EUROPEAN COMMISSION
Community Research

SEVENTH FRAMEWORK PROGRAMME

AMELI
Advanced Methodology for European Laeken Indicators

Deliverable 4.2

Robust Methodology for Laeken Indicators

Version: 2011

Beat Hulliger, Andreas Alfons, Peter Filzmoser, Angelika Meraner, Tobias Schoch and Matthias Templ

The project **FP7–SSH–2007–217322 AMELI** is supported by European Commission funding from the Seventh Framework Programme for Research.

http://ameli.surveystatistics.net/

Contributors to deliverable 4.2

Chapter 1: Beat Hulliger and Tobias Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 2: Beat Hulliger and Tobias Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 3: Andreas Alfons, Matthias Templ, Peter Filzmoser, and Josef Holzer, Vienna University of Technology.

Chapter 4: Beat Hulliger and Tobias Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 5: Beat Hulliger and Tobias Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 6: Beat Hulliger and Tobias Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 7: Beat Hulliger and Tobias Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 8: Matthias Templ, Alexander Kowarik, Peter Filzmoser, Vienna University of Technology.

Chapter 9: Beat Hulliger and Tobias Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 10: Beat Hulliger and Tobias Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 11: Matthias Templ, Karel Hron, Peter Filzmoser, Vienna University of Technology.

Chapter 12: Angelika Meraner, Peter Filzmoser, and Matthias Templ, Vienna University of Technology.

Main responsibility

Beat Hulliger, University of Applied Sciences Northwestern Switzerland.

Evaluators

Internal experts: Ralf Münnich, Christian Bruch, Tobias Enderle, Jan-Philipp Kolb, and Stefan Zins

Aim and Objectives of Deliverable 4.2

Indicators and in particular the Laeken indicators are vulnerable to outliers. Outliers may not only bias the indicators but also introduce a high additional variability. Therefore, outliers and other deviations from theoretical distributions may undermine the quality of indicators thoroughly. Robust procedures remain stable in those situations but are more complex to handle. The workpackage on robustness of the AMELI project developed robust imputation procedures, in particular for multivariate data, including detection of outlying and influential observations and small area estimation procedures. The procedures were implemented in the statistical programming language R and corresponding packages were developed. The evaluation of the robustness of classical indicators, in particular the Laeken indicators, and of the new robust procedures will be described in Deliverable D7.1. Quality measures of the robustness of indicators and of the impact of robust procedures were developed in Workpackage 4 and implemented in the Workpackages 6 and 7. As a result of the analysis carried out on the robustness of procedures in Workpackage 4, 6, and 7 recommendations are formulated in Deliverable D7.1 for the use of indicators in what concerns robustness issues. Workpackage 4 was only possible due to an intensive and fruitful collaboration between Technical University of Vienna and University of Applied Sciences Northwestern Switzerland. The discussions on contaminations and the subsequent implementation in the simulation environments of the AMELI project have brought research on robustness in survey sampling to a new level. The insight that was possible due to the simulations and the new methods developed would not have been possible without the support of the European Union through the Socio-economic Sciences and Humanities programme of the 7th Framework Programme for Research and Development. The workpackage contributors are greatful for the support of the European Union and for the additional support granted from their respective institutions.

Contents

Chapter 1

Introduction

The workpackage on robustness could profit from the development work in the EUREDIT (cf. EUREDIT, 2003) project of the 5th Framework for Research Technology Development (RTD) of the European Union. The work of EUREDIT resulted in two important papers on multivariate outlier detection and imputation: Béguin and Hulliger (2004) and Béguin and Hulliger (2008). In that project, research concentrated more on the issues of business surveys and also used real data to test the methods, but, in contrast, could not draw on extensive simulation framework like the one developed in AMELI. On the other hand, the EUREDIT concepts of true data, raw data and imputed data as well as certain of the criteria to evaluate the success of editing and imputation were also used in AMELI. The EDIMBUS manual on editing and imputation of cross sectional business surveys was another important source for the AMELI project since it clarified the importance of an overall design of data preparation (EDIMBUS, 2007).

The robustness problems posed by SILC survey data are radically different when they are tackled from a univariate point of view, i.e. starting with the equivalized disposable income, or when they are regarded as a multivariate problem with possibly outlying income components. AMELI researched both avenues and the structure of this deliverable reflects this: Part I is dedicated to univariate robust methods and part II treats multivariate robust methods. Most of the social inclusion indicators called Laeken indicators are inherently robust because the estimands are robust functionals of the population distribution. This holds for the Median and the poverty threshold as well as for the At-risk-of-poverty-rate (ARPR) and for the Relative-Poverty-Median-Gap. These measures are defined through the median, which is outlier robust and a functional of the population distribution with bounded influence of outliers. These measures are not investigated in depth in the AMELI project but sometimes serve as references and counterexamples to non-robust behaviour.

Though the population mean is not part of the social cohesion indicators of the EU, it is an important indicator simply because its relation with the total disposable income of a population or a part of the population. In addition it is an indicator which is used in many other contexts and therefore its robustness and robusfitication are treated to a certain extent in Chapter 2.

