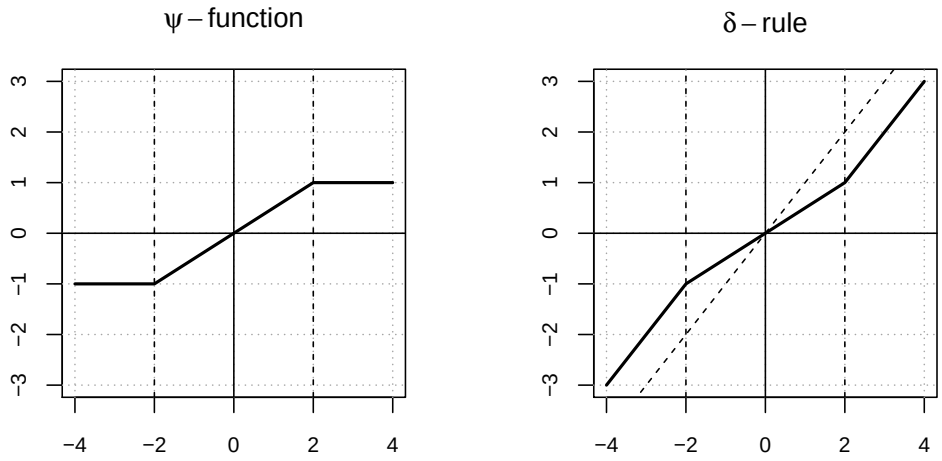


Figure 6.2.: Limited translation rule, δ, and the corresponding ψ-function are shown for the prior variance $A = 1$ and tuning constant $k = 1$. Observe that the rule δ shrinks observations $|y| \leq 2$ towards zero, i.e. "away" from the 45°-line.

term that obtains for the squared Huber ψ-function in robust statistics. This is no surprise since the limited translation rule $\delta_i^{EM,k}$ resembles the Huber ψ-function.

Table 6.1.: Relative savings loss (RSL) of the limited translation rule

RSL	0	0.05	0.1	0.2	0.3	0.4	0.5
$k/\sqrt{B}$	∞	1.464	1.180	0.869	0.763	0.525	0.405

Numerical values are from Efron and Morris (1971, 809).

In return for sacrificing some of the possible Bayes ensemble savings, the limited translation rule protects against large individual risks that are experienced with the Bayes rule. The numbers in Table 6.1 illustrate this behavior. For instance, let $B = 1$. For the choice $k = 1.46$, $\delta_i^{EM,k}$ incurs a relative savings loss of 0.05 in comparison with the Bayes estimator. On the other hand, the relative minimax risk equals $k^2 = 2.14$, which is in sharp contrast to the unbounded RRR of the Bayes rule.

6.3. Empirical Bayes approach and the James–Stein rule

We continue to assume that the hierarchical Gaussian model of Section 6.2 applies, but now the hyperparameter A (i.e., variance specification of prior π_0) is supposed *unknown*. Therefore, the Bayes rule δ_i^* in (6.8) cannot be used as a regular estimator. However, the unknown quantity can be replaced by a suitable estimator (Efron and Morris, 1973b, 117), i.e., we attempt to estimate A [or equivalently, $B = 1/(1+A)$] from the data. The resulting estimator is then called

an *empirical Bayes* estimator. Under the prior distribution in (6.5), $\|\mathbf{y}\|^2 = s$ is a sufficient statistics for B with marginal distribution $S \sim (1/B)\chi^2_n$, where χ^2_n denotes the d.f. of a chi-square distribution with n degrees of freedom. For $n \geq 3$, the following unbiased estimate of B,

$$\mathbb{E}\left[\frac{n-2}{S}\right] = \frac{1}{A+1} \tag{6.19}$$

is available (Efron and Morris, 1973b, 117–118), $\mathbb{E}$ denoting expectation w.r.t. the χ^2_n-distribution. Replacing the unknown $B = 1/(A+1)$ with the estimate $(n-2)/\|\mathbf{y}\|^2$ in (6.8), gives the natural estimator

$$\delta_i^{JS}(\mathbf{y}) = \left(1 - \frac{n-2}{\|\mathbf{y}\|^2}\right) y_i, \quad i = 1, \ldots, n, \tag{6.20}$$

which is the *celebrated James–Stein* (JS) rule; see James and Stein (1961). For $n \leq 2$, δ_i^{JS} is not defined. The estimator for $\boldsymbol{\theta} \in \mathbb{R}^n$ shall be denoted by $\boldsymbol{\delta}^{JS} = (\delta_1^{JS}, \ldots, \delta_n^{JS})^T$, having suppressed the dependency on $\mathbf{y}$. Observe from (6.20) that the rule for the ith coordinate pools together the information from all $i = 1, \ldots, n$ observations. This is a surprising realization since the r.v.'s Y_i were taken to be independent under the model. If $\|\mathbf{y}\|^2 \gg 0$ (i.e., when $\mathbf{y}$ deviates strongly from the a priori mean which is equal to the origin), then the term $(n-2)/\|\mathbf{y}\|^2$ in (6.20) is close to zero, thus $\boldsymbol{\delta}^{JS}$ mimics the m.l.e. On the other hand, if $\|\mathbf{y}\|$ is close to zero, then the (Bayes) prior assumption is compatible with the observed data, and $\boldsymbol{\delta}^{JS}$ shrinks the coordinates towards the origin. At its heart the JS rule is a shrinkage device to reduce variance at the expense of introducing a little bias.

It is clear that not knowing the parameter A beforehand, but having to estimate it, incurs a loss in precision. By Thm. 1 of Efron and Morris (1973b), the Bayes risk of $\boldsymbol{\delta}^{JS}$ versus the Gaussian prior π_0 (see Eq. 6.5) is

$$r\big(\pi_0, \boldsymbol{\delta}^{JS}\big) = 1 - \frac{n-2}{n} B, \tag{6.21}$$

and in terms of relative savings loss,

$$\mathrm{RSL}\big(\pi_0, \boldsymbol{\delta}^{JS}\big) = \frac{2}{n}. \tag{6.22}$$

Note that the Bayes risk is the same for every coordinate $i = 1, \ldots, n$. By comparing (6.9) with (6.21), it is apparent that for moderate to large n, the JS estimator is almost as good as the Bayes rule in terms of Bayes risk. Furthermore, when $n \geq 3$, the Bayes risk of $\boldsymbol{\delta}^{JS}$ is smaller than the corresponding risk of the m.l.e. Moreover, the JS estimator is minimax, implying that $\mathrm{RRR}(\pi_0, \boldsymbol{\delta}^{JS}) = 0$.

The expression for $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{JS})$ is rather involved. However, using Stein's unbiased estimate of risk (Stein's Theorem, see Thm. 6.1) we can obtain an upper bound to $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{JS})$. The proposition we shall give is probably not a new one, but

we are not aware of a paper where the result has already been published.

Proposition 6.3. *Suppose* $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I}_n)$ *for some* $\boldsymbol{\theta} \in \mathbb{R}^n$. *The risk of the rule* $\boldsymbol{\delta}^{JS}$ *satisfies*

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^{JS}) \leq \frac{2}{n} + \frac{(n-2)\|\boldsymbol{\theta}\|^2}{n(n-2) + n\|\boldsymbol{\theta}\|^2}.$$

Proof. Under the hypothesis $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I}_n)$, $Z = \|\mathbf{Y}\|^2$ has a non-central chi-square distribution with n degrees of freedom and non-centrality parameter equal to $\lambda = \|\boldsymbol{\theta}\|^2$. Consider the r.v. $J \sim \text{Poisson}(\lambda/2)$, then conditionally on $J = j$ we have, $(Z \mid J = j) \sim \chi^2_{n+2j} \stackrel{d}{=} \text{Gamma}(n/2 + j, 2)$ and unconditionally Z has the non-central chi-square distr. as required; see (A.39). In terms of notation, $\mathbb{E}_\lambda$ and $\mathbb{E}_\Gamma$ refer to, respectively, the Poisson law with parameter λ and the gamma law. Now, by Stein's theorem, we obtain

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^{JS}) = \frac{1}{n} - \frac{(n-2)^2}{n}\mathbb{E}_{\boldsymbol{\theta}}\frac{1}{\|\mathbf{Y}\|^2}$$

where

$$\mathbb{E}_{\boldsymbol{\theta}}\frac{1}{\|\mathbf{Y}\|^2} = \mathbb{E}_\lambda\left[\mathbb{E}_\Gamma\left(\frac{1}{Z} \mid J = j\right)\right] = \mathbb{E}_\lambda\left[\frac{1}{n-2+2J}\right] \tag{A}$$

and by Jensen's inequality

$$\geq \frac{1}{\mathbb{E}_\lambda[n-2+2J]} = \frac{1}{n-2+\|\boldsymbol{\theta}\|^2},$$

which implies the result. Note: the second equality in (A) obtains since $(1/Z \mid J = j)$ has an inverse gamma distr. with mean given in (A.37). ■

Remark. The assumption underlying Proposition 6.3, that $\|\mathbf{Y}\|^2$ has a non-central chi-square distribution with non-centrality parameter $\|\boldsymbol{\theta}\|^2 > 0$, is crucial as this hypothesis makes it possible that the components θ_i may truly deviate from the origin.

Positive-part rule

The JS rule has the peculiar property in connection with small values of $\|\mathbf{y}\|^2$ that the term $(1 - (n-2)/\|\mathbf{y}\|^2)$ in (6.20) can be negative. If it is negative, the behavior of δ_i^{JS} is weird insofar that it does not shrink toward the origin, but kind of reflects the estimate onto the opposite Cartesian plane (here, opposite means w.r.t. the ordinate). A simple remedy is to replace negative values by zero. The resulting estimator (see Baranchik, 1964, who attributes the proposal to C. Stein) is called *positive-part* JS-estimator as it truncates the shrinkage factor at zero,

$$\delta_i^{c+}(\mathbf{y}) = \left(1 - \frac{c}{\|\mathbf{y}\|^2}\right)^+ y_i, \quad i = 1, \ldots, n,$$

where $(u)^+ = \max\{0, u\}$ and c is a constant. The slight generalization of the positive-part rule, where c can be chosen by the statistician, is discussed in Berger (1982). Let $\boldsymbol{\delta}^{c+} = (\delta_1^{c+}, \ldots, \delta_n^{c+})^T$; we write $\boldsymbol{\delta}^+$ to mean the canonical rule with c equal to $n-2$.

The positive-part JS estimator dominates $\boldsymbol{\delta}^{JS}$ in terms of RSL, which follows from Thm. 1 of Berger (1982); the actual expression of RSL is rather involved and shall not be presented here. What is more important is that the risk improvement over $\boldsymbol{\delta}^{JS}$ can be substantial (Berger, 1982, 362). Also, restricting c to be not larger than $4 = 2(n-2)$ ensures that $\boldsymbol{\delta}^{c+}$ is minimax.

Remarks *(miscellaneous)***.** Neither $\boldsymbol{\delta}^{c+}$ nor $\boldsymbol{\delta}^+$ is admissible since these rules are not analytic; see Brown (1971) who establishes that analyticity is a necessary condition. Efron and Morris (1973b, Sect. 5) approximate $\boldsymbol{\delta}^+$ by a truncated Bayes rule versus a prior h on B (i.e., three-stage model). If h is a point prior that puts all its mass at the location 1, one obtains a positive-part rule. Their calculations show that it is very difficult to dominate $\boldsymbol{\delta}^+$. Shao and Strawderman (1994) found classes of estimators that dominates the positive-part JS rule, but they were unable to demonstrate admissibility of any rule in their class. Their estimators are of far more general form than $\boldsymbol{\delta}^{c+}$ and are considerably smoother in the neighborhood of the origin. Further, Shao and Strawderman (1994) indicate that a candidate rule has to "wiggle" sufficiently about the estimator to be improved. Yet, the longstanding problem of finding admissible estimators that dominate the positive-part JS rule seems (as far as we know) to be open; recently, Maruyama and Strawderman (2005) gave necessary conditions for domination of the positive-part JS rule. *Is the availability of an admissible rule that dominates the positive-part JS rule a showstopper for the further course of discussion?* No. Although the positive-part JS rule is not admissible, Bock (1988) showed that there does not exist a rule whose unbiased estimator of risk is *everywhere less* than the positive-part JS rule. Moreover, Shao and Strawderman (1994) indicate that (at least for their class of rules) the maximal improvement over the positive-part JS rule is quite small. In the light of this finding, we feel comfortable to stick with positive-part JS-type rules.

6.3.1. Component risk

We have seen that the JS rule does well in terms of RSL (ensemble risk). This formidable achievement is because it concentrates all its attention to average (or total) squared error loss. Clearly, this behavior is key to guarantee a reduction in ensemble risk, but on the other hand it is also the reason why the rule may do very poorly in estimating a single component θ_i. This problem is particularly emphasized for θ_i's with unusually large values. The prime reason for such poor behavior is, obviously, that the JS rule shrinks atypically large θ_i's too much;

hence, it increases the component-wise risk for such θ_i's.

For the JS rule, Baranchik (1964, Eq. 2.4.3) proved an expression for the *component-wise risk*. Efron and Morris (1972, 132) refer to Branchik's result and show that the component-wise risk is maximized for fixed $\|\boldsymbol{\theta}\|^2$ at $\theta_1^2 = \|\boldsymbol{\theta}\|^2$. In words, the maximum occurs if one coordinate equals $\|\boldsymbol{\theta}\|^2$ while the remaining $n-1$ coordinates are zero. The result is summarized in Proposition 6.4; the proof is ours. We have devised our own proof because it is considerably simpler than the one given in Baranchik (1964, 16). Our point of attack relies on three observations: (i) the maximum risk occurs, for a given $\boldsymbol{\theta}$, at $\theta_1 = \|\boldsymbol{\theta}\|^2$ [see above]; hence, w.l.o.g. we assume that $Y_1 \sim \mathcal{N}(\|\boldsymbol{\theta}\|, 1)$, which implies that Y_1^2 has a non-central chi-square distribution with non-centrality parameter $\|\boldsymbol{\theta}\|^2$. The remaining $n-1$ coordinates have a central chi-square distr.; (ii) Stein's theorem simplifies the computation of the risk, but the major ingredient is (iii) a theorem of Lukacs (1955), see Thm. A.8, on independence of (certain functions of) two gamma r.v.'s with the same scale parameter. This independence result enables us to split up the expected value of the product of two independent r.v.'s into the product of two expectations.

Proposition 6.4. *Suppose* $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I}_n)$ *for some* $\boldsymbol{\theta} \in \mathbb{R}^n$. *For the rule* $\delta_i^{JS}(\mathbf{y})$, $i = 1, \ldots, n$, *we have*

$$R(\theta_i, \delta_i^{JS}) = 1 + (n-2)\mathbb{E}_\lambda[\Lambda], \tag{6.23}$$

where

$$\Lambda = \frac{2 - n + 2nJ}{(n - 2 + 2J)(n + 2J)}$$

and J *is a Poisson r.v. with mean* $\lambda = \|\boldsymbol{\theta}\|^2/2$.

Proof. We may write $\delta_i^{JS}(\mathbf{y}) = Y_i - g_i(\mathbf{y})$, where $g_i(\mathbf{y}) = y_i(n-2)/\|\mathbf{y}\|^2$, for all $i = 1, \ldots, n$. Since

$$g_i'(\mathbf{y}) = \frac{n-2}{\|\mathbf{y}\|^4}\big(\|\mathbf{y}\|^2 - 2y_i^2\big)$$

we obtain by application of Stein's theorem

$$\begin{aligned} R(\theta_i, \delta_i^{JS}) &= 1 + (n-2)^2\mathbb{E}_\theta \frac{Y_i^2}{\|\mathbf{Y}\|^4} - 2(n-2)\mathbb{E}_\theta\left[\frac{1}{\|\mathbf{Y}\|^4}\big(\|\mathbf{Y}\|^2 - 2Y_i^2\big)\right] \\ &= 1 + (n^2-4)\mathbb{E}_\theta \frac{Y_i^2}{\|\mathbf{Y}\|^4} - 2(n-2)\mathbb{E}_\theta \frac{1}{\|\mathbf{Y}\|^2}. \end{aligned} \tag{A}$$

To attack (A), we need some preparation. W.l.o.g. we shall assume (by the symmetry of δ_i^{JS} w.r.t. to the coordinates $i = 1, \ldots, n$) that Y_1^2 has a non-central chi-square distr. with non-centrality parameter $\lambda = \|\boldsymbol{\theta}\|^2/2$, formally $Y_1^2 \sim \chi_1^2(\lambda)$; see Section A.2. Conditionally, we have, see (A.39),

$$Z_{1+2j} := (Y_1^2 \mid J = j) \sim \chi_{1+2j}^2 \stackrel{d}{=} \mathrm{Gamma}(1/2 + j, 2), \qquad j = 1, 2, \ldots,$$

where $J \sim \mathrm{Poisson}(\lambda/2)$, the expectation of which is denoted by $\mathbb{E}_\lambda$. The re-

maining $n-1$ coordinates satisfy $Y_j^2 \overset{\text{ind.}}{\sim} \chi_1^2$, $j=2,\ldots,n$. Put $S=\sum_{j=2}^n Y_j^2$ and observe that $S \sim \chi_{n-1}^2 \overset{d}{=} \mathrm{Gamma}(n/2-1/2,2)$. We denote expectation w.r.t. a gamma law by $\mathbb{E}_\Gamma$. Hence, we may write

$$\frac{Y_i^2}{\|\mathbf{Y}\|^4} = \frac{1}{Y_1^2+S}\frac{Y_1^2}{Y_1^2+S}$$

and applying Lukacs' theorem, see Thm. A.8 (which applies since Y_1^2 *conditionally* on $J=j$ is the gamma r.v. Z_{1+2j}), and using Thm. A.9 yields

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{Y_1^2+S}\frac{Y_1^2}{Y_1^2+S}\right] &= \mathbb{E}_\lambda\left[\mathbb{E}_\Gamma\left(\frac{1}{Z_{1+2j}+S}\frac{Z_{1+2j}}{Z_{1+2j}+S} \mid J=j\right)\right] \\ &= \mathbb{E}_\lambda\left[\mathbb{E}_\Gamma\left(\frac{1}{Z_{1+2j}+S} \mid J=j\right)\mathbb{E}_\Gamma\left(\frac{Z_{1+2j}}{Z_{1+2j}+S} \mid J=j\right)\right]. \quad \text{(B)}\end{aligned}$$

By application of (A.35), we have $(Z_{1+2j}+S) \sim \mathrm{Gamma}(n/2+j,2)$; hence, $1/(Z_{1+2j}+S)$ has an inverse gamma distr. with mean, see (A.37),

$$\mathbb{E}_\Gamma\left[\frac{1}{Z_{1+2j}+S} \mid J=j\right] = \frac{1}{n-2+2j}. \quad \text{(C)}$$

Also, by (A.36) and (A.32), we have

$$\mathbb{E}_\Gamma\left[\frac{Z_{1+2j}}{Z_{1+2j}+S} \mid J=j\right] = \frac{1+2j}{n+2j}. \quad \text{(D)}$$

Substituting (C) and (D) into (B) gives

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{Y_i^2}{\|\mathbf{Y}\|^4}\right] = \mathbb{E}_\lambda\left[\frac{1+2J}{(n+2J)(n-2+2J)}\right]$$

and since (C) is seen to be equal to $\mathbb{E}_{\boldsymbol{\theta}}[1/\|\mathbf{Y}\|^2]$, we have on collecting terms,

$$\begin{aligned}&1+(n^2-4)\mathbb{E}_\lambda\left[\frac{1+2J}{(n+2J)(n-2+2J)}\right] - 2(n-2)\mathbb{E}_\lambda\left[\frac{1}{n-2+2J}\right] \\ &\quad = 1+(n-2)\mathbb{E}_\lambda\left[\frac{2-n+2nJ}{(n-2+2J)(n+2J)}\right],\end{aligned}$$

and the assertion follows. ■

What else can we say about the maximum component risk? Let us regard Λ in Proposition 6.4 as a function of J and n, i.e. $\Lambda(J,n)$. A Taylor series expansion of $\Lambda(J,n)$ around $n=0$ yields

$$\frac{1}{J(2J-2)} + \frac{(2J-1)(J^2-J-1)n}{4(J-1)^2J^2} + \ldots$$

which is unfortunate since the r.v. J still appears in the denominator. However, for large n, we may use the first two terms of the Laurent series approximation[4]

[4] Wolfram Alpha came to our help...

(i.e., the series expansion at $n = \infty$),

$$\Lambda(J, n) = \frac{2J - 1}{n} - \frac{8(J^2 - J)}{n^2} + \mathcal{O}\left(\frac{1}{n^3}\right),$$

which then yields (since the Poisson r.v. J has mean and variance equal to $\|\boldsymbol{\theta}\|^2/2$),

$$\mathbb{E}_\lambda\big[\Lambda(J, n)\big] \approx \frac{\|\boldsymbol{\theta}\|^2 - 1}{n} - \frac{2\|\boldsymbol{\theta}\|^4}{n^2}. \tag{6.24}$$

It is easy to see that the expression on the r.h.s. of (6.24) is maximized when $\|\boldsymbol{\theta}\|^2$ equals $n/4$ (for large n). Beyond this rather crude result, the approximation is too weak for uncovering further details.

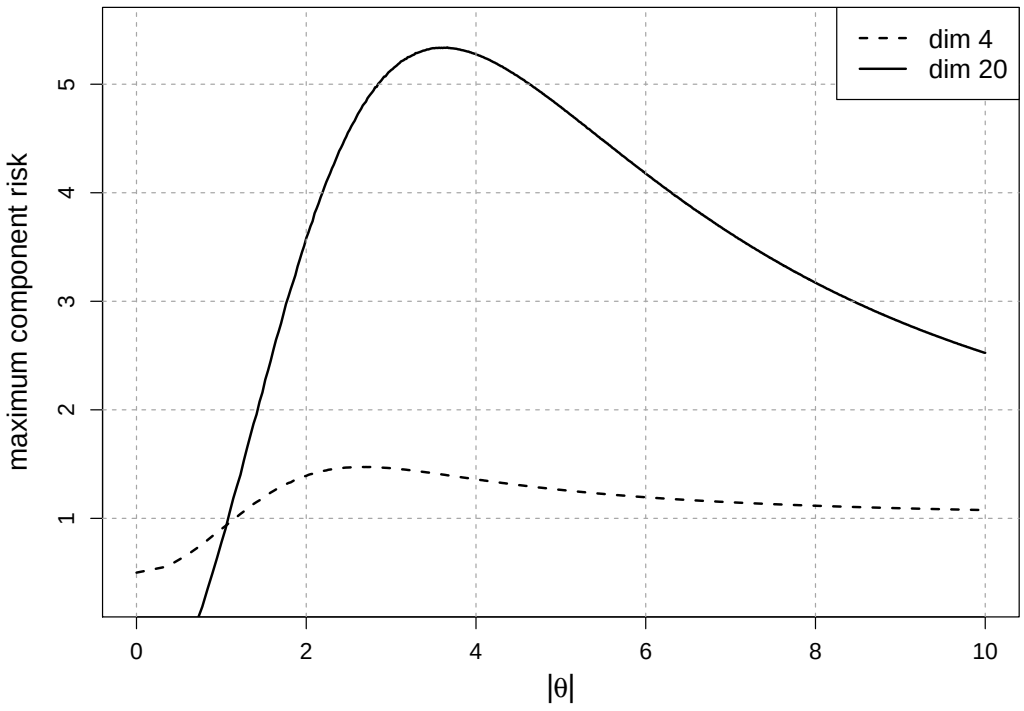


Figure 6.3.: Maximum component risk of the James–Stein estimator as a function of $\|\boldsymbol{\theta}\|$, shown for two dimensions $n = 4$ and $n = 20$. [note: $|\theta|$ in the labeling of the abscissa stands for the Euclidean norm $\|\boldsymbol{\theta}\|$]

In order to learn more about the maximum component risk, we resort to numerical computations. The behavior of the maximum component risk in (6.23) as a function of $\|\boldsymbol{\theta}\|$ is illustrated for two problems of different size (i.e., $n = 4$ and $n = 20$) in Figure 6.3. The numerical values in the display are computed by means of Monte Carlo integration of (6.23), using 10^5 replications for each value of $\|\boldsymbol{\theta}\|$; the $\|\boldsymbol{\theta}\|$-values are taken equidistant in the interval $[0, 10]$. The maximum component risk for the JS rule as function of $\|\boldsymbol{\theta}\|$ in Figure 6.3 increases from $n/2$ (not visible for $n = 20$ because of a numerical artifcat; the fact that the curve attains the value $n/2$ at 0 is discussed in Efron and Morris, 1972,

132) to a maximum and thereafter decreases asymptotically to 1. Let us focus on the case $n = 4$. The maximum risk of (roughly) 1.5 is located at approximately $\|\boldsymbol{\theta}\| = 2.25$; hence, the maximum that can possibly be experienced is 1.5 times larger than the minimax risk. The situation is even worse for larger dimensional problems as the case $n = 20$ illustrates. In general, for large n, the maximum risk for an individual coordinate over all $\boldsymbol{\theta}$ can be as large as $n/4$ as Efron and Morris (1972, 130) have worked out through numerical computations. In general, the JS rule tends to *grossly misestimate* some individual components.

6.3.2. "Estimated" limited translation rule

We have seen that the JS rule and its positive-part analog concentrate all attention on average (or total) squared error loss without any concerns for the effects on individual coordinates. Therefore, these rules may do very poorly in estimating those θ_i with unusually large values.

It was shown in the Bayes case that compromise methods are available which capture most of the ensemble savings of the Bayes rule while protecting against unusual individual components. The best known such compromise estimator is the *limited translation rule* $\delta_i^{EM,k}$ of Efron and Morris (1971) [see Formula (6.17)], where $k \geq 0$ denotes the "tuning" constant that determines the amount of translation from the m.l.e. This rule seeks a trade-off between the Bayes rule and the m.l.e. Clearly, rule $\delta_i^{EM,k}$ cannot be applied in the current context since it depends on the unknown parameter B. However, Efron and Morris (1972, 132) show that we may consider *estimating* the limited translation rule (pursuing the analogy of having obtained the JS rule by estimating the unknown B in the Bayes rule). The resulting rule is

$$\hat{\delta}_i^{EM,k}(\mathbf{y}) = \left(1 - \frac{n-2}{\|\mathbf{y}\|^2} \min\left\{1, \frac{k\|\mathbf{y}\|^2}{y_i\sqrt{n-2}}\right\}\right) y_i, \qquad i = 1, \ldots, n, \tag{6.25}$$

having replaced the unknown B by its moment-based estimator $(n-2)/\|\mathbf{y}\|^2$. Since $y_i/\|\mathbf{y}\| \leq 1$, for all $i = 1, \ldots, n$, it suffices to consider k in the interval $[0, \sqrt{n-2}]$. It is easy to see that $\hat{\delta}_i^{EM,k}$ is the compromise between the JS rule and the m.l.e. Note that we write $\hat{\delta}_i^{EM,k}$ to distinguish it from the (Bayes) limited translation rule; for the n-variate problem, let $\hat{\boldsymbol{\delta}}^{EM,k} = (\hat{\delta}_1^{EM,k}, \ldots, \hat{\delta}_n^{EM,k})^T$, having suppressed the dependency on $\mathbf{y}$.

How does rule $\hat{\delta}_i^{EM,k}$ perform in terms of risk? Efron and Morris (1972) derived an expression of the Bayes risk for this rule; see Theorem 6.2, below. Their result, however, requires numerical integration. We give another result that is easier to work with (see Corollary 6.1). For the moment, we shall study the numerical performance of rule $\hat{\delta}_i^{EM,k}$ and present the technical details later. Table 6.2 shows a set of values for the "tuning constant" k in order to attain a given

relative efficiency versus the JS rule. It is evident that the values for k depend on the problem size n (this and further peculiarities of the tabulated numbers, e.g. the case $n = \infty$, will be discussed later). Suppose that $n = 3$. From Table 6.2, we see that rule $\hat{\delta}_i^{EM,k}$ with $k = 0.785$ achieves 99% relative efficiency versus the JS rule.

So far we have discussed the behavior in terms of Bayes risk (or RSL). It is equally important to consider also the maximum possible harm that could be encountered using rule $\hat{\delta}_i^{EM,k}$ instead of the minimax rule δ_i^0 (relative to the potential Bayes risk improvement over δ_i^0, see RRR-measure). In the Bayes case (with A known; see Section 6.2.2), the maximum component risk was shown to be no greater than $1 + k^2/B$. We would therefore choose k small enough to keep the maximum component risk at an acceptable value. On the other hand, larger values of k increase the savings over the m.l.e.

In case of the JS rule, the situation is very similar. Although a closed-form expression for the maximum component risk of the JS rule may be nice to have (but difficult to compute), it is not not really needed as Efron and Morris (1972, 135) have suggest a rule of thumb: the maximum component risk is $\leq 1 + D_k^2(\omega)$, where D_k is defined in (6.30), reported in Table 6.2, and will be discussed later. We shall illustrate the implication of the rule of thumb by way of example. Consider a problem of size $n = 10$. Suppose we are willing to sacrifice 10% of the savings vs. the JS rule in order to protect against extreme component risks. From Table 6.2 we read off $D_k = 0.974$; hence, the maximum component risk is at most $1 + D_k^2 = 1.95$ in contrast to 2.93 which results for the JS rule [the max. component risk of the JS rule is from Table 1 in Efron and Morris (1972)]. In other words, the maximum component risk of $\hat{\delta}_i^{EM,k}$ is 50.2% smaller than the one of the JS rule. In problems of size $n > 30$, the rule of thumb suggested by Efron and Morris (1972) is very accurate which can be inferred from the values of their Table 1. In very small problems (i.e. $n < 10$), the resulting numbers are rather crude approximations.

Overall, the behavior of rule $\hat{\delta}_i^{EM,k}$ is striking since one can achieve an acceptable maximum component risk while retaining most of the savings over the m.l.e.

Risk of the estimated limited translation rule

We shall consider estimating rules of more general form than (6.25) which are defined as

$$\delta_i^h(\mathbf{y}) = \left(1 - \frac{n-2}{\|\mathbf{y}\|^2} h\left((n-2)\frac{y_i}{\|\mathbf{y}\|^2}\right)\right) y_i, \quad i = 1, \ldots, n, \tag{6.26}$$

where $h : \mathbb{R} \to \mathbb{R}$ is a given function, and is called "relevance" function by Efron and Morris (1972, 132). It measures the relevance of the ensemble to the

Table 6.2.: Values of the maximal translation from the m.l.e. (i.e., "tuning" constant k) in order to attain a predetermined level of efficiency w.r.t. the JS rule (computed for several dimensions n)

efficiency	$n=3$	$n=4$	$n=5$	$n=10$	$n=50$	$n=100$	$n=171$	$n=\infty$
99%	0.785	1.042	1.206	1.577	1.943	1.995	2.016	2.048
95%	0.632	0.823	0.939	1.184	1.405	1.435	1.447	1.465
90%	0.536	0.692	0.784	0.974	1.137	1.159	1.167	1.180
85%	0.469	0.602	0.679	0.836	0.968	0.985	0.992	1.002
80%	0.415	0.531	0.597	0.731	0.841	0.855	0.861	0.869

Note: The data are computed using Formula (6.30) and Corollary 6.1.

component risk and adapts or influences the component accordingly. For such generalized rules, Efron and Morris (1972) established the following result.

Theorem 6.2 (Efron and Morris, 1972)**.** *Consider rule δ_i^h defined in (6.26), then for all $i=1,\dots,n$, the component-wise Bayes risk versus prior π_0 is*

$$\begin{aligned} r_i(\pi_0,\delta_i^h) &= \big[1-\omega_n(h)\big]r_i(\pi_0,\delta_i^0)+\omega_n(h)r_i(\pi_0,\delta_i^{JS}) \\ &= 1-\omega_n(h)\frac{n-2}{n}B, \end{aligned} \tag{6.27}$$

where

$$\omega_n(h) = 1-\mathbb{E}_W\Big[\big\{1-h\big((n-2)W\big)\big\}^2\Big],$$

with

$$W\sim \mathrm{Beta}\left(\frac{3}{2},\,\frac{n-1}{2}\right).$$

Proof. See Thm. 4.1 in Efron and Morris (1972); the proof is rather technical and is not given here. ■

Remarks. (i) From (6.27) it is seen that the Bayes risk of δ_i^h can be represented as a *mixture of the risks* of the m.l.e. and the JS rule. More importantly, the mixing constants ω_n are independent of B and are identical for all coordinates.

(ii) Observe from (6.27) that the risk is expressed in terms of the expectation w.r.t. the $\mathrm{Beta}(\frac{3}{2},\,\frac{p-1}{2})$ distribution. Therefore, numerical computations are rather involved. We have worked out an explicit formula for the risk of the limited translation rule which simplifies computation considerably; see Corollary 6.1, below.

Theorem 6.2 provides a useful device for computing the Bayes risk of generalized rules defined in (6.26). In what follows, we give a *complementary* result that builds on Stein's lemma; see Theorem 6.3. Depending on the particular

form of the relevance function h in (6.26), either Theorem 6.2 or 6.3 is easier to work with.

Theorem 6.3. *Consider the hypotheses of Thm. 6.2. Let $h : \mathbb{R} \to \mathbb{R}$ be a known relevance function, which is supposed to be a.e. differentiable with derivative h' such that $\mathbb{E}_{\boldsymbol{\theta}}|h'(\mathbf{Y})| < \infty$. For rules of the form*

$$\delta_i^h(\mathbf{y}) = \left(1 + \frac{n-2}{\|\mathbf{y}\|^2} h\left((n-2)\frac{y_i^2}{\|\mathbf{y}\|^2}\right)\right) y_i, \qquad i = 1, \dots, n,$$

the component-wise Bayes risk can be written as

$$r(\pi_0, \delta_i^h) = 1 + B\big(p e_{U^*} - e_U\big),$$

where

$$e_{U^*} = \mathbb{E}_{U^*}\left[(n-2)h^2\big(U^*\big) + 2h\big(U^*\big) + \frac{4(n^2+4)}{3} h'\big(U^*\big)\right]$$

and

$$e_U = \mathbb{E}_U\Big[2(n-2)h'\big(U\big) + h\big(U\big)\Big],$$

with

$$\frac{U}{n-2} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{n-1}{2}\right) \quad \textit{and} \quad \frac{U^*}{n-2} \sim \mathrm{Beta}\left(\frac{3}{2}, \frac{n-1}{2}\right).$$

Proof. Observe that δ_i^h can be written as $Y_i + g(\mathbf{Y})$ for all $i = 1, \dots, n$, where

$$g(\mathbf{Y}) = -(n-2)h\big((n-2)U\big)\frac{Y_i}{\|\mathbf{Y}\|^2}$$

and r.v. $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$. By Stein's lemma, we have

$$R(\theta_i, \delta_i^h) = \mathbb{E}_{\boldsymbol{\theta}}\big\|Y_i + g(\mathbf{Y}) - \theta_i\big\|^2 = 1 + \mathbb{E}_{\boldsymbol{\theta}}\big[g^2(\mathbf{Y}) + 2\nabla_i g(\mathbf{Y})\big], \quad i = 1, \dots, g,$$

where ∇_i denotes the partial derivative w.r.t. Y_i. Hence,

$$r_i(\pi_0, \delta_i^h) = 1 + \mathbb{E}_{\pi_0}\mathbb{E}_{\boldsymbol{\theta}}\big[g^2(\mathbf{Y}) + 2\nabla_i g(\mathbf{Y})\big], \quad i = 1, \dots, n. \tag{A}$$

Let $U = Y_i^2/\|\mathbf{Y}\|^2$ and $V = \|\mathbf{Y}\|^2 B/2$ (regarded as functions of $\mathbf{Y}$). Observe that

$$\mathbb{E}_{\pi_0}\mathbb{E}_{\boldsymbol{\theta}}\big[g^2(\mathbf{Y})\big] = \frac{B(n-2)^2}{2}\mathbb{E}_U\mathbb{E}_V\left[h^2\big((n-2)U\big)\frac{U}{V}\right]$$

and under the model, we know that

$$U \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{n-1}{2}\right) \quad \text{and} \quad V \sim \mathrm{Gamma}\left(\frac{n}{2}, 1\right) \tag{B}$$

with U and V being independent (see Efron and Morris, 1972, Eq.'s A.3 and A.4). Hence,

$$\mathbb{E}_{\pi_0}\mathbb{E}_{\boldsymbol{\theta}}\big[g^2(\mathbf{Y})\big] = \frac{B(n-2)^2}{2}\mathbb{E}_U\mathbb{E}_V\left[h^2\big((n-2)U\big)\frac{U}{V}\right]$$

$$= \frac{B(n-2)^2}{2}\mathbb{E}_U\Big\{U\,\mathbb{E}_V\Big[\frac{1}{V}h^2\big((n-2)U\big)\mid U\Big]\Big\}$$

and since $1/V$ has an inverse Gamma distr. with mean $2/(n-2)$, we have

$$\begin{aligned}&= B(n-2)\mathbb{E}_U\Big[Uh^2\big((n-2)U\big)\Big]\\ &= Bn(n-2)\mathbb{E}_{U^*}\Big[h^2\big((n-2)U^*\big)\Big], \quad \text{with} \quad U^*\sim \mathrm{Beta}\Big(\frac{3}{2},\frac{n-1}{2}\Big). \qquad \text{(C)}\end{aligned}$$

The last equality follows from (B) and because the $\mathrm{Beta}(\alpha,\beta)$ distr. with p.d.f., see (A.31),

$$f(u\mid\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)}u^{\alpha-1}\big(1-u\big)^{\beta-1},$$

where $\mathrm{B}(\alpha,\beta)$ is the beta function, has the property

$$f(u\mid\alpha+1,\beta) = u\,\frac{\mathrm{B}(\alpha,\beta)}{\mathrm{B}(\alpha+1,\beta)}f(u\mid\alpha,\beta) = u\,f(u\mid\alpha,\beta)\frac{\alpha+\beta}{\alpha}, \qquad \text{(*)}$$

where we have used identities (A.28) and (A.29). Next, we have

$$\begin{aligned}&\mathbb{E}_{\pi_0}\mathbb{E}_{\boldsymbol{\theta}}\big[\nabla_i g(\mathbf{Y})\big]\\ &\quad= -\frac{B(n-2)}{2}\mathbb{E}_{\pi_0}\mathbb{E}_{\boldsymbol{\theta}}\Big[(n-2)h'\big((n-2)U\big)U'\frac{Y_i}{V}+h\big((n-2)U\big)\Big(\frac{1}{V}-\frac{Y_i}{V^2}V'\Big)\Big],\end{aligned}$$

which, using

$$V' = \nabla_i Y = BY_i \qquad \text{and} \qquad U' = \nabla_i U = \frac{B}{2}\Big(\frac{2Y_i}{V}-\frac{Y_i^2}{V^2}V'\Big) = \frac{B}{V}\Big(Y_i-\frac{BY_i^3}{2V}\Big),$$

yields

$$\begin{aligned}\mathbb{E}_{\pi_0}\mathbb{E}_{\boldsymbol{\theta}}\big[\nabla_i g(\mathbf{Y})\big] &= -B(n-2)\mathbb{E}_U\Big\{\mathbb{E}_V\Big[\frac{1}{V}\Big((n-2)h'\big((n-2)U\big)(1-U^2)\\ &\qquad+\frac{1}{2}h\big((n-2)U\big)(1-U)\Big)\mid U\Big]\Big\}\\ &= -2B\mathbb{E}_U\Big[(n-2)h'\big((n-2)U\big)(1-U^2)\\ &\qquad+\frac{1}{2}h\big((n-2)U\big)(1-U)\Big]\\ &= -B\mathbb{E}_U\Big[2(n-2)h'\big((n-2)U\big)+h\big((n-2)U\big)\Big]\\ &\qquad+2B(n-2)\mathbb{E}_U\Big[U^2h'\big((n-2)U\big)\Big]+B\mathbb{E}_U\Big[Uh\big((n-2)U\big)\Big]. \qquad \text{(D)}\end{aligned}$$

Now, we treat the summands in (D) separately. Using the technique in (*), we obtain

$$B\mathbb{E}_U\Big[Uh\big((n-2)U\big)\Big] = nB\mathbb{E}_{U^*}\Big[h\big((n-2)U^*\big)\Big], \qquad \text{(E)}$$

with r.v. U^* defined in (C). By the same technique again, and applying identities

in (A.28) and (A.29) recursively, we get

$$2B(n-2)\mathbb{E}_U\Big[U^2h'\big((n-2)U\big)\Big] = \frac{2Bn(n^2+4)}{3}\mathbb{E}_{U^*}\Big[h'\big((n-2)U^*\big)\Big], \qquad \textbf{(F)}$$

with r.v. U^* given in (C).
Collecting the remaining terms in (D), together with (C), (E), and (F), we have in (A)

$$r(\pi_0, \delta_i^h) = 1 + e_{U^*} - e_U,$$

where

$$e_{U^*} = Bn\mathbb{E}_{U^*}\left[(n-2)h^2\big((n-2)U^*\big) + 2h\big((n-2)U^*\big) + \frac{4(n^2+4)}{3}h'\big((n-2)U^*\big)\right]$$

and

$$e_U = B\mathbb{E}_U\Big[2(n-2)h'\big((n-2)U\big) + h\big((n-2)U\big)\Big]$$

the r.v.'s U and U^* are defined, respectively, in (B) and (C). ■

Both Theorems 6.2 and 6.3 are useful tools for evaluating the Bayes risk of a generalized rule defined by some relevance function h. For the limited translation rule, $\hat\delta_i^{EM,k}$, i.e. the rule with relevance function $h_k(u) = \min\{1,\, k/\sqrt{u}\}$, Theorem 6.2 proves easier to work with. The following corollary of Theorem 6.2 establishes an explicit formula for the Bayes risk of rule $\hat\delta_i^{EM,k}$ defined in (6.25).

Corollary 6.1. *Consider the hypotheses of Theorem 6.2. Fix k such that $0 \le k \le \sqrt{n-2}$. The Bayes risk of rule $\hat\delta_i^{EM,k}$ can be written as*

$$r\big(\pi_0, \hat\delta_i^{EM,k}\big) = 1 - \omega_n(k)\frac{n-2}{n}B, \qquad i = 1, \dots, n, \tag{6.28}$$

with, letting $K = k/\sqrt{n-2}$,

$$\begin{aligned}\omega_n(k) = {} & I_{K^2}\left(\frac{3}{2},\, \frac{n-1}{2}\right) - nK^2\left[1 - I_{K^2}\left(\frac{1}{2},\, \frac{n-1}{2}\right)\right] \\ & + 2KC_n\big(1-K^2\big)^{(n-1)/2},\end{aligned} \tag{6.29}$$

where $I_x(a,b)$ denotes the regularized incomplete beta function and

$$C_n = \frac{2\Gamma\big(n/2+1\big)}{\sqrt{\pi_0}\Gamma\big(n/2+1/2\big)}.$$

Proof. Let $f_{[\alpha,\beta]}$ denote the p.d.f. of the $\mathrm{Beta}(\alpha,\beta)$ distr. with c.d.f. $F_{[\alpha,\beta]}(u) = I_u(\alpha,\beta)$; see Section A.2. The assertion of the corollary follows by evaluation of (see Thm. 6.2)

$$\begin{aligned}\omega_n(k) &= 1 - \mathbb{E}_{[\alpha,\beta]}\Big\{\big[1 - h_k\big((n-2)U\big)\big]^2\Big\} \\ &= 2\mathbb{E}_{[\alpha,\beta]}\Big\{h_k\big((n-2)U\big)\Big\} - \mathbb{E}_{[\alpha,\beta]}\Big\{h_k^2\big((n-2)U\big)\Big\},\end{aligned} \qquad \textbf{(A)}$$

where $U \sim \mathrm{Beta}\big(\frac{3}{2}, \frac{n-1}{2}\big)$ and $\mathbb{E}_{[\alpha,\beta]}$ denotes expectation (here, with $\alpha = \frac{3}{2}$ and $\beta = \frac{n-1}{2}$). For the p.d.f. $f_{[\alpha,\beta]}$, it is easy to see using identities (A.28) and (A.29) that

$$f_{[\alpha-1,\beta]}(u)\frac{\alpha-1+\beta}{\alpha-1} = \frac{1}{u} f_{[\alpha,\beta]}(u). \tag{*}$$

Put $K = k/\sqrt{n-2}$. For the third term on the r.h.s. of (A), we have

$$\begin{aligned}
\mathbb{E}_{[\alpha,\beta]}\Big[h_k^2\big((n-2)U\big)\Big] &= \mathbb{E}_{[\alpha,\beta]}\left[\min\left\{1,\ \frac{k^2}{(n-2)U}\right\}\right] \\
&= K^2 \int_{K^2}^1 \frac{1}{u} f_{[\alpha,\beta]}(u)\mathrm{d}u + \int_0^{K^2} f_{[\alpha,\beta]}(u)\mathrm{d}u \\
&= \frac{K^2(\alpha-1+\beta)}{\alpha-1}\int_{K^2}^1 f_{[\alpha-1,\beta]}(u)\mathrm{d}u + F_{[\alpha,\beta]}\big(K^2\big) \\
&= \frac{K^2(\alpha-1+\beta)}{\alpha-1}\Big(1 - F_{[\alpha-1,\beta]}\big(K^2\big)\Big) + F_{[\alpha,\beta]}\big(K^2\big)
\end{aligned}$$

which, for $\alpha = \frac{3}{2}$ and $\beta = \frac{n-1}{2}$, yields

$$= nK^2\Big(1 - F_{\left[\frac{1}{2}, \frac{n-1}{2}\right]}\big(K^2\big)\Big) + F_{\left[\frac{3}{2}, \frac{n-1}{2}\right]}\big(K^2\big). \tag{B}$$

Next, consider the second term on the r.h.s. of (A). By a similar technique to the one in (*), we get

$$\begin{aligned}
\frac{1}{\sqrt{u}} f_{[\alpha,\beta]}(u) &= \frac{\mathrm{B}(\alpha-\frac{1}{2},\beta)}{\mathrm{B}(\alpha,\beta)}\, f_{\left[\alpha-\frac{1}{2},\beta\right]}(u) = \frac{\Gamma(\alpha-\frac{1}{2})\Gamma(\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha-\frac{1}{2}+\beta)\Gamma(\alpha)\Gamma(\beta)}\, f_{\left[\alpha-\frac{1}{2},\beta\right]}(u) \\
&= \frac{\Gamma\big(\alpha-\frac{1}{2}\big)\Gamma\big(\alpha+\beta\big)}{\Gamma\big(\alpha-\frac{1}{2}+\beta\big)\Gamma\big(\alpha\big)}\, f_{\left[\alpha-\frac{1}{2},\beta\right]}(u),
\end{aligned} \tag{**}$$

hence,

$$\begin{aligned}
\mathbb{E}_{[\alpha,\beta]}\Big[h_k\big((n-2)U\big)\Big] &= \mathbb{E}_{[\alpha,\beta]}\left[\min\left\{1,\ \frac{k}{\sqrt{(n-2)U}}\right\}\right] \\
&= K\int_{K^2}^1 \frac{1}{\sqrt{u}} f_{[\alpha,\beta]}(u)\mathrm{d}u + \int_0^{K^2} f_{[\alpha,\beta]}(u)\mathrm{d}u \\
&= K\frac{\Gamma\big(\alpha-\frac{1}{2}\big)\Gamma\big(\alpha+\beta\big)}{\Gamma\big(\alpha-\frac{1}{2}+\beta\big)\Gamma\big(\alpha\big)}\int_{K^2}^1 f_{\left[\alpha-\frac{1}{2},\beta\right]}(u)\mathrm{d}u + F_{[\alpha,\beta]}\big(K^2\big)
\end{aligned}$$

then, by substitution of $\alpha = \frac{3}{2}$ and $\beta = \frac{n-1}{2}$ and using (**), we obtain

$$= C_n K\Big(1 - F_{\left[1, \frac{n-1}{2}\right]}\big(K^2\big)\Big) + F_{\left[\frac{3}{2}, \frac{n-1}{2}\right]}\big(K^2\big). \tag{C}$$

where

$$C_n = \frac{2\Gamma(n/2+1)}{\sqrt{\pi_0}\Gamma(n/2+1/2)}.$$

By substitution of (B) and (C) into (A) [and using identity $I_x(1,b) = 1-(1-x)^b$], the assertion of the corollary follows. ■

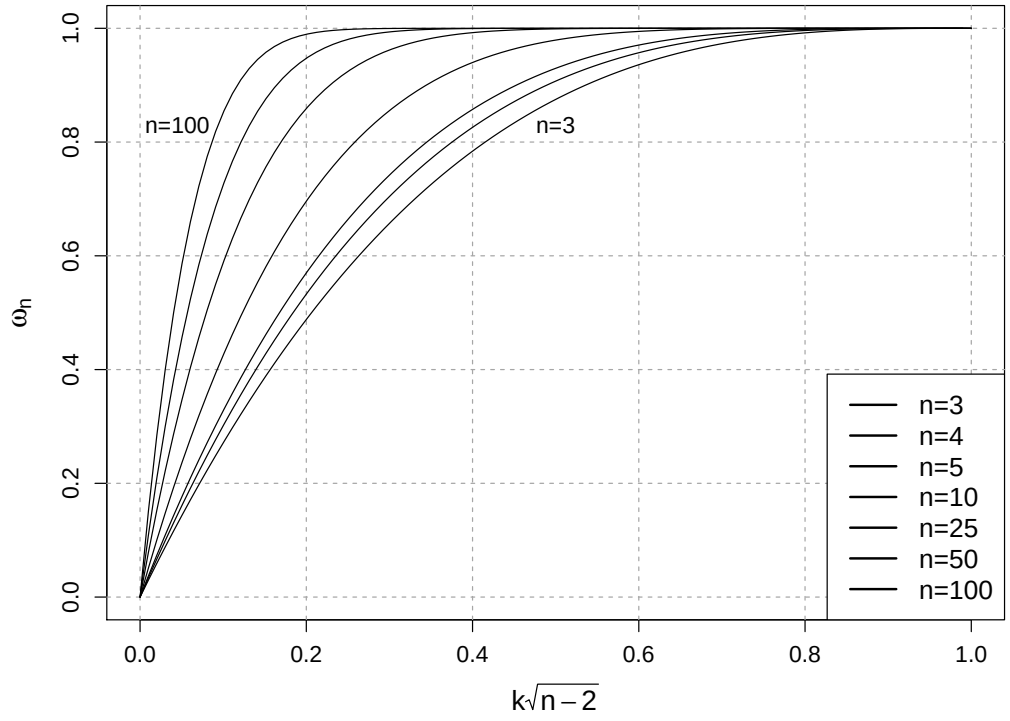


Figure 6.4.: Graph of ω_n as a function of k (scaled by the factor $\sqrt{n-2}$, for ease of comparison) and n; the curves for $n = 3, 4, 5, 10, 25, 50, 100$ are shown from right to left.

Remarks. (i) Function $\omega_n(k)$ in (6.29) is plotted in Figure 6.4. It is strictly increasing from 0 at $k = 0$ to 1 at $k = \sqrt{n-2}$, therefore it has an inverse function $D_k(\omega)$ defined by (see Efron and Morris, 1972, 134)

$$D_k(\omega) = D \qquad \Leftrightarrow \qquad \omega_n(k) = \omega. \tag{6.30}$$

Since Corollary 6.1 provides an explicit expression for $\omega_n(k)$, we can easily obtain the value of k that solves $D_k(\omega)$ given ω [i.e., root finding problem in one dimension]. This device is useful since ω determines the efficiency of the limited translation rule in terms of Bayes risk.

(ii) For $n > 171$, the value of k can be approximated using the numbers for $n = \infty$ in Table 6.2. In general, the numerical computation in `R` with the formulas in Corollary 6.1 break down when $n > 172$ [the Γ-function implemented in `R:stats` is only defined for arguments < 171.6]. Consider the case $n = 171$. If we use the $(n = \infty)$-approximation instead of the computed numbers for $n = 171$, the error incurred is at most 1.6%. We feel pretty save that this level of accuracy is sufficient in practical applications; if not, we should look for an alternative representation of the term C_n [e.g.,

Stirling's approximation for the Γ-function].

(iii) The numbers in Table 6.2 for $n = \infty$ correspond to the Bayes case; this was pointed out by Efron and Morris (1972, 135) who argue that for $n \to \infty$, we get perfect information on the value B (hence, the empirical Bayes situation becomes an intrinsic Bayes problem). Therfore, the tabulated numbers can be obtained using (6.18).

(iv) The data presented in Table 6.2 are in line with those given in Efron and Morris (1972, Table 1); differences emerge only in the 3rd decimal place [note: Efron and Morris obtained their numbers (presumably) by means of numerical integration; the details of their approach is not discussed in the paper]. The fact that our computations match those of Efron and Morris can be seen as numerical evidence that Corollary 6.1 is correct.

Relaxing some of previously introduced simplifications

In the discussion so far, we have tacitly introduced some simplifying assumptions. Although these assumptions eased the discussion, the resulting methods are hardly usable in practical applications. In what follows, we drop some assumptions an extend the concepts to cover more general situations. The reader will immediately verify that abandoning the playground of the simplified world and heading towards more practically useful methods does not introduce any difficulty in terms of concepts, but comes at the price of algebraic laboriousness.

6.3.3. Letting the data choose the origin: Estimation of location

First, we drop the assumption that the prior location μ is known to be zero. Yet, we keep the assumption that $D = 1$ is known. The goal is to let the data chose the origin for us by shrinking toward a central location of the n observed values y_i. A natural choice is the grand mean $\bar{y} = \sum_{i \leq n} y_i$. Of course, we can choose any origin we want by subtracting arbitrary constants from the data. The resulting JS-type rule for estimating θ_i, which is due to D. V. Lindley (see Brandwein and Strawderman, 1990, 359), is given by

$$\delta_i^{JS,\bar{y}}(\mathbf{y}) = \bar{y} + \left(1 - \frac{n-3}{\|\mathbf{y} - \bar{y}\mathbf{1}_n\|^2}\right)\big(y_i - \bar{y}\big), \quad i = 1, \ldots, n, \tag{6.31}$$

where $\mathbf{1}_n$ denotes the n-vector of ones. An alternative formulation of the rule is

$$\delta_i^{JS,\bar{y}}(\mathbf{y}) = y_i - \hat{B}(y_i - \bar{y}),$$

where

$$\hat{B} = \frac{n-3}{\|\mathbf{y} - \bar{y}\mathbf{1}_n\|^2}.$$

Let $\boldsymbol{\delta}^{JS,\bar{y}} = (\delta_1^{JS,\bar{y}}, \ldots, \delta_n^{JS,\bar{y}})^T$; having suppressed the dependency on $\mathbf{y}$. For this estimator to be well-defined, we require that $n \geq 4$. The price of letting the data

choose the origin instead of assuming $\mu = 0$ is an increase of the risk. Efron and Morris (1973b, 126) show that $\mathrm{RSL}(\pi, \boldsymbol{\delta}^{JS,\bar{y}}) = 3/n$, which in comparison with the relative saving loss of the JS rule given by $2/n$, reveals that the price will be negligible when n is large.

In its core, rule $\delta_i^{JS,\bar{y}}$ is essentially a JS rule, but translated by $\bar{y}$. Therefore, the drawbacks we encountered in case of the JS rule apply here as well. Namely, the shrinkage factor employed in $\delta_i^{JS,\bar{y}}$ can possibly take negative values, resulting in a weird behavior of the rule. A simple remedy is to consider a *positive-part rule* in exactly the same manner as we did for the JS rule. The RSL of the resulting positive-part rule is rather complicated and can be seen in Efron and Morris (1973b, Eq. 7.4). The second issue concerns the fact that the JS rule and also $\delta_i^{JS,\bar{y}}$ may do very poorly in estimating the θ_i's with unusually large values and therefore result in large component risk. In Section 6.3.2, we argued that the estimated *limited translation rule* is a compromise rule that protects against excessively large component-wise risk at relatively low costs in terms of ensemble risk. Clearly, such a compromise rule can also be formulated on the basis of $\delta_i^{JS,\bar{y}}$. The last comment concerns the choice of prior location; here, we assume that the r.v.'s Θ_i share the grand mean as common location; it will be seen in the next paragraph that we can easily replace the grand mean by a linear regression relationship.

6.3.4. Dropping the assumption of equal variances

So far we have assumed that the variances satisfy the identity $D_i \equiv D$, with D known. Now, we shall assume that

- the D_i's are allowed to be different (but known values) for the $i = 1, \ldots, n$ coordinates,
- the Θ_i's are independent r.v. which are modelled by the linear regression relationship, $\mathbf{x}_i^T \boldsymbol{\beta}$, where $\mathbf{x}_i$ is a q-vector of known auxiliary variables and $\boldsymbol{\beta} \in \mathbb{R}^q$; the Θ_i's share a common but unknown a priori variance $A \geq 0$.

Therefore, the hierarchical model now writes

$$\text{(sampling model)} \quad (Y_i \mid \Theta_i = \theta_i) \overset{\text{ind.}}{\sim} \mathcal{N}(\theta_i,\, D_i), \tag{6.32}$$

$$\text{(prior distr. } \pi_0) \qquad (\Theta_i \mid \boldsymbol{\beta}, A) \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\beta},\, A), \quad \text{for all } i = 1, \ldots, n. \tag{6.33}$$

Remarks. (i) Throughout the discussion, we suppose the $\mathbf{x}_i$'s to be such that $\sum_{i \leq n} \mathbf{x}_i \mathbf{x}_i^T$ has full column rank q, $q \leq n - 2$. Also, we write $\mathbf{X}$ to mean the $(n \times q)$ matrix given by $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$.

(ii) The model defined by (6.32) and (6.33) coincides with our definition of the Fay–Herriot model (see Definition 6.1) except that the constants b_i that pre-multiply the "area-level random effect" are taken to be $b_i \equiv 1$ for ease of discussion. This assumption is not really consequential, but simplifies notation; if the b_i's are indeed not all equal to one, it is straightforward to modify the estimating rules accordingly.

(iii) The Gaussian prior distribution π_0 embodies our prior beliefs that the Θ_i's share a common / parent distribution. We shall stick to this hypothesis. In particular, we do not consider group-specific distributions and the like; for approaches of this kind, we refer to Efron and Morris (1973a) on applying shrinkage estimators to separate sets versus combined estimators, George (1986b) on shrinkage towards any number of subspaces, and George (1986a) on group-wise or clustered shrinkage estimation.

Although our distributional assumptions postulate a common prior variance A, we shall for the moment assume that the Θ_i's satisfy the law $\Theta_i \,|\, \boldsymbol{\beta}, A_i \sim \mathcal{N}(\mathbf{x}_i^T\boldsymbol{\beta}, A_i)$, $i = 1, \ldots, n$, where the A_i's are possibly different from each other. Clearly, this assumption is not meaningful to an empirical Bayes approach, as the A_i's are not estimable having observed only one realization, y_i, on the ith coordinate (except in some special, but empirically unrealistic cases, e.g. when $A_i \propto D_i$). A formal Bayes argument shows – with A_i and $\boldsymbol{\beta}$ regarded as known quantities – that the posterior mean for the ith component, hence the Bayes rule under squared error loss, is (see Morris, 1983, 48)

$$\mathbf{x}_i^T\boldsymbol{\beta} + (1 - B_i)(y_i - \mathbf{x}_i^T\boldsymbol{\beta}), \qquad i = 1, \ldots, n, \tag{6.34}$$

where $(1 - B_i)$ is the shrinkage factor and

$$B_i = \frac{D_i}{D_i + A_i}. \tag{6.35}$$

This Bayes estimator is optimal under any symmetric loss function when A_i and $\boldsymbol{\beta}$ are known. Now, a reasonable empirical Bayes estimator obtains by adopting the form of the Bayes rule in (6.34), but replacing the unknown quantities by estimates and restricting attention to the case of a common-prior variance, $A_i \equiv A$. The resulting rule is given by

$$\delta_i^{UV}(\mathbf{y}) = \mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \left(1 - \frac{D_i}{D_i + \hat{A}^+}\right)(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}), \qquad i = 1, \ldots, n, \tag{6.36}$$

where

$$\hat{A}^+ = \max(0, \hat{A}) \tag{6.37}$$

and the tuple $(\hat{\boldsymbol{\beta}}, \hat{A})$ is the solution to the system of estimating equations

$$\left.\begin{aligned} \sum_{i=1}^{n} \frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{D_i + A} \mathbf{x}_i = \mathbf{0} \\ \sum_{i=1}^{n} \frac{1}{w_i(A)} \left(A + D_i - \frac{n}{n-q} \big(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\big)^2 \right) = 0 \end{aligned}\right\}. \tag{6.38}$$

The superscript "UV" in the notation of δ_i^{UV} is an abbreviation for unequal variances. The facilitation of possibly different D_i's is the distinctive feature of the new rule (not the translation in terms of linear regression relationship versus the grand mean). This marks also the difference to the JS rule, as will be seen below.

The solution to the first estimating equation in (6.38) results in the generalized least squares estimator $\hat{\boldsymbol{\beta}}$. The second estimating equation in (6.38), which is motivated by Morris (1983, 53), is more interesting in that it is a *general form* which produces a couple of estimators of A [or $A(\mathbf{y})$ if we wish to highlight the dependence on $\mathbf{y}$] through the choice of the w_i's and q's. The choices of (w_i, q) relate to the estimators of A in the following way:

(C1) M.l.e.: let $w_i \equiv (D_i + A)^2$ and put $q = 0$, see Efron and Morris (1975, 314), Fay and Herriot (1979, Appendix A2), and Morris (1983, 53),

(C2) (Approximate) reduced m.l.e.: let $w_i \equiv (D_i + A)^2$ and $q = \operatorname{rank}(\mathbf{X})$; see Morris (1983, 53),

(C3) Fay–Herriot estimator: let $w_i \equiv D_i + A$ and $q = \operatorname{rank}(\mathbf{X})$; see Fay and Herriot (1979, 271–2).

Remarks. (i) The first remark concerns our choice of the constants $b_i \equiv 1$; see Definition 6.1. Again, we point out that this choice is not consequential but simplifies notation. All estimators of A can be modified in a straightforward manner to account for possibly different b_i's.

(ii) Cases (C1) and (C2) form a group of estimators, insofar that they rely explicitly on the Gaussian distributional assumption. This is not the case for the method-of-moments estimator due to Fay and Herriot (1979), which is consistent as $n \to \infty$ under more general distributional assumptions. From a mathematical point of view, the major difference between (C1, C2) and (C3) is the "weighting": (C1, C2) employ a weighting by $(D_i + A)^2$ whereas (C3) uses $D_i + A$. Prasad and Rao (1990, 165) suggested another simplified moment-type estimator, which actually features a closed-form solution; see also Rao (2003, 118).

(iii) For (C1) and (C2), it is not necessary to require that (6.37) is enforced as the m.l.e. and reduced m.l.e. are non-negative by definition. For the

moment-based estimator (C3) on the other hand, imposing (6.37) prevents us from ending up with a situation similar to the negative shrinkage factor encountered with the JS rule (viz. positive-part rule).

(iv) The reduced (or restricted or residual) m.l.e. in (C2) accounts for the loss of degrees of freedom due to estimation of $\boldsymbol{\beta}$. Method (C2) is an approximation of the reduced m.l.e. and is due to Morris (1983, 53). The, so to speak, fully-fledged reduced m.l.e. approach obtains as follows (see e.g. Rao, 2003, 101): put $\mathbf{P}(A) = \mathbf{V}^{-1}[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}]$, where $\mathbf{V}^{-1}$ is a diagonal matrix with elements $D_i/(D_i + A)$; then the reduced m.l.e. estimator of A solves the estimating equation

$$\operatorname{tr}\{\mathbf{P}(A)\} - \mathbf{y}^T\mathbf{P}(A)\mathbf{y} = 0. \tag{6.39}$$

(v) What estimator of A is best? We shall answer this question with regard to relative efficiency using large sample arguments (i.e., taking $n \to \infty$ and holding q fix). Rao (2003, 120) shows that the chain

$$AV(\hat{A}_{rmle}) = AV(\hat{A}_{mle}) \leq AV(\hat{A}_{FH}) \leq AV(\hat{A}_{PR}) \tag{6.40}$$

holds, where expressions of the asymptotic variance, AV, for $\hat{A}_{mle}$ and $\hat{A}_{rmle}$ [which is the solution to (6.39)] are given in Rao (2003, 120), see also Datta and Lahiri (2000). The asymptotic variance of the Fay–Herriot estimator, $\hat{A}_{FH}$, is given in Datta, Rao, and Smith (2005, 195); $AV(\hat{A}_{PR})$ is given in Prasad and Rao (1990, 167). The fact that the m.l.e. (and reduced m.l.e.) prove to be the most efficient estimators in (6.40) is no surprise; on the other hand, the usefulness of these estimators depends heavily on the appropriateness of the Gaussian distributional assumption.

(vi) The rule δ_i^{UV} defined in (6.36) is a *generalization* of the JS rule. This can be seen if we let $\mathbf{x}_i \equiv 1$ and write μ instead of $\boldsymbol{\beta}$. Then, by the first equation in (6.38), we have

$$\hat{\mu} = \sum_{i=1}^{n} \frac{y_i}{D_i + A} \Big/ \sum_{i=1}^{n} \frac{1}{D_i + A},$$

which is a weighted mean of the y_i's, not the arithmetic mean. Therefore, δ_i^{UV} does not coincide with the JS rule defined in (6.31) unless $D_i \equiv D$. Though, even if $D_i \equiv D$, δ_i^{UV} does not (exactly) coincide with the JS rule for any of the given estimators of A. This has been pointed out by Efron and Morris (1973b, 128) for the case of the m.l.e. If we nevertheless wish to obtain a rule that coincides with the JS rule when all D_i's are equal, we have to, loosely speaking, correct for the loss of degrees of freedom of the m.l.e. of A. More precisely, under the model $\Theta_i \sim \mathcal{N}(0, A)$, Efron and Morris (1973b, 128) propose an estimator that is based on the law $Y_i^2 \overset{\text{ind.}}{\sim} (A + D_i)\chi^2_{d_i}$, i.e. a chi-square distribution with d_i degrees of freedom. The resulting rule

features coordinate-specific estimates, say $\tilde{A}_i$ and $\tilde{B}_i$, (cf. Eq. 6.35) and indeed reduces to the JS rule if all D_i are equal. On the other hand, the rule shows the adverse property of potentially negative shrinking factors; but the authors counter this disadvantage by taking, say, $\tilde{B}_i^+$, as the nearest point in $[0, 1]$ to $\tilde{B}_i$. We conjecture that the estimator $\tilde{A}_i$ of B. Efron and C. Morris can be generalized to the case of a linear regression relationship, but we have not worked out the details.

Another way to generalize the JS rule is to consider a transformation and define $\tilde{y}_i = \sqrt{D_i} y_i$ and $\tilde{\theta}_i = \sqrt{D_i}\theta_i$; then, apply the JS rule to the transformed data, and finally transform back to the original coordinates. This approach applies the same shrinkage factor to all coordinates (Efron and Morris, 1973b, 127), which is unappealing since we would prefer a method that applies stronger shrinkage to those coordinates with large D_i.

Limited translation rule

In a large number of papers, the rule δ_i^{UV} defined in (6.36) is regarded as "the" estimating rule under the Fay–Herriot model; see e.g. Prasad and Rao (1990). Yet, this is not rule that Fay and Herriot (1979) actually proposed. They indeed were fully aware of the inferior behavior that δ_i^{UV} shows with regard to component-wise risk. Therefore, they "applied" the *limited translation rule* to δ_i^{UV} in the very same manner Efron and Morris (1972) proposed to do on grounds of ad hoc reasoning. The resulting rule is – suppressing the dependence on y – given by (Fay and Herriot, 1979, 271)

$$\hat{\delta}_i^{EM,k} = \begin{cases} \delta_i^{UV} & \text{if } \delta_i^{UV} \leq \left| y_i - k\sqrt{D_i} \right|, \\ \delta_i^{UV} - k\sqrt{D_i} & \text{if } \delta_i^{UV} < y_i - k\sqrt{D_i}, \\ \delta_i^{UV} + k\sqrt{D_i} & \text{if } \delta_i^{UV} > y_i + k\sqrt{D_i}, \end{cases}$$

where k is a "tuning constant" chosen by the analyst.

Behavior of the estimators in the presence of outliers

Suppose the Fay–Herriot model and the rule δ_i^{UV}, $i = 1, \ldots, n$, defined in (6.36). This rule relies on the tuple of unknown parameters $(\boldsymbol{\beta}, A)$, which are subject to estimation; the corresponding estimating equations are given in (6.38). Now, we could go through the effort expressing the estimators of $\boldsymbol{\beta}$ and A as functionals of some d.f. F and then study the influence function (see Definition 2.5). Instead of doing so, it suffices to note that the estimator of $\boldsymbol{\beta}$ is the generalized least squares estimator; hence, it has an unbounded influence function w.r.t. to y_i and also the row of the design matrix, $\mathbf{x}_i$ (see e.g. Hampel et al., 1986, chap. 6.2 and 6.3). From (6.38) it is easy to see that the estimators of A also have an unbounded influence function. Furthermore, the residual, $y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$, enters the formula of the rule δ_i^{UV} in an "unprotected fashion"—therefore, influential

observations in both y_i and $\mathbf{x}_i$ may heavily influence δ_i^{UV}. Altogether, we conclude that Principle 1 (Hampel's qualitative robustness) is not fulfilled, neither is Principle 2 (quantitative robustness). Therefore, we shall study robust estimators under the Fay–Herriot model in what follows.

Before we turn to the discussion of robust methods, we shall study the *qualitative* behavior of rule δ_i^{UV} in the presence of contaminated y-data. To this end, let the observed data be denoted by $\mathbf{y} = (y_1, \ldots, y_n)^T$. In addition, we suppose another set of observed values of size n, $\mathbf{y}^*$, which contains at least one *influential* outlier. Let $\hat{A}(\mathbf{y})$ denote the m.l.e. of A (or any of the non-robust estimators in C1–C3, see above) based on the data $\mathbf{y}$; the concomitant estimate $\hat{\boldsymbol{\beta}}$ in the estimator of A is taken w.l.o.g. to be the generalized least squares estimate. Likewise, $\hat{A}(\mathbf{y}^*)$ is the estimate based on $\mathbf{y}^*$. Except in very rare cases, we observe inflated estimates in the presence of contamination, formally

$$\hat{A}(\mathbf{y}^*) > \hat{A}(\mathbf{y}),$$

sometimes even $\hat{A}(\mathbf{y}^*) \gg \hat{A}(\mathbf{y})$. What is the impact of contaminated data on the estimating rules? For ease of disposition, we have reprinted the rule defined in (6.36) [assuming that $\hat{A} = \max(0, \hat{A})$] in a slightly modified representation,

$$\delta_i^{UV}(\mathbf{y}) = \frac{D_i}{D_i + \hat{A}(\mathbf{y})}\mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\mathbf{y}) + \left(1 - \frac{D_i}{D_i + \hat{A}(\mathbf{y})}\right)y_i, \qquad i = 1, \ldots, n. \tag{6.36}$$

With contaminated data, $\hat{A}(\mathbf{y}^*)$ appears in place of $\hat{A}(\mathbf{y})$ in (6.36), and leads ceteris paribus to smaller shrinkage factors $D_i/(D_i + \hat{A}(\mathbf{y}^*))$ compared with uncontaminated data. As a result, contamination decreases the amount of shrinkage and thus the rule tends ceteris paribus to favor the direct estimates y_i. Moreover, less shrinkage implies a loss in relative savings over the m.l.e. (i.e., sacrifice of efficiency). In the extreme case, $\hat{A}(\mathbf{y}^*)$ is so large that the shrinkage factors are virtually zero; hence, the rule almost or effectively coincides with the m.l.e. y_i. From this point of reasoning it is obvious that the estimate of A plays a major role in the behavior of rule δ_i^{UV}. Somewhat surprisingly, Ghosh et al. (2008) propose to use the m.l.e. of A in their "robust" empirical Bayes estimator under the Fay–Herriot model.

What about the generalized least squares estimator $\hat{\boldsymbol{\beta}}$ in (6.36)? It plays only a subordinate role in the rule. Although $\hat{\boldsymbol{\beta}}(\mathbf{y}^*)$ is complicit in the inflated estimate produced by $\hat{A}(\mathbf{y}^*)$, its appearance in the rule barely matters (for better or for worse). This is especially pronounced if $\hat{A}(\mathbf{y}^*) \gg \hat{A}(\mathbf{y})$ since then the rule virtually coincides with the m.l.e. In such cases, $\hat{\boldsymbol{\beta}}$ can practically take arbitrary values—it does not matter as the respective shrinkage factor is virtually zero.

6.4. Robust estimation and prediction under the Fay–Herriot model

The goal of this section is to show how we can, in the light of the above theoretical discussion,

(i) obtain *robust estimates* of the model parameters in the presence of outlying and influential observations;

(ii) get *robust predictions* of the area-level means, given a robust model fit.

Robust James–Stein rule

It makes sense to begin with a discussion of the robustification of the James–Stein (JS) rule. To this end, consider (once again) the simple hierarchical model: $(Y_i \,|\, \Theta_i = \theta_i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_i, D)$ and $\Theta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, A)$, $i = 1, \ldots, n$. The variance D is supposed known; variance A is unknown. Stein (1981, 1145) observed (refering to Efron and Morris, 1971, 1972) that if any of the θ_i's happen to fall substantially outside the central location (towards which we want to shrink), the James–Stein estimator will collapse back to Y_i (resp., the realization y_i) and will offer little improvement over the m.l.e. (i.e., considerable loss in relative savings). He points out that this is likely to happen when the Y_i's have a long-tailed empirical distribution instead of the assumed Gaussian distribution [Stein tacitly assumes a symmetric distribution of the Y_i's]. For such cases, he suggested a modification of the limited translation rule. Let $a \wedge b$ denote the minimum of a and b, and let

$$Z_i = |Y_i|, \quad i = 1, \ldots, n, \quad \text{and} \quad Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)},$$

where $Z_{(i)}$ is the ith order statistics of Z_i. For a large fraction k of n, which must be chosen by the statistician, Stein (1981, 1145) defines the coordinate-wise rule

$$\delta_i^{JSr}(\mathbf{Y}) = \left(1 - \frac{(k-2)D \min\left\{1,\ Z_{(k)}/|Y_i|\right\}}{\sum_{j=1}^n Y_j^2 \wedge Z_{(k)}^2}\right) Y_i, \qquad i = 1, \ldots, n.$$

Stein (1981) calls the estimator a trimmed estimator; other authors call it a truncated estimator (see e.g. Dey and Berger, 1983). In robust statistics, we would rather call it a winsorized estimator as it is based on the winsorized sum of squares, $\sum_{j=1}^n Y_j^2 \wedge Z_{(k)}^2$ or when we take the term $(k-2)$ in the numerator into account, it is a winsorized estimator of the unknown variance A given by

$$\frac{1}{k-2} \sum_{j=1}^n Y_j^2 \wedge Z_{(k)}^2.$$

Unlike the estimator $\|\mathbf{Y}\|^2/(n-2)$ [which is incorporated in the JS rule], the winsorized variance estimator does not suffer from inflation under long-tailed

distribution (provided k is chosen appropriately). Indeed, Stein (1981) shows that for long-tailed distributions, rule δ_i^{JSr} is more efficient than the regular JS rule. One difficulty in dealing with rule δ_i^{JSr} is the choice of the fraction k. However, Dey and Berger (1983) show that a sensible choice can be derived with regard to the improvement in Bayes risk over the m.l.e. via an adaptive estimation procedure. Further, in case the prior location μ is different from zero, Angers and Berger (1991, 46–47) suggest to estimate it by the trimmed mean.

The robustification of the James–Stein rule discussed so far is established by merely ad hoc arguments (i.e., through replacing classical estimates by robust estimates), but it proves useful when the data follow a long-tailed empirical distribution. In order to tackle the robustification of estimators under the Fay–Herriot model, it will be beneficial to seize on our discussion of the Hodges–Lehmann theory rather than following ad hoc procedures.

To begin with, we revisit the simplest model $Y = \Theta + E$ where r.v. $E \sim \mathcal{N}(0,1)$ and an arbitrary prior distribution P is assumed to be placed on r.v. Θ. Our discussion of the Hodges–Lehmann theory (see Section 6.2.1) showed that an interesting class of rules is given by $\delta = y + \psi(y)$, where $\psi \equiv \nabla \log g$ with marginal distribution $g \equiv \Phi * P$. Unfortunately, an exact expression for the compromise rules formulated in Principles PI and PII is not known. However, we have pointed out that, for a given $\epsilon \in (0,1)$, the *least favorable prior* distribution P_ϵ implies the *approximate* least favorable marginal distribution

$$\hat{g}_\epsilon^*(y) = \begin{cases} (1-\epsilon) g_0(y_0) \exp\big[k(y-y_0)\big] & \text{for } y < y_0, \\ (1-\epsilon) g_0(y) & \text{for } y_0 \le y \le y_1, \\ (1-\epsilon) g_0(y_1) \exp\big[-k(y-y_1)\big] & \text{for } y_1 < y, \end{cases} \tag{6.41}$$

where y_0 and y_1 (s.t. $y_0 < y_1$) mark the boundaries of the interval defined as

$$\left| \frac{g_0'(y)}{g_0(y)} \right| \le k$$

and k is a constant, more precisely $k(\epsilon)$, that is determined such that $\int g_\epsilon(y)\mathrm{d}y = 1$. The distribution in (6.41) refers to the *general case*, where g_0 is any suitable base marginal p.d.f. (see Marazzi, 1980, 15). In our case, the sampling model is Gaussian and we elicit a Gaussian base prior, therefore $g_0 \equiv \Phi * \Phi$ which implies (expressed in logarithms for ease of display)

$$\log \hat{g}_\epsilon(y) = \begin{cases} ky + k^2/2 + \text{const.} & \text{for } y < -k, \\ -y^2/2 + \text{const.} & \text{for } |y| \le k, \\ -ky + k^2/2 + \text{const.} & \text{for } y > k, \end{cases} \tag{6.42}$$

where $\text{const} = \log(1-\epsilon) + (\log 2 + \log \pi)/2$, and π denotes the area of the unit circle. Observe that $\hat{g}_\epsilon$ in (6.42) is – up to some additive constants (which are

irrelevant when it comes to estimation) – equal to function ρ_k, which is the ρ-function accompanying the Huber ψ-function.

Fay–Herriot model

Exploiting the properties of the Gaussian distribution as a location-scale family, it is straightforward to apply the methods to the Fay–Herriot model. Given the known q-vectors of auxiliary variables $\mathbf{x}_i$ and the constants D_i (variances of the direct estimator, $D_i > 0$), we have

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sqrt{A + D_i}\, E_i, \qquad \text{with r.v.'s } E_i \overset{\text{ind.}}{\sim} \hat{G}_\epsilon, \quad i = 1, \ldots, n, \tag{6.43}$$

where

$$\hat{G}_\epsilon(e) = \int_{-\infty}^{e} \hat{g}_\epsilon(t)\mathrm{d}t, \tag{6.44}$$

$\hat{g}_\epsilon$ is defined in (6.42), and $\boldsymbol{\beta} \in \mathbb{R}^q$ and $A \in \mathbb{R}^+$ are unknown parameters. Write $\mathbf{Y}$ to mean the r.v. $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ with realizations $\mathbf{y} = (y_1, \ldots, y_n)^T$. When the parameters $\boldsymbol{\beta}$ and A *were* indeed known, the rule under the least favorable prior is of the form

$$\delta_i^{EM,k}(\mathbf{y}) = y_i - \frac{D_i}{\sqrt{A + D_i}} \psi_k\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{A + D_i}}\right), \qquad i = 1, \ldots, n, \tag{6.45}$$

which coincides with the limited translation rule (LTR) due to Efron and Morris (1971, cf.) that was suggested by Fay and Herriot (1979).

Remark. The rule in (6.45) is an immediate implication of the location-scale characteristics of the underlying distribution. From the above theory, we know that the ψ-function obtains, for any function g, as the logarithmic derivative $\psi_g \equiv \nabla \log g$. Consider the function $\hat{g}_\epsilon$ defined (6.42) and suppose $\hat{g}_\epsilon(y/\sigma)$ for some $\sigma > 0$; then,

$$\nabla \log\left(\frac{1}{\sigma}\hat{g}_\epsilon(y/\sigma)\right) = \frac{\hat{g}'_\epsilon(y/\sigma)}{\sigma \hat{g}_\epsilon(y/\sigma)} = \frac{1}{\sigma}\psi_k(y/\sigma),$$

hence, we get the rule in (6.45). The results in Marazzi (1990, 106; see also Marazzi, 1985) imply the rule

$$\delta_i(\mathbf{y}) = y_i - D_i \psi_k\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{A + D_i}\right), \qquad i = 1, \ldots, n,$$

which is unappealing as the argument of the function ψ_k is not properly decorrelated. For this reason, we shall not consider the rule due to A. Marazzi.

Of course $\delta_i^{EM,k}$ in (6.45) is not of any practical value unless $\boldsymbol{\beta}$ and A are actually known. Therefore, Fay and Herriot (1979) attempted (so did Efron and Morris, 1972) to estimate the unknown parameters from the data in the spirit of empirical Bayes theory. There is nothing wrong with that theory. How-

ever, it is crucial to have a *clear understanding* about the true data generating mechanism prior to estimation. Fay and Herriot (1979) suggested the application of the limited translation rule (LTR) as some ad hoc (but fully operational) measure to limit exceedingly large component risks, a situation that is usually experienced with the generalization of the James–Stein method. Yet, R.E. Fay and R.A. Herriot had the hierarchical Gaussian model in mind when it comes to estimation. As a consequence, (and this is consistent with their assumptions) they suggested the weighted least squares estimator for $\boldsymbol{\beta}$ and came up with their famous moment-type estimator for A.

Our argument is different, insofar that the limited translation rule is not introduced as some "afterthought" but is regarded as an integral element of the underlying theory. Recall that under the hierarchical model, it is the prior distribution that governs the behavior of the estimators of $\boldsymbol{\beta}$ and A. From this point of view, an outlying y-value is taken to be caused by an "outlying" or "bad" realization of the r.v. Θ whose d.f. is the prior distribution. Consequently, the focus was on the prior distribution. Since we did have rather strong beliefs in the prior distribution, but were not willing to entirely rely on these beliefs, the Hodges–Lehmann theory motivated rules that provide some robustness against prior misspecification. Unfortunately, the optimal rule under Principles PI and PII is not manageable, but the approximation under that results from a least favorable prior distribution turned out to be the limited translation rule.

Let us come back to the estimation of the unknown parameters $\boldsymbol{\beta}$ and A. It should be clear from our discussion so far that the prior elicitation process is key to estimation under the Fay–Herriot model. In particular, the LTR and the least favorable prior P_ϵ *imply each other*, and this has a fundamental consequence: Under prior P_ϵ, the marginal distribution of $\mathbf{Y}$ is not a convolution of two Gaussian d.f.'s, but is instead given by

$$m(\mathbf{y} \mid \boldsymbol{\beta}, A) = \prod_{i=1}^{n} \frac{1}{\sqrt{A + D_i}} \hat{g}_\epsilon \left(\frac{y_i - \mathbf{y}_i^T \boldsymbol{\beta}}{\sqrt{A + D_i}} \right), \tag{6.46}$$

where $\hat{g}_\epsilon$ is defined in (6.42). Since empirical Bayes-type estimators of $\boldsymbol{\beta}$ and A directly derive from the marginal, the fact that $m(\mathbf{y} \mid \boldsymbol{\beta}, A)$ is not the product of Gaussians (unless $k \to \infty$, which shall be ruled out) has far reaching consequences. None of the estimators proposed in literature (under the Gaussian sampling model and prior) for $\boldsymbol{\beta}$ or A is the m.l.e. under (6.46); therefore, these estimators are not efficient. This applies to the weighted least squares estimator of $\boldsymbol{\beta}$, as well as to the estimators of A due to Fay and Herriot (1979), Prasad and Rao (1990), Datta et al. (2005), etc.

What estimator is efficient under the marginal distribution in (6.46)? The joint M-estimator of $\boldsymbol{\beta}$ and A, which will be discussed below. Using the limited translation rule together with non-robust standard estimators for $\boldsymbol{\beta}$ and A corresponds to *stop halfway down the road*. Though, we should act consistently. If

we do have doubts in the appropriateness of the Gaussian prior, we are highly recommended to apply the limited translation rule *and* estimate the parameters robustly. If the Gaussian prior is not in doubt, the limited translation rule is superfluous.

6.4.1. Robust model fit

6.4.1.1. M-estimators

Under the marginal distribution in (6.46), joint M-estimates of $\boldsymbol{\beta}$ and A can be obtained via minimization of the negative log-likelihood function (omitting constant terms)

$$-\log \mathcal{L}(\boldsymbol{\beta}, A \mid \mathbf{y}) = \frac{1}{2}\sum_{i=1}^{n} \log(A + D_i) - \sum_{i=1}^{n} \log \hat{g}_\epsilon\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sqrt{A + D_i}}\right), \tag{6.47}$$

where minimization is w.r.t. $\boldsymbol{\beta}$ and A. As we have pointed out, $-\log \hat{g}_\epsilon$ coincides (up to some constants) with ρ_k, the ρ-function accompanying the Huber ψ-function. It can be useful to consider a more general class of ρ-functions. Let

$$\rho : \mathbb{R} \to [0, \infty) \tag{6.48}$$

be such that

(i) ρ is a differentiable convex function that satisfies $\rho(0) = 0$, and

(ii) there is a constant k such that

$$0 < \lim_{|u| \to \infty} \frac{\rho(u)}{|u|} = k < \infty \text{ holds.}$$

Under the more general definition of the ρ-function, one considers minimization of the objective function (on the set $\mathbb{R}^q \times \mathbb{R}^+$)

$$Q(\boldsymbol{\beta}, A) = \frac{1}{n}\sum_{i=1}^{n} \left\{ \left[\rho\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sqrt{A + D_i}}\right) + \kappa \right] \sqrt{A + D_i} \right\}, \tag{6.49}$$

instead of (6.47); $\kappa > 0$ is a constant. The objective function in (6.49) is a straightforward adaption of the objective function under the linear regression model; see Huber (1981, 179). Taking partial derivatives of $Q(\boldsymbol{\beta}, A)$ w.r.t $\boldsymbol{\beta}$ and A, we get the following characterization of the minimization problem

$$\left.\begin{aligned} \sum_{i=1}^{n} \psi(r_i) \frac{\mathbf{x}_i}{\sqrt{A + D_i}} &= \mathbf{0} \\ \sum_{i=1}^{n} \frac{\chi(r_i)}{A + D_i} &= 0 \end{aligned}\right\}, \tag{6.50}$$

where

$$r_i = \frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{A + D_i}}, \qquad i = 1, \ldots, n, \tag{6.51}$$

with (see Huber, 1981, 177–78)

$$\psi(u) = \rho'(u) \tag{6.52}$$

$$\chi(u) = u\psi(u) - \rho(u). \tag{6.53}$$

The most important special case satisfying (6.52) and (6.53) is the following configuration

$$\psi(u) = \psi_k(u) \qquad \text{and} \qquad \chi(u) = 0.5[\psi_k^2(u) - \kappa], \tag{6.54}$$

where κ is a consistency correction term under the Gaussian core model, and is defined as

$$\kappa = \mathbb{E}\psi_k^2(u), \tag{6.55}$$

$\mathbb{E}$ denoting expectation w.r.t. the standard Gaussian distribution. The choice of functions in (6.54) refers to "Proposal 2" of Huber (1964) and implies the following system of M-estimator estimating equations for the parameter tuple $(\boldsymbol{\beta}, A)$,

$$\left.\begin{aligned} \sum_{i=1}^{n} \psi_k(r_i) \frac{\mathbf{x}_i}{\sqrt{A + D_i}} &= \mathbf{0} \\ \sum_{i=1}^{n} \frac{\psi_k^2(r_i) - \kappa}{A + D_i} &= 0 \end{aligned}\right\}. \tag{6.56}$$

The M-estimates will be denoted by $(\hat{\boldsymbol{\beta}}, \hat{A})$.

Remarks. (i) Our choice of ψ-functions is not consequential; any bounded, odd, and monotone ψ-function will do fine. It can be useful to choose different tuning constants, say k_1 and k_2, for the ψ_k-functions in the system of estimating equations in (6.56).

(ii) The M-estimator defined in (6.56) obtains as a special case of the RML method proposed by Sinha and Rao (2009) for the basic unit-level model. This fact has been known for quite some time (see e.g. Schoch, 2011a). Recently, S. Warnholz has worked out the details in his Ph.D. thesis (Warnholz, 2016b, chap. 3.3.1) and implemented the method in the `R` package `saeRobust` (Warnholz, 2016a). To see that the method of Sinha and Rao (2009) applies, let $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$, and put $\mathbf{r} = (r_1, \ldots, r_n)^T = \mathbf{U}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, where $\mathbf{U}$ is a diagonal matrix whose elements are equal to the diagonal elements of the model's covariance matrix, $\boldsymbol{\Sigma}(A)$ [highlighting the fact that $\boldsymbol{\Sigma}$ depends on the unknown variance parameter A]. Let $\boldsymbol{\psi}(\mathbf{r}) = (\psi(r_1), \ldots, \psi(r_n))^T$, where ψ is the Huber ψ-function (neglecting the dependency on the tuning constant for ease of dis-

play). Sinha and Rao (2009) obtain estimates of $\boldsymbol{\beta}$ and A as solutions to the system of estimating equations (see their Eq. 16)

$$\left.\begin{aligned} \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{U}^{1/2}\boldsymbol{\psi}(\mathbf{r}) &= \mathbf{0} \\ \boldsymbol{\psi}^T(\mathbf{r})\mathbf{U}^{1/2}\boldsymbol{\Sigma}^{-2}\mathbf{U}^{1/2}\boldsymbol{\psi}(\mathbf{r}) - \operatorname{tr}\left\{\kappa\boldsymbol{\Sigma}^{-1}\right\} &= 0 \end{aligned}\right\}, \tag{6.57}$$

where κ is a consistency correction term. Under the Fay–Herriot model, the system in (6.57) simplifies considerably since $\boldsymbol{\Sigma} = \operatorname{diag}_{i=1,\ldots,n}\{A + D_i\}$. Hence, we have

$$\boldsymbol{\Sigma}^{-1} = \operatorname{diag}_{i=1,\ldots,n}\big\{(A + D_i)^{-1}\big\} \qquad \text{and} \qquad \mathbf{U} = \operatorname{diag}_{i=1,\ldots,n}\big\{A + D_i\big\},$$

and using these expressions in (6.57), we get (6.56).

(iii) The M-estimator of A defined in (6.56) is a robustification of the m.l.e., see method (C1) discussed subsequent to (6.38). Likewise, we can formulate a robustification of the approximate reduced m.l.e. (REML) suggested by Morris (1983); see method (C2). To this end, let r_i be given in (6.51); then, the approximate REML-type M-estimator of A is the solution to the equation

$$\sum_{i=1}^{n} \frac{\psi_k^2(r_i) - \kappa(n-q)/n}{A + D_i} = 0, \tag{6.58}$$

where κ is defined in (6.55) and $q = \operatorname{rank}(\mathbf{X})$.

(iv) A robustification of the method-of-moments estimator of A proposed by Fay and Herriot (1979) [see estimator (C3) subsequent to (6.38)] obtains as the solution to the estimating equation

$$\sum_{i=1}^{n} \psi_k^2(r_i) - \kappa(n-q) = 0, \tag{6.59}$$

where r_i is given in (6.51) and κ is defined in (6.55).

(v) In a similar vein, we may also write down an M-estimator that derives from the moment-based estimator proposed by Prasad and Rao (1990) and the REML estimator in (6.39).

Existence of the M-estimator

Some of the problems with existence and convergence proofs in the context of M-estimators arise when the ψ-functions of the regression and scale estimators are totally unrelated (Huber, 1981, 176). These problems can be bypassed if we consider configurations where the ψ-functions are naturally related. This is the case when the Huber ψ-functions in system (6.56) satisfy $k_1 = k_2 = k$ (i.e., when we do not consider separate robustness tuning constants for different estimating equations although we could). In fact, this is the setup which is

typically used in practice.

Before we continue our argument, it proves useful to recall the regression problem. The M-estimator of regression (with scale σ) can be obtained from the following minimization problem (Huber, 1981, 179)

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^q, \sigma\in\mathbb{R}^+}{\arg\min}\ \bar{Q}(\boldsymbol{\beta}, \sigma)$$

with

$$\bar{Q}(\boldsymbol{\beta}, \sigma) = \frac{1}{n}\sum_{i=1}^{n}\left[\rho_k\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right) + \kappa\right]\sigma, \tag{6.60}$$

where κ and ρ_k are defined above. Observe the similarities of $\bar{Q}(\boldsymbol{\beta}, A)$ and $Q(\boldsymbol{\beta}, A)$ in (6.49). Let $f_i : \mathbb{R} \to \mathbb{R}$ be the linear predictor $\mathbf{x}_i^T\boldsymbol{\beta}$; differentiation of $f_i(t)$ w.r.t. t is denoted by $\dot{f}_i$. Huber (1981, 177–178) assumes that the tuple $(\boldsymbol{\beta}, \sigma)$ depends linearly on some real t. Let the summands in (6.60) be (omitting the index i)

$$q = \kappa\sigma + \rho_k\left(\frac{y - f}{\sigma}\right)\sigma, \tag{6.61}$$

then Huber (1981, 178) shows that the second-order derivative of q w.r.t. t is given by

$$\ddot{q} = \rho_k''\left(\frac{y - f}{\sigma}\right)\left[\frac{y - f}{\sigma}\dot{\sigma} + \dot{f}\right]^2\frac{1}{\sigma} \geq 0 \tag{6.62}$$

which proves that $\bar{Q}(\boldsymbol{\beta}, \sigma)$ is convex in $(\boldsymbol{\beta}, \sigma)$ [he also points out that the result holds even if ρ_k is not everywhere differentiable]. Suppose some local minimum of $\bar{Q}(\boldsymbol{\beta}, \sigma)$ exists. Then, convexity implies that the attained local minimum is also a global minimum (see e.g. Boyd and Vandenberghe, 2004, Ch. 4.2.2).

We are now in the position to apply the insights gained from regression M-estimation to the M-estimator under the Fay–Herriot model defined in (6.56). For the FH model, it is easy to see that we have

$$q_i = \kappa\sigma_i + \rho_k\left(\frac{y_i - f_i}{\sigma_i}\right)\sigma_i, \qquad i = 1, \ldots, n, \tag{6.63}$$

in place of (6.61), where $\sigma_i = \sqrt{A + D_i}$, $i = 1, \ldots, n$. The corresponding objective function $Q(\boldsymbol{\beta}, A)$ is defined in (6.49).

Proposition 6.5. *Suppose the system defined in (6.56), then the solution $(\hat{\boldsymbol{\beta}}, \hat{A})$ to (6.56) is unique provided it exists.*

Proof. To fix ideas, let $a, b, c : \mathbb{R} \to \mathbb{R}$ denote arbitrary continuous twice a.e. (Lebesgue) differentiable functions. Suppose that b is convex and non-increasing, and c is concave. The composition $a = b \circ c$ is convex on its domain because it satisfies for a.e. $x \in \mathbb{R}$,

$$a''(x) = b''\big[c(x)\big]c'(x)^2 + b'\big[c(x)\big]c''(x) \geq 0. \tag{A}$$

Write $q_i(\boldsymbol{\beta}, \sigma)$ to mean q_i in (6.63) and observe that it is a non-increasing function in its second argument. Consider $q_i\big(\boldsymbol{\beta}, \tau(s)\big)$, where $\tau : u \mapsto \sqrt{u}$ is obviously a concave function on $\mathbb{R}^+$. Formula (A) applied to $q_i\big(\boldsymbol{\beta}, \tau(s)\big)$ implies that q_i is convex in $\big(\boldsymbol{\beta}, \tau(s)\big)$. This also holds if we take $s_i = \sqrt{A + D_i}$, for all $i = 1, \ldots, n$. Moreover, since convexity is preserved under non-negative weighted addition, $Q(\boldsymbol{\beta}, A) = \sum_{i \le n} q_i$ is convex in $(\boldsymbol{\beta}, A)$. This implies that a solution to $\min\{Q(\boldsymbol{\beta}, A) : \boldsymbol{\beta} \in \mathbb{R}^q,\, A \in \mathbb{R}^+\}$ is a global minimum, provided a minimum exists. ■

Remarks. (i) The convexity result related to Formula (A) can be proved without assuming differentiability of the functions a, b, and c; see Boyd and Vandenberghe (2004, Sect. 3.2.4).

(ii) The argument related to (A) in the proof may seem like an "unnecessary complication". Why can't we compute the partial derivatives of q_i right away? We could do so, however, the square root in the term $\sqrt{A + D_i}$ complicates this approach considerably.

(iii) An alternative demonstration of (6.62) is as follows. Let f denote the linear predictor (omitting the index i), where f is regarded as function of a real t. Unlike Huber's assumptions, the σ is not assumed to depend linearly on some real t. The first-order partial derivatives of the summands in (6.60) are (omitting the index i)

$$\frac{\partial}{\partial \sigma} q = \kappa - \rho_k'\left(\frac{y-f}{\sigma}\right)\left(\frac{y-f}{\sigma}\right) + \rho_k\left(\frac{y-f}{\sigma}\right), \qquad \frac{\partial}{\partial f} q = -\rho_k'\left(\frac{y-f}{\sigma}\right)\dot{f},$$

with second-order derivatives,

$$\frac{\partial^2}{\partial \sigma^2} q = \frac{1}{\sigma}\rho_k''\left(\frac{y-f}{\sigma}\right)\left(\frac{y-f}{\sigma}\right)^2, \qquad \frac{\partial^2}{\partial \sigma \partial f} q = \rho_k''\left(\frac{y-f}{\sigma}\right)\frac{\dot{f}}{\sigma}\frac{y-f}{\sigma}$$
$$\frac{\partial^2}{\partial f^2} q = \frac{1}{\sigma}\rho_k''\left(\frac{y-f}{\sigma}\right)(\dot{f})^2 - \rho_k'\left(\frac{y-f}{\sigma}\right)\ddot{f},$$

hence, the Hessian matrix satisfies (since $\ddot{f} \equiv 0$ under the linear model)

$$\frac{1}{\sigma}\rho_k''\left(\frac{y-f}{\sigma}\right)\begin{bmatrix} \dot{f} \\ (y-f)/\sigma \end{bmatrix}\begin{bmatrix} \dot{f} \\ (y-f)/\sigma \end{bmatrix}^T \ge 0, \tag{6.64}$$

where the last inequality is implied by $\rho_k''(u) \ge 0$ for all $u \in \mathbb{R}$ and positive definiteness of the quadratic form in (6.64).

Proposition 6.6. *Consider the M-estimator defined in (6.56) and suppose that*

(i) $\sum_{i \le n} \mathbf{x}_i \mathbf{x}_i^T$ *has full column rank,*

(ii) not all terms $y_i - \mathbf{x}_i^T \boldsymbol{\beta}$, $i = 1, \ldots, g$, *are zero (this excludes perfect prediction),*

(iii) the D_i's (i.e., variances of the direct estimators) satisfy $D_i > 0$, $i = 1, \ldots, n$,

(iv) the parameter space of the model is to the form $\Theta = \Theta_{\boldsymbol{\beta}} \times \Theta_A$, where $\Theta_{\boldsymbol{\beta}}$ is a compact and convex subset of $\mathbb{R}^q$; $\Theta_A = [0, a]$, where $a > 0$ is a constant that can be chosen arbitrarily large.

(v) the tuning constant k is such that $k > 0$.

Then, the M-estimate $(\hat{\boldsymbol{\beta}}, \hat{A})$ defined as the solution of the system in (6.56) exists.

Proof. Convexity of the objective function $Q(\boldsymbol{\beta}, A)$ [see Proposition 6.5 and hypothesis (iv)] implies that the first-order conditions, i.e. the system of estimating equations in (6.56), are sufficient for optimality (see e.g. Boyd and Vandenberghe, 2004, chap. 3.1.3). Let $r_i(\boldsymbol{\beta}, A)$ denote r_i given in (6.51), and define the map

$$\mathbf{f}_1(\boldsymbol{\beta}, A) = \left(\sum_{i=1}^{n} u\big(r_i(\boldsymbol{\beta}, A)\big) \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sqrt{A + D_i}} \right)^{-1} \sum_{i=1}^{n} u\big(r_i(\boldsymbol{\beta}, A)\big) \frac{\mathbf{x}_i y_i}{\sqrt{A + D_i}},$$

where $u(r) = \psi_k(r)/r$. Observe that on the domain Θ, $\mathbf{f}_1$ is well-defined under hypotheses (i) and (ii). Let f_2 denote the function of the fixed-point equation

$$\sum_{i=1}^{n} \frac{\kappa}{\sqrt{A + D_i}} = \sum_{i=1}^{n} \frac{\psi_k^2\big(r_i(\boldsymbol{\beta}, A)\big)}{\sqrt{A + D_i}} \qquad \Leftrightarrow \qquad f_2(\boldsymbol{\beta}, A) = A.$$

Function f_2 is well-defined on Θ under hypotheses (i) and (iii). Consider the map

$$\mathbf{T}(\boldsymbol{\beta}, A) = \begin{bmatrix} \mathbf{f}_1(\boldsymbol{\beta}, A) \\ f_2(\boldsymbol{\beta}, A) \end{bmatrix} \tag{A}$$

and observe that the M-estimator defined in (6.56) can equivalently be defined as the solution $(\hat{\boldsymbol{\beta}}, \hat{A})$ to the fixed-point equations

$$\mathbf{T}(\hat{\boldsymbol{\beta}}, \hat{A}) = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{A} \end{bmatrix}. \tag{B}$$

It is easy to see that $\mathbf{T}$ is continuous on Θ. By Brouwer's fixpoint theorem (see e.g. Ortega and Rheinboldt, 1970, Thm. 6.3.2), we conclude that (B) has a fixed point. ■

Remarks. (i) Together, Propositions 6.5 and 6.6 assert (under minimal conditions) the existence of a unique M-estimate as the solution to (6.56).

(ii) The key to Proposition 6.6 is hypothesis (iii). Since all D_i's are strictly positive, the M-estimate is properly defined even if the unknown variance parameter A is truly equal to zero. This property enables us to take the parameter space of A, Θ_A, equal to a closed interval of the positive reals that includes zero, namely, $[0, a]$, where $a > 0$ can be chosen as large as needed.

This together with the assumption of a compact and convex parameter space $\Theta_{\boldsymbol{\beta}}$ implies that $\Theta = \Theta_{\boldsymbol{\beta}} \times \Theta_A$ is a compact and convex set. These assumptions [together with a continuous mapping] are sufficient conditions for Brouwer's fixpoint theorem to apply.

(iii) Is the assumption that all D_i's satisfy $D_i > 0$ restrictive? No, not at all. If a particular D_i were indeed equal to zero, there are good reasons to remove the respective area from the model as $D_i = 0$ implies that the direct estimator has infinite precision (hence, there is no need for SAE estimation).

Computation of $\hat{\boldsymbol{\beta}}$ and $\hat{A}$

Solving the estimating equation of $\boldsymbol{\beta}$ defined in (6.56) essentially boils down to computing an M-estimator of regression. Early algorithms on robust regression have been studied in Huber (1973, sec. 8), but the major contributions are due to Dutter (1977a,b); credit goes also to A. Marazzi for developing effective implementations in the `ROBETH` library (see Marazzi, 1993, sec. 2). Two candidate algorithms proved to be particularly useful: one that modifies the residuals (`H`-algorithm), the other one modifies the weights (`W`-algorithm) in each iteration step.

Our choice is a variant of the `W`-algorithm. Let $\{\boldsymbol{\beta}^{\{t\}} \in \mathbb{R}^q : t = 0, 1, 2, \ldots\}$ be a sequence, $\boldsymbol{\beta}^{\{0\}}$ denoting the starting value. We write $r_i^{\{t\}}$ to mean r_i defined in (6.51) with $\boldsymbol{\beta}^{\{t\}}$ substituted for $\boldsymbol{\beta}$. Define the weight function

$$u(r) = \frac{\psi_k(r)}{r} \qquad \text{for } r \in \mathbb{R}, \tag{6.65}$$

having suppressed the dependency on k. We take $\boldsymbol{\beta}^{\{0\}}$ to be the weighted least squares estimate (with weights equal to $1/\sqrt{D_i}, i = 1, \ldots, n$). For fixed A and all $t = 1, 2 \ldots$, updated estimates are computed by

$$\boldsymbol{\beta}^{\{t+1\}} = \left(\sum_{i=1}^{n} u\big(r_i^{\{t\}}\big) \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sqrt{A + D_i}} \right)^{-1} \sum_{i=1}^{n} u\big(r_i^{\{t\}}\big) \frac{\mathbf{x}_i y_i}{\sqrt{A + D_i}}. \tag{6.66}$$

The updating scheme in (6.66) is terminated when $\|\boldsymbol{\beta}^{\{t+1\}} - \boldsymbol{\beta}^{\{t\}}\|_2$ is smaller than some predetermined number (e.g. 10^{-5}). If the updating procedure terminates within a finite number of steps, we take $\boldsymbol{\beta}^{\{t+1\}}$ to be the M-estimate $\hat{\boldsymbol{\beta}}$, otherwise we report failure of convergence.

For numerical computations, the updating rule in (6.66) is not well suited (because of the matrix inversion); therefore, we shall consider the well-known QR-decomposition algorithm to solve the (iterated) least squares problem in (6.66). For ease of discussion, we shall neglect the iterative nature of the updating rule in (6.66) and focus on computing of $\boldsymbol{\beta}$ (omitting the superscripts $\{t\}$ and $\{t+1\}$).

Put

$$\tilde{\mathbf{x}}_i = \left(\frac{u(r_i)}{A + D_i}\right)^{1/2} \mathbf{x}_i, \qquad \tilde{y}_i = \left(\frac{u(r_i)}{A + D_i}\right)^{1/2} y_i$$

and define the $(n \times q)$ matrix $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1^T, \ldots, \tilde{\mathbf{x}}_n^T)^T$ and the n-vector $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T$. The QR-decomposition (Gentle, 2007, 188–189 and 226) of matrix $\tilde{\mathbf{X}}$ is defined as

$$\tilde{\mathbf{X}} = \mathbf{QR}, \tag{6.67}$$

where $\mathbf{R} = (\mathbf{R}_1^T, \mathbf{0}^T)^T$ with an $(q \times q)$ upper triangular matrix $\mathbf{R}_1$ and an $(n \times n)$ orthogonal matrix $\mathbf{Q}$. Matrix $\mathbf{Q}$ features the partition $(\mathbf{Q}_1, \mathbf{Q}_2)^T$, where $\mathbf{Q}_1$ is an $(n \times q)$ matrix whose columns are orthogonal. With this, the decomposition in (6.67) can be expressed more economically as $\tilde{\mathbf{X}} = \mathbf{Q}_1\mathbf{R}_1$; hence, the linear system of interest, $\tilde{\mathbf{X}}\boldsymbol{\beta} = \tilde{\mathbf{y}}$, can be written as

$$\mathbf{R}_1\boldsymbol{\beta} = \mathbf{Q}_1^T\tilde{\mathbf{y}} \quad \Leftrightarrow \quad \boldsymbol{\beta} = \mathbf{R}_1^{-1}\mathbf{Q}_1^T\tilde{\mathbf{y}} \tag{6.68}$$

which is easy to solve since $\mathbf{R}_1$ is a $(q \times q)$ triangular matrix. In the iterative updating scheme, we consider solving (6.68) for each $\boldsymbol{\beta}^{\{t+1\}}$, $t = 1, 2, \ldots$

Computation of $\hat{A}$

Although the M-estimator of A in (6.56) shows similarities with Proposal 2 of Huber (1964), we cannot directly relate to Huber's proposal for numerical computations since the estimating equation for A depends on the possibly different D_i's. However, it is easy to see that we can use a modification of the Fisher scoring algorithm discussed in Rao (2003, 119) for computing the m.l.e. of A. Let $\{A^{\{t\}} \in \mathbb{R}^+ : t = 0, 1, 2, \ldots\}$ be a sequence, $A^{\{0\}}$ denoting the starting value which is taken to be $A^{\{0\}} = 0$. We write $r_i^{\{t\}}$ to mean r_i defined in (6.51) with $A^{\{t\}}$ substituted for A. For fixed $\boldsymbol{\beta}$ and all $t = 1, 2 \ldots$, updated estimates are obtained by

$$A^{\{t+1\}} = \max\left(0,\, A^{\{t\}} + \frac{S}{J}\right) \tag{6.69}$$

where

$$S = \frac{1}{2}\sum_{i=1}^{n} \frac{1}{A^{\{t\}} + D_i}\left(\frac{\psi_{k_2}^2\left(r_i^{\{t\}}\right)}{\kappa} - 1\right)$$

$$J = \frac{1}{2}\sum_{i=1}^{n} \frac{1}{(A^{\{t\}} + D_i)^2}.$$

The updating scheme in (6.69) is terminated if $|A^{\{t+1\}} - A^{\{t\}}|$ is smaller than a predetermined number (e.g. 10^{-5}). If the updating procedure terminates within a finite number of steps, we take $A^{\{t+1\}}$ to be the M-estimate $\hat{A}$, otherwise we report failure of convergence.

The updating rule in (6.69) can also be applied (in a slightly modified version) to compute the robustification of the approximate REML-type estimator and the

Fay–Herriot estimator defined in, respectively, (6.58) and (6.59).

Algorithm

For the $\boldsymbol{\beta}$-updating rule in (6.66) to work, it is assumed that A is fixed; and vice versa, $\boldsymbol{\beta}$ is fixed in the computation of (6.69). The main algorithm *alternates* between (6.66) and (6.69), and updates trial values for $\hat{\boldsymbol{\beta}}$ and $\hat{A}$.

For regression M-estimators with monotone ψ-function, Dutter (1977a) proved convergence of the `W`-algorithm. His proof and a similar convergence proof for the `H`-algorithm are discussed in Huber (1981, chap. 7.8). P.J. Huber shows that the M-estimator objective function $\bar{Q}(\boldsymbol{\beta}, \sigma)$ defined in (6.60) is decreased at each (alternating) updating step. Does decreasing the objective function eventually imply that we attain the minimum? Yes, it does—under ideal circumstances. Though, to rigorously prove this claim, Huber (1981, 188–192) goes through quite some effort in order to rule out an accumulation point at $\sigma = 0$. Such an accumulation point prevents the algorithm from proper convergence since it renders the objective function undefined. The M-estimator under the Fay–Herriot model given in (6.56), on the other hand, does not suffer from an accumulation point problem (even if A equals exactly zero) provided the D_i's are strictly positive. In this regard, we conjecture that a convergence proof for the `H`- and / or `W`-algorithm under the Fay–Herriot model could be obtained under technically less restrictive conditions than Huber (1981, 188–92).

M-estimator rule

The M-estimator type rule of the ith coordinate is given by

$$\hat{\delta}_i^{EM,k} = y_i - \frac{D_i}{\sqrt{\hat{A} + D_i}} \psi_k\left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}}{\sqrt{\hat{A} + D_i}}\right), \qquad i = 1, \ldots, n. \tag{6.70}$$

This rule derives from the rule defined in (6.45) by the substitution of the M-estimates $\hat{\boldsymbol{\beta}}$ and $\hat{A}$ for the unknown parameters $\boldsymbol{\beta}$ and A. The tuning constant k associated with the ψ-function in (6.70) can be chosen different from the tuning constants used in the M-estimators of the parameters $\boldsymbol{\beta}$ and A. Here, the tuning constant controls the amount of translation from y_i whence it limits the maximum component risk. As a rule, the smaller we choose k, the less shrinkage is applied.

6.4.1.2. GM-estimators

In Section 6.4.1.1, we have shown that the joint M-estimator of the parameters $(\boldsymbol{\beta}, A)$ is the m.l.e. under the marginal distribution in (6.46), which follows from

model [for any given $\epsilon \in (0,1)$]

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sqrt{A + D_i} E_i, \qquad i = 1, \ldots, n,$$

with $E_i \overset{\text{ind.}}{\sim} \hat{G}_\epsilon$, where $\hat{G}_\epsilon$ is defined in (6.44). The vectors $\mathbf{x}_i$, $i = 1, \ldots, n$, constituting the design matrix were regarded as known constants.

Here, we shall assume that the *non-constant* elements of the vectors $\mathbf{X}_i$ [i.e., design vector without the regression intercept and other categorical variables] are *stochastic* and have law $\mathbf{X}_i \overset{\text{i.i.d.}}{\sim} H$, $i = 1, \ldots, n$, with density h w.r.t. the Lebesque measure. In order to account for possibly long-tailed distributions, we assume that ($|\mathbf{A}|$ denoting the determinant of matrix $\mathbf{A} \in \mathbb{R}^{q \times q}$)

$$h(\mathbf{x}_i) = |\boldsymbol{\Omega}_x|^{-1/2} h_0\big(\boldsymbol{\Omega}_x^{-1/2}(\mathbf{x}_i - \boldsymbol{\mu}_x)\big), \qquad i = 1, \ldots, n,$$

where $\boldsymbol{\mu}_x \in \mathbb{R}^q$ is an unknown location vector, $\boldsymbol{\Omega}_x$ is an unknown positive semi-definite $(q \times q)$ scatter matrix of rank q, and h_0 is the density of a spherically symmetric distribution (the characteristic generator; Fang, Kotz, and Ng, 1990, 28–29). The distribution H is said to be an elliptically symmetric or elliptically contoured distribution function. Under the additional assumption that the r.v.'s E_i are conditionally independent of $\mathbf{X}_i$, $i = 1, \ldots, n$ (viz. Hampel et al., 1986, 308), the *joint* marginal distribution of $(\mathbf{y}, \mathbf{X})$ is

$$m(\mathbf{y}, \mathbf{X} \mid \boldsymbol{\beta}, A, \boldsymbol{\mu}_x, \boldsymbol{\Omega}_x) = \prod_{i=1}^{n} m_0(y_i, \mathbf{x}_i \mid \boldsymbol{\beta}, A) \tag{6.71}$$

where

$$m_0(y_i, \mathbf{x}_i \mid \boldsymbol{\beta}, A, \boldsymbol{\mu}_x, \boldsymbol{\Omega}_x) = \frac{1}{\sqrt{A + D_i}} \hat{g}_\epsilon\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{A + D_i}}\right) h(\mathbf{x}_i).$$

The joint GM-estimator of the parameters $(\boldsymbol{\beta}, A)$ obtains as the m.l.e. under the joint marginal distribution. For the moment, we shall assume that the location and scatter of the r.v.'s $\mathbf{X}_i$, $\boldsymbol{\mu}_x$ and $\boldsymbol{\Omega}_x$, are known. Denote by $d^2 : \mathbb{R}^q \to \mathbb{R}^+$ the squared Mahalanobis distance function, which is defined as

$$d^2(\mathbf{u}) = (\mathbf{u} - \boldsymbol{\mu}_x)^T \boldsymbol{\Omega}_x^{-1} (\mathbf{u} - \boldsymbol{\mu}_x), \qquad \text{for } \mathbf{u} \in \mathbb{R}^q. \tag{6.72}$$

Let $\omega : \mathbb{R} \to [0, 1]$ be a weight function. For instance, we may take

$$\omega(u) = \frac{\psi_{k_x}(u)}{u}, \tag{6.73}$$

where ψ_{k_x} is the Huber ψ-function indexed by the tuning constant k_x. Another weight function obtains if we replace ψ_{k_x} by the Tukey bisquare function ψ_b (or ψ_{b_x} if we wish to highlight that the tuning constant b refers to downweighting in the design space; see Section 2.3.1 for a definition of ψ_b).

The modified Mallows-type regression GM-estimator of $(\boldsymbol{\beta}, A)$ is defined as

the solution to the system of estimating equations

$$\left.\begin{aligned} \sum_{i=1}^{n} \omega\big(d(\mathbf{x}_i)\big)\psi_k(r_i)\frac{\mathbf{x}_i}{\sqrt{A+D_i}} &= \mathbf{0} \\ \sum_{i=1}^{n} \frac{\omega\big(d(\mathbf{x}_i)\big)\psi_k^2(r_i) - \kappa}{A+D_i} &= 0 \end{aligned}\right\}, \tag{6.74}$$

where κ is defined in (6.55).

Remarks. (i) In practical applications, the location $\boldsymbol{\mu}_x$ and scatter matrix $\boldsymbol{\Omega}_x$ are unknown. We can get estimated distances, $\hat{d}(\mathbf{x}_i)$, by substituting robust estimates, $\hat{\boldsymbol{\mu}}_x$ and $\hat{\boldsymbol{\Omega}}_x$, for the unknown parameters in (6.72). It is important that robust estimates are substituted (see below), for otherwise the estimated distances may suffer from two problems: masking and swamping (see e.g. Rousseeuw and van Zomeren, 1990).[5]

(ii) Whether the weight function in (6.73) is based on the Huber or the Tukey bisquare ψ-function can make a great difference. By its monotonicity, the Huber ψ-function ensures a positive $\omega(u)$ for all $u \in \mathbb{R}$. Hence, any observation contributes irrespective how far away it is from the bulk of data. This is not the case for the Tukey bisquare ψ-function, which by its redescending nature implies $\omega(u) = 0$ for all $|u| > b$. Thus, observations with Mahalanobis distances larger than b are assigned a weight of zero [i.e., are discarded].

(iii) We are free to choose different (monotone) ψ-functions in the definition of the estimating equations in (6.74).

(iv) The estimator defined as the solution to the system (6.74) derives from the Mahalanobis-distance (MD) approach to Mallows-type GM-regression estimators (see e.g. Maronna and Yohai, 1981, Sect. 5). In place of MD, one may use other "distance measures". In fact, early GM-estimators relied on influence measures other than MD. The (original) Mallows-type estimator (Mallows, 1973) builds on the following weight function

$$\omega(\mathbf{x}_i) = \frac{\psi_k(u_i)}{u_i}, \qquad \text{where} \qquad u_i = \big(\mathbf{x}_i^T \hat{\mathbf{A}}^{-1}\mathbf{x}_i\big)^{1/2}$$

with

$$\hat{\mathbf{A}} = \frac{\bar{\kappa}}{n}\sum_{i=1}^{n} \omega^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T \qquad \text{and} \qquad \bar{\kappa} = \frac{\kappa}{n}\sum_{i=1}^{n}\omega(\mathbf{x}_i).$$

[5] *Masking:* Outliers can distort the classical estimates of mean and covariance in such a way (e.g. via inflating the covariance matrix) that they do not get necessarily large values of the Mahalanobis distance. *Swamping:* Outliers can distort the classical estimates of mean and covariance in such a way that observations which are consistent with the majority of the data get large values for the Mahalanobis distance.

It is easily recognized that the u_i's are closely related to the diagonal elements of the "hat" matrix, say h_i, in the least squares problem. Actually, the Schweppe-type estimator (Handschin, Kohlas, Fiechter, and Schweppe, 1975) uses the h_i's directly; see also Krasker and Welsch (1982). The major advantage of our MD-based approach is that the resulting estimator has nice invariance properties under linear transformations of the $\mathbf{x}_i$'s; this is not the case for the original Mallows-type estimator.

(v) Another remark concerns the estimator of A in (6.74). The questions is whether the ω-weighting also applies to the estimating equation of A or not. Let us consider the (standard) Mallows-type regression GM-estimator of the tuple $(\boldsymbol{\beta}, \sigma)$, $\sigma > 0$ denoting the regression scale. In this case, Maronna et al. (2006, 148) do not consider such a weighting, but others do, see e.g. Samarov and Welsch (1982), Krasker and Welsch (1982) or Marazzi (1986). Let ρ_k be a ρ-function defined in (6.48) and write $\omega(\mathbf{x}_i)$ instead of $\omega(d(\mathbf{x}_i))$ for ease of discussion. The aforementioned standard estimator can be regarded as the (unique) minimizer of the optimality criterion (see e.g. Samarov and Welsch, 1982, 412–13)

$$\kappa\sigma + \sum_{i=1}^{n} \sigma\omega(\mathbf{x}_i)\rho_k\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right), \tag{6.75}$$

and taking the partial derivate w.r.t. σ leads to the estimating equation for σ,

$$\frac{1}{n}\sum_{i=1}^{n} \omega(\mathbf{x}_i)\left[\psi_k^2\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right) - \kappa\right] = 0.$$

Indeed, referring to (6.75), it is evident that the estimating equation for σ should include the ω-weighting. The very same argument applies also to the system of estimating equations in (6.74).

(vi) The estimator of A defined as the solution to (6.74) can be seen as a robustification of the m.l.e. It is straightforward to derive a GM-type approximate REML-estimator, a GM-type FH-estimator, etc. see Section 6.4.1.1.

Properties of the GM-estimator

The properties of the GM-estimator (existence and uniqueness) are similar to the ones we discussed in case of the M-estimator; see Section 6.4.1.1. Consistence and asymptotic normality can be proved along the lines of Maronna and Yohai (1981, Sect. 5).

Robust estimates of $d(\mathbf{x}_i)$, respectively, $\boldsymbol{\mu}_x$ and $\boldsymbol{\Omega}_x$

We shall assume that the realizations $\mathbf{x}_i$ consist of only non-constant measurements over the index set $\{1, \ldots, n\}$. This implies that the intercept term and other categorical variables in the model's design matrix must be dropped; the

so reduced design matrix is assumed to have dimension $(n \times q)$ and full column rank. With regard to the dimension q of the vectors $\mathbf{x}_i$, two cases can be distinguished.

Case $q = 1$. The estimated distances are defined as

$$\hat{d}(x) = \frac{|x - \operatorname{med}_{i=1,\dots,n}\{x_i\}|}{\operatorname{mad}_{i=1,\dots,n}\{x_i\}}, \qquad \text{for } x \in \mathbb{R}, \tag{6.76}$$

where mad is the median absolute deviation from the median (scaled to be consistent at the Gaussian core model) and med denotes the sample median.

Case $q > 1$. We consider two approaches.

(i) *M-estimator of location and scatter.* Let $\hat{d}_i = d(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\Omega}}_x)$ denote the Mahalanobis distance of $\mathbf{x}_i$, $i = 1, \dots, n$, given the trial values $\hat{\boldsymbol{\mu}}_x$ and $\hat{\boldsymbol{\Omega}}_x$. The joint M-estimator of location and scatter obtains directly from the marginal distribution (6.71) and is the solution $(\hat{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\Omega}}_x)$ to the system of estimating equations (cf. Maronna, 1976, 52)

$$\left.\begin{aligned} \sum_{i=1}^{n} w_1(\hat{d}_i)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x) &= \mathbf{0} \\ \frac{1}{n}\sum_{i=1}^{n} w_2(\hat{d}_i)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x)^T &= \hat{\boldsymbol{\Omega}}_x \end{aligned}\right\}, \tag{6.77}$$

where w_1 and w_2 are known real-valued functions (not necessarily equal). When $dw_2(d)$ is a monotone nondecreasing function of $d \in \mathbb{R}^+$, the solution of the system in (6.77) exists and is unique (Maronna, 1976, Thm.'s 2, 3, and 4). Huber (1981, chap. 8.6 and 8.8) studies a slightly more general definition of monotone M-estimators, which is not consequential for practical purposes, and proves consistency of such monotone M-estimators.

(ii) *Estimation of covariance matrix elements through robust variances*. The simplest approach is to apply a robust location estimator to each coordinate and a robust estimator of covariance to each pair of variables. Unfortunately, the resulting dispersion matrix of such pairwise robust estimators may lack affine equivariance and positive definiteness (Huber, 1981, 203). However, the gained reduction in computation time may offset the drawbacks.

Gnanadesikan and Kettenring (1972) proposed a robust pairwise covariance based on the identity

$$\operatorname{cov}(y, x) = \frac{1}{4ab}\big(\operatorname{var}(ax + by) - \operatorname{var}(ax - by)\big),$$

where a and b are arbitrary constants. When one replaces var by S^2, where S is a robust estimator of scale (e.g., trimmed standard deviation), and

standardizes appropriately, one obtains a kind of robust correlation, ρ. Scaling ρ by $S(x)$ and $S(y)$ yields a robust estimator of covariance that is scale equivariant and behaves at the bivariate Gaussian model like the regular covariance estimator (Huber, 1981, 203). Moreover, the resulting robust scatter matrix is consistent if the underlying distribution is elliptically contoured (Maronna et al., 2006, 225–26).

To overcome (some of) the drawbacks of the pairwise estimation approach, Maronna and Zamar (2002) proposed a modification of the Gnanadesikan–Kettenring procedure that ensures a positive definite matrix and "approximately equivariant" estimates of location and scatter. This method is called orthogonalized Gnanadesikan–Kettenring (OGK) estimator. The OGK method also provides a robust estimate of location. Affine equivariant estimators of $(\boldsymbol{\mu}_x, \boldsymbol{\Omega}_x)$ are necessary for the GM-type regression estimators to be affine equivariant (Maronna et al., 2006, 150).

Remark. The so to speak "weak part" of the joint M-estimator of $(\boldsymbol{\mu}_x, \boldsymbol{\Omega}_x)$ is the estimate $\hat{\boldsymbol{\Omega}}_x$. When $\boldsymbol{\mu}_x$ is supposed known, Maronna (1976, 63–64) shows that the asymptotic breakdown point (BP) of the M-estimator of $\boldsymbol{\Omega}_x$ with monotone w_2-function is bounded from above by $BP \leq 1/(q+1)$. From this it is evident that the BP can become dangerously low for large q. The BP of the OGK-algorithm is considerably better, though worse than the BP achievable with specific high breakdown-point estimators; see Maronna et al. (2006, chs. 6.4 and 6.5). Yet, for our purposes, breakdown-point considerations are of minor concern. Much more important is computational stability. In this respect, joint M-estimators and OGK are reasonable choices.

The most popular algorithm to solve the M-estimator estimating equations in (6.77) is an iterated fixed-point algorithm (see e.g. Huber, 1981, chap. 8.11). Dümbgen, Nordhaus, and Schumacher (2016) proposed a new algorithm which builds on a second-order Taylor series expansion of the target functional and devises a partial Newton–Raphson procedure. Their method proves to be considerably faster than the fixed-point method in a lot of cases, and is implemented in the `R` package `fastM` (Dümbgen, Nordhaus, and Schumacher, 2014). We use their method to compute M-estimators of location and scatter. Regarding OGK, we apply the algorithm implemented in the `R` package `rrcov` (Todorov, 2016; Todorov and Filzmoser, 2009). Both methods, M-estimator and OGK, can be specifically tuned to provide a good variance–robustness trade-off. Our experiences show that both method do a reasonably good job in "plain vanilla" mode; nonetheless, we recommend to tune them slightly "pessimistically" (i.e. sacrificing some efficiency).

Algorithm and implementation

The algorithm for computing GM-estimators under the FH model derives di-

rectly from the regression GM-estimator; see e.g. Samarov and Welsch (1982) or Marazzi (1986). In fact, the numerical methods are identical to the algorithm discussed for the M-estimator (see Section 6.4.1.1), except that the ω-weighting (outliers in the design space) is included. Put the other way round, M-estimators are GM-estimators with $\omega \equiv 1$.

GM-estimator rule

The GM-estimator limited translation rule of the ith coordinate is

$$y_i - \frac{D_i}{\sqrt{\hat{A} + D_i}} \psi_k \left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}}{\sqrt{\hat{A} + D_i}} \right) \omega\big(\hat{d}(\mathbf{x}_i)\big), \qquad i = 1, \ldots, n, \tag{6.78}$$

where the ω-function is defined in (6.73), and $\hat{d}(\mathbf{x}_i)$ is the estimated Mahalanobis distance; see Formula (6.72). This rule differs from the M-estimator type estimating rule in (6.70) by the multiplicative weighting term $\omega\big(\hat{d}(\mathbf{x}_i)\big)$. The ω-weight takes values close to zero [or even equal to zero when the ψ-function in Formula (6.73) is a redescending ψ-function; e.g., Tukey bisquare] if the configuration of the $\mathbf{x}_i$-variables for the ith coordinate deviates strongly from the central model. In such situations, the shrinkage applied by rule in (6.78) towards location $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ is very minute (or even absent). Such behavior is desirable when the overall model does not fit well for a particular coordinate.

A note of caution

For the GM-type rule in (6.78) to work properly, it is important that the design variables are roughly symmetrically distributed prior to estimation because both, the OGK-method and the joint M-estimator of location and scatter, rely on the assumption of elliptically contoured distributions. We therefore recommend to apply variable-specific symmetrizing transformations when the univariate empirical distributions are noticeably skewed. Our experience shows that the logarithmic transformation (more general, Box–Cox transformation) achieves a sufficient degree of symmetry in a large number of practical applications.

6.4.2. Robust inference

For practical applications, we need an estimator of the mean square prediction error (MSPE) as a measure of uncertainty associated with estimating the area-level means. A large amount of research has been dedicated to analytic MSPE estimation methods in the last three decades, among which second-order approximations proved to be particularly useful; see e.g. Prasad and Rao (1990), Datta and Lahiri (2000) or Datta et al. (2005). Besides analytic methods, resampling MSPE-estimators (e.g., jackknife or bootstrap) are commonly used; see e.g. Rao (2003, chap. 9.2) or Gershunskaya, Jiang, and Lahiri (2009, sec. 7) for a summary. However, the mentioned MSPE-estimators should not be used in case

of contaminated data because they are inherently non-robust.

Since the Fay–Herriot model can be regarded as a degenerate MLM with block-diagonal covariance matrix, we may use the robust EBLUP methods that were originally developed under the unit-level model and apply them appropriately modified. In this vein, Warnholz (2016b, chap. 3.5.2) suggested a modification of the conditional MSPE-estimator due to Chambers, Chandra, and Tzavidis (2011), which can be expressed in pseudo-linear form (i.e., expressed as a weighted sum of sample values). The (pseudo-) linearity of the MSPE-estimator simplifies analytic computations considerably, albeit there is a price to pay: the estimator is an approximation to the true MSPE and tends to underestimate; it shows also a rather high variability. The underestimation results because the estimator ignores the contribution arising from the variability in the estimation of the variance parameter; a disadvantage that is not encountered with the modified method due to Chambers et al. (2014). The method proposed by S. Warnholz also suffers from underestimation.

A more promising approach obtains from an adaption of the parametric bootstrap (see Hall and Maiti, 2006, sec. 2.5) that has been proposed by Sinha and Rao (2009) under the unit-level model; see also Section 5.5. A major objection against bootstrap procedures has been the time-consuming computational effort required to compute the estimates. However, this objection does not really apply here because our software implementation is fast enough to be useful in practical applications. A parametric bootstrap estimate of the MSPE (for each area $i = 1, \ldots, n$) can be obtained as follows. Let $\hat{\delta}_i$ denote the estimate / prediction of the ith coordinate (or area-level mean) θ_i, $i = 1, \ldots, n$. The $\hat{\delta}_i$'s are based on the parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{A}$. The goal is to compute

$$\mathrm{MSPE}(\hat{\delta}_i) = \mathbb{E}\{\hat{\delta}_i - \theta_i\}^2, \qquad i = 1, \ldots, n, \tag{6.79}$$

by a parametric bootstrap, which is specified in three steps.

1. Given $\hat{\boldsymbol{\beta}}$ and $\hat{A}$, draw realizations e_i^* from r.v. $E_i^* \sim \mathcal{N}(0, \hat{A})$ for all $i = 1, \ldots, n$, and let one batch of bootstrap samples be defined as the vector $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)^T$, where
$$y_i^* = \mathbf{X}\hat{\boldsymbol{\beta}} + e_i^*, \tag{6.80}$$
$\mathbf{X}$ denoting the design matrix.

2. Generate $r = 1, \ldots, b$ batches of bootstrap samples $\{\mathbf{y}^{*[1]}, \mathbf{y}^{*[2]}, \ldots, \mathbf{y}^{*[b]}\}$ using model (6.80); the design matrix $\mathbf{X}$ is the same in all bootstrap samples. For each generated bootstrap sample $\mathbf{y}^{*[r]}$ ($r = 1, \ldots, b$), compute sample-specific parameter estimates $\hat{\boldsymbol{\beta}}^{[r]}$ and $\hat{A}^{[r]}$ and compute also the $\hat{\delta}_i^{[r]}$'s, $i = 1, \ldots, n$.

3. For each $i = 1, \ldots, n$, the bootstrap estimate of $\mathrm{MSPE}(\hat{\delta}_i)$ in (6.79) is com-

puted by

$$\frac{1}{b}\sum_{r=1}^{b}\big(\hat{\delta}_i^{[r]} - \hat{\delta}_i\big)^2. \tag{6.81}$$

Remarks. (i) The method tends to slightly underestimate the true MSPE. The underestimation results mainly because the uncertainty of estimating $\boldsymbol{\beta}$ has not been taken into account.

(ii) Warnholz (2016b, chap. 3.5.1) has also studied a parametric bootstrap.

6.5. Simulation

In order to study the behavior of the proposed method in more detail, we consider a small model-based simulation study using the `simFrame` (Alfons, 2014; Templ and Filzmoser, 2010) simulation environment. The data were generated by the model

$$Y_i = \beta_0 + \beta_1 X_i + V_i + E_i, \tag{6.82}$$

where $\beta_0 = \beta_1 = 2$,

$$X_i \sim \mathcal{N}(1, 1), \qquad V_i \sim \mathcal{N}(0, D_i), \qquad \text{and} \qquad E_i \sim \mathcal{N}(0, A = 2). \tag{6.83}$$

The D_i's denote the variances of the direct estimators, and are generated from the folded Gaussian distribution, i.e. using $D_i = |\tilde{D}_i|$ with $\tilde{D}_i \sim \mathcal{N}(0.1, 1)$. The sample size consists of $n = 100$ units generated under the model in (6.82). The Monte Carlo study is based on 1000 replications. For the purpose of comparison, we included the following additional estimating methods in the simulation: `eblupFH` (with variance estimators: reduced m.l.e., `REML`, and the variance estimator proposed by Fay and Herriot, `FH`; both methods are implemented in package `sae` due to Molina and Marhuenda, 2015), and `rfh` (i.e., Sinha–Rao-type M-estimator under the FH model) in package `saeRobust` (Warnholz, 2016a). In total, the following methods are studied:

(i) `FH` and `REML` in package `sae` (Molina and Marhuenda, 2015),

(ii) `M`-estimator (see Section 6.4.1.1),

(iii) two GM-estimators based on estimated robust Mahalanobis distances (using the OGK-method for computing robust estimates of location and scatter; see Section 6.4.1.2); the two GM-estimators differ in terms of the weight functions to downweight influential observations in the design space, namely,

  (a) estimator `GM` is based on the Huber weight function,

  (b) estimator `GM2` uses the Tukey bisquare weight function.

[*Note*: both weight functions are tuned such that they achieve 95% efficiency for an estimator of location at the standard Gaussian distribution; Huber: $k = 1.345$, Tukey bisquare: $b = 4.685$]

(iv) M-estimator `rfh` (denoted `RFH` in our simulation) in package `saeRobust` (Warnholz, 2016a).

The suffixes `_b0`, `_b1`, `_b2`, and `_var` appended to the names of the methods (see below) refer to, respectively, the parameters β_0, β_1, β_2, and the variance parameter A.

For the evaluation of the simulation results, the following numerical robust and non-robust criteria are computed (a rigorous definition of the criteria is given in Appendix B):

(i) `mean` and `var` (variance) of the estimates,

(ii) `med` (median) and `mad` (scaled median absolute deviations from the median) of the estimates,

(iii) percentage of `bias(%)` and `mse(%)` (mean squared error) of the estimates w.r.t to the true value,

(iv) `mare(%)` (maximum relative absolute error from the true value, in percent).

Table 6.3.: Simulation criteria for the scenario of uncontaminated data

	mean	med	var	mad	bias(%)	rmse(%)	mare(%)
a) intercept							
MLE_b0	2.001	2.002	0.018	0.134	0.062	0.917	23.616
M_b0	2.001	2.001	0.019	0.141	0.050	0.936	22.363
GM_b0	2.002	2.003	0.019	0.142	0.081	0.936	22.158
b) slope							
MLE_b1	2.004	2.004	0.017	0.130	0.189	0.859	21.840
M_b1	2.004	2.006	0.017	0.127	0.216	0.866	21.878
GM_b1	2.006	2.006	0.018	0.130	0.319	0.903	22.086
c) variance							
MLE_var	0.981	0.971	0.067	0.252	-1.908	6.722	89.286
M_var	0.941	0.930	0.086	0.305	-5.941	8.991	119.173
GM_var	0.941	0.927	0.086	0.303	-5.852	8.958	119.941

See text for explanations.

6.5.1. Uncontaminated data

For uncontaminated data, we expect the robust methods (with tuning constant $k = 1.345$) to behave like the m.l.e. but with a slightly increased variance (i.e., ef-

ficiency penalty). This is indeed the case as can be seen from Table 6.3. The relative mean square error (`mse(%)`) is slightly lower for `MLE` compared with method `M` (for all estimated parameters). Observe that the relative bias (`bias(%)`) is negative for all variance estimators. This behavior is expected since all variance estimators reported in Table 6.3 are biased by design (they do not explicitly account for the reduction of degrees of freedom that results from estimating the regression parameters). The results for the method `GM2` are not shown in Table 6.3 as they are identical (or virtually identical) with the numerical criteria of estimator `GM`.

The loss in efficiency that results when robust methods are applied to uncontaminated data is comparatively small. With regard to the huge gains obtainable in terms of efficiency and bias reduction when the data are indeed subject to contamination (see below), such small losses can be worthwhile.

6.5.2. Outliers in the response variable

It is seen from the simulation model defined in (6.82) that the r.v.'s V_i and E_i, more precisely their underlying parameters, are not identifiable. In this regard, it is sufficient to consider a contamination scheme that affects only the response variable Y_i. The contaminated samples are generate as follows.

(i) Draw a sample of $n = 100$ observations from the model defined in (6.82).

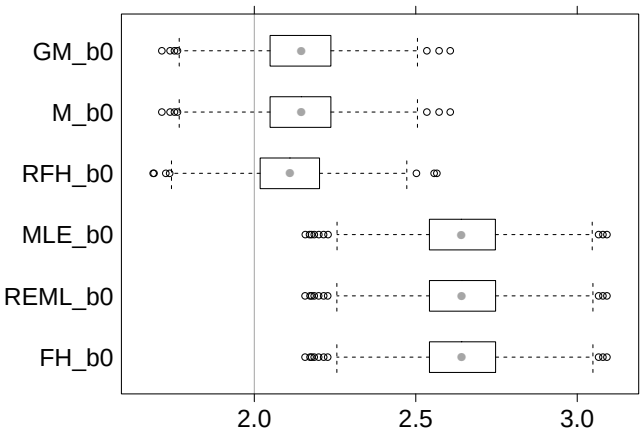
(ii) Randomly select 5% of the observations from the generated sample data, and replace the response variable of the selected entities by a realization drawn from $\mathcal{N}(15, 1)$.

The contamination scheme in use is characterized by the mechanism "outlying completely at random"; see Hulliger et al. (2011, chap. 4.2.).

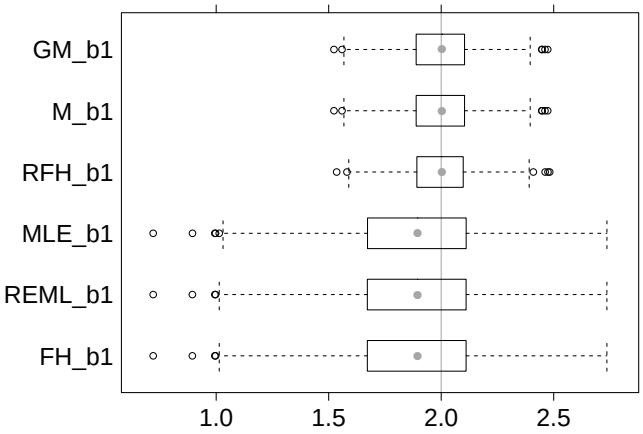
The impact of a contaminated response variable on the estimators and estimands is shown in Figure 6.5 and can be inferred from the numerical criteria reported in Table 6.4. The major findings can be summarized as follows.

- The non-robust estimation methods, `FH`, `MLE`, and `REML`, suffer greatly from contamination. In particular, the estimates of the intercept β_0 and the variance parameter A are heavily biased whereas the estimated slope β_1 show only a small bias. The dispersion of all three estimated parameters is noticeably inflated when compared with the robust methods. None of the methods, `FH`, `MLE`, and `REML`, stands out; they behave virtually the same under contamination (as was to be expected).

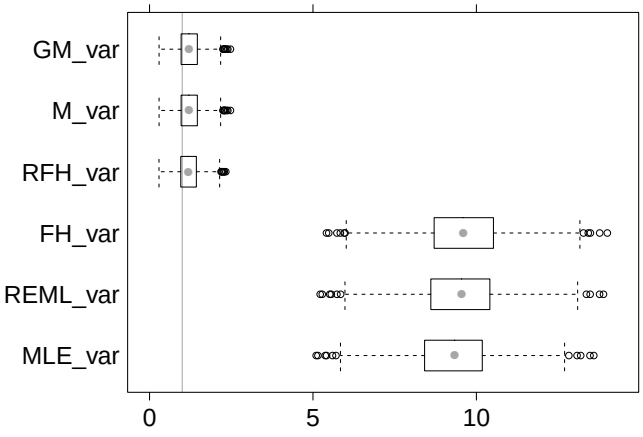
- The robust methods `RFH` (due to Warnholz, 2016b) and `M` are conceptionally identical. The resulting tiny differences in terms of the simulation criteria between the two methods are supposedly due to slight variations in the software implementations, but these differences are not consequential.



(a) Intercept



(b) Slope



(c) Variance

Figure 6.5.: Impact of contaminating the response variable shown for different estimation methods and estimands (5% contamination from $\mathcal{N}(15, 1)$, sample size $n = 100$, number of Monte Carlo replications: 1000; the vertical line marks the true value; see text for further details).

Table 6.4.: Simulation criteria for the scenario of contaminated responses

	mean	med	var	mad	bias(%)	rmse(%)	mare(%)
a) intercept							
FH_b0	2.642	2.642	0.024	0.152	32.088	21.778	54.597
REML_b0	2.642	2.642	0.024	0.152	32.088	21.778	54.597
MLE_b0	2.641	2.641	0.024	0.153	32.069	21.753	54.599
RFH_b0	2.109	2.110	0.019	0.136	5.435	1.541	28.266
M_b0	2.141	2.146	0.020	0.138	7.038	1.984	30.336
GM_b0	2.141	2.146	0.020	0.138	7.038	1.984	30.336
b) slope							
FH_b1	1.882	1.895	0.107	0.328	-5.886	6.023	64.000
REML_b1	1.882	1.895	0.107	0.328	-5.886	6.023	64.000
MLE_b1	1.882	1.896	0.107	0.328	-5.877	6.016	64.002
RFH_b1	1.999	2.003	0.022	0.151	-0.058	1.111	24.109
M_b1	1.997	2.004	0.025	0.160	-0.149	1.235	23.827
GM_b1	1.997	2.004	0.025	0.160	-0.149	1.235	23.827
c) variance							
REML_var	9.534	9.547	1.901	1.347	853.401	7472.839	1288.886
FH_var	9.595	9.598	1.746	1.354	859.511	7562.025	1300.571
MLE_var	9.324	9.339	1.833	1.324	832.385	7111.777	1259.185
RFH_var	1.199	1.185	0.121	0.349	19.938	16.030	133.261
M_var	1.224	1.205	0.130	0.364	22.373	18.016	147.424
GM_var	1.224	1.205	0.130	0.364	22.373	18.016	147.424

See text for explanations.

- The GM-estimator behaves exactly like method `M`.

In summary, when the response variable is subject to contamination, robust methods can – without surprise – do much better in terms of bias and variability than their non-robust counterparts.

6.5.3. Influential observations in the design space

The model defined in (6.82) with only one single non-constant independent variable was appropriate to study the impact of outliers in the response variable. Here, we consider a model with two independent variables which is defined as

$$Y_i = \beta_0 + \mathbf{x}_i^T(\beta_1,\, \beta_2)^T + V_i + E_i, \qquad i = 1, \ldots, n, \tag{6.84}$$

where $\beta_0 = \beta_1 = \beta_2 = 2$, and $\mathbf{x}_i = (x_{i1}, x_{i2})^T$ is drawn from the bivariate Gaussian distribution with variances equal to one and (Pearson) correlation coefficient equal to -0.6. We have chosen correlated predictor variables since such a scenario is – from the perspective of practical applications – much more realistic than assuming a diagonal covariance matrix for the $\mathbf{x}_i$'s. All other parameters are unchanged compared with the situation studied under the simulation model

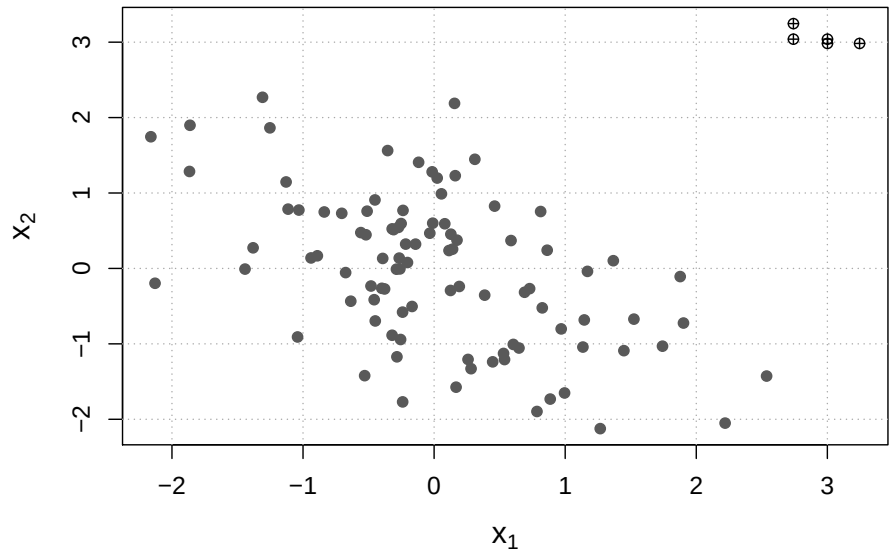


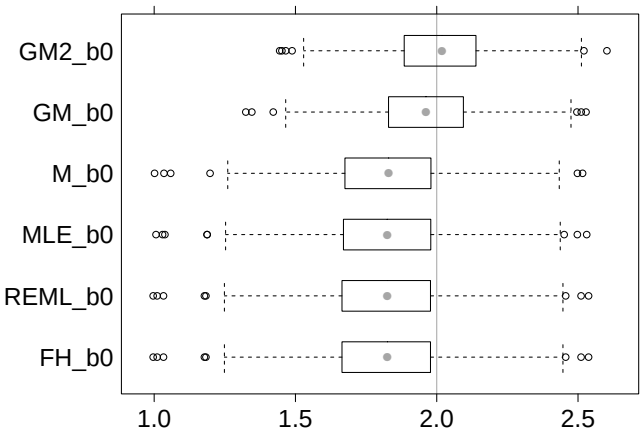
Figure 6.6.: Scatter plot of one instance of a generated design matrix (n observations, two variables, x_1 and x_2); 5% of the observations are denominated as outliers (at random) and then shifted to the upper right corner (outliers are flagged by the symbol $\oplus$).

in (6.82); notably, $n = 100$ (sample size) and the number of Monte Carlo replications equals 1000.

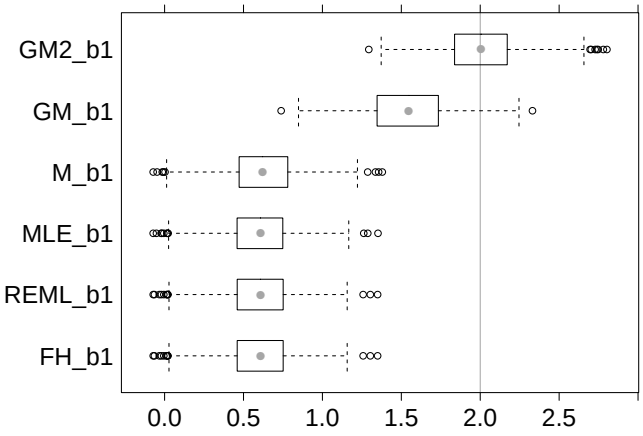
Under the model in (6.84), we want to study the behavior of the estimators in presence of outliers in the design space (i.e., influential observations with a high leverage). Therefore, the design matrix (of the non-constant measurements) $(\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ in (6.84) is contaminated as follows: 5% of the design vectors $\mathbf{x}_i$ are selected at random and replaced by the vector $(z_1, z_2)^T$, where the realizations z_1 and z_2 are drawn independently from $\mathcal{N}(3, 0.01)$. Figure 6.6 shows a scatter plot of the two x-variables for one batch of contaminated data; the outliers are located in the top right corner. The projections of the outlying observations onto the abscissa and the ordinate, respectively, reveal that the projected outliers are not far from the bulk of data. However, the configuration of the outliers in the 2-dimensional space is chosen such that it can have a heavy influence on the regression estimates.

The simulated impact of a contaminated design matrix on the estimators is shown in Figures 6.7 and 6.8, respectively, for the coefficients β_0, β_1, β_2, and the variance parameter A. The numerical criteria are documented in Table 6.5. The major findings can be summarized as follows:

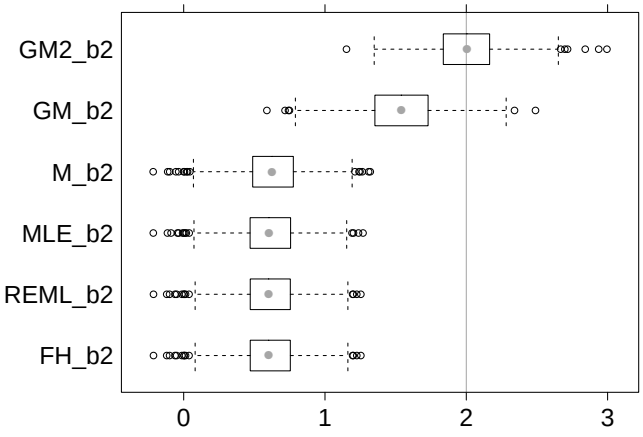
- The non-robust estimation methods, `FH`, `MLE`, and `REML`, are hopelessly biased. The bias is particularly pronounced for estimators of the parameters β_1, β_2 and the variance A, which is self-explanatory since we deliberately



(a) Intercept



(b) Coefficient β_1



(c) Coefficient β_2

Figure 6.7.: Impact of contaminating the design matrix shown for different estimation methods and estimands (5% contamination, sample size: $n = 100$, number of Monte Carlo replications: 1000; the vertical line marks the true value; see text for further details).

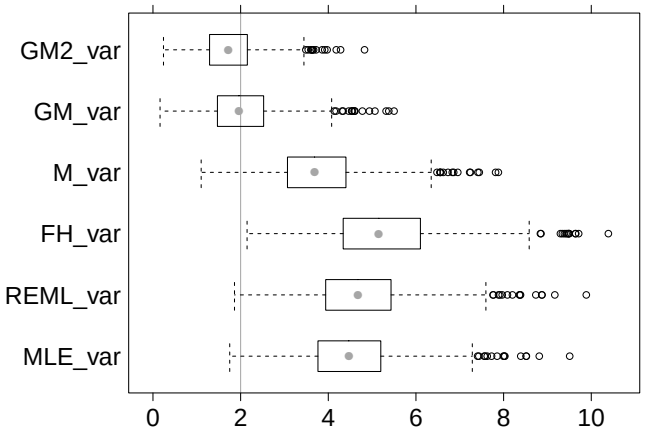


Figure 6.8.: Impact of contaminating the design matrix shown for different estimators of the model variance A (5% contamination, sample size: $n = 100$, number of Monte Carlo replications: 1000; the vertical line marks the true value; see text for further details).

contaminated the covariates vector $\mathbf{x}_i$. Estimates of the intercept parameter β_0, though they are no "directly" affect by contamination, still get influenced through the overall (biased) model fit.

- The M-estimator (see method `M`), though it is robust w.r.t. outliers in the response variable, is biased as much as the methods `FH`, `MLE`, and `REML` under the current contamination regime. This rather bad performance was to be expected as M-estimators are not protected against influential observations in the design space.

- The two GM-estimators under the Fay–Herriot model, `GM` and `GM2`, do much better in terms of bias and variability of the estimates than any other studied methods. The estimators differ from each other in terms of the weight function applied to downweight influential observation in the design space. This difference is also reflected in the behavior of the estimators as can be seen in Figures 6.7 and 6.8 (and also Table 6.5). Much insight into the behavior is gained when we compare how much downweighting the methods apply to the 5% influential observations. In case of the `GM` method (with Huber weight function), outlying observations receive on average a weight of 0.178 [quantiles: $Q_{25} = 0.16$ and $Q_{75} = 0.19$]. The `GM2` method on the other hand applies a weighting based on the Tukey bisquare ψ-function which, by its redescending nature, assigns a value of zero to observations that are far from the bulk of data. This is indeed the behavior we encounter for an outlier placement as shown in Figure 6.6. The outlying observations receive a weight that is virtually zero. Hence, the `GM2` method does not depend on those observations, and remains therefore (almost) unaffected.

- Method `GM2` tends to underestimate the variance slightly; see Figure 6.8. This behavior is expected since the variance estimator of method `GM2` is biased by design (it does not explicitly account for the reduction of degrees of freedom that results from estimating the regression parameters). A simple remedy is to consider the GM-type reduced m.l.e. discussed in the Remark of Section 6.4.1.2.

In summary, GM-estimators do much better in terms of bias and variability than both M-estimators and the regular estimators under the Fay–Herriot model when the design matrix is subject to contamination.

Table 6.5.: Simulation criteria for the scenario of a contaminated design matrix

	mean	med	var	mad	bias(%)	rmse(%)	mare(%)
a) intercept β_0							
FH_b0	1.825	1.826	0.056	0.236	-8.768	4.341	50.176
REML_b0	1.825	1.826	0.056	0.236	-8.768	4.341	50.176
MLE_b0	1.825	1.825	0.055	0.229	-8.762	4.277	49.673
M_b0	1.828	1.830	0.053	0.228	-8.592	4.130	49.929
GM_b0	1.963	1.962	0.037	0.196	-1.825	1.939	33.783
GM2_b0	2.012	2.019	0.035	0.186	0.615	1.763	30.117
b) coefficient β_1							
FH_b1	0.603	0.608	0.050	0.215	-69.857	100.108	103.612
REML_b1	0.603	0.608	0.050	0.215	-69.857	100.108	103.612
MLE_b1	0.604	0.608	0.050	0.215	-69.819	99.984	103.623
M_b1	0.624	0.621	0.053	0.227	-68.791	97.298	103.617
GM_b1	1.541	1.547	0.074	0.287	-22.959	14.216	63.076
GM2_b1	2.006	2.005	0.063	0.248	0.283	3.133	40.145
c) coefficient β_2							
FH_b2	0.604	0.602	0.049	0.211	-69.820	99.950	110.684
REML_b2	0.604	0.602	0.049	0.211	-69.820	99.950	110.684
MLE_b2	0.604	0.604	0.049	0.208	-69.778	99.816	110.752
M_b2	0.625	0.626	0.053	0.216	-68.762	97.196	110.795
GM_b2	1.540	1.540	0.073	0.278	-23.004	14.245	70.561
GM2_b2	2.005	2.005	0.062	0.243	0.259	3.078	49.763
d) variance							
FH_var	5.321	5.152	1.732	1.268	432.095	2040.071	938.990
REML_var	4.770	4.677	1.332	1.101	377.017	1554.447	888.665
MLE_var	4.555	4.469	1.231	1.058	355.498	1386.806	850.790
M_var	3.759	3.688	1.090	0.996	275.935	870.277	688.054
GM_var	2.060	1.962	0.675	0.783	106.016	179.800	450.079
GM2_var	1.767	1.722	0.437	0.638	76.749	102.536	382.691

See text for explanations.

6.5.4. Robust prediction

So far, we have discussed the behavior of robust parameter estimates in the presence of contamination. As far as small area estimation is concerned, computing robust parameter estimates is only an intermediate step on the way to robust predictions of area-specific means. However, robust prediction is an easy task once the parameters have been estimated. Moreover, it is evident that biased parameter estimates lead to biased and (in most cases) inefficient predictions.

Nonetheless, the influence that outliers or excessive variability in the data can exert on the predicted values and the associated MSPE estimates can be substantial. To really understand when robust methods are advantageous, we shall study robust prediction and MSPE estimation in the context of real sample data; see the case studies, below.

Table 6.6.: Speed comparison

implementation	mean(time)	med(time)	max(time)	mean(# iterations)
sae	1.820	1.744	3.830	4.990
saeRobust	483.242	476.805	655.065	16.890
saeRobust(maxIter = 2)	383.167	380.318	528.749	[hold fix]
rsae	1.208	1.192	1.716	2.370

Time/ duration is measured milliseconds; the measurements are obtained using the `microbenchmark` functionality of the `R` package `microbenchmark` with 100 replications (Mersmann et al., 2015). The average number of iterations refers to the number of iterations until convergence for β is achieved.

6.5.5. Other aspects

In this section, we compare the implementations of the M-estimators in the packages `saeRobust` (Warnholz, 2016a) and `rsae` [our implementation] in terms of computation time. We restrict attention to computing M-estimators with tuning constant $k = 1.345$, using the simulation setup outlined in (6.82–6.83) with a sample size of $n = 100$. Both algorithms are tuned such that they declare convergence when the iteratively updated estimates differ by less than 10^{-5} (termination rule).

The measurements of computation time are obtained with the help of the functionality in the `R` package `microbenchmark` (Mersmann et al., 2015). The measurements are reported in Table 6.6 and can be summarized as follows.

- Our implementation, see `rsae` in Table 6.6, is more than 400 times faster than the `rfh`-method in `saeRobust`. This is surprising since the majority of the code base of `saeRobust` is written in C++.
- The major reason for the poor performance of the `rfh`-method in `saeRobust` seems to be a poorly tuned updating algorithm, which requires on

average 16.9 iterations until convergence for the parameter β (compared with 2.4 in the case of `rsae`). However, the number of iterations is not the only cause for poor performance as the numerical criteria do not radically improve when we restrict the maximal allowed iterations (`maxIter`) to 2; see the numbers of method `saeRobust(maxIter=2)`.

- To be fair, it must be pointed out that the `rfh`-method implemented in `saeRobust` provides tools to fit the Fay–Herriot (FH) model and some spatial and/or temporal extensions of the FH model. The implementation chosen in `saeRobust` seems to be such (as far as we understand it) that the FH model is computed as a special case of the more general methods that deal with spatial / temporal correlations. Our implementation, on the other hand, is a highly specialized algorithm that focuses exclusively on the FH model; consequently, it is expected to do better in this special case. Nonetheless, we are quite surprised by the huge differences in computation time.

6.6. Case studies

In this section, we apply the robust methods under the Fay–Herriot model to the following three SAE applications.

- Toxoplasmosis prevalence estimates for cities in El Salvador; method: robust generalized James–Stein estimation
- Small area estimates of average expenditures for milk in the U.S.; method: M-estimator
- District-level estimates of crop yield for paddy in the State of Uttar Pradesh (India); method: GM-estimator

6.6.1. Case study: Toxoplasmosis prevalence estimates for cities in El Salvador

We shall study an instructive application of the methods to the toxoplasmosis data in Efron and Morris (1975, sec. 3). These data are a now famous showcase of empirical Bayes estimation. The baseline data consist of toxoplasmosis prevalence measurements for 5171 individuals in El Salvador (collected in 1963–64). The individual data are not made available. Instead, B. Efron and C. Morris work with the aggregated sample data on toxoplasmosis prevalence rates (i.e. means), X_i, and the standard deviation of the X_i's, denoted by $\sqrt{D_i}$, for $i = 1, \ldots, 36$ $(= n)$ cities; see Table 6.7. The goal is to obtain a set of prevalence estimates for the cities which is more efficient than the set of X_i's. To this end, Efron and Morris (1975, 314) suggested the empirical Bayes estimator under the hierarchical model

$$\begin{aligned} (X_i \mid \theta_i) &\overset{\text{ind.}}{\sim} \mathcal{N}(\theta_i, D_i), \qquad i = 1, \ldots, 36, \\ \theta_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, A), \qquad i = 1, \ldots, 36, \end{aligned}$$

$A \in \mathbb{R}^+$ being an unknown parameter, the a priori mean equals zero; see Efron and Morris (1975, 314) for a justification of the model specification. Since the variance parameter A is unknown, one can attempt to estimate it from the data. Efron and Morris (1975) discuss the m.l.e. of A, but in the end, they use a different estimator, which has the (theoretical) advantage that it reduces to the James–Stein estimator when $D_i \equiv D$. Their alternative estimator produces separate estimates $\hat{A}_i$ for each city (see Table 6.7; the $\hat{B}_i$ are given by $\hat{B}_i = D_i/(\hat{A}_i + D_i)$). However, the authors point out that the difference between the m.l.e. and their alternative estimator is minor in the case of the toxoplasmosis data [p. 314].

Under the assumption that the origin is the location towards which we want

to shrink (i.e. the prior mean), the generalized James-Stein type estimate[6] is computed as follows (where `sqrtDi`$:= \sqrt{D_i}$ and `Xi`$:= X_i$)

Table 6.7.: Toxoplasmosis data

i	X_i	$\sqrt{D_i}$	δ_i	$\hat{A}_i$	$\hat{B}_i$
1	0.2930	0.3040	0.0350	0.0120	0.8820
2	0.2140	0.0390	0.1920	0.0108	0.1020
3	0.1850	0.0470	0.1590	0.0109	0.1430
4	0.1520	0.1150	0.0750	0.0115	0.5090
5	0.1390	0.0810	0.0920	0.0112	0.3360
6	0.1280	0.0610	0.1000	0.0110	0.2210
7	0.1130	0.0610	0.0880	0.0110	0.2210
8	0.0980	0.0870	0.0620	0.0113	0.3700
9	0.0930	0.0490	0.0790	0.0109	0.1540
10	0.0790	0.0410	0.0700	0.0109	0.1120
11	0.0630	0.0710	0.0450	0.0111	0.2790
12	0.0520	0.0480	0.0440	0.0109	0.1480
13	0.0350	0.0560	0.0280	0.0110	0.1920
14	0.0270	0.0400	0.0240	0.0108	0.1070
15	0.0240	0.0490	0.0200	0.0109	0.1540
16	0.0240	0.0390	0.0220	0.0108	0.1020
17	0.0140	0.0430	0.0120	0.0109	0.1220
18	0.0040	0.0850	0.0030	0.0112	0.3590
19	-0.0160	0.1280	-0.0070	0.0116	0.5640
20	-0.0280	0.0910	-0.0170	0.0113	0.3920
21	-0.0340	0.0730	-0.0240	0.0111	0.2910
22	-0.0400	0.0490	-0.0340	0.0109	0.1540
23	-0.0550	0.0580	-0.0440	0.0110	0.2040
24	-0.0830	0.0700	-0.0600	0.0111	0.2730
25	-0.0980	0.0680	-0.0720	0.0111	0.2620
26	-0.1000	0.0490	-0.0850	0.0109	0.1540
27	-0.1120	0.0590	-0.0890	0.0110	0.2100
28	-0.1380	0.0630	-0.1060	0.0110	0.2330
29	-0.1560	0.0770	-0.1070	0.0112	0.3140
30	-0.1690	0.0730	-0.1200	0.0111	0.2910
31	-0.2410	0.1060	-0.1280	0.0114	0.4680
32	-0.2940	0.1790	-0.0830	0.0118	0.7190
33	-0.2960	0.0640	-0.2250	0.0111	0.2380
34	-0.3240	0.1520	-0.1140	0.0117	0.6470
35	-0.3970	0.1580	-0.1330	0.0117	0.6650
36	-0.6650	0.2160	-0.1400	0.0119	0.7890

Data: see Table 3 in Efron and Morris (1975, 314).

```
> rfh(Xi ~ 0, ~I(sqrtDi^2), data = toxo)

Call:
rfh(formula = Xi ~ 0, var = ~I(sqrtDi^2), data = toxo)

Method: James-Stein-type estimator
```

[6] The "classical" James-Stein estimator assumes that all variances D_i are equal to, say, D. This is not the case under the above model.

```
Location: [constrained to zero]
Variance estimate (adjusted for 0 df): 0.0122
```

The m.l.e. of A equals 0.0122 (rounded to the 4th decimal digit) and coincides with the value reported in Efron and Morris (1975, 314). The printed supplement "`(adjusted for 0 df)`" in the output of the variance estimate points out that the estimate is an m.l.e. of variance. Observe from the above function call that the zero on the r.h.s. of the first formula object (i.e., `Xi ~ 0`) constrains the location to zero. If we let the data choose the location, we obtain a location estimate of -0.0179, which is indeed close to zero (see the subsequent chunk of code).

```
> rfh(Xi ~ 1, ~I(sqrtDi^2), data = toxo)

Call:
rfh(formula = Xi ~ 1, var = ~I(sqrtDi^2), data = toxo)

Method: James-Stein-type estimator
Coefficient(s):
location
 -0.0179
Variance estimate (adjusted for 0 df): 0.0124
```

[*Side note*: If we wish to constrain the location to an arbitrary number, say, μ, the first formula object in the call of `rfh` should be `I(Xi`$-\mu$`) ~ 0`]

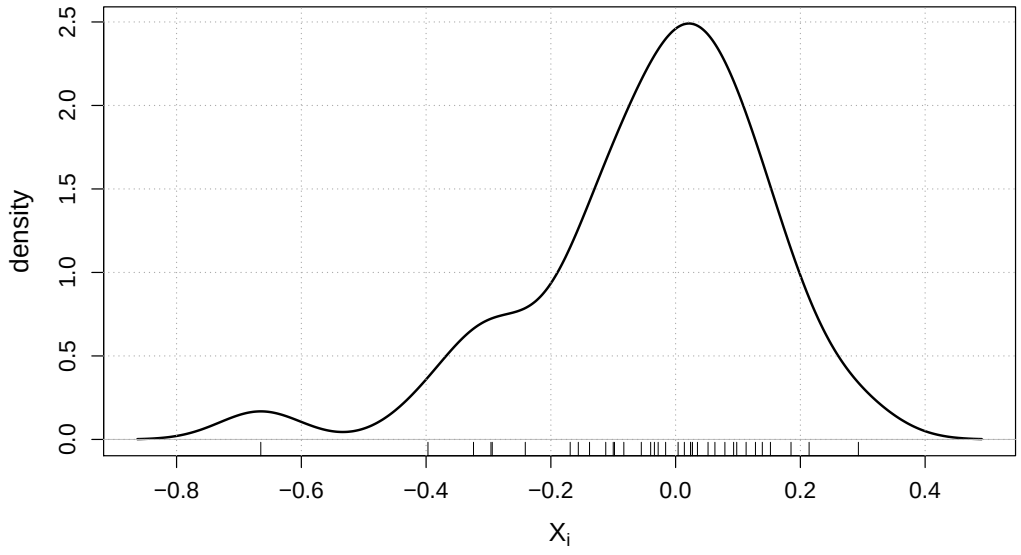


Figure 6.9.: Kernel density estimate of the variable X_i (data: from Table 6.7).

Limiting excessive component risks

The toxoplasmosis prevalence rates X_i are very nearly normally distributed; see the density estimate in Figure 6.9. Though, the density estimate is slightly

skewed to the left and shows a small bump at location -0.665. Overall, the distributional assumptions of the hierarchical Gaussian model are not really violated. Still, a slight robustification can be worthwhile. Robust estimates can be computed with function `rfh` by specifying `method="M"` and a robustness tuning constant `k` in the function call. In Figure 6.10, robust estimates of the variance parameter A for different values of the tuning constant k are shown (k is taken at equidistant points in the interval $[0.5, 4]$).

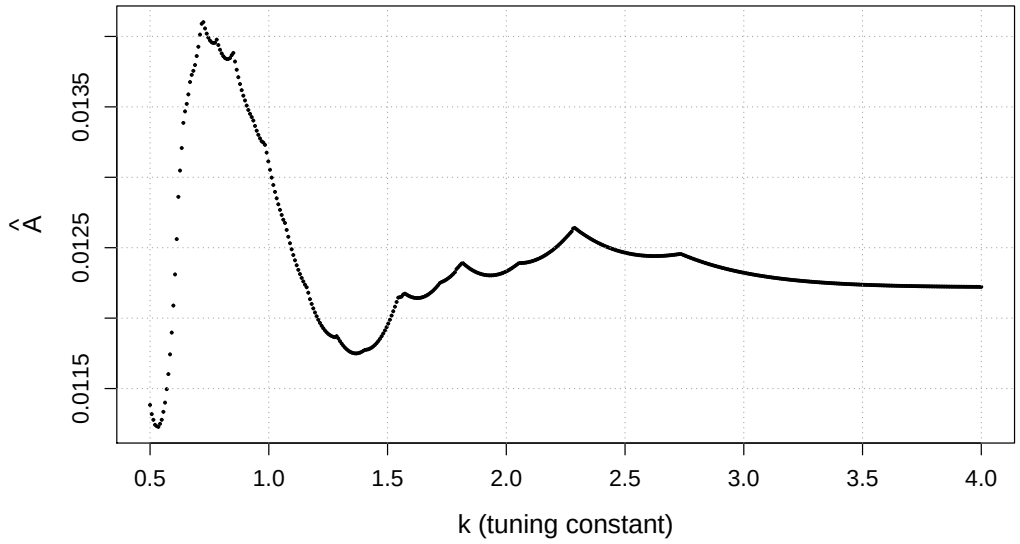


Figure 6.10.: Estimates of the variance parameter A for different values of the tuning constant k; k is taken at equidistant points in the interval $[0.5, 4]$.

From Figure 6.10, we recognize two interesting choices of k: The first choices lies in close proximity of the "famous constant" at $k = 1.345$; the second choice corresponds to $k \approx 0.5$. Tuning constants smaller than unity are in general doubtful, since the corresponding estimators do not even "trust" data in the range of plus/minus one standard deviation. The local minimum of $\hat{A}$ is attained at 0.0117 for $k = 1.37$ and is subtly smaller (about 4%) than the m.l.e. Why are we interest in a particularly small estimate of A? Small values are preferred for the rule at hand,

$$\delta_i = \left(1 - \frac{D_i}{\hat{A} + D_i}\right) X_i + \frac{D_i}{\hat{A} + D_i} 0, \qquad i = 1, \ldots, n, \tag{6.85}$$

where the second summand is actually zero (the "0" represents the prior mean), but was kept for pedagogical reasons. Now, we observe from (6.85) that for smaller $\hat{A}$, the prior mean receives more weight compared with the m.l.e. X_i; hence, more shrinkage towards zero takes place. Further, more shrinkage implies (ceteris paribus) a smaller ensemble risk and therefore a higher overall efficiency. A useful measure of the ensemble risk is empirical Bayes risk. Let δ_i denote the rule in (6.85), and put $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^T$. The empirical Bayes risk of

rule $\boldsymbol{\delta}$ (as a function of parameter A) is defined *conditionally* on the X_i's as (see Efron and Morris, 1975, 315)

$$\begin{aligned}\mathrm{EBrisk}(\boldsymbol{\delta}, A) &= \frac{1}{nD_0}\sum_{i=1}^{n}\mathbb{E}\big[(\delta_i-\theta_i)^2 \mid X_1,\ldots,X_n\big] \\ &= \frac{1}{nD_0}\sum_{i=1}^{n}\Big\{\delta_i^2+(1-B_i)\big[D_i-2\delta_i X_i+(1-B_i)X_i^2\big]\Big\},\end{aligned} \tag{6.86}$$

where $D_0=(1/n)\sum_{i\le n}D_i$, $B_i=D_i/(A+D_i)$, having used

$$\mathbb{E}\big[\theta_i^2 \mid X_i\big]=\mathbb{V}\big[\theta_i \mid X_i\big]+\big(\mathbb{E}\big[\theta_i \mid X_i\big]\big)^2=D_i(1-B_i)+(1-B_i)^2X_i^2$$

under the posterior distribution

$$\theta_i \mid X_i \overset{\text{ind.}}{\sim} \mathcal{N}\Big((1-B_i)X_i,\, D_i(1-B_i)\Big), \qquad i=1,\ldots,n.$$

In the definition of EBrisk, the δ_i's are regarded as fixed numbers. By how much do the values of EBrisk differ between the m.l.e. $\hat{A}=0.0122$ and the M-estimate of A equal to 0.0117 (for $k=1.37$)? The respective ensemble risk values are 0.389 and 0.383. So, the M-estimate achieves a slightly higher efficiency, but the difference is almost unnoticeable. Can we thus conclude that robust estimation is not needed, here? First, it was to be expected that robust estimates of A do not turn the world upside down in view of the distribution of the X_i's shown in Figure 6.9. Second, the toxoplasmosis data has been well-chosen by B. Efron and C. Morris to show the effectiveness of their estimators under the Gaussian assumption. In conclusion, M-estimation does not really pay off; however, we shall now demonstrate that the limited translation rule is superior in terms of component-wise risk, while it maintains much of the Bayes rule's gains in ensemble risk over the m.l.e.

Given the m.l.e. of A, we compute the rule in (6.85), i.e. EBLUP and also the limited translation rule (LTR).

```
> mle <- rfh(Xi ~ 0, ~I(sqrtDi^2), data = toxo)
> pmle <- predict(mle, k = 1.345)
```

The LTR (which is part of the prediction object `pmle`) limits shrinkage towards the origin for all coordinates which satisfy $|X_i/\sqrt{\hat{A}+D_i}|>k=1.345$, $\hat{A}$ denoting the m.l.e. of A. Figure 6.11 shows the "pull-in" behavior of the generalized James–Stein rule defined (6.85) and the LTR. It is apparent from the visual display that the LTR applies less shrinkage to observations further away from the origin.

The LTR limits the translation from the m.l.e. and therefore limits the maximum component risk that could be experienced. But this feature comes at some price in terms of ensemble risk. The LTR with $k=1.345$ has an EB risk of 0.618 compared with 0.390 in case of the generalized James–Stein rule (which corre-

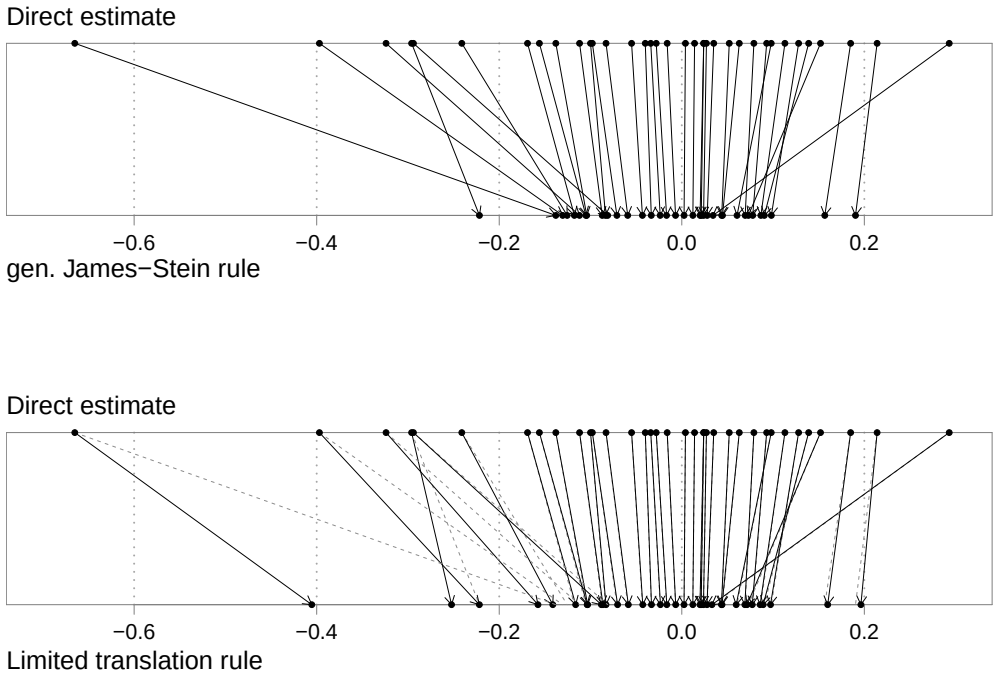


Figure 6.11.: Shrinkage or "pull-in" behavior of the generalized James–Stein rule and the limited translation rule; the direct estimates refer to the X_i's discussed in the text.

Table 6.8.: Empirical Bayes (EB) risk of the limited translation rule (for different choices of the tuning constant k; toxoplasmosis data)

k	0.5	1.0	1.5	2.0	2.5	3.0	∞
EB risk	1.194	0.787	0.562	0.445	0.395	0.390	0.390

Note: EB risk is computed with Formula (6.86).

sponds to $k = \infty$); see Table 6.8. Using the LTR instead of the generalized JS rule results in an increase of EB risk by almost 59%. At first sight, the increase appears to be large. However, compared with the conditional risk of the m.l.e., which is 1.875 [and can be obtained from Formula (6.86) with δ_i replaced by X_i], the LTR still preserves much of the gains in EB risk.

When robustness matters

Consider the toxoplasmosis prevalence rates X_i discussed above, but now we shall suppose that observation X_{33} has been changed (displaced) from -0.296 to -0.665; see Figure 6.12. Despite the displacement of observation X_{33}, the density plot still looks rather "well-behaved". For ease of notation, we do not introduce some special notation for the modified data, but rather write $X_1, \ldots, X_n$ to mean

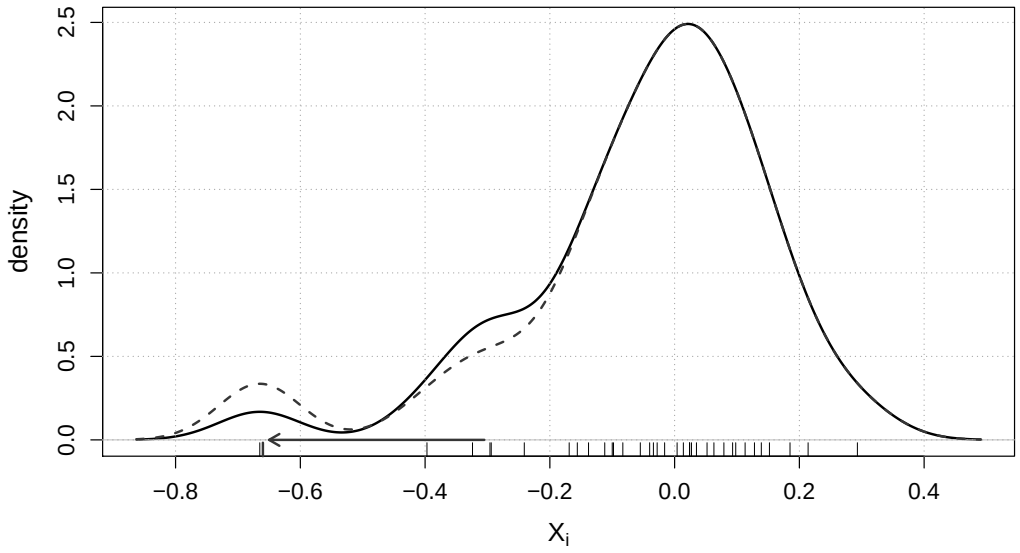


Figure 6.12.: Kernel density estimates of the variables X_i (solid line) and $\tilde{X}_i$ (dashed line); variable $\tilde{X}_i$ coincides with the original variable X_i except that observation X_{33} has been changed (displaced) from -0.296 to -0.665.

the modified sample.

Given the modified data, we computed robust M-estimates of the variance parameter A for a series of tuning constants k; see Figure 6.13. From the visual display, we recognize that a broad interval of tuning constants (i.e., $k \in [1,\, 2.5]$) leads to estimates which are considerably smaller than the m.l.e.; the m.l.e., $\hat{A}_{mle}$, obtains for a tuning constant $k \approx 4$ (see right edge of the graph). Let the M-estimates of A be denoted by $\hat{A}_{M,k}$, where subscript k indicates the tuning constant k. We shall consider the following estimating rules

$$\text{m.l.e. LTR:} \quad X_i - \frac{D_i}{\sqrt{\hat{A}_{mle} + D_i}} \psi_k \left(\frac{X_i}{\sqrt{\hat{A}_{mle} + D_i}} \right),$$

$$\text{M-est. LTR:} \quad X_i - \frac{D_i}{\sqrt{\hat{A}_{M,k} + D_i}} \psi_k \left(\frac{X_i}{\sqrt{\hat{A}_{M,k} + D_i}} \right),$$

$$\text{M-est.:} \quad X_i - \frac{X_i}{\hat{A}_{M,k} + D_i}.$$

In rule M-est. LTR, tuning constant k is used for both, computation of $\hat{A}_{M,k}$ and to specify the behavior of the limited translation rule. In case of rule m.l.e. LTR, the tuning constant k controls the behavior of the LTR. The EB risks of the three estimating rules evaluated at the modified toxoplasmosis data are shown in Figure 6.14. It is apparent from the visual display that rule M-est. attains the smallest values for EB risk, but it does not limit the maximal component risk. For this reason, we shall not study rule M-est. in more detail. The two

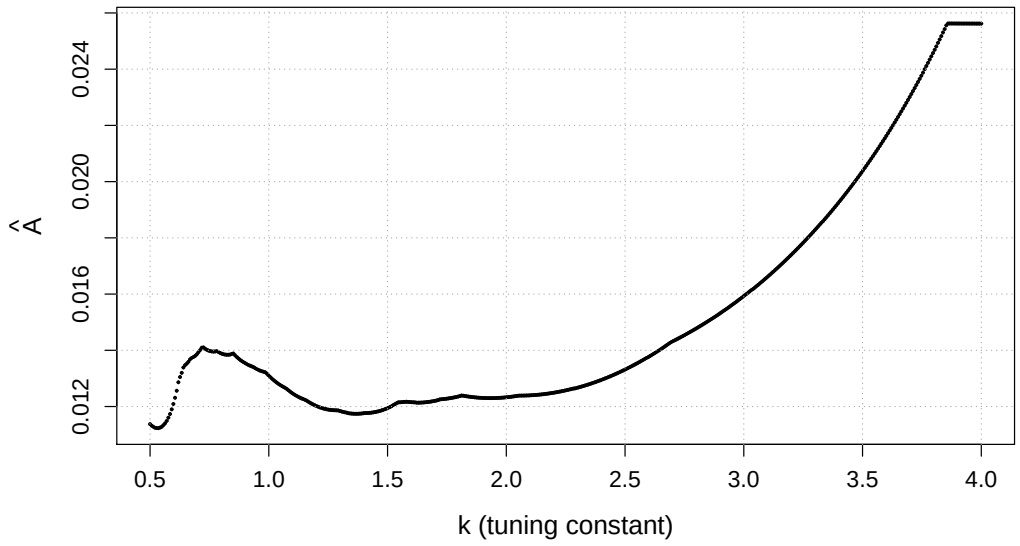


Figure 6.13.: Estimates of the variance parameter A for different values of the tuning constant k, using the modified toxoplasmosis data; k is taken at equidistant points in the interval $[0.5, 4]$.

LTR-type rules (m.l.e. LTR and M-est. LTR), on the other hand, seek a trade-off between ensemble and component risk.

The EB risk curves of m.l.e. LTR, M-est., and M-est. LTR as functions of the tuning constant k overlap in such a manner that the curves span an interesting region; see the shaded area in Figure 6.14 between the vertices A and D. The shaded area represents the region where rule M-est. LTR (with $1.4 \leq k \leq 3.8$) dominates rule m.l.e. LTR in terms of EB risk (while limiting the maximal amount of translation).

Two particular locations on the boundary of the shaded area in Figure 6.14 are noteworthy. Point C marks the location where rule M-est. LTR attains its minimum EB risk. Point B indicates the configuration where rule M-est. LTR (with $k = 1.74$) attains the EB risk of the generalized JS rule (see horizontal line at 0.526). Although the generalized JS rule and rule M-est. LTR are identical in terms of EB risk in Point B, the maximum component risk experienced with rule M-est. LTR is much smaller compared with the generalized JS rule.

What choice of k is *optimal*? It depends on our "utility function", which we shall take to be the weighted average (for positive real values a_1, a_2 s.t. $a_1 + a_2 = 1$)

$$a_1 \times \text{[ensemble risk]} + a_2 \times \text{[component risk]}.$$

When ensemble risk is considered more important than component risk ($a_1 \gg a_2$), a choice of k in the neighborhood of Point C is beneficial. If, on the other hand, ensemble *and* component risk matter in equal shares, the choice $k = 1.74$ (see Point B) may be advantageous as it maintains EB risk of the generalized JS rule while it is superior in terms of component risk.

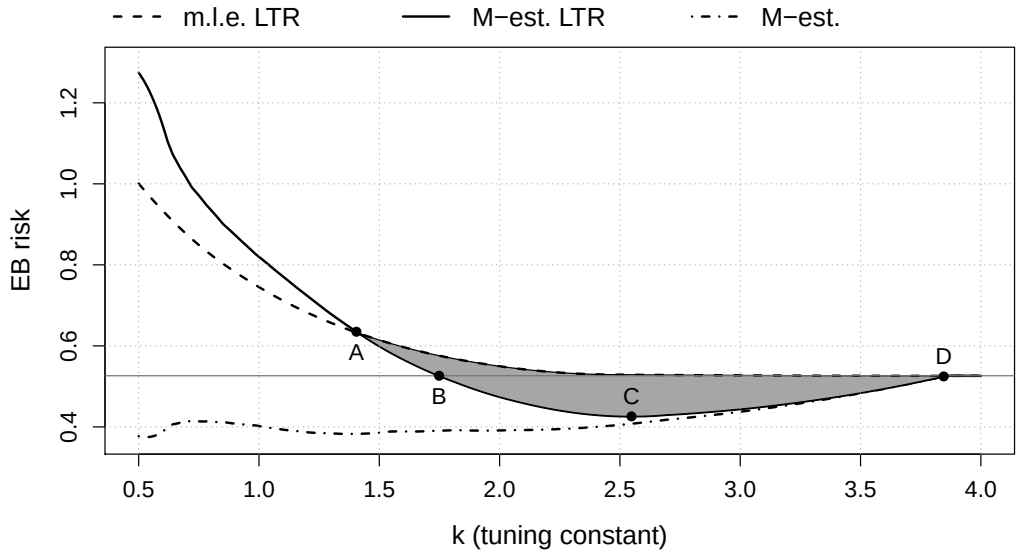


Figure 6.14.: Empirical Bayes risk of three estimating rules evaluated at the modified toxoplasmosis data; the rules are: (i) limited translation rule (LTR) with m.l.e. of A [denoted by m.l.e. LTR], (ii) LTR with M-estimate of A [M-est. LTR], and (iii) empirical Bayes rule using M-estimate of A [M-est.]. All rules are computed over a set of tuning constants k.

6.6.2. Case study: Small area estimates of average expenditures for milk in the U.S.

The U.S. Bureau of Labor Statistics computed weekly consumer expenditures on various items, goods, and services for publication areas (i.e., small areas) throughout the U.S. in 1989. Arora and Lahiri (1997) modeled data on average expenditure for fresh milk in $n = 43$ small areas using hierarchical Bayes methods. The area-specific sample sizes range from 95 to 633 in the small areas, and the coefficients of variation (CV) of the direct estimates range from 0.074 to 0.341. Arora and Lahiri (1997) classified the 43 small areas into eight major areas (the division into groups is not given in their paper). You and Chapman (2006) studied the same data, but proposed a different division of the areas into groups: they consider four major areas (identified via variable `MajorArea`, see below). Zimmermann (2018, chap. 3.5.3) also studied the data, and he specified the following Fay–Herriot type model for the direct estimates of milk, y_i,

$$y_i = \beta_0 + \beta_2 \mathbb{1}\{i \in A_2\} + \beta_3 \mathbb{1}\{i \in A_3\} + \beta_4 \mathbb{1}\{i \in A_4\} + e_i, \qquad i = 1, \ldots, 43,$$

where the $\mathbb{1}$'s are dummy variables that take the value one if observation i is in areas A_2, A_3, or A_4; otherwise the indicators are equal to zero. For reasons of parameter identification, the dummy variable referring to area A_1 has been dropped from the model equation. The data are available in the `R` package `sae`

(Molina and Marhuenda, 2015).[7]

Zimmermann (2018, 51–52) did an in-depth data analysis and noticed on the basis of Cook's distance and some other influence diagnostics that the areas no. 4 and 11 behave slightly differently compared with the majority of areas (see Figure 3.4 in Zimmermann, 2018). Since the Shapiro–Wilk test did not reject the null hypothesis of a Gaussian distribution at the 5% level of significance, he considered to fit the above model.

The influential-value diagnostics studied by Th. Zimmermann are an indispensable instrument to detect deviations from the underlying model assumptions. However, the model at hand is rather special insofar that it contains only dummy variables as predictors. This implies that influential observations in the model's design space (detected via regression-type model diagnostics) are rather evidence that the imposed area structure reveals or generates outliers in the *response* variable. In fact, the "hat" matrix is given by [$\mathbf{J}_n$ denoting the $(n \times n)$ matrix of ones]

$$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \text{blockdiag}\Big(\tfrac{1}{7}\mathbf{J}_7,\ \tfrac{1}{7}\mathbf{J}_7,\ \tfrac{1}{11}\mathbf{J}_{11},\ \tfrac{1}{18}\mathbf{J}_{18}\Big).$$

Observe that pre-multiplying the response variable [regarded as an (43×1) vector] by the "hat" matrix is equivalent to computing the means of the milk expenditures in each of the four major areas. From this it is seen that the diagnostic measures that derive from the "hat" matrix actually refer to "problems" in the response variable.

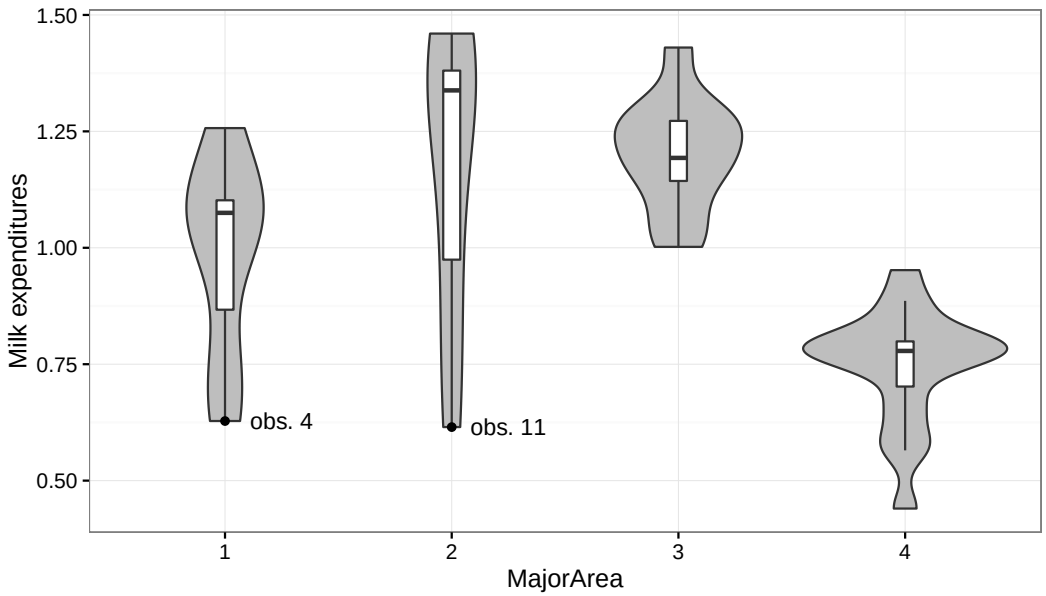


Figure 6.15.: Box-/violin plot of milk expenditures in the four major areas. Observations 4 and 11 are rather atypical within their respective major area.

[7] The data are obtained using `sae::data(milk)`; then, we redefine variable `MajorArea` as a `factor`, `milk$MajorArea <- factor(milk$MajorArea)`. The standard deviation of the direct estimates y_i is recorded in variable `SD`.

Figure 6.15 shows box-/ violin plots of milk expenditures in the four major areas. From the plot, we recognize that the areas embedded in major area no. 2 have a comparatively larger dispersion. It is also apparent that observations no. 4 and 11 mark the lower end of their respective area-specific distributions. Especially observation no. 11 is considerably smaller (an "inlier") compared with the bulk of data in major area 2. In view of those two atypical observations, we shall compute M-estimates of the model parameters. First, we consider the m.l.e. which is computed as follows.

```
> mle <- rfh(yi ~ MajorArea, ~ I(SD^2), data = milk)

Call:
rfh(formula = yi ~ MajorArea, var = ~I(SD^2), data = milk)

Method: MLE

Coefficient(s):
(Intercept)   MajorArea2   MajorArea3   MajorArea4
      0.968        0.128        0.227       -0.243

Variance estimate (adjusted for 0 df): 0.015
```

M-estimates of the model parameter are obtained as follows. Observe that we set the robustness tuning constant k equal to 2.5, which leads to a rather mild robustification. This choice is justified as only a handful of observations is supposed to deviate from the central model (see Figure 6.15).

```
> mest <- rfh(yi ~ MajorArea, ~ I(SD^2), data = milk,
+             method = "M", k = 2.5)

Call:
rfh(formula = yi ~ MajorArea, var = ~I(SD^2), data = milk,
   k = 2.5, method = "M")

Method: M-estimation

Coefficient(s):
(Intercept)   MajorArea2   MajorArea3   MajorArea4
      1.051        0.238        0.156       -0.286

Variance estimate (adjusted for 0 df): 0.016
```

Next, we consider prediction of area-level means by means of EBLUP and the limited translation rule (robLTR) with $k = 1.345$. For reasons of comparison, bootstrap estimates of the MSPE are computed for EBLUP and robLTR (although we could have used the analytic second order approximation to MSPE estimation for EBLUP). Note that the bootstrap estimates are computed using a rather larger number of bootstrap replicates (`reps = 5000`) in order to ensure

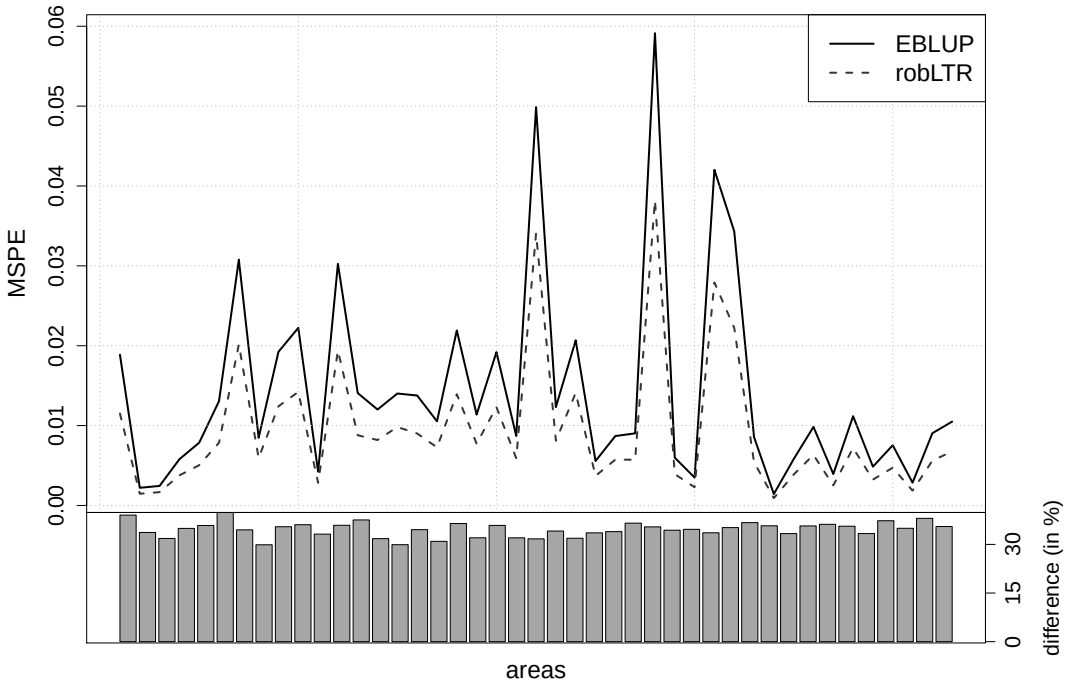


Figure 6.16.: Milk expenditure estimates: Mean square prediction error estimates of the EBLUP and the limited translation rule on the basis of M-estimates. The lower part shows the relative difference: $100\% \times [1 - \mathrm{MSPE}(robLTR)/\mathrm{MSPE}(EBLUP)]$.

a high degree of numerical accuracy. [*Remark:* Computing 5000 bootstrap replicates takes less than 3 seconds on a single-core computer.]

```
> pmle <- predict(mle, mse="bootstrap", reps = 5000)
> pmest <- predict.rfhestimate(mest1, mse="bootstrap", reps = 5000,
+                                    k = 1.345)
```

Figure 6.16 shows the MSPE estimates for all 43 small areas and for both methods: EBLUP vs. the limited translation rule (with M-parameter estimates). The lower part of the figure shows the relative difference between the two methods. It is remarkable that the MSPE estimates of robLTR are on average more than 30% smaller than the MSPE estimates of EBLUP. In conclusion, the M-estimator rule `robLTR` is remarkably more efficient than EBLUP.

6.6.3. Case study: District-level estimates of crop yield for paddy in the State of Uttar Pradesh (India)

Figures on crop area and crop production form the backbone of the system of agricultural statistics in India. Though, it is all too common that estimates of crop production are not available for districts at the desired level of precision. In some instances, estimates are completely unavailable. This is the situation Sud, Bhatia, Chandra, and Srivastava (2011) encountered when estimating crop yield for paddy in the State of Uttar Pradesh. Only 58 of the 70

districts have collected sample data on crop yield [see also Chandra, Salvati, and Sud (2011) who used a similar data set]. Some of the 58 district-level direct estimates show rather high variability. Therefore, Sud et al. (2011) considered to estimate the district-level Fay–Herriot model,

$$yield_i = \beta_0 + \beta_1 HH\_F_i + \beta_2 HH\_SIZE_1 + e_i, \qquad i = 1, \ldots, 58,$$

where

- $yield_i$ denotes average yield for paddy crop (direct estimates using sample data collected in 2009/2010; the variance of the direct estimators is denoted by `vd`);
- HH_F_i is the share female population of marginal household (Populations Census 2001);
- HH_SIZE_i denotes average household size (Population Census 2001).

The data are documented in Table C.1 (Appendix). Sud et al. (2011) estimate the above model; we apply a log-transformation to the independent variables prior to estimation (this removes some of the distributional skewness).

The m.l.e. parameter estimates are computed as follows.

```
> mle <- rfh(yield ~ I(log(HH_F)) + I(log(HH_SIZE)), ~vd, dat)
> mle

Call:
rfh(formula = yield ~ I(log(HH_F)) + I(log(HH_SIZE)), var = ~vd,
    data = dat)

Method: MLE

Coefficient(s):
    (Intercept)     I(log(HH_F))  I(log(HH_SIZE))
          56434             1644           -29761

Variance estimate (adjusted for 0 df): 11755616.42
```

The `summary` method applied to the object `mle` shows (among other things) the estimated coefficients together with the std. err., t-values, etc.

```
                     beta  std.error    tvalue       pvalue
(Intercept)      56434.429 14249.3765  3.960484 7.479796e-05
I(log(HH_F))      1644.125   800.7359  2.053267 4.004667e-02
I(log(HH_SIZE)) -29760.755  7762.3633 -3.833981 1.260858e-04
```

We observe that the coefficient of variable $\log(HH\_F)$ is barely significantly different from zero at the 5% level of significance. When we drop district no. 24 from the data and re-estimate the model with the reduced data, the estimate

for variable $\log(HH\_F)$ equals 625, not 1644. On closer inspection (not shown), it is seen that the partial regression effect between the response variable and variable $\log(HH\_F)$ is dominated by observation 24. In view of this, we shall compute GM-estimates.

```
> mest <- rfh(yield ~ I(log(HH_F)) + I(log(HH_SIZE)), ~vd, dat,
+       method = "GM", k = 1.345, k_x = 1.345)
> mest

Call:
rfh(formula = yield ~ I(log(HH_F)) + I(log(HH_SIZE)), var = ~vd,
    data = dat, k = 1.345, method = "GM", k_x = 1.345)

Method: GM-estimation

Coefficient(s):
    (Intercept)     I(log(HH_F))  I(log(HH_SIZE))
          57010             1072           -27361

Variance estimate (adjusted for 1 df): 8086645.601
```

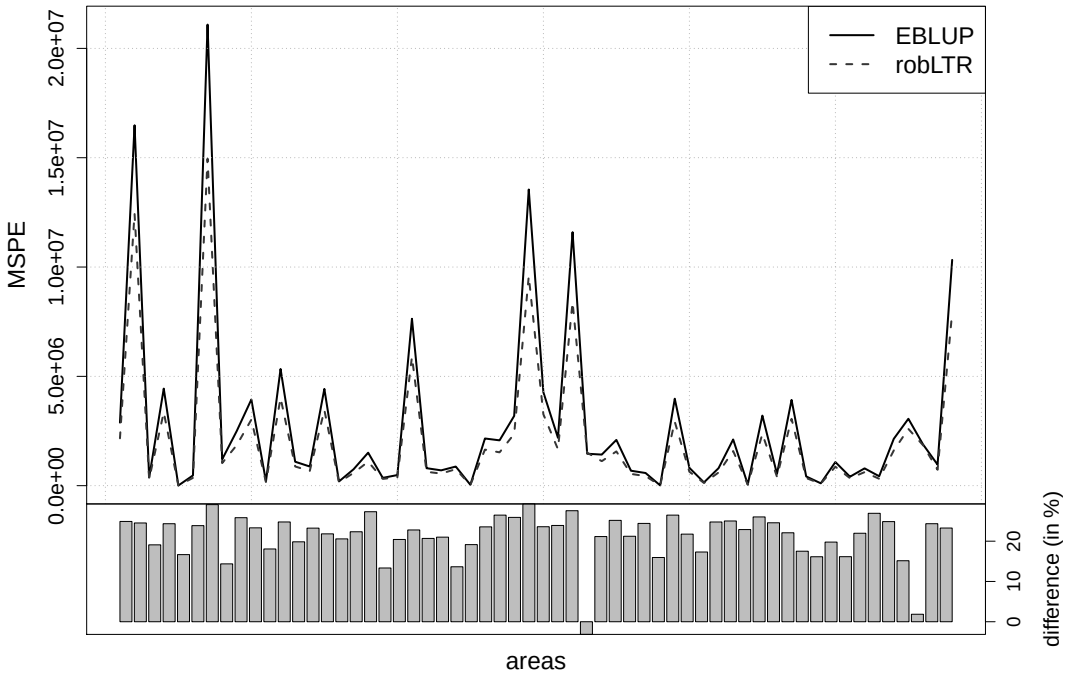


Figure 6.17.: Crop yield for paddy: Mean square prediction error estimates of the EBLUP and the limited translation rule on the basis of GM-estimates. The lower part shows the relative difference: $100\% \times [1 - \mathrm{MSPE}(robLTR)/\mathrm{MSPE}(EBLUP)]$.

The GM-estimate of the coefficient associated with variable $\log(HH\_F)$ is considerably smaller in comparison with the m.l.e and it is not significant at the 5% level (not shown). In addition, the GM-estimate of the variance parameter is distinctly smaller than the m.l.e. of variance. The advantage of using the GM-

estimate instead of the m.l.e. will become apparent when we consider estimates of the mean squared prediction error (MSPE).

Next, we compute the EBLUP and the predictions with the limited translation rule (LTR, with $k = 1.345$). For both methods, estimates of the MSPE are obtained from a parametric bootstrap with 5000 replications. We have computed bootstrap estimates for the EBLUP for reasons of comparison.

```
> pmle <- predict(mle, mse = "bootstrap", reps = 5000)
> pmest <- predict(mest, mse = "bootstrap", k = 1.345, reps = 5000)
```

Figure 6.17 shows the MSPE estimates for each of the 58 districts and for both methods: EBLUP vs. the limited translation rule (with GM-estimates). The lower part of the figure shows the relative difference between the two methods. It is apparent from the visual display that the MSPE estimates of rule robLTR are smaller in all but one district. It is remarkable that the MSPE estimates of robLTR are on average 21.4% smaller than the MSPE estimates of EBLUP. In conclusion, the GM-estimator LTR rule is remarkably more efficient than EBLUP.

6.7. Summary and discussion

M-estimators under the Fay–Herriot model can be obtained in a number of ways. One particular approach is to modify the proposal of Sinha and Rao (2009); this has been done by Warnholz (2016b). Our approach is different. We adopt a robust Bayes perspective and derive robust empirical Bayes estimators.

Bayesian robustness

Consider the following hierarchical Bayes (2-stage) Gaussian model

$$\left.\begin{array}{ll} \text{(sampling model)} & (Y_i \mid \Theta_i = \theta_i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_i,\, D), \\ \text{(prior distr. } P_0\text{)} & (\Theta_i \mid A) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\, A), \quad \text{for all } i = 1, \ldots, n \end{array}\right\}, \tag{6.87}$$

where the parameters $A \in \mathbb{R}^+$ and $D \in \mathbb{R}^+$ are supposed known. The hierarchical model is a *special case* of the Fay–Herriot model. For ease of discussion, we shall stick with the model and assume that $D \equiv 1$. The Bayes rule under model (6.87) and average squared error loss is $\delta_i^* = (1 - B)y_i$, where $B = 1/(A + 1)$. This rule achieves considerable gains over the m.l.e. $\delta_i^0 = y_i$, $i = 1, \ldots, n$ in terms of Bayes risk (or ensemble risk), but it may suffer from relatively large component-wise risk. The m.l.e., on the other hand, is minimax and limits the maximal component risk that could be experienced.

It is natural to ask whether compromise rules exist. The central idea is that we might be willing to be worse than the m.l.e. in terms of the maximum component risk only in proportion to potential gains obtainable in using prior information. Such compromise rules limit the component-wise risk, but still offer substantial gains over the m.l.e. in terms of ensemble risk.

In the specification of model (6.87), the sampling-model stage is *not* completely under our control. As users of the model, we are given the direct estimates (of the mean) and their variances. In this regard, the Gaussian sampling model has not been chosen solely on grounds of convenience, but is actually one of very few sensible choices. What distributional specification would otherwise make sense to model the direct estimates (means) and their variances? If, for instance, the within-area units had a skewed distribution, the sampler should have computed an estimate of location other than the arithmetic mean (e.g. median). Though, in view of the central limit theorem, the Gaussian distribution may still be appropriate. We shall stick with the Gaussian sampling model. The Gaussian prior distribution P_0 in (6.87), on the other hand, is quite often chosen for reasons of convenience. Unlike the sampling model, the choice of prior distribution is fully under our control (and the specification can be tested in light of the data).

For the situation that we do have *rather strong beliefs* in the prior distribution P_0 but are *not willing to entirely rely* on such information, Hodges and Lehmann (1952) suggested a formal compromise between the Bayes and the minimax rule.

The compromise rules δ_ϵ obtain from the restricted minimax principle (or equivalently, the restricted Bayes principle). Let $\epsilon \in [0,1]$ be a given number, and suppose some class of eligible estimating rules $\mathscr{D}$. The restricted minimax principle is as follows.

Find δ_ϵ such that

$$\sup_{P\in\mathcal{P}_\epsilon} r(P,\delta_\epsilon) = \inf_{\delta\in\mathscr{D}} \sup_{P\in\mathcal{P}_\epsilon} r(P,\delta) \tag{6.88}$$

where, for arbitrary d.f. H supported on $\mathbb{R}$,

$$\mathcal{P}_\epsilon = \big\{P \,:\, P(y) = (1-\epsilon)P_0(y) + \epsilon H(y)\big\}.$$

Since we have some faith in the Gaussian model, we chose the standardized Gaussian c.d.f. Φ as our base prior distribution P_0. Thus, the elments in the class of suitable prior distributions $\mathcal{P}_\epsilon$ are mixture distributions, composed of Φ and an *arbitrary* c.d.f. H. Conceptionally, our framework differs from the robust Bayesian view taken in Datta and Lahiri (1995) insofar that they propose a *specific* heavy-tailed prior, namely a Cauchy prior for outlying areas and a scale mixture of normals prior on the non-outlying areas. Under our approach, the prior distribution is not "hand picked"—rather it follows from the above principle. The solution to the restricted minimax principle in terms of the prior distribution is the *least favorable* (l.f.) prior distribution $P_\epsilon \in \mathcal{P}_\epsilon$ for every choice of $\epsilon \in [0,1]$. The l.f. is sufficiently pessimistic such that the Bayes solution will be minimax. The Cauchy prior of Datta and Lahiri (1995) appears yet to be an overly pessimistic choice.

What estimating rule is implied by the least favorable prior distribution? Let us first note that the formal Bayes rule δ_i^* vs. a well-behaved prior P can be written as $\delta_i = y_i + \psi(y)$ with $\psi \equiv \nabla \log(\Phi * P)$, where $*$ denotes convolution. For rules of the form $y_i + \psi(y_i)$, where function ψ depends on the choice of P, Marazzi (1980), Bickel (1980), and Levit (1980) independently discovered that the Bayes risk vs. prior P can be expressed as

$$r(P,\delta) = 1 - \mathcal{I}(\Phi * P), \tag{6.89}$$

where $\mathcal{I}$ denotes the Fisher information of location. In view of (6.89), the solution of (6.88) is, for every $\epsilon \in [0,1]$, equivalent to the minimization of $\mathcal{I}(g)$ on the set (Marazzi, 1980)

$$\mathcal{G}_\epsilon = \{g : g(y) = (1-\epsilon)(\Phi * P_0)(y) + \epsilon(\Phi * H)(y),\ H \text{ arbitrary}\}.$$

The result to this minimization is the least favorable *marginal* p.d.f., say, $g_\epsilon \in \mathcal{G}_\epsilon$, and implies rule $\delta_\epsilon(y_i) = y_i + g_\epsilon'(y_i)/g_\epsilon(y_i)$. Unfortunately, this minimization problem is extremely difficult to solve explicitly; the major obstacle is the convolution $\Phi * H$ in the definition of $\mathcal{G}_\epsilon$. In fact, closed-form representations of g_ϵ and the corresponding rule δ_ϵ are not known. Bickel and Collins (1983) proved that

g_ϵ is symmetric and that it concentrates (granted some regularity conditions) its mass on a countable set of isolated points (i.e. discrete distribution), possibly including $\{\pm\infty\}$. Efron and Morris (1971) have pointed out (through clever guess work and computer simulation) that rule δ_ϵ *oscillates*; see also Marazzi (1985) who managed to approximate the rule numerically.

Limited translation rule: An approximate solution

The oscillating behavior of rule δ_ϵ is a rather unpleasant property (at least from the point of application). Since a closed-form representation of the rule is unknown anyway, we are wondering whether the rule could be *approximated* in a fashion consistent with the underlying theory. To this end, consider the set

$$\mathcal{G}_\epsilon^* = \{g : g(y) = (1-\epsilon)(\Phi * P_0)(y) + \epsilon H(y),\ H \text{ arbitrary}\}$$

and observe that this set is similar to $\mathcal{G}_\epsilon$ except that here, we have H in place of $\Phi * H$. Therefore, $\mathcal{G}_\epsilon^*$ can be regarded as an approximation to $\mathcal{G}_\epsilon$. More importantly, minimization of $\mathcal{I}$ on the set $\mathcal{G}_\epsilon^*$ [and our choice $P_0 \equiv \Phi$] can be computed explicitly (see Marazzi, 1980). Using the minimax results of Huber (1964), we obtain an *approximation* of the least favorable marginal distribution g_ϵ, defined as (expressed in logarithms for ease of display)

$$\log \hat{g}_\epsilon(y_i) = \begin{cases} ky_i + k^2/2 + \text{const.} & \text{for } y_i < -k, \\ -y_i^2/2 + \text{const.} & \text{for } |y_i| \le k, \\ -ky_i + k^2/2 + \text{const.} & \text{for } y_i > k, \end{cases} \tag{6.90}$$

which is – up to the additive but unimportant constant – equal to the ρ-function accompanying the Huber ψ-function; k is a constant that is determined as a function $k(\epsilon)$. The rule corresponding to $\hat{g}_\epsilon$ is $y_i + \psi(y_i)$ with $\psi(y_i) = \nabla \log \hat{g}_\epsilon(y_i)$. This rule coincides with the *limited translation rule* (LTR) of Efron and Morris (1971) who established the rule on grounds of ad hoc arguments; the rigorous mathematical argument is due to A. Marazzi and P. Bickel (and also B. Levit and J. Berger).

The LTR behaves at the center of the data like the Bayes rule, but it limits the maximum translation from the m.l.e. by k for observations that are far from the prior mean zero (i.e. inhibits shrinkage). This behavior is a direct implication of the form of marginal p.d.f. $\hat{g}_\epsilon$, which is Gaussian in the center and has exponential tails. In return for sacrificing some of the possible Bayes ensemble savings, the LTR protects against large individual risks that are experienced with the Bayes rule.

The empirical Bayes case

We continue to assume that the hierarchical Gaussian model applies, but now the variance parameter A is supposed unknown. Clearly, we cannot use the

Bayes rule. Following Efron and Morris (1972), we can attempt to estimate A from the data. Under the hierarchical Gaussian model, $(n-2)/\|\mathbf{y}\|^2$ is an estimate of $B = 1/(1+A)$ which can be substituted for the unknown A to yield the celebrated *James–Stein* (JS) rule (see James and Stein, 1961). At its heart the JS rule is a shrinkage device to reduce variance at the expense of introducing a little bias. It is superior in terms of ensemble risk compared with the m.l.e. (when $n \geq 3$), but the rule may do very poorly in estimating a single component (in particular when $|\theta_i|$ is unusually large).

Following Efron and Morris (1972), we may attempt to *estimate* the limited translation rule (pursuing the analogy of having obtained the JS rule by estimating the unknown variance parameter). The resulting estimated LTR is a compromise of the JS rule and the m.l.e. It maintains most of the ensemble savings while being protected against unusual individual components.

Use of the estimated limited translation rule

In a figurative sense, the James–Stein rule pulls the coordinate-specific estimation problems together and treats them as a joint location–scale problem; see model (6.87). The "mechanics" behind this procedure may be called "coupling and shrinkage" of the separate problems. It is key to achieve gains in ensemble risk. When we are concerned about excessive risks for individual coordinates, the estimated LTR acts as some kind of "decoupling device" as it shrinks only those coordinates that are considered "close enough". As a consequence, LTR is a compromise between the JS rule and the m.l.e.

Two further comments are in order. First, LTR (and also the estimated LTR) behaves as if the prior distribution were actually the least favorable prior (not the Gaussian prior). Second, (and related to the first observation) the decoupling implied by LTR only applies to the "prediction stage", not the stage of estimation. This may lead to a rather problematic situation: Suppose that limiting coordinate-wise risk matters to us. Therefore, we plan to use the LTR. But, how can we justify to estimate the unknown parameters (in particular, A) still under the Gaussian prior assumption instead of the least favorable prior? Put another way: even though we know that shrinkage will be limited for coordinates with excessively large values, how can we justify not to account for this at the stage of estimation? This question remains unanswered. Beyond this point the problem remains open, and further research is required to get a deeper understanding.

Heavy-tailed distributions

When the empirical data are supposed to have a heavy-tailed distribution, the Gaussian prior distribution is obviously not an appropriate choice. Instead, we may consider the least favorable prior distribution which implies the p.d.f. of the marginal distribution in (6.90). In the empirical Bayes case, the parame-

ters are unknown and must be estimated. We have shown that the m.l.e. of parameter A, given $i = 1, \ldots, n$ i.i.d. distributed r.v.'s with marginal p.d.f. $\hat{g}_\epsilon$, is the M-estimator (in case of the simplified hierarchical model and also the Fay–Herriot model). Under the heavy-tailed prior distribution, LTR (with M-estimates substituted for the unknown parameters) is the optimal choice. If in addition, the auxiliary variables in the specification of the Fay–Herriot model are supposed to be outlier-prone (influential values in the model's design space), we have suggested a generalized M-estimator (GM).

The proposed method has been studied in three case studies. The findings show that significant gains are achievable over EBLUP in terms of mean squared prediction error when the data have heavier tails than the Gaussian distribution.

7. Conclusion and outlook

The demand for reliable statistics has been growing over the past decades as more and more decision makers (in businesses and politics) request evidence-based knowledge. In this context, sample surveys have long been used as cost-effective means for data collection. Such data have been effectively used to provide suitable statistics not only for the population targeted by the survey, but also for a variety of domains of interest. For small areas, however, the area-specific sample is typically not large enough to produce a direct estimate with reliable precision. Therefore, indirect or specific small-area estimation methods have been employed that borrow strength either from other small areas, over time or both. Indirect methods are much more dependent on the underlying modelling assumptions. For this reason, small area estimation methods are referred to as model-based estimating methods. Most of the (traditional) direct estimators, on the other hand, are merely model-assisted estimators in the sense that they are asymptotically model independent. Though, neither class of estimators is completely model free, and as consequence, both classes may suffer from model misspecification (although to varying degrees). Therefore, robustness considerations become of vital importance as the standard, non-robust estimators can be severely biased (or have inflated variance) in the presence of atypical or outlying observations.

Robust estimation under linear assisting model

There is a large body of literature on robust estimation under the linear assisting model. Our contribution is a compilation of robust domain estimators under the linear assisting model. The compiled list of estimators may help others to figure out what estimator is appropriate for their needs.

The generalized regression (GREG) estimator is the “flagship” in the class of randomization-assisted estimators, though it is not design-consistent. Under mild assumptions, the design bias of the GREG estimator vanishes asymptotically as the sample size grows without bounds. Therefore, the GREG estimator is said to be asymptotically design consistent (ADU). The contribution of the assisting model to the estimator vanishes as the sample size grows (irrespective whether the model holds or not). This property is an intrinsic characteristic of ADU estimators and has become the cornerstone of the model-assisted sampling paradigm. Though, we would not go so far as to say – as some advocates of model-assisted sampling do – that model-assisted estimator are “robust”. Nevertheless, ADUness can be seen as a measure to limit the (potentially harmful)

influence of models. Our contribution is the notion of *strongly* design consistent estimators. Surprisingly, we require only mild regularity assumptions on the design, and the additional assumption that the study variable is nonnegative. The restriction to nonnegative variables is quite natural in the context of ratio estimation of a mean or total. Our results contribute to the understanding of ratio estimators and the class of QR-estimators.

Robust estimation under the basic unit-level model

The basic unit-level model has (since it is a special case of the family of mixed linear models, MLM, with a block-diagonal covariance matrix) unlike location-scale or regression models no nice invariance structure. This implies that the model parameters cannot be estimated consistently in the presence of contamination; there is an unavoidable asymptotic bias. For the maximum likelihood estimator, the bias can be arbitrarily large. Richardson and Welsh (1995) suggested two types of M-estimators, RML 1 and 2, under MLM's with block-diagonal covariance matrix that limit the potential bias. The existing algorithms to compute robust estimates, however, proved to be notoriously instable; see Richardson (1995, chap. 6.5). Later developments also suffered from serious convergence issues; see Chaubey and Venkateswarlu (2002).

In the context of small area estimation, Sinha and Rao (2009) proposed an approximation to RML 2 that has since inspired further developments. The approximation suggested by Sinha and Rao (2009) is not sufficient to fix the major numerical problems encountered when fitting robust MLM's. The methods of Sinha–Rao, though, has another drawback: it uses an improper decorrelation of the residuals which in turn makes it difficult to chose appropriate robustness tuning constants.

Our contribution is a robust procedure that derives from the RML 2 method, but it uses a different parametrization. It has superior numerical properties and does not suffer from failure of convergence. The method is implemented in the `R` package `rsae`; see Schoch (2014) and can be obtained from `CRAN`. In addition, we suggested a method for the prediction of the area-level means which is computationally much less demanding than the proposal of Sinha and Rao (2009).

Our simulation results on the behavior of the method are convincing: The loss in efficiency that results when robust methods are applied to uncontaminated data is comparatively small. With regard to the huge gains obtainable (in terms of efficiency and bias reduction) when the data are indeed subject to contamination, such small losses can be worthwhile. In the presence of contamination, bias and MSE of the M-estimator are much smaller compared with the m.l.e. Contamination of the model error affects the robust estimates of the area-level variance very little, since the contamination affects the diagonal elements of the covariance matrix, but not the off-diagonal elements. Contamination of the

area-specific random effects, on the other hand, affects both diagonal and off-diagonal elements of the variance. When both components are contaminated, the effects on the estimates are the combination of the effects of contaminating the components one at a time.

Robust estimation under the Fay–Herriot model

It has been known to the SAE community at least since the paper of Sinha and Rao (2009) that the methods suggested in this paper also apply to obtain robust estimators and predictors under the Fay–Herriot (FH). Hence, the robustification can be tackled using the M-estimator theory for MLM's developed by Richardson and Welsh (1995) or the approximated method due to Sinha and Rao (2009). The details of this approach have been written up in the Ph.D. thesis of Warnholz (2016b, chap 3.3).

Our approach for obtaining robust estimators and predictors is different. We go back to the roots of the FH model, that is James–Stein estimation. We pick up this line of argument, adopt a robust Bayes view and derive a robust empirical Bayes method that creates a direct link to the seminal work of Huber (1964), but under a slightly more general model. From this, we obtain M-estimators under the FH model. Technically, the resulting M-estimator under the FH model coincides largely with the Sinha–Rao proposal applied to the FH model. Nevertheless, our major contribution is to put the derivation of M-estimators under the area-level model onto a formal, theoretical foundation. Via this theoretical motivation, we can view the existing estimators and their behavior from a new perspective. In addition to M-estimators, we also introduce the class of generalized regression M-estimator (GM) under the Fay–Herriot model.

Our model-based simulation study shows that the loss in efficiency that results when robust are applied to uncontaminated data is comparatively small. On the other hand, the gains obtainable (in terms of efficiency and bias reduction) when the data are indeed subject to area-level contamination, are huge for both the M- and the GM-estimator. Things look different in case of influential observations in the model's design space. The M-estimator, although robust w.r.t. outliers in the response variable, is biased as much as the non-robust methods. This rather bad performance was to be expected as M-estimators are not protected against influential observations in the design space. Therefore, GM-estimators proof to be an indispensable tool in the presence of influential observations in the design space of the model.

A. Background material

A.1. Results from real analysis

In this section, we state some well-known results from real analysis, which are given without proof. To this end, let us fix some notation. By $[a, b]$ and (a, b) we denote, respectively, the closed and open interval, where $a, b \in \mathbb{R}$ and $a < b$. Let $\mathrm{B}_r(\mathbf{x})$ denote the open ball of radius r with center $\mathbf{x} \in \mathbb{R}^n$, $n \geq 1$. The Lebesgue measure on $\mathbb{R}$ is denoted by λ; likewise, we write λ_n to mean the Lebesgue measure on the n-dimensional Euclidean space. Unless otherwise stated, the notions of "measurability" and "almost everywhere" are understood w.r.t. Lebesgue measure. $L^1(\mathbb{R}^n)$ denotes the equivalence class of absolutely Lebesgue integrable function on $\mathbb{R}^n$. Let A, such that $A \subset \mathbb{R}$, be a measurable set. We choose to write Lebesgue integration of the (integrable) function f over the set A like a Riemann integral, $\int_A f(y)\mathrm{d}y$, instead of $\int_A f(y)\mathrm{d}\lambda(y)$ or $\int_A f\mathrm{d}\lambda$.

First, we introduce the notion of an almost everywhere (a.e.) differentiable function (see e.g. Tao, 2011, chap. 1.6).

Definition A.1 (Almost everywhere differentiable function)**.** *A function $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$ is said to be almost everywhere differentiable if it is (pointwise) differentiable at almost every point $x \in [a, b]$ with derivative $f'(x)$ if the limit*

$$f'(x) = \lim_{y \to x,\, y \in [a,b] \setminus \{x\}} \frac{f(y) - f(x)}{y - x} \tag{A.1}$$

exists.

For functions $f(x)$ that satisfy Definition A.1, the set of numbers, whose elements are in $[a, b]$ such that $f(x)$ is not differentiable in x, formally,

$$\mathcal{Z} = \{x : x \in [a, b],\ f(x) \text{ is not pointwise differentiable in } x\} \tag{A.2}$$

has Lebesgue measure zero. Without loss of generality (w.l.o.g.), we put $f'(x) = 0$ for all $x \in \mathcal{Z}$.

The following definition introduces the notion of an absolutely continuous (a.c.) function on $[a, b]$. This forms the basis for our first result in this section, the fundamental theorem of (Lebesgue) calculus; see below.

Definition A.2 (Absolutely continuous function)**.** *A function $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$ is said to be absolutely continuous on $[a, b]$ if for any $\epsilon > 0$ there exists $\delta > 0$ such*

that

$$\sum_{i=1}^{N} \big|f(b_i) - f(a_i)\big| < \epsilon \qquad \textit{whenever} \qquad \sum_{i=1}^{N} (b_i - a_i) < \delta, \tag{A.3}$$

and the intervals (a_i, b_i)*,* $i = 1, \ldots, N$ *are disjoint.*

We define

$$\mathrm{AC}[a,b] := \big\{ f : [a,b] \to \mathbb{R} \ : \ f \text{ is absolutely continuous on } [a,b] \big\}. \tag{A.4}$$

The space of *locally* absolutely continuous function on $\mathbb{R}$ is

$$\mathrm{AC}_{\mathrm{loc}}(\mathbb{R}) := \big\{ f : \mathbb{R} \to \mathbb{R} \ : \ f \in \mathrm{AC}[a,b] \text{ for every } a < b \big\}. \tag{A.5}$$

From the Definition A.2 (see also Stein and Shakarchi, 2005, 127), it is clear that absolutely continuous functions are uniformly continuous, hence a fortiori continuous. Moreover, if f is a.c. on a bounded interval, then it is also of bounded variation on the same interval. Also, if $F(x) = \int_{[a,x]} f(y)\mathrm{d}y$ with integrable f, then F is a.c. In the subsequent lemma, we give some important results on the characterization of absolutely continuous functions. The main focus of the lemma lies on the composition of two a.c. functions since such compositions are in general not a.c.

Lemma A.1. *Let* f, g *denote real-valued functions on some interval of* $\mathbb{R}$*. In what follows, absolute continuity of a function will always refers to the set* $[a,b]$*.*

(i) If the functions f *and* g *are a.c. on* $[a,b]$*, then so are* $f+g$*,* fg*, and* f/g *(provided* g *is nonzero).*

(ii) Let f *be a strictly increasing continuous function on* $[a,b]$*, then* f *is a.c. if it takes the set* $\{x : f'(x) = +\infty\}$ *to a set measure of zero. Also, the inverse function* f^{-1} *is a.c. if the set* $\{x : f'(x) = 0\}$ *has measure zero.*

(iii) Let $f \in \mathrm{AC}[a,b]$ *be a monotone function and let* $g \in \mathrm{AC}[c,d]$ *where* $[c,d]$ *contains* $f([a,b])$*. Then* $g \circ f$ *is a.c.*

(iv) Let $f \in \mathrm{AC}[a,b]$ *and let* g *be a Lipschitz continuous function on* $[c,d]$ *where* $[c,d]$ *contains* $f([a,b])$*. Then* $g \circ f$ *is a.c.*

Proof. See Cor. 5.3.3. in Bogachev (2006) for a proof of (i). The assertion in (ii) is a result due to M.A. Zareckiĭ, see Bogachev (2006, 389); for (iii) see ibid. p. 391; the result in (iv) is due to G.M. Fichtenholz, see Bogachev (2006, 391). ■

The following result is on integration by parts for a.c. functions.

Lemma A.2 (Integration by parts)**.** *Let* $f, g \in \mathrm{AC}[a,b]$*, then*

$$\int_{[a,b]} f'(x)g(x)\mathrm{d}x = f(b)g(b) - f(a)g(a) - \int_{[a,b]} f(x)g'(x)\mathrm{d}x. \tag{A.6}$$

Proof. See Corollary 5.4.3 in Bogachev (2006). ■

The integration by parts formula also holds for functions $f, g : \mathbb{R} \to \mathbb{R}$ that are a.c. on bounded intervals (i.e. locally a.c. functions). Recall that $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ is our notation for the extended real line. Let $f, g \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$, then (see Bogachev, 2006, Ex. 5.8.43)

$$\int_{\overline{\mathbb{R}}} f'(x)g(x)\mathrm{d}x = -\int_{\overline{\mathbb{R}}} f(x)g'(x)\mathrm{d}x. \tag{A.7}$$

Furthermore, absolutely continuous functions form an important class of functions since they satisfy the fundamental theorem of calculus, as can be seen from the following theorem.

Theorem A.1 (Fundamental theorem of calculus)**.** *Suppose F is a.c. on $[a, b]$. Then its derivative F' exists a.e. (Lebesgue) and is integrable. Moreover,*

$$F(x) - F(a) = \int_{[a,x]} F'(y)\mathrm{d}y, \qquad \textit{for all } a \leq x \leq b. \tag{A.8}$$

By selecting $x = b$, we get $F(b) - F(a) = \int_{[a,b]} F'(y)\mathrm{d}y$. Conversely, if f is integrable on $[a, b]$, then there exists an a.c. function F such that $F'(x) = f(x)$ a.e. (Lebesgue), and we may take $F(x) = \int_{[a,x]} f(y)\mathrm{d}y$.

Proof. See e.g. Thm. 3.3.11 in Stein and Shakarchi (2005). ■

For multivariable functions, the following result due to H.L. Lebesgue can be seen as a generalization of the fundamental theorem of calculus to multivariable functions.

Theorem A.2 (Lebesgue's differentiation theorem)**.** *Let $f \in L^1(\mathbb{R}^n)$. Then, for a.e. $\mathbf{x} \in \mathbb{R}^n$*

$$\lim_{r\to 0} \frac{1}{\lambda_n\big(\mathrm{B}_r(\mathbf{x})\big)} \int_{\mathrm{B}_r(\mathbf{x})} \big|f(\mathbf{y}) - f(\mathbf{x})\big|\mathrm{d}\mathbf{y} = \mathbf{0} \tag{A.9}$$

and in particular, it follows that for a.e. $\mathbf{x} \in \mathbb{R}^n$,

$$\lim_{r\to 0} \frac{1}{\lambda_n\big(\mathrm{B}_r(\mathbf{x})\big)} \int_{\mathrm{B}_r(\mathbf{x})} f(\mathbf{y})\mathrm{d}\mathbf{y} = f(\mathbf{x}). \tag{A.10}$$

Proof. See e.g. Thm. 3.1.3 in Stein and Shakarchi (2005). ■

In fact, a more general version of Lebesgue's differentiation theorem holds under the weaker hypothesis that f is only locally integrable, i.e. $f \in L^1_{loc}(\mathbb{R}^n)$ (which is intuitively clear since differentiation is a local property); see Thm. 3.1.14 in Stein and Shakarchi (2005).

One of the most important advantages of the Lebesgue integration theory are, from a mathematical point of view, the nice results on convergence, namely P. Fatou's lemma on convergence of the integral of a sequence of nonnegative functions and H.L. Lebesgue's theorem on dominated convergence (the third

result, B. Levi's theorem on monotone convergence, is not shown here).

Lemma A.3 (Fatou). *Suppose* $\{f_n,\, n \geq 1\}$ *is a sequence of measurable functions with* $f_n \geq 0$. *If* $\lim_{n\to\infty} f_n(x) = f(x)$ *for a.e.* x, *then*

$$\int f(x)\mathrm{d}x \leq \liminf_{n\to\infty} \int f_n(x)\mathrm{d}x. \tag{A.11}$$

Proof. See e.g. Lem. 1.1.7 in Stein and Shakarchi (2005). ■

Theorem A.3 (Lebesgue's dominated convergence). *Let* $g \in L^1(\mathbb{R})$. *Suppose* $\{f_n,\, n \geq 1\}$ *is a sequence of measurable functions,* $f_n : A \subseteq \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, *such that* $f_n(x) \to f(x)$ *a.e.* x, *as* $n \to \infty$. *If* $|f_n(x)| \leq g(x)$ *for all* n *in the index set of the sequence, then*

$$\lim_{n\to\infty} \int_A |f_n(x) - f(x)|\mathrm{d}x = 0 \tag{A.12}$$

and consequently

$$\lim_{n\to\infty} \int_A f_n(x)\mathrm{d}x = \int_A f(x)\mathrm{d}x. \tag{A.13}$$

Proof. See e.g. Thm. 1.1.13 in Stein and Shakarchi (2005). ■

Another important result, due to G. Fubini, gives conditions when it is possible to compute a double integral using iterated integrals. To this end, let $f(\mathbf{x}, \mathbf{y})$ be a function on $\mathbb{R}^n \times \mathbb{R}^m$, $n, m \in \mathbb{N}_+$. The *slice* of f corresponding to $\mathbf{y} \in \mathbb{R}^m$ is the function $f^{\mathbf{y}}$ of the $\mathbf{x} \in \mathbb{R}^n$ variable, given by $f^{\mathbf{y}}(\mathbf{x}) = f(\mathbf{x}, \mathbf{y})$. Likewise, the slice of f for a fixed $\mathbf{x} \in \mathbb{R}^n$ is $f^{\mathbf{x}}(\mathbf{y}) = f(\mathbf{x}, \mathbf{y})$.

Theorem A.4 (Fubini). *Suppose* $f(\mathbf{x}, \mathbf{y})$ *is integrable on* $\mathbb{R}^n \times \mathbb{R}^m$, $n, m \in \mathbb{N}_+$. *Then, for a.e.* $\mathbf{y} \in \mathbb{R}^m$:

(i) the slice $f^{\mathbf{y}}$ *is integrable on* $\mathbb{R}^n$,

(ii) the function defined by $\int_{\mathbb{R}^n} f^{\mathbf{y}}(\mathbf{x})\mathrm{d}\mathbf{x}$ *is integrable on* $\mathbb{R}^m$,

(iii) it holds that

$$\int_{\mathbb{R}^m} \left(\int_{\mathbb{R}^n} f(\mathbf{x}, \mathbf{y})\mathrm{d}\mathbf{x} \right) \mathrm{d}\mathbf{y} = \int_{\mathbb{R}^{n+m}} f. \tag{A.14}$$

Proof. See e.g. Thm. 2.3.1 in Stein and Shakarchi (2005). ■

Remarks. (i) The term $\int_{\mathbb{R}^{n+m}} f$ in assertion (iii) of Fubini's theorem means integration over the product $\mathbb{R}^{n+m} = \mathbb{R}^n \times \mathbb{R}^m$, where a point in $\mathbb{R}^{n+m}$ takes the form $(\mathbf{x}, \mathbf{y})$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$.

(ii) Clearly, the theorem is symmetric in $\mathbf{x}$ and $\mathbf{y}$ so that we may also conclude that the slice $f^{\mathbf{x}}$ is integrable on $\mathbb{R}^m$ for a.e. $\mathbf{x}$.

(iii) Fubini's theorem states that the integral of f on $\mathbb{R}^{n+m}$ can be computed by iterating lower-dimensional integrals (and the integrations can be taken in any order).

Another useful result is on change of variables.

Theorem A.5 (Change of variables)**.** *Let Ω_1 and Ω_2 denote open sets in $\mathbb{R}^n$, $n \geq 1$. Assume that the map $\Phi : \Omega_1 \to \Omega_2$ is a bijection in the class $\mathcal{C}^1$ whose inverse map is also in $\mathcal{C}^1$. Assume that f is a Lebesgue measurable map on Ω_2. Then $f \circ \Phi$ is Lebesgue measurable on Ω_1 and*

$$\int_{\Omega_2} f(\mathbf{y})\mathrm{d}\mathbf{y} = \int_{\Omega_1} \big(\Phi(\mathbf{x})\big)\big|J(\mathbf{x})\big|\mathrm{d}\mathbf{x}, \tag{A.15}$$

where $J(\mathbf{x})$ denotes the Jacobian of f at $\mathbf{x} \in \Omega_1$. Formula (A.15) is valid in two senses: If $f \geq 0$, then it is true without further qualification. In general, $f \in L^1(\Omega_2)$ if and only if $f \circ \Phi|J| \in L^1(\Omega_1)$, and then the formula is valid.

Proof. See Jones (1993, 502). ■

Lipschitz continuous functions

In what follows, we focus on a particularly important class of Lipschitz continuous functions, which is defined as follows.

Definition A.3 (Lipschitz continuous function)**.** *A function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz (continuous) if*

$$\big|f(y) - f(x)\big| \leq M|y - x| \tag{A.16}$$

for some real-valued M and all $x, y \in \mathbb{R}$.

The class of Lipschitz continuous functions on $[a, b]$ is denoted by $\mathrm{Lip}[a, b]$. The smallest M in the above definition is called the (best) Lipschitz constant of f. Also, f is Lipschitz on $U \subset \mathbb{R}$ if and only (i) if it is a.c. on U and (ii) if it has a bounded derivate, $\sup_{x \in U} |f'(x)| < \infty$; see Stein and Shakarchi (2005, 152). For a map $f : \mathbb{R}^p \to \mathbb{R}^q$, the above definition, obviously, generalizes to

$$\big\|f(\mathbf{y}) - f(\mathbf{x})\big\| \leq M\|\mathbf{y} - \mathbf{x}\| \qquad \text{for some real } M \text{ and all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^p. \tag{A.17}$$

The following lemma states some nice properties of Lipschitz continuous functions.

Lemma A.4. *Let $f, g \in \mathrm{Lip}[a, b]$ and denote the Lipschitz constant of f and g, respectively, by $L(f)$ and $L(g)$. Then we have that*

(i) for any real constant α, $\alpha f \in \mathrm{Lip}[a, b]$ and $L(\alpha f) = |\alpha| \cdot L(f)$,

(ii) $(f + g) \in \mathrm{Lip}[a, b]$ and $L(f + g) \leq L(f) + L(g)$.

Let $f, g \in \mathrm{Lip}[a, b]$. In addition, we assume that f and g are bounded functions, i.e. there exist real-valued constants C_f and C_g such that $f(x) \leq C_f$ and $g(x) \leq C_g$ for all $x \in [a, b]$, then

(iii) the product $(fg) \in \mathrm{Lip}[a,b]$ *and* $L(fg) \leq C_f L(g) + C_g L(f)$.

Let $f \in \mathrm{Lip}[a,b]$ *and* $|f(x)| \geq \epsilon > 0$ *for all* $x \in [a,b]$*, then*

(iv) the quotient $1/f$ *is Lipschitz continuous on* $[a,b]$ *and* $L(1/f) \leq L(f)/\epsilon^2$.

Let $f \in \mathrm{Lip}[a,b]$ *and* $g \in \mathrm{Lip}[c,d]$ *where* $[c,d]$ *contains* $f\big([a,b]\big)$*, then*

(v) the composition $(f \circ g) \in \mathrm{Lip}[a,b]$ *and* $L(f \circ g) \leq L(f) \cdot L(g)$.

Proof. See Prop. 1.5.2 for (i) and (ii), Prop. 1.5.3 for (iii) and (iv), and Prop. 1.2.2 for assertion (v) in Weaver (1999). ■

On the convolution

For $f, g \in L^1(\mathbb{R}^n)$, $n \geq 1$, convolution is defined as

$$(f * g)(x) := \int_{\mathbb{R}^n} f(x-y)g(y)\mathrm{d}y. \tag{A.18}$$

With this we have.

Theorem A.6 (Convolution)**.** *Let* $f, g \in L^1(\mathbb{R}^n)$*. Then, the function*

$$(f * g)(x) = \int_{\mathbb{R}^n} f(x-y)g(y)\mathrm{d}y \tag{A.19}$$

is defined for a.e. x *and is integrable. Moreover,* $f * g = g * f$ *a.e. In addition,*

$$\|f * g\|_{L^1} \leq \|f\|_{L^1}\|g\|_{L^1}. \tag{A.20}$$

Proof. See Thm. 3.9.2 in Bogachev (2006). ■

Apart from convolutions of L^1-functions, one can consider convolutions of measures. This will prove useful in order to prove that $(f * g) \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$ provided $f, g \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$. Here, we restrict attention to the measurable space $(\mathbb{R}, \mathcal{A})$. We continue to write λ to mean the Lebesgue measure. Let μ and ν denote σ-finite measures on $(\mathbb{R}, \mathcal{A})$. The product measure is denoted by $\mu \otimes \nu$. Following Bogachev (2006, 207–8), the convolution of μ and ν, denoted by $\mu * \nu$, is defined as the measure on $\mathbb{R}$ that is the image of $\mu \otimes \nu$ on $\mathbb{R} \times \mathbb{R}$ under the mapping $(x, y) \mapsto x + y$, that is

$$\begin{aligned}(\mu * \nu)(B) &= \int_{\mathbb{R}^2} \mathbb{1}_B(x+y)\mu(\mathrm{d}x)\nu(\mathrm{d}y) \\ &= \int_{\mathbb{R}} \mu(B-y)\nu(\mathrm{d}y) = \int_{\mathbb{R}} \nu(B-x)\mu(\mathrm{d}x),\end{aligned} \tag{A.21}$$

where $\mathbb{1}_B$ denotes the indicator function. Now, if μ is absolutely continuous w.r.t. to the Lebesgue measure λ (denoted by $\mu \ll \lambda$), the Radon–Nikodym theorem proves that there exits a function $m : \mathbb{R} \to (0, \infty)$ which is integrable w.r.t. λ

(in fact this holds for more general measures, not only λ; see Thm. 3.2.2 in Bogachev (2006)) such that

$$\mu(E) = \int_E m(x)\lambda(\mathrm{d}x) \qquad \text{for all} \quad E \in \mathcal{A}, \tag{A.22}$$

which can also be written as $d\mu = m\mathrm{d}\lambda$. Likewise, we assume that $d\nu = n\mathrm{d}\lambda$. Hence,

$$(\mu * \nu)(B) = \int_{\mathbb{R}} m(B-y)n(y)\lambda(\mathrm{d}y), \tag{A.23}$$

where the r.h.s. is a Lebesgue integral. Comparing (A.23) with (A.18), we see that convolution of measures, provided the measures satisfy $\mu, \nu \ll \lambda$, coincides with convolution of L^1-functions. From Exercise 5e) in Rudin (1987, 175), we have that $\mu * \nu \ll \lambda$ if $\mu \ll \lambda$. In other words, for such measures, the convolution $\mu * \nu$, which coincides with convolution of L^1-functions, is absolutely continuous w.r.t. λ, which then translates to that $m * n$ is a.c. on $\mathbb{R}$. This finding is wrapped up in the subsequent proposition (for ease of referencing).

Proposition A.1. *Let $f, g \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$, then $(f * g) \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$.*

Our last result concerning convolutions is a theorem on differentiation under the integral sign. We denote by $C^\infty(\mathbb{R})$ the space of (continuous) infinitely differentiable functions (i.e., partial derivatives of all orders exist and are continuous) on $\mathbb{R}$. We restrict attention to functions on $\mathbb{R}$; the original result in Lang (1993) considers the more general case of $\mathbb{R}^p$.

Theorem A.7. *Let $f \in L^1(\mathbb{R})$ and let $g \in C^\infty(\mathbb{R})$ be a function with compact support. Then $f * g$ is $C^\infty(\mathbb{R})$ and for the n^{th}-order derivative, we have*

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}(f * g)(x) = f * \frac{\mathrm{d}^n}{\mathrm{d}x^n} g(x). \tag{A.24}$$

Proof. See Thm. 2.3 in Lang (1993, 227). ■

A candidate function in $C^\infty(\mathbb{R})$ is ϕ, i.e., the p.d.f. of the standard Gaussian distribution. This fact follows from the following observation: ϕ is a special case of the Gaussian function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = a \exp\left(-\frac{(x-b)^2}{2c^2} \right), \quad \text{with fixed parameters } a, b, c \in \mathbb{R}, \tag{A.25}$$

which is a real analytic function, i.e. a function that is locally represented by a convergent power series (which coincides with the Taylor series expansion), hence it is infinitely (continuously) differentiable; see Rudin (1976, 172–4).

A.2. Probability distributions

Without attempting to be comprehensive, we give a more or less lose collection of important results on some probability distributions and their interrelations. The goal of this section is twofold: (i) fix notation and parametrization for probability distributions which exhibit many different representations and (ii) state some results for ease of referencing.

Note: Independence of two r.v.'s X and Y is denoted by $X \perp\!\!\!\perp Y$.

Beta, incomplete beta, and gamma function

Let $\mathrm{B}(\alpha, \beta)$ and $I_x(\alpha, \beta)$ denote, respectively, the beta and the regularity incomplete beta function; the functions are defined as (see e.g. Abramowitz and Stegun, 1972, Eq.'s 6.2.1 and 6.6.1),

$$\mathrm{B}(\alpha, \beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t \tag{A.26}$$

and

$$I_x(\alpha, \beta) := \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t \tag{A.27}$$

for $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}_+$, and $x \in [0, 1]$. Of particular importance are the identity (Abramowitz and Stegun, 1972, Eq. 6.2.2)

$$\mathrm{B}(\alpha, \beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{A.28}$$

and the following recurrence relation for the gamma function,

$$\Gamma(z+1) = z\Gamma(z), \tag{A.29}$$

where, for $z \in \mathbb{R}_+$,

$$\Gamma(z) := \int_0^\infty x^{z-1}\exp(-x)\mathrm{d}x. \tag{A.30}$$

More generally, Γ could also be defined in terms of Euler's integral representation for complex-valued z such that $\Re[z] > 0$; see Abramowitz and Stegun (1972, Eq. 6.1.15).

Beta distribution

The p.d.f. of the beta distribution is given by (Johnson, Kotz, and Balakrishnan, 1995, 210–11)

$$f(x \mid \alpha, \beta) := \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}\big(1-x\big)^{\beta-1}, \qquad 0 < x < 1, \tag{A.31}$$

where $\alpha > 0$ and $\beta > 0$ are shape parameters; $\mathrm{B}(\alpha, \beta)$ denotes the beta function. The c.d.f. is denoted by $\mathrm{Beta}(\alpha, \beta)$. If $X \sim \mathrm{Beta}(\alpha, \beta)$, the first two central

moments are (Johnson et al., 1995, 217)

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \mathbb{V}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{A.32}$$

Gamma distribution

A r.v. X has a (two-parameter) gamma distribution, denoted by $\mathrm{Gamma}(k,\theta)$, if its p.d.f. is of form (Johnson, Kotz, and Balakrishnan, 1994, 337)

$$f(x \mid k,\theta) := \frac{1}{\Gamma(k)\theta^k} x^{k-1}\exp(-x/\theta), \qquad x \in (0,\infty), \tag{A.33}$$

where $k > 0$ is a shape parameter and $\theta > 0$ is a scale parameter; Γ denotes the gamma function. The gamma distribution features also a shape–rate parametrization. We shall not use this representation. The first two central moments are

$$\mathbb{E}[X] = k\theta \qquad \text{and} \qquad \mathbb{V}[X] = k\theta^2. \tag{A.34}$$

An important feature of the gamma distribution is its *reproductive property* (Johnson et al., 1994, 340). Let X_1 and X_2 be independent r.v. such that $X_1 \sim \mathrm{Gamma}(k_1,\theta)$ and $X_2 \sim \mathrm{Gamma}(k_2,\theta)$, then

$$(X_1 + X_2) \sim \mathrm{Gamma}(k_1+k_2,\theta) \tag{A.35}$$

and

$$\frac{X_1}{X_1+X_2} \sim \mathrm{Beta}(k_1+k_2,\theta). \tag{A.36}$$

The gamma distribution is related to the chi-square distribution: let $v \in \mathbb{N}_+$, then $\mathrm{Gamma}(v/2,2) \overset{d}{=} \chi^2_v$, i.e. the chi-square distribution with v degrees of freedom (Johnson et al., 1994, 338). Conversely, if $X \sim \chi^2_v$ and $c \in \mathbb{R}_+$ is some constant, then $cX \sim \mathrm{Gamma}(v/2, 2c)$.

For two independent gamma r.v.'s with a common scale parameter $\theta > 0$, we have the following nice result.

Theorem A.8 (Lukacs). *Suppose $X \sim \mathrm{Gamma}(a,\theta)$ and $Y \sim \mathrm{Gamma}(b,\theta)$, with $a,b,\theta > 0$, such that $X \perp\!\!\!\perp Y$. Put $U = X+Y$, $V = X/Y$, and $W = Y/(X+Y)$, then*

(i) $U \perp\!\!\!\perp V$ and

(ii) $U \perp\!\!\!\perp W$.

Proof. For (i), see Thm. 1 in Lukacs (1955); assertion (ii) follows from his Eq. (1.3) and the discussion on p. 321. ■

Another important observation concerns the distribution of the inverse of a gamma r.v. Suppose $X \sim \mathrm{Gamma}(k,\theta)$. Let $Y = 1/X$. Then, Y has an inverse gamma distr., formally $\text{Inv-}\mathrm{Gamma}(k,1/\theta)$. The first two moments of the inverse

gamma distr. are given by

$$\mathbb{E}[Y] = \frac{1}{\theta(k-1)} \qquad \text{and} \qquad \mathbb{V}[Y] = \frac{1}{\theta^2(k-1)^2(k-2)}. \tag{A.37}$$

Non-central chi-square distribution

A r.v. X has a non-central chi-square distribution with v degrees of freedom and non-centrality parameter λ, denoted by $\chi^2_v(\lambda)$, if its p.d.f. is of form (Johnson et al., 1995, 436)

$$f(x \mid v, \lambda) := \sum_{i=0}^{\infty} \frac{(\lambda/2)^i}{i!} \exp(-\lambda/2) f_{\chi^2}(x \mid v + 2i), \tag{A.38}$$

where $f_{\chi^2}(x \mid v)$ denotes the p.d.f. of the central chi-square distribution with v degrees of freedom. From this representation it is seen that $f(x \mid v, \lambda)$ is expressed as Poisson-weighted mixture of central χ^2_v p.d.f.'s. Stated differently: let $v > 0$ and $\lambda > 0$ be fixed and let $J \sim \text{Poisson}(\lambda/2)$. Conditionally, we have

$$(Z \mid J = j) \sim \chi^2_{v+2j}, \tag{A.39}$$

hence $Z \sim \chi^2_v(\lambda)$. It is possible to derive $\chi^2_v(\lambda)$ by a process of induction using the relation (Johnson et al., 1995, 436)

$$\chi^2_v(\lambda) = \chi^2_1(\lambda) + \chi^2_{v-1}. \tag{A.40}$$

Let $X \sim \chi^2_v(\lambda)$. The first two central moments are (Johnson et al., 1995, 447)

$$\mathbb{E}[X] = v + \lambda \qquad \text{and} \qquad \mathbb{V}[X] = 2(v + 2\lambda). \tag{A.41}$$

The non-central chi-square distributions satisfies the following reproductive property (Johnson et al., 1995, 450): let $X_1 \sim \chi^2_{v_1}(\lambda_1)$ and $X_2 \sim \chi^2_{v_2}(\lambda_2)$, then $(X_1 + X_2) \sim \chi^2_{v_1+v_2}(\lambda_1 + \lambda_2)$.

A.3. Results from probability theory

Let $(\Omega, \mathcal{F}, P)$ be a fixed probability space.

Theorem A.9. *If $\{X_i, 1 \leq i \leq n\}$ are independent r.v.'s and $\{f_i, 1 \leq i \leq n\}$ are Borel measurable functions, then $\{f_i(X_i), 1 \leq i \leq n\}$ are independent r.v.'s.*

Proof. See e.g. Thm. 3.3.1 in Chung (2001). ■

The next result is the Borel–Cantelli lemma.

Lemma A.5 (Borel–Cantelli lemma)**.** *Let $\{E_n, n \geq 1\}$ be a sequence of arbitrary*

events on $(\Omega, \mathcal{A}, P)$. *If*

$$\lim_{n\to\infty} \sum_n \mathbb{P}\{E_n\} < \infty \tag{A.42}$$

then

$$\mathbb{P}\big\{E_n \textit{ infinitely often}\big\} = 0. \tag{A.43}$$

Proof. See e.g. Thm. 4.2.1 in Chung (2001). ■

The term "infinitely often" will be abbreviated by the shorthand notation "i.o.". Now, the power of the Borel–Cantelli lemma is that it provides the following necessary and sufficient criterion of convergence almost surely (a.s.). Denote by $\{X_i, i \geq 1\}$ a sequence of r.v and let $\epsilon > 0$. Then $X_n \xrightarrow{\text{a.s.}} 0$ if and only if $\mathbb{P}\{|X_i| > \epsilon \text{ i.o.}\} = 0$ for all $\epsilon > 0$; see Chung (2001, Thm. 4.2.2).

Another indispensable tool is Kronecker's lemma, which is a result about the relationship between convergence of infinite sums and convergence of sequences.

Lemma A.6 (Kronecker's lemma)**.** *Let* $\{x_i, i \geq 1\}$ *be a sequence of real numbers and denote by* $\{a_i, i \geq 1\}$ *a monotone sequence of positive real numbers,* $0 < a_1 \leq a_2 \leq \cdots \leq a_n \uparrow \infty$ *as* $n \to \infty$*. Suppose that the series* $\lim_{n\to\infty} \sum_{i=1}^n x_i$ *converges. Then*

$$\frac{1}{a_n} \sum_{i=1}^n a_i x_i \to 0. \tag{A.44}$$

Proof. See Chung (2001, Lem. 5.4.1). ■

The next result is called the generalized Dini–Abel theorem. Since it is not well known, we give a detailed proof.

Lemma A.7 (generalized Abel–Dini Thm.; Petrov, 1969)**.** *Let* $\{b_i, i \geq 1\}$ *denote a sequence of nonnegative real numbers,* $B_n = \sum_{i=1}^n b_i$ *s.t.* $B_n \to \infty$ *as* $n \to \infty$*. Then the series*

$$\lim_{n\to\infty} \sum_n \frac{b_n}{B_n \gamma(B_n)} \tag{A.45}$$

converges for all $\gamma \in \Gamma$.

Proof. Let $n_0 \in \mathbb{N}_+$ be s.t. $B_{n_0} > 0$ and $\gamma(B_{n_0}) > 0$. It is easy to see that, for all $\gamma \in \Gamma$, the series $\lim_{n\to\infty} \sum_n 1/(n\gamma(n))$ converges; this implies the convergence of the (improper Riemann-) integral

$$I_{B_{n_0}} = \int_{B_{n_0}}^{\infty} \frac{\mathrm{d}x}{x\gamma(x)}. \tag{A}$$

By the mean-value theorem, we have for all $n > n_0$

$$\int_{B_{n-1}}^{B_n} \frac{\mathrm{d}x}{x\gamma(x)} = (B_n - B_{n-1})c_n, \tag{B}$$

where $c_n \in \mathbb{R}_+$ which is bounded by

$$\frac{1}{B_n \gamma(B_n)} \leq c_n \leq \frac{1}{B_{n-1}\gamma(B_{n-1})}. \tag{C}$$

Observe that $B_n - B_{n-1} = b_n$ hence (A) together with (B) and (C) imply

$$I_{B_{n_0}} = \sum_{n=n_0+1}^{\infty} \int_{B_{n-1}}^{B_n} \frac{\mathrm{d}x}{x\gamma(x)} = \sum_{n=n_0+1}^{\infty} (B_n - B_{n-1})c_n \geq \sum_{n=n_0+1}^{\infty} \frac{b_n}{B_n \gamma(B_n)}$$

and the assertion of the lemma follows since $I_{B_{n_0}}$ converges. ■

B. Simulation criteria

Let θ^* denote the parameter to be estimated at the true data y_i. We write $\hat{\theta}_k$ to mean the kth estimate of θ^*, given $k = 1, \ldots, r$ replications. Inspired by Hulliger et al. (2011, chap. 6.2), we define the following numerical criteria:

1. average of the point-estimates,

$$mean_\theta \equiv avg_\theta = \frac{1}{r} \sum_{k=1}^{r} \hat{\theta}_k \tag{B.1}$$

2. variance of the point-estimates,

$$var_\theta = \frac{1}{r-1} \sum_{k=1}^{r} \left(\hat{\theta}_k - avg_\theta\right)^2 \tag{B.2}$$

3. bias of the point estimates,

$$bias = \frac{1}{r} \sum_{k=1}^{r} \left(\hat{\theta}_k - \theta^*\right) \tag{B.3}$$

   or relative bias

$$bias(\%) = 100 \cdot \frac{bias}{\theta^*}\% \tag{B.4}$$

4. median point-estimate,

$$med_\theta = \mathrm{med}_k\left\{\hat{\theta}_k,\ k = 1, \ldots, r\right\} \tag{B.5}$$

5. median absolute deviation about the median point-estimate (cf. Rousseeuw and Croux, 1993, p.1273),

$$mad \equiv mad_\theta = 1.4826 \cdot \mathrm{med}_k\left\{|\hat{\theta}_k - med_\theta|,\ k = 1, \ldots, r\right\}, \tag{B.6}$$

6. median error of point estimates, (cf. Richardson and Welsh, 1995, p.1436),

$$mede = \mathrm{med}_k\left\{\hat{\theta}_k - \theta^*,\ k = 1, \ldots, r\right\}, \tag{B.7}$$

7. root mean square error of the point estimates,

$$rmse = \sqrt{(1/r) \sum_{k=1}^{r} \left(\hat{\theta}_k - \theta^*\right)^2} \tag{B.8}$$

or relative root mean square error of the point-estimates,

$$rmse(\%) = 100 \cdot \frac{rmse}{\theta^*} \tag{B.9}$$

8. median absolute error of the point estimates (cf. Richardson and Welsh, 1995, p.1436),

$$medae = 1.4826 \cdot \mathrm{med}_k\big\{|\hat{\theta_k} - \theta^*|,\ k = 1, \ldots, r\big\} \tag{B.10}$$

9. maximum absolute relative error, [This measure may be useful to assess the sensitivity of an estimator to the presence of influential units in the sample; Beaumont and Alavi (2004, p.12)],

$$mare(\%) = \max_k \big\{|(\hat{\theta}_k - \theta^*)/\theta^*|,\ k = 1, \ldots, r\big\}. \tag{B.11}$$

C. Case studies

C.1. Biomass data

In this section, we document some of the R code that underlies the computations discussed in Section 5.8. First, we load the required packages and the data from package JoSAE.

```
> library(rsae); library(JoSAE); library(lattice)
```

Next, the loaded data need some modifications (rename the variables etc.).

```
> data(JoSAE.domain.data); sample <- JoSAE.sample.data
> colnames(sample) <- c("sampleID", "domainID", "biomass", "canopy")
> sample[,"domainID"] <- factor(sample[,"domainID"], 1:14,
+ paste0("D", 1:14))
> data(JoSAE.sample.data); populationdata <- JoSAE.domain.data
> populationdata[,"intercept"] <- rep(1, 14)
> names(populationdata) <- c("domainID", "N", "canopy")
```

The data.frame populationdata has the following content

```
   domainID       N    canopy intercept
1         1  105267 108.15832         1
2         2  202513  77.34845         1
3         3  134156  94.26035         1
4         4  193807  86.64053         1
5         5 1379945  84.87776         1
6         6  176731  77.66091         1
7         7  474615  71.40756         1
8         8  442280  65.50692         1
9         9  495568  81.65170         1
10       10  520141  80.04376         1
11       11  230756  92.17368         1
12       12   83441  82.38918         1
13       13   57858  63.28690         1
14       14  905387  66.04283         1
```

where variable canopy contains the area-level population means of the auxiliary variable canopy; N is the population size (in each area). The variable intercept is needed, together with variable canopy, for the prediction of the area-level means; hence, we define

```
> population <- populationdata[, c(4,3)]
```

C.2. Toxoplasmosis data

Read in the toxoplasmosis data from Table 3 in Efron and Morris (1975) as `data.frame` `toxo`.

```
toxo <- read.table(
+ i; x; d; delta; a; k; b
+ 1; .293; .304; .035; .0120; 1334.1; .882
+ 2; .214; .039; .192; .0108; 21.9; .102
+ 3; .185; .047; .159; .0109; 24.4; .143
+ 4; .152; .115; .075; .0115; 80.2; .509
+ 5; .139; .081; .092; .0112; 43.0; .336
+ 6; .128; .061; .100; .0110; 30.4; .221
+ 7; .113; .061; .088; .0110; 30.4; .221
+ 8; .098; .087; .062; .0113; 48.0; .370
+ 9; .093; .049; .079; .0109; 25.1; .154
+ 10; .079; .041; .070; .0109; 22.5; .112
+ 11; .063; .071; .045; .0111; 36.0; .279
+ 12; .052; .048; .044; .0109; 24.8; .148
+ 13; .035; .056; .028; .0110; 28.0; .192
+ 14; .027; .040; .024; .0108; 22.2; .107
+ 15; .024; .049; .020; .0109; 25.1; .154
+ 16; .024; .039; .022; .0108; 21.9; .102
+ 17; .014; .043; .012; .0109; 23.1; .122
+ 18; .004; .085; .003; .0112; 46.2; .359
+ 19; -.016; .128; -.007; .0116; 101.5; .564
+ 20; -.028; .091; -.017; .0113; 51.6; .392
+ 21; -.034; .073; -.024; .0111; 37.3; .291
+ 22; -.040; .049; -.034; .0109; 25.1; .154
+ 23; -.055; .058; -.044; .0110; 28.9; .204
+ 24; -.083; .070; -.060; .0111; 35.4; .273
+ 25; -.098; .068; -.072; .0111; 34.2; .262
+ 26; -.100; .049; -.085; .0109; 25.1; .154
+ 27; -.112; .059; -.089; .0110; 29.4; .210
+ 28; -.138; .063; -.106; .0110; 31.4; .233
+ 29; -.156; .077; -.107; .0112; 40.0; .314
+ 30; -.169; .073; -.120; .0111; 37.3; .291
+ 31; -.241; .106; -.128; .0114; 68.0; .468
+ 32; -.294; .179; -.083; .0118; 242.4; .719
+ 33; -.296; .064; -.225; .0111; 31.9; .238
+ 34; -.324; .152; -.114; .0117; 154.8; .647
+ 35; -.397; .158; -.133; .0117; 171.5; .665
+ 36; -.665; .216; -.140; .0119; 426.8; .789,
+ sep = ";" )
```

Rename the columns of the `data.frame` `toxo`.

```
names( toxo ) <- c( "i", "Xi", "sqrtDi", "delta", "Ai", "Ki", "Bi" )
```

Define the function to compute the empirical Bayes risk.

```
EBrisk <- function( A, delta, Di, Xi ){
+   n <- length( delta )
```

```
+   D0 <- mean( Di )
+   Bi <- Di / ( A + Di)
+   sum( delta^2 + (1-Bi) * ( Di - 2*delta*Xi + (1-Bi)*Xi^2 ) ) / (n*D0)
+ }
```

C.3. Crop yield data for paddy in the State of Uttar Pradesh

The data (see Table C.1) are downlodable from the Sample Survey Resource Server maintained by the Indian Agricultural Statistics Research Institute in New Dehli; see http://sample.iasri.res.in/ssrs/SAE_Using_R.pdf.

Table C.1.: Crop yield data for paddy in the State of Uttar Pradesh

District	yield	nd	vd	HH_F	HH_SIZE
1	19575.0	10.0	6511409.0	4015.0	6.3
2	23483.3	6.0	23239128.0	9155.0	6.7
3	19441.7	12.0	2001095.0	12700.0	6.5
4	17700.0	8.0	8709740.0	13474.0	6.6
5	17250.0	8.0	268926.8	7925.0	6.7
6	10850.0	4.0	2204171.0	10281.0	6.6
7	16800.0	4.0	27181677.0	6980.0	6.1
8	17417.9	14.0	4113834.0	25771.0	6.4
9	12418.8	8.0	6030969.0	16420.0	6.4
10	10482.5	4.0	8171308.0	10277.0	6.5
11	12125.0	10.0	1391514.0	9336.0	6.5
12	14018.8	8.0	10023130.0	2163.0	6.6
13	12721.4	14.0	3667241.0	7492.0	6.5
14	13510.7	14.0	3169703.0	12903.0	6.5
15	14937.5	8.0	8876054.0	4562.0	6.5
16	18862.5	12.0	1379737.0	4875.0	6.4
17	14975.0	16.0	2898957.0	5357.0	6.3
18	15986.2	20.0	4389845.0	9107.0	6.1
19	19286.1	18.0	2028998.0	17407.0	6.1
20	12842.9	14.0	2344970.0	10832.0	5.8
21	17331.2	8.0	13348221.0	7142.0	5.7
22	19505.6	18.0	3102449.0	7359.0	5.8
23	8880.0	5.0	2831658.0	12284.0	6.5
24	34050.0	4.0	3437501.0	27842.0	6.5
25	15462.5	4.0	615573.6	2068.0	6.2
26	23716.7	6.0	5583780.0	4077.0	6.1
27	21200.0	8.0	5259643.0	3146.0	6.0
28	15375.0	8.0	7046690.0	5280.0	5.8
29	8887.5	4.0	19079337.0	3780.0	6.2
30	14612.0	10.0	8633866.0	5658.0	5.9
31	16303.6	14.0	5600813.0	8936.0	6.2
32	15450.0	8.0	17310434.0	6167.0	6.0
33	19465.0	20.0	5204632.0	39002.0	6.7
34	18667.9	14.0	4307989.0	10470.0	5.8
35	16379.2	12.0	5454831.0	7184.0	6.1

(to be continued)

Crop yield data for paddy in the State of Uttar Pradesh *(continued)*

District	yield	nd	vd	HH_F	HH_SIZE
36	17691.7	12.0	2790868.0	6299.0	6.9
37	16608.9	18.0	2527936.0	14993.0	6.3
38	14714.3	14.0	327126.5	2707.0	6.2
39	15075.0	4.0	8118966.0	2446.0	6.2
40	11975.0	10.0	3070124.0	2629.0	6.6
41	16981.2	16.0	1206812.0	3699.0	6.4
42	12828.6	14.0	3020820.0	2703.0	6.6
43	14267.9	14.0	5347319.0	5385.0	6.6
44	13318.8	8.0	716242.9	3314.0	6.8
45	21690.0	10.0	6941439.0	3806.0	6.5
46	12163.9	18.0	2397597.0	7306.0	6.8
47	19342.9	14.0	8284146.0	6096.0	6.8
48	8363.9	18.0	2162576.0	4864.0	7.2
49	11957.1	28.0	1036936.0	13539.0	7.3
50	9820.0	10.0	3720622.0	14498.0	7.4
50	9820.0	10.0	3720622.0	14498.0	7.4
51	7029.2	12.0	2132824.0	7507.0	7.3
52	16990.0	20.0	3043116.0	17876.0	7.1
53	10858.3	18.0	2083432.0	13337.0	7.1
54	12000.0	10.0	5551114.0	7941.0	7.0
55	17665.0	10.0	7377556.0	28651.0	7.3
56	6693.3	6.0	5872979.0	15224.0	8.4
57	15625.0	10.0	3348231.0	9018.0	6.9
58	15283.3	6.0	16396272.0	2116.0	5.8

D. Consistency correction term

We have shown in Section 5.6.2 that it is sufficient to restrict attention to the bivariate Gaussian distribution, instead of $\tilde{\mathbf{r}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{U}_i^{-1/2}\mathbf{\Omega}_i\mathbf{U}_i^{-1/2})$, for all $i = 1, \ldots, g$. Therefore, let $(X, Y)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ with

$$\mathbf{V} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \qquad \text{where} \quad 1 < \rho < 1. \tag{D.1}$$

We remark that bounding ρ away from elements in the set $\{-1, 1\}$ is not necessary for the definition of the bivariate normal law but will be convenient. The bivariate normal law with centered first moments and covariance matrix in (D.1) implies the conditional distribution $(Y \mid X = x) \sim \mathcal{N}(\rho x, 1 - \rho^2)$ with p.d.f.

$$f(y \mid x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\bigg(- \frac{(y - \rho x)^2}{2(1-\rho^2)} \bigg), \tag{D.2}$$

which enables us to express the p.d.f. of the above bivariate normal as

$$f(x, y) = \phi(x) f(y \mid x). \tag{D.3}$$

The consistency correction term for σ_u^2 is defined as

$$\begin{aligned} \mathbb{E}_{f(x,y)}\big[\psi_k(X)\psi_k(Y)\big] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi_k(x)\psi_k(y) f(x, y) \mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \psi_k(x) \underbrace{\bigg(\int_{-\infty}^{\infty} \psi_k(y) f(y \mid x) \mathrm{d}y \bigg)}_{= A, \text{ say}} \phi(x) \mathrm{d}x. \end{aligned} \tag{D.4}$$

First, we work on A. Let

$$\tau = \sqrt{1 - \rho^2}, \qquad \underline{k} = \frac{-k - x\rho}{\tau}, \quad \text{and} \quad \overline{k} = \frac{k - x\rho}{\tau}. \tag{D.5}$$

We have,

$$\begin{aligned} A \equiv \int_{-\infty}^{\infty} \psi_k(y) f(y \mid x) \mathrm{d}y = \frac{k}{\tau}\int_{-\infty}^{k} \phi\bigg(\frac{y - x\rho}{\tau}\bigg)\mathrm{d}y + \frac{k}{\tau}\int_{k}^{\infty} \phi\bigg(\frac{y - x\rho}{\tau}\bigg)\mathrm{d}y \\ + \frac{1}{\tau}\int_{-k}^{k} y\phi\bigg(\frac{y - x\rho}{\tau}\bigg)\mathrm{d}y, \end{aligned} \tag{D.6}$$

where

$$\frac{k}{\tau}\int_{-\infty}^{k} \phi\bigg(\frac{y - x\rho}{\tau}\bigg)\mathrm{d}y = k\int_{-\infty}^{\underline{k}} \phi(t)\mathrm{d}t = k\Phi(\underline{k}), \tag{D.7}$$

and, likewise,

$$\frac{k}{\tau}\int_{k}^{\infty}\phi\Big(\frac{y-x\rho}{\tau}\Big)\mathrm{d}y = k\int_{\overline{k}}^{\infty}\phi(t)\mathrm{d}t = k\big(1-\Phi(\overline{k})\big). \tag{D.8}$$

By identity $\phi'(u) = -u\phi(u)$, we obtain

$$\begin{aligned}\frac{1}{\tau}\int_{-k}^{k} y\phi\Big(\frac{y-x\rho}{\tau}\Big)\mathrm{d}y &= \int_{\underline{k}}^{\overline{k}}(t\tau+x\rho)\phi(t)\mathrm{d}t = \tau\int_{\underline{k}}^{\overline{k}} t\phi(t)\mathrm{d}t + x\rho\int_{\underline{k}}^{\overline{k}}\phi(t)\mathrm{d}t \\ &= -\tau\int_{\underline{k}}^{\overline{k}}\phi'(t)\mathrm{d}t + x\rho\big(\Phi(\overline{k})-\Phi(\underline{k})\big) \\ &= \tau\big(\phi(\underline{k})-\phi(\overline{k})\big) + x\rho\big(\Phi(\overline{k})-\Phi(\underline{k})\big). \end{aligned}\tag{D.9}$$

On collecting the term terms, we get

$$\begin{aligned}A(x) \equiv \int_{-\infty}^{\infty}\psi_k(y)f(y\mid x)\mathrm{d}y &= x\rho\big(\Phi(\overline{k})-\Phi(\underline{k})\big) + \tau\big(\phi(\underline{k})-\phi(\overline{k})\big) \\ &\quad + k\big(1+\Phi(\underline{k})-\Phi(\overline{k})\big). \end{aligned}\tag{D.10}$$

Now, we may write

$$\begin{aligned}&E_{f(x,y)}\big[\psi_k(X)\psi_k(Y)\big] \\ &= k\int_{-\infty}^{-k}A(x)\phi(x)\mathrm{d}x + \int_{-k}^{k}xA(x)\phi(x)\mathrm{d}x + k\int_{k}^{\infty}A(x)\phi(x)\mathrm{d}x.\end{aligned}$$

One may have the impression that, using the the integrals of Gaussian functions (Owen, 1980),

$$\begin{aligned}\int\phi(x)\phi(a+bx)\mathrm{d}x &= \frac{1}{\sqrt{1+b^2}}\phi\Big(\frac{a}{\sqrt{1+b^2}}\Big)\Phi\Big(\sqrt{1+b^2}x+\frac{ab}{\sqrt{1+b^2}}\Big)+C \\ \int x\Phi(a+bx)\mathrm{d}x &= \frac{b^2x^2-a^2-1}{2b^2}\Phi(a+bx)+\frac{bx-a}{2b^2}\phi(a+bx)\Big]+C \\ \int x^2\Phi(a+bx)\mathrm{d}x &= \frac{b^3x^3+a^3+3a}{3b^3}\Phi(a+bx)+\frac{b^2x^2-abx+a^2+2}{3b^3}\phi(a+bx)+C \\ \int x\phi(x)\Phi(a+bx)\mathrm{d}x &= \frac{b}{\sqrt{1+b^2}}\phi\Big(\frac{a}{\sqrt{1+b^2}}\Big)\Phi\Big(x\sqrt{1+b^2}+\frac{ab}{\sqrt{1+b^2}}\Big) \\ &\quad -\phi(x)\Phi(a+bx)+C\end{aligned}$$

we can obtain a "reasonable" expression for $E_{f(x,y)}[\psi_k(X)\psi_k(Y)]$. Though, we did not succeed. From our point of view, the "building blocks" of Gaussian integrals are already "too messy". Hopefully, this attempt to solve the problem may help or inspire someone else with his/ her endeavor.

Bibliography

ABRAMOWITZ, M. AND I. STEGUN (1972): *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Washinton (D.C.): US Government Printing Office, 10th ed.

ALFONS, A. (2014): *simFrame: Simulation framework*, [R package version 0.5.3].

ANDERSON, E., Z. BAI, C. BISCHOF, L. S. BLACKFORD, J. DEMMEL, J. DONGARRA, J. D. CROZ, A. GREENHAUM, S. HAMMARLING, A. MCKENNEY, AND D. SORENSEN (2000): *LAPACK Users' Guide*, Philadelphia: Society for Industrial and Applied Mathematics (SIAM), 3rd ed.

ANDREWS, D. F., P. J. BICKEL, F. R. HAMPEL, P. J. HUBER, W. H. ROGERS, AND J. W. TUKEY (1972): *Robust Estimates of Location: Survey and Advances*, Princeton: Princeton University Press.

ANGERS, J.-F. AND J. O. BERGER (1991): "Robust Hierarchical Bayes Estimation of Exchangeable Means," *The Canadian Journal of Statistics*, 19, 39–56.

ARORA, V. AND P. LAHIRI (1997): "On the superiority of the Bayesian method over the BLUP in small area estimation problems," *Statistica Sinica*, 4, 1053–1063.

BARANCHIK, A. J. (1964): "Multiple regression and estimation of the mean of a multivariate normal distribution," Tech. rep., Stanford University, Technical Report no. 51.

BATTESE, G. E., R. M. HARTER, AND W. A. FULLER (1988): "An error component model for prediction of county crop areas using," *Journal of the American Statistical Association*, 83, 28–36.

BEAUMONT, J.-F. AND A. ALAVI (2004): "Robust Generalized Regression Estimation," *Survey Methodology*, 30, 195–208.

BEAUMONT, J.-F. AND L.-P. RIVEST (2009): "Dealing with outliers in survey data," in *Sample Surveys: Theory, Methods and Inference*, ed. by D. Pfeffermann and C. Rao, Amsterdam: Elsevier, vol. 29A of *Handbook of Statistics*, chap. 11, 247–280.

BEDNARSKI, T. AND S. ZONTEK (1996): "Robust estimation of parameters in a mixed unbalanced model," *The Annals of Statistics*, 24, 1493–1510.

BERGER, J. O. (1982): "Estimation in continuous exponential families: Bayesian estimation suject to risk restrictions and inadmissibility results," in *Statistical Decision Theory and Related Topics*, ed. by S. S. Gupta and J. O. Berger, New York: Academic Press, vol. 3, 109–141.

——— (1985): *Statistical Decision Theory and Bayesian Analysis*, New York: Springer, 2nd ed.

——— (1994): "An overview of robust Bayesian analysis," *Test*, 3, 5–124.

BERKHIN, P. E. AND B. Y. LEVIT (1980): "Second-order asymptotically minimax estimates for the mean of a normal population," *Problems Inform. Transmission*, 16, 212–229, [original paper in Russian: Probl. Peredachi Inf., 16:3 (1980), 60–79].

BICKEL, P. J. (1980): "Minimax estimation of the mean of a normal distribution subject to doing well at a point," Research report, University of California, Berkeley.

——— (1981): "Minimax estimation of the mean of a normal distribution when the parameter space is restricted," *The Annals of Statistics*, 9, 1301–1309.

——— (1983): "Minimax estimation of the mean of a normal distribution subject to doing well at a point," in *Recent Advances in Statistics. Festschrift for H. Chernoff*, ed. by H. Rizvi and D. Siegmund, New York: Academic Press, 511–528.

——— (1984): "Parameter robustness: small biases can be worthwhile," *The Annals of Statistics*, 12, 864–879.

BICKEL, P. J. AND J. R. COLLINS (1983): "Minimizing Fisher Information over Mixtures of Distributions," *Sankhyā. Series A*, 45, 1–19.

BINDER, D. A. (1983): "On the variances of asymptotically normal estimators from complex surveys," *International Statistical Review*, 51, 279–292.

BIRCH, J. AND R. MYERS (1982): "Robust Analysis of Covariance," *Biometrics*, 38, 699–713.

BLACKFORD, L. S., A. PETITET, R. POZO, K. REMINGTON, R. C. WHALEY, J. DEMMEL, J. DONGARRA, I. DUFF, S. HAMMARLING, G. HENRY, M. HEROUX, L. KAUFMAN, AND A. LUMSDAINE (2002): "An updated set of basic linear algebra subprograms (BLAS)," *ACM Transactions on Mathematical Software*, 28, 135–151.

BOCK, M. E. (1988): "Shrinkage estimators: pseudo-Bayes rules for normal mean vectors," in *Statistical Decision Theory and Related Topics*, ed. by S. S. Gupta and J. O. Berger, New York: Springer, 281–298.

BOGACHEV, V. I. (2006): *Measure Theory*, vol. 1, New York: Springer.

BOYD, S. AND L. VANDENBERGHE (2004): *Convex Optimization*, Cambridge: Cambridge University Press.

BRANDWEIN, A. C. AND W. E. STRAWDERMAN (1990): "Stein Estimation: The Spherically Symmetric Case," *Statistical Science*, 5, 356–369.

BREIDENBACH, J. (2015): *JoSAE: Functions for some Unit-Level Small Area Estimators and their Variances*, [R package version 0.2-3].

BREIDENBACH, J. AND R. ASTRUP (2012): "Small area estimation of forest attributes in the Norwegian National Forest Inventory," *European Journal of Forest Research*, 131, 1255–1267.

BREIDT, F. J. AND J. D. OPSOMER (2000): "Local polynomial regression estimators in survey sampling," *The Annals of Statistics*, 28, 1026–1053.

BRENT, R. P. (1973): *Algorithms for Minimization without Derivatives*, Englewood Cliffs, NJ: Prentice-Hall.

BREWER, K. R. W. (1979): "A class of robust sampling designs for large-scale surveys," *Journal of the American Statistical Society*, 74, 911–915.

——— (1999): "Design-based or Prediction Inference? Stratified Random vs. Stratified Balanced Sampling," *International Statistical Review*, 67, 35–47.

——— (2005): "Anomalies, probings, insights: Ken Foreman's role in the sampling inference controversy of the late 20th century," *Australian and New Zealand Journal of Statistics*, 47, 385–399.

——— (2013): "Three controversies in the history of survey sampling," *Survey Methodology*, 39, 249–262.

BROWN, L. D. (1971): "Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems," *Annals of Mathematical Statistics*, 42, 855–904.

BURGARD, J. P., R. MÜNNICH, AND T. ZIMMERMANN (2014): "The Impact of Sampling Designs on Small Area Estimates for Business Data," *Journal of Official Statistics*, 30, 749–771.

CASSEL, C.-M., C.-E. SÄRNDAL, AND J. H. WRETMAN (1976): "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations," *Biometrika*, 63, 615–620.

——— (1977): *Foundations of Inference in Survey Sampling*, New York: Wiley.

CHAMBERS, R. (1986): "Outlier Robust Finite Population Estimation," *Journal of the American Statistical Association*, 81, 1063–1069.

CHAMBERS, R., H. CHANDRA, N. SALVATI, AND N. TZAVIDIS (2014): "Outlier Robust Small Area Estimation," *Journal of the Robust Statistical Society. Series B*, 76, 47–69.

CHAMBERS, R., H. CHANDRA, AND N. TZAVIDIS (2011): "On bias-robust mean squared error estimation for pseudo-linear small area estimators," *Survey Methodology*, 37, 153–170.

CHAMBERS, R. AND N. TZAVIDIS (2006): "M-quantile models for small area estimation," *Biometrika*, 92, 255–268.

CHAMBERS, R. L., A. H. DORFMAN, AND T. E. WERLY (1993): "Bias Robust Estimation in Finite Populations Using Nonparametric Calibration," *Journal of American Statistical Association*, 88, 268–277.

CHAMBERS, R. L. AND A. W. DORFMAN (2003): "Transformed Variables in Survey Sampling," Working paper, southampton statistical science research institute, University of Southampton.

CHANDRA, H., N. SALVATI, AND U. C. SUD (2011): "Disaggregate-level estimates of indebtedness in the state of Uttar Pradesh in India: an application of small-area estimation technique," *Journal of Applied Statistics*, 38, 2413–2432.

CHANDRA, T. AND A. GOSWAMI (1992): "Cesàro uniform integrability and the strong laws of large numbers," *Sankhyā, Series A*, 54, 215–231.

CHATRCHI, G. (2012): *Robust estimation of variance components in small area estimation*, Ottawa, master thesis, Carleton University.

CHAUBEY, Y. P. AND K. VENKATESWARLU (2002): "Robust estimators for the one-way variance component model," in *Advances on Methodological and Applied Aspects of Probability and Statistics*, ed. by N. Balakrishnan, New York: Taylor & Francis, chap. 14, 241–260.

CHAUDHURI, A. AND H. STENGER (2005): *Survey Sampling : Theory and Methods*, Statistics: Textbooks and Monographs, Boca Raton (FL): Chapman & Hall.

CHAUVET, G. (2014): "A note on the consistency of the Narain-Horvitz-Thompson estimator," Tech. rep., unpublished manuscript, arXiv: 1412.2887v1.

CHEN, P. AND S. SUNG (2016): "A strong law of large numbers for nonnegative random variables and applications," *Statistics and Probability Letters*, 118, 80–86.

CHUNG, K. L. (2001): *A Course in Probability Theory*, London: Academic Press, 3rd ed.

CLARKE, B. AND C. MILNE (2004): "Small Sample Bias Correction for Huber's Proposal-2 Scale M-Estimator," *Australian and New Zealand Journal of Statistics*, 46, 649–56.

COLLINS, J. R. (1976): "Robust Estimation of a Location Parameter in the Presence of Asymmetry," *The Annals of Statistics*, 4, 68–85.

——— (1977): "Upper Bounds on Asymptotic Variances of M-Estimators of Location," *The Annals of Statistics*, 5, 646–657.

COPT, S. AND M.-P. VICTORIA-FESER (2006): "High Breakdown Inference in the Mixed Linear Model," *Journal of the American Statistical Association*, 101, 292–300.

——— (2009): "Robust prediction in mixed linear models," Tech. rep., University of Geneva.

CSÖRGÖ, S., K. TANDORI, AND V. TOTIK (1983): "On the strong law of large numbers for pairwise independent random variables." *Acta Mathematica Hungarica*, 42, 319–330.

DATTA, G. S. (2009): "Model-Based Approach to Small-Area Estimation," in *Sample Surveys: Theory, Methods and Inference*, ed. by D. Pfeffermann and C. Rao, Amsterdam: Elsevier, vol. 29B of *Handbook of Statistics*, chap. 32, 251–288.

DATTA, G. S. AND P. LAHIRI (1995): "Robust Hierarchical Bayes Estimation of Small Are Characteristics in the Presence of Covariates and Outliers," *Journal of Multivariate Analysis*, 54, 310–328.

——— (2000): "A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems," *Statistica Sinica*, 10, 613–627.

DATTA, G. S., J. N. K. RAO, AND D. SMITH (2005): "On measuring the variability of small area estimators under a basic area level model," *Biometrika*, 92, 183–196.

DAVID, H. A. AND H. N. NAGARAJA (2003): *Order Statistics*, Hoboken: John Wiley and Sons, 3rd ed.

DAVIES, P. L. AND U. GATHER (2005): "Breakdown and groups," *The Annals of Statistics*, 33, 977–988.

——— (2007): "The breakdown point: examples and counterexamples," *REVSTAT: Statistical Journal*, 5, 1–17.

DEMIDENKO, E. (2004): *Mixed Models: Theory and Applications*, Hoboken: John Wiley & Sons.

DEVILLE, J.-C. (1999): "The Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques," *Survey Methodology*, 25, 193–203.

DEY, D. K. AND J. O. BERGER (1983): "On Truncation of Shrinkage Estimators in Simultaneous Estimation of Normal Means," *Journal of the American Statistical Association*, 78, 865–869.

DONOHO, D. L. AND P. J. HUBER (1983): "The notion of breakdown point," in *A Festschrift for Erich L. Lehmann: In honor of his sixty-fifths birthday*, ed. by P. J. Bickel, K. A. Doksum, and J. L. Hodges, Belmont, CA: Wadsworth, 157–184.

DUCHESNE, P. (1999): "Robust calibration estimators," *Survey Methodology*, 25, 43–56.

DÜMBGEN, L., K. NORDHAUS, AND H. SCHUMACHER (2014): *fastM: Fast Computation of Multivariate M-estimators*, [R package version 0.0-2].

——— (2016): "New Algorithms for M-Estimation of Multivariate Scatter and Location," *Journal of Multivariate Analysis*, 144, 200–2017.

DURRETT, R. (2010): *Probability: Theory and Examples*, Cambridge: Cambridge University Press, 4th ed.

DUTTER, R. (1977a): "Algorithms for the Huber estimator in multiple regression," *Computing*, 18, 167–176.

——— (1977b): "Numerical solution of robust regression problems: Computational aspects, a comparison," *Journal of Statistical Computing and Simulation*, 5, 207–238.

EFRON, B. AND C. MORRIS (1971): "Limiting the Risk of Bayes and Empirical Bayes Estimators–Part I: The Bayes Case," *Journal of the American Statistical Association*, 66, 807–815.

——— (1972): "Limiting the Risk of Bayes and Empirical Bayes Estimators–Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, 67, 130–139.

——— (1973a): "Combining Possibly Related Estimation Problems [with discussion]," *Journal of the Royal Statistical Society. Series B*, 35, 379–421.

——— (1973b): "Stein's Estimation Rule and Its Competitors–An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117–130.

——— (1975): "Data Analysis Using Stein's Estimator and its Generalizations," *Journal of the American Statistical Association*, 70, 311–319.

ESTEVAO, V. M., M. A. HIDIROGLOU, AND C.-E. SÄRNDAL (1995): "Methodological principles for a generalized estimation system at Statistics Canada," *Journal of Official Statistics*, 11, 181–204.

ETEMADI, N. (1983a): "On the laws of large numbers for nonnegative random variables," *Journal of Multivariate Analysis*, 13, 187–193.

——— (1983b): "Stability of Sums of Weighted Nonnegative Random Varibles," *Journal of Multivariate Analysis*, 13, 361–365.

FAHRMEIR, L. AND H. KAUFMANN (1985): "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models," *The Annals of Statistics*, 13, 342–368.

FANG, K.-T., S. KOTZ, AND K.-W. NG (1990): *Symmetric Multivariate and Related Distributions*, Monographs on Statistics and Applied Probability, London: Chapman and Hall / CRC Press.

FAY, R. E. AND R. A. HERRIOT (1979): "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, 74, 269–277.

FELLER, W. (1968): *An Introduction to Probability Theory and its Applications*, vol. 1, New York: John Wiley & Sons, 3rd ed.

——— (1971): *An Introduction to Probability Theory and its Applications*, vol. 2, New York: John Wiley & Sons, 3rd ed.

FELLNER, W. (1986): "Robust estimation of variance components," *Technometrics*, 28, 51–60.

FULLER, W. A. (1991): "Simple estimators for the mean of skewed populations," *Statistica Sinica*, 1, 137–158.

——— (2009): *Sampling Statistics*, Hoboken, NJ: Wiley.

GENTLE, J. E. (2007): *Matrix Algebra: Theory, Computations, and Applications in Statistics*, New York: Springer.

GEORGE, E. (1986a): "Combining Minimax Shrinkage Estimators," *Journal of the American Statistical Association*, 81, 437–445.

——— (1986b): "Minimax Multiple Shrinkage Estimation," *The Annals of Statistics*, 14, 188–205.

GERSHUNSKAYA, J., J. JIANG, AND P. LAHIRI (2009): "Resampling Methods in Surveys," in *Sample Surveys: Inference and Analysis*, ed. by D. Pfeffermann and C. Rao, Amsterdam: Elsevier, vol. 29B of *Handbook of Statistics*, chap. 28, 121–151.

GHOSH, M., T. MAITI, AND A. ROY (2008): "Influence functions and robust Bayes and empirical Bayes small area estimation," *Biometrika*, 95, 573–585.

GNANADESIKAN, R. AND J. KETTENRING (1972): "Robust estimates, residuals, and outlier detection with multiresponse data," *Biometrics*, 28, 81–124.

GODAMBE, V. (1982): "Estimation in Survey Sampling: Robustness and Optimality," *Journal of the American Statistical Association*, 77, 393–406.

GODAMBE, V. P. (1960): "An Optimum Property of Regular Maximum Likelihood Estimation," *The Annals of Mathematical Statistics*, 31, 1208–1211.

GOLUB, G. AND C. VAN LOAN (1996): *Matrix Computations*, London: The Johns Hopkins University Press, 3rd ed.

GWET, J.-P. (1997): "Robust Statistical Inference in Survey Sampling," Ph.D. thesis, Carleton University, Department of Mathematics and Statistics.

GWET, J.-P. AND L.-P. RIVEST (1992): "Outlier Resistant Alternatives to the Ratio Estimator." *Journal of the American Statistical Association*, 87, 1174–1182.

HALL, P. AND T. MAITI (2006): "On parametric bootstrap methods for small area prediction," *Journal of the Royal Statistical Society. Series B*, 68, 221–238.

HAMPEL, F. R. (1968): "Contributions to the theory of robust estimation," Ph.D. thesis, Department of Statistics, University of California, Berkley.

——— (1974): "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, 69, 383–393.

HAMPEL, F. R., E. M. RONCHETTI, P. J. ROUSSEEUW, AND W. A. STAHEL (1986): *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley & Sons.

HANDSCHIN, E., E. KOHLAS, A. FIECHTER, AND F. SCHWEPPE (1975): "Bad Data Analysis for Power System State Estimation," *IEEE Transactions on Power Apparatus and Systems*, 2, 329–337.

HANSEN, M. H., W. N. HURWITZ, AND W. G. MADOW (1953): *Sample Survey Methods and Theory*, New York: Wiley.

HARTLEY, H. O. AND J. N. K. RAO (1967): "Maximum Likelihood Estimation for the Mixed Analysis of Variance," *Biometrika*, 54, 93–108.

HARVILLE, D. A. (1977): "Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72, 320–338.

HEDLIN, D., H. FALVEY, R. CHAMBERS, AND P. KOKIC (2001): "Does the Model Matter for GREG Estimation? A Business Survey Example," *Journal of Official Statistics*, 17, 527–544.

HERITIER, S., E. CANTONI, S. COPT, AND M.-P. VICTORIA-FESER (2009): *Robust Methods in Biostatistics*, New York: John Wiley & Sons.

HETTMANSPERGER, T. P. AND J. W. MCKEAN (1998): *Robust Nonparametric Statistical Methods*, London: Arnold, 2nd ed.

HIDIROGLOU, M. AND P. LAVALLÉE (2009): "Sampling and Estimation in Business Surveys," in *Sample Surveys: Theory, Methods and Inference*, ed. by D. Pfeffermann and C. Rao, Amsterdam: Elsevier, vol. 29A of *Handbook of Statistics*, chap. 11, 441–470.

HIDIROGLOU, M. A. AND Z. PATAK (2004): "Domain Estimation Using Linear Regression," *Survey Methodology*, 30, 67–78.

HODGES, J. L. AND E. L. LEHMANN (1952): "The use of previous experience in reaching statistical decisions," *Annals of Mathematical Statistics*, 23, 396–407.

HORVITZ, D. G. AND D. J. THOMPSON (1952): "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

HUBER, P. J. (1964): "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101.

——— (1973): "Robust regression: asymptotics, conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821.

——— (1981): *Robust Statistics*, New York: John Wiley & Sons.

——— (1983): "Minimax Aspects of Bounded-Influence Regression," *Journal of the American Statistical Association*, 78, 66–72.

HULLIGER, B. (1991): "Nonparametric M-estimation of a population mean," Ph.D. thesis, ETH Zurich, Nr. 9443.

——— (1995): "Outlier Robust Horvitz-Thompson Estimators," *Survey Methodology*, 21, 79–87.

——— (1999): "Simple and robust estimators for sampling," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, American Statistical Association, 54–63.

HULLIGER, B., A. ALFONS, C. BRUCH, P. FILZMOSER, M. GRAF, J.-P. KOLB, R. LEHTONEN, D. LUSSMANN, A. MERANER, R. MÜNNICH, M. MYRSKYLÄ, D. NEDYALKOVA, J. SEGER, T. SCHOCH, M. TEMPL, M. VALASTE, A. VEIJANEN, , AND S. ZINS (2011): *Report on the analysis of the simulation results*,

Deliverable 7.1, AMELI: Advanced methodology for European Laeken indicators, research project funded by the European Commission, FP7-SSH-2007-217322.

ISAKI, C. T. AND W. A. FULLER (1982): "Survey Design Under the Regression Superpopulation Model," *Journal of the American Statistical Association*, 77, 89–96.

JAMES, W. AND C. STEIN (1961): "Estimation With Quadratic Loss," in *Proceeding of the Fourth Berkeley Symposium of Mathematical Statistics and Probability*, Berkeley: University of California Press, vol. 1, 361–379.

JIANG, J., Y. LUAN, AND Y.-G. WANG (2007): "Iterative estimating equations: Linear convergence and asymptotic properties," *The Annals of Statistics*, 35, 2233–2260.

JIONGO, V. D., D. HAZIZA, AND P. DUCHESNE (2013): "Controlling the bias of robust small-area estimators," *Biometrika*, 100, 843–858.

JOHNSON, N. L., S. KOTZ, AND N. BALAKRISHNAN (1994): *Continuous Univariate Distributions*, vol. 1, New York: John Wiley & Sons, 2nd ed.

——— (1995): *Continuous Univariate Distributions*, vol. 2, New York: John Wiley & Sons, 2nd ed.

JONES, F. (1993): *Lebesgue Integration on Euclidean Space*, London / New York: Jones and Bartlett Publishers.

JUREČKOVÁ, J. AND P. K. SEN (1996): *Robust Statistical Procedures. Asymptotics and Interrelations*, John Wiley & Sons.

KARLBERG, F. (2000): "Population total prediction under a lognormal superpopulation model," *Metron*, 58, 53–80.

KORCHEVSKY, V. (2010): "On the applicability conditions of the strong law of large numbers for sequences of independent random variables," *Vestnik St. Petersburg University: Mathematics*, 43, 217–219.

——— (2015): "A generalization of the Petrov strong law of large numbers," *Statistics and Probability Letters*, 104, 102–108.

KRASKER, W. S. AND R. E. WELSCH (1982): "Efficient Bounded-Influence Regression Estimation," *Journal of the American Statistical Association*, 77, 595–604.

KUCZMASZEWSKA, A. (2016): "Convergence rate in the Petrov SLLN for depedent random variables," *Acta Mathematica Hungarica*, 148, 56–72.

LAHIRI, P. (2003): "On the impact of bootstrap in survey sampling and small area estimation," *Statistical Science*, 18, 199–210.

LAHIRI, P. AND J. N. K. RAO (1995): "Robust Estimation of Mean Squared Error of Small Area Estimators," *Journal of the American Statistical Association*, 90, 758–766.

LANG, S. (1993): *Real and Functional Analysis*, New York: Springer, 3rd ed.

LEE, H. (1991): "Model-Based Estimators That Are Robust to Outliers," in *Proceedings of the 1991 Annual Research Conference, Bureau of the Census 178-202. Washington, DC, Department of Commerce.*

LEHMANN, E. L. AND G. CASELLA (1998): *Theory of point estimation*, New York: Springer, 2nd ed.

LEHTONEN, R. (2011): "A short note on extended GREG family estimators for domains," in *Official Statistics: Methodology and Applications: In Honour of Daniel Thorburn*, ed. by M. Carlson, H. Nyqvist, and M. Villani, Stockholm: Department of Statistics, Stockholm University, chap. 10, 107–116.

LEHTONEN, R., C.-E. SÄRNDAL, AND A. VEIJANEN (2003): "The Effect of Model Choice in Estimation for Domains, Including Small Domains," *Survey Methodology*, 29, 33–44.

——— (2005): "Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains," *Statistics in Transition*, 7, 649–674.

LEHTONEN, R. AND A. VEIJANEN (1998): "Logistic generalized regression estimator," *Survey Methodology*, 24, 51–55.

——— (2009): "Design-based Methods of Estimation for Domains and Small Areas," in *Sample Surveys: Inference and Analysis*, ed. by D. Pfeffermann and C. Rao, Amsterdam: Elsevier, vol. 29B of *Handbook of Statistics*, chap. 31, 219–249.

LEVIT, B. Y. (1979): "On the Theory of the Asymptotic Minimax Property of Second Order," *Theory of Probability and its Applications*, 24, 435–437.

——— (1980): "On Asymptotic Minimax Estimates of the Second Order," *Theory of Probability and its Applications*, 25, 552–598, [original paper in Russian: Teor. Veroyatnost. i Primenen., 25:3 (1980), 561–576].

LINDLEY, D. V. AND A. F. M. SMITH (1972): "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society. Series B*, 34, 1–41.

LINDSTROM, M. J. AND D. M. BATES (1988): "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, 83, 1014–1022.

LUKACS, E. (1955): "A Characterization of the Gamma Distribution," *The Annals of Mathematical Statistics*, 26, 319–324.

MALLOWS, C. L. (1973): "On Some Topics in Robustness," unpublished memorandum, Bell Telephone Laboratories, Murray Hill (NJ).

——— (1978): "Minimizing an Integral (Problem 78.4)," *SIAM Review*, 20, 183.

MARAZZI, A. (1980): "Robust Bayesian Estimation for the Linear Model," Research report, ETH Zürich, Fachgruppe für Statistik.

——— (1985): "On Constrained Minimization of the Bayes Risk for the Linear Model," *Statistics and Decisions*, 3, 277–296.

——— (1986): "On the Numerical Solutions of Bounded Influence Regression Problems," in *COMPSTAT, 7th Symposium held at Rome 1986*, ed. by F. de Antoni, N. Lauro, and A. Rizzi, Vienna, Proceedings in Computational Statistics, 114–119.

——— (1990): "Restricted minimax credibility: Two special cases," *Bulletin of the Swiss Association of Actuaries*, 1, 101–114.

——— (1993): *Algorithms, Routines, and S Functions for Robust Statistics: The FORTRAN Library ROBETH with an interface to S-PLUS*, New York: Chapman & Hall.

MARONNA, R. AND R. ZAMAR (2002): "Robust estimates of location and dispersion for high-dimensional datasets," *Technometrics*, 44, 307–317.

MARONNA, R. A. (1976): "Robust M-Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51–67.

MARONNA, R. A., D. MARTIN, AND V. J. YOHAI (2006): *Robust Statistics: Theory and Methods*, Chichester: John Wiley.

MARONNA, R. A. AND V. J. YOHAI (1981): "Asymptotic behavior of general M-estimates for regression and scale with random carriers," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 58, 7–20.

MARUYAMA, Y. AND W. E. STRAWDERMAN (2005): "Necessary Conditions for Dominating the James–Stein Estimator," *Annals of the Institute of Statistical Mathematics*, 57, 157–165.

MERSMANN, O., C. BELEITES, R. HURLING, AND A. FRIEDMAN (2015): *microbenchmark: Accurate timing functions*, [R package version 1.4-2.1].

MILLER, J. J. (1977): "Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance," *The Annals of Statistics*, 5, 746–765.

MOLINA, I. AND Y. MARHUENDA (2015): *sae: Small Area Estimation*, [R package version 1.1].

MORRIS, C. N. (1983): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–55.

MÜLLER, C. H. AND S. UHLIG (2001): "Estimation of Variance Components with High Breakdown Point and High Efficiency," *Biometrika*, 88, 353–366.

MÜNNICH, R. AND J. P. BURGARD (2012): "On the Influence of Sampling Design on Small Area Estimates," *Journal of the Indian Society of Agricultural Statistics*, 66, 145–156.

MÜNNICH, R., S. GABLER, M. GANNINGER, J. BURGARD, AND J.-P. KOLB (2012): *Stichprobenoptimierung und Schätzung im Zensus 2011*, no. 21 in Statistik und Wissenschaft, Wiesbaden: Statistisches Bundesamt.

MÜNNICH, R., S. ZINS, A. ALFONS, C. BRUCH, P. FILZMOSER, M. GRAF, B. HULLIGER, J.-P. KOLB, R. LEHTONEN, D. LUSSMANN, A. MERANER, M. MYRSKYLÄ, D. NEDYALKOVA, T. SCHOCH, M. TEMPL, M. VALASTE, AND A. VEIJANEN (2011): *Policy Recommendations and Methodological Report*, Deliverable 10.1, AMELI: Advanced methodology for European Laeken indicators, research project funded by the European Commission, FP7-SSH-2007-217322.

ORTEGA, J. M. AND W. C. RHEINBOLDT (1970): *Iterative Solution of Nonlinear Equations in Several Variables*, New York: Academic Press.

OWEN, D. (1980): "A table of normal integrals," *Communications in Statistics: Simulation and Computation*, B9, 389–419.

PETROV, V. V. (1969): "On the strong law of large numbers," *Theory of Probability and its Applications*, 14, 183–192.

——— (1975): *Sums of Independent Random Variables*, Berlin: Springer-Verlag.

——— (2009): "On the strong law of large numbers for nonnegative random variables," *Theory of Probability and its Applications*, 53, 346–349.

PRASAD, N. AND J. N. K. RAO (1990): "The Estimation of the Mean Squared Error of Small-Area Estimators," *Journal of the American Statistical Association*, 85, 163–171.

PRÁŠKOVÁ, Z. AND P. K. SEN (2009): "Asymptotics in Finite Population Sampling," in *Sample Surveys: Theory, Methods and Inference*, ed. by D. Pfeffermann and C. Rao, Amsterdam: Elsevier, vol. 29N of *Handbook of Statistics*, chap. 40, 389–522.

PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING, AND B. P.FLANNERY (1986): *Numerical Recipes in Fortran 77: The art of scientific computing*, Cambridge, UK: Cambridge University Press.

RAO, J. (2003): *Small Area Estimation*, Hoboken: Wiley.

RAO, J. AND I. MOLINA (2015): *Small Area Estimation*, Hoboken: Wiley, 2nd ed.

RAO, J. N. K. (1994): "Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage," *Journal of Official Statistics*, 10, 153–165.

RICHARDSON, A. M. (1995): "Some problems in estimation in mixed linear models," Ph.D. thesis, Australian National University.

——— (1997): "Bounded Influence Estimation in the Mixed Linear Model," *Journal of the American Statistical Association*, 92, 151–161.

RICHARDSON, A. M. AND A. H. WELSH (1995): "Robust Restricted Maximum Likelihood in Mixed Linear Models," *Biometrics*, 51, 1429–1439.

ROBINSON, P. (1982): "On the Convergence of the Horvitz–Thompson Estimator," *Australian Journal of Statistics*, 24, 234–238.

ROBINSON, P. M. AND C. E. SÄRNDAL (1983): "Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling," *Sankhyā. The Indian Journal of Statistics. Series B*, 45, 240–248.

ROCKE, D. M. (1983): "Robust statistical analysis of interlaboratory studies," *Biometrika*, 70, 421–431.

——— (1991): "Robustness and balance in the mixed model," *Biometrics*, 47, 303–309.

ROUSSEEUW, P. J. (1984): "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

ROUSSEEUW, P. J. AND C. CROUX (1993): "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, 88, 1273–1283.

ROUSSEEUW, P. J. AND A. M. LEROY (1987): *Robust Regression and Outlier Detection*, (Wiley Series in Probability and Statistics), Hoboken: John Wiley & Sons.

ROUSSEEUW, P. J. AND K. VAN DRIESSEN (2006): "Computing LTS Regression for Large Data Sets," *Data Mining and Knowledge Discovery*, 12, 29–45.

ROUSSEEUW, P. J. AND B. C. VAN ZOMEREN (1990): "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–639.

ROUSSEEUW, P. J. AND V. YOHAI (1984): "Robust Regression by Means of S Estimators," in *Robust and Nonlinear Time Series Analysis*, ed. by J. Franke, W. Härdle, and R. Martin, New York: Springer, 256–274.

ROYALL, R. M. (1970): "On finite population sampling theory under certain regression models," *Biometrika*, 57, 377–387.

RUBIN-BLEUER, S. AND I. SCHIOPU-KRATINA (2005): "On the Two-phase framework for joint model and design-based inference," *The Annals of Statistics*, 33, 2789–2810.

RUDIN, W. (1976): *Principles of Mathematical Analysis*, New York: McGraw-Hill, 3rd ed.

——— (1987): *Real and Complex Analysis*, New York: McGraw-Hill, 3rd ed.

SAMAROV, A. AND R. E. WELSCH (1982): "Computational Procedures for Bounded-Influence Regression," in *COMPSTAT, 5th Symposium held at Toulouse 1982*, ed. by H. Caussinus, P. Ettinger, and R. Tomassone, Vienna, Proceedings in Computational Statistics, 412–418.

SÄRNDAL, C.-E. (1980): "On π-inverse weighting versus best linear unbiased weighting in probability sampling," *Biometrika*, 67, 639–650.

——— (2011): "Models in Survey Sampling," in *Official Statistics: Methodology and Applications: In Honour of Daniel Thorburn*, ed. by M. Carlson, H. Nyqvist, and M. Villani, Stockholm: Department of Statistics, Stockholm University, chap. 2, 15–27.

SÄRNDAL, C.-E., B. SWENSSON, AND J. WRETMAN (1992): *Model Assisted Survey Sampling*, New York: Springer.

SÄRNDAL, C.-E. AND R. L. WRIGHT (1984): "Cosmetic Form of Estimators in Survey Sampling," *Scandinavian Journal of Statistics*, 11, 146–156.

SCHAIBLE, W. L. (1996): *Indirect Estimators in U.S. Federal Programs*, New York: Springer.

SCHMID, T. AND R. MÜNNICH (2014): "Spatial robust small area estimation," *Statistical Papers*, 55, 653–670.

SCHMID, T., N. TZAVIDIS, R. MÜNNICH, AND R. CHAMBERS (2016): "Outlier-Robust Small-Area Estimation," *Scandinavian Journal of Statistics*, 43, 806–826.

SCHOCH, T. (2011a): "The Robust Basic Unit-Level Small Area Model: A Simple and Fast Fisher-Scoring Algorithm for Large Datasets," in *Proceedings of the Conference on New Technologies and Techniques in Statistics (NTTS)*, EUROSTAT, Brussels: EUROSTAT.

——— (2011b): "Robust High Breakdown Point Estimation in Unit-Level SAE Models," in *Proceedings of the Conference on Small Area Estimation (SAE)*, Trier, August 11–13: SAE, 72–78.

——— (2011c): *rsae: Robust Small Area Estimation*, [R package version 0.1-2].

——— (2012): "Robust Unit-Level Small Area Estimation: A Fast Algorithm for Large Data," *Austrian Journal of Statistics*, 41, 243–265.

——— (2014): *rsae: Robust Small Area Estimation*, [R package version 0.1-5].

SEARLE, S. (1971): *Linear Models*, New York: John Wiley & Sons.

——— (1987): *Linear Models for Unbalanced Data*, New York: John Wiley & Sons.

SEARLE, S., G. CASELLA, AND C. E. MCCULLOCH (1992): *Variance Components*, Hoboken: John Wiley & Sons.

SEARLS, D. T. (1966): "An Estimator for a Population Mean which Reduces the Effect of Large True Observations," *Journal of the American Statistical Association*, 61, 1200–1204.

SEN, P. K. (1988): "Asymptotics in Finite Population Sampling," in *Sampling*, ed. by P. R. Krishnaiah and C. R. Rao, Amsterdam: Elsevier, vol. 6 of *Handbook of Statistics*, chap. 12, 291–331.

SEN, P. K. AND J. M. SINGER (1993): *Large Sample Methods in Statistics: An Introduction with Applications*, Boca Raton (FL): Chapman & Hall / CRC Press.

SHAO, P. Y.-S. AND W. E. STRAWDERMAN (1994): "Improving on the James–Stein Positive-Part Estimator," *The Annals of Statistics*, 22, 1517–1538.

SINHA, S. K. AND J. N. K. RAO (2009): "Robust small area estimation," *The Canadian Journal of Statistics*, 37, 381–399.

SMALL, C. G. AND J. WANG (2003): *Numerical Methods for Nonlinear Estimating Equations*, Oxford: Oxford University Press.

STAHEL, W. A. AND A. H. WELSH (1997): "Approaches to robust estimation in the simplest variance components model," *Journal of Statistical Planning and Inference*, 57, 295–319.

STEIN, C. M. (1981): "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135–1151.

STEIN, E. M. AND R. SHAKARCHI (2005): *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, Princeton Lectures in Analysis, vol. 3, Princeton: Princeton University Press.

SUD, U. C., V. K. BHATIA, H. CHANDRA, AND A. K. SRIVASTAVA (2011): "Crop Yield Estimation at District Level by Combining Improvement of Crop Statistics Scheme Data and Census Data," in *Wye City Group on Statistics on Rural Development and Argiculture Household Income*, Rio de Janeiro.

SUGDEN, R. A. AND T. M. F. SMITH (1984): "Ignorable and Informative Designs in Survey Sampling Inference," *Biometrika*, 71, 495–506.

TAM, S. M. (1988): "Some Results on Robust Estimation in Finite Population Sampling," *Journal of the American Statistical Association*, 83, 248–248.

TAO, T. (2011): *An Introduction to Measure Theory*, Providence (RI): American Mathematical Society.

TEMPL, A. A. M. AND P. FILZMOSER (2010): "An Object-Oriented Framework for Statistical Simulation: The R Package simFrame." *Journal of Statistical Software*, 37, 1–36.

THOMPSON, M. E. (1997): *Theory of Survey Samples*, London: Chapman & Hall.

TODOROV, V. (2016): *rrcov: Scalable Robust Estimators with High Breakdown Point*, [R package version 1.4-3].

TODOROV, V. AND P. FILZMOSER (2009): "An Object Oriented Framework for Robust Multivariate Analysis," *Journal of Statistical Software*, 32, 1–47.

TUKEY, J. W. (1977): *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

WALK, H. (2005): "Strong laws of large numbers by elementary Tauberian arguments," *Monatshefte für Mathematik*, 144, 329–346.

WARNHOLZ, S. (2016a): *saeRobust: Robust Small Area Estimation*, [R package version 0.1.0].

——— (2016b): "Small Area Estimation Using Robust Extensions to Area Level Models," Ph.D. thesis, Freie Universität Berlin, FB Wirtschaftswissenschaften, Berlin.

WEAVER, N. (1999): *Lipschitz Algebras*, River Edge (NJ): World Scientific Publications Co.

WELLMANN, J. (1994): "Robuste statistische Verfahren und Ausreisseridentifikation beim Modell der Einfachklassifikation mit zufälligen Effekten," Ph.D. thesis, Department of Statistics, University of Dortmund, Dortmund.

——— (2000): "Robustness of an S-Estimator in the One-Way Random Effects Model," *Biometrical Journal*, 42, 215–221.

WELLMANN, J. AND U. GATHER (2003): "Identification of outliers in a one-way random effects model," *Statistical Papers*, 44, 335–348.

WELSH, A. H. AND A. M. RICHARDSON (1997): "Approaches to the Robust Estimation of Mixed Models," in *Robust Inference*, ed. by G. Maddala and C. Rao, Amsterdam: Elsevier Science, vol. 13 of *Handbook of Statistics*, chap. 13, 343–384.

WELSH, A. H. AND E. RONCHETTI (1998): "Bias-calibrated estimation from sample surveys containing outliers," *Journal of the Royal Statistical Society, Series B*, 60, 413–428.

WEST, M. (1984): "Outlier Models and Prior Distributions in Bayesian Linear Regression," *Journal of the Royal Statistical Society. Series B*, 46, 431–439.

WRIGHT, R. L. (1983): "Finite population sampling with multivariate auxiliary information," *Journal of the American Statistical Association*, 78, 879–884.

WU, C. (2003): "Optimal calibration estimators in survey sampling," *Biometrika*, 90, 937–951.

WU, C.-F. (1981): "Asymptotic theory of nonlinear least squares estimation," *The Annals of Statistics*, 9, 501–513.

YOU, Y. AND B. CHAPMAN (2006): "Small area estimation using area level models and estimated sampling variances," *Survey Methodology*, 32, 97–103.

YOU, Y. AND J. N. K. RAO (2002): "A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights," *The Canadian Journal of Statistics*, 30, 431–439.

ZIMMERMANN, T. (2018): "The interplay between sampling design and statistical modelling in small area estimation," Ph.D. thesis, Universität Trier, Fachbereich IV, Trier.