The most critical social inclusion indicator in terms of outlier sentsitivity is the Quintile Share Ratio (QSR), sometimes also called S80/S20. The QSR is designed to measure

inequality of income distribution and therefore must, by definition, be sensitive to large incomes. The conflict between the sensitivity of the population functional to outliers, which is an relevant objective of any inequality measure, and the inherent non-robustness is discussed in Part I. The Quintile Share Ratio has a very valuable property: it is simple to explain. The simplicity of the explanation is particularly important when the discussion on social inclusion should be carried to a large part of the society and this is exactly the purpose of these indicators. The Gini-indicator, which has a much longer history in measuring distribution inequality, lacks the simplicity of the Quintile Share Ratio. Nevertheless it has been extensively studied and it is interesting to understand the different behaviour of the Quintile Share Ratio and the Gini-indicator.

Two lines of robustification have been investigated in the AMELI project. One borrows strength from a parametric model of the tail distribution. It is treated in Chapter 3. The second is non-parametric in spirit and builds on the experiences with the estimation of population means and tries to find an optimal trade-off between variance and bias. This robustification of the Quintile Share Ratio is developed in Chapter 4. An important issue is the robustification of estimates which are derived for small areas or domains with the help of models. These methods can be used to obtain better predictions of individual incomes and they may be used to derive social inclusion indicators for small areas or domains. Therefore, the robustification must already be used when estimating the models that lead to predicted income. First methods have been developed in these areas and are described in Chapter 5.

Part II describes the problems and the methods encountered when outliers in the income components should be detected and replaced by values which are less harmful to the social inclusion indicators. Only very few methods can stand the complexity of the data of SILC. For example, the approach to consider the income components as compositions of the equivalized disposable income did not prove viable mainly because of the zero-inflated distributions of the components. Other methods fail due to the missing values contained in the data. The evaluation of these methods is mainly described in Deliverable HULLIGER et al. (2011). Intermediate simulations which were carried out for the development of the methods are reported here.

Bibliography

Béguin, C. and **Hulliger, B.** (**2004**): *Multivariate Outlier Detection in Incomplete Survey Data: the Epidemic Algorithm and Transformed Rank Correlations*. Journal of the Royal Statistical Society, Series A, 167 (2), pp. 275–294.

Béguin, C. and **Hulliger, B.** (**2008**): *The BACON-EEM Algorithm for Multivariate Outlier Detection in Incomplete Survey Data*. Survey Methodology, Vol. 34, No. 1, pp. 91–103.

EDIMBUS (**2007**): EDIMBUS, 2007. Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys, O. Luzi, M. Di Zio, A. Manzari, T. De Waal, J. Pannekoek, J. Hoogland, C. Tempelman, B. Hulliger, D. Kilchmann.

EUREDIT (**2003**): *Towards Effective Statistical Editing and Imputation Strategies – Findings of the Euredit project (Volume 1)*. Technical report, EUREDIT, http://www.cs.york.ac.uk/euredit/.

Hulliger, B., **Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Graf, M.**, **Kolb, J.-P.**, **Lehtonen, R.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**): *Report on the simulation results*. Technical report, AMELI deliverable D7.1.
URL http://ameli.surveystatistics.net/

Part I

Robust Univariate Methods

Chapter 2

Basic Robust Univariate Estimators

In this chapter, we introduce some basic, univariate location estimators that are robust w.r.t. outlier contamination. These estimators serve as building blocks for more sophisticated estimators and are usefull for comparison purposes.

Finite population parameters are often very sensitive to the presence of outliers in the population. The most prominent example is the mean of a finite population: it depends on all observations in the population and therefore must also be influenced by extreme observations. This is to be contrasted to model (infinite population) parameters, which are usually insensitive to outliers. This basic difference is also reflected in the distinction between representative and non-representative outliers introduced by (Chambers, 1986). Because of this difference of estimands, the problem of outlier robustness is therefore different for finite and infinite populations (Hulliger, 1991, 1995; Beaumont and Alavi, 2004). As noted by Chambers (1986), it is the sampling error (or the prediction error in a model-based framework) of an estimator which must be insensitive to outliers in finite populations and not necessarily the estimator itself. Nevertheless, and in particular for social inclusion indicators, much attention should be given to the question whether a sensitive estimand is really needed. The Laeken indicators have striken a good choice there in that the poverty indicators are robust estimands. However, it is inevitable to use a sensitive estimand for inequality, as is the Quintile Share Ratio.

Suppose a sample s (with fixed sample size $n < N$) has been drawn from the finite population $U = \{1, \ldots, N\}$ according to the sampling design $p(S)$. Denote by y_i, $i \in U$ the variable of interest. In the matter at hand, we are interested in estimating the population mean $\bar{y} = \sum_U y_i/N$. To each sampled element in s is attached a weight w_i that reflects the sample inclusion probabilities π_i (and probably weight-adjustments such as calibration, non-response corrections and the like). Notably, we may have that $w_i = 1/\pi_i$ where $\pi_i = \sum_{s:i\in s} p(s)$, $i = 1, \ldots, N$ and in any case we assume that $\sum_{i\in s} w_i = N$. The weighted mean writes

$$T_M = \frac{\sum_{i\in s} w_i y_i}{\sum_{i\in s} w_i}. \tag{2.1}$$

Because we assume that $w_i = 1/\pi_i$, T_M is the Hajek estimator (cf. Hulliger, 1999). We do not study pure Horvitz-Thompson estimators for the population mean, since they are

rarely used in practice. It is obvious that the weighted mean can be severely influenced by both representative and non-representative outliers.

2.1 Weighted Sample Median

The weighted sample median may be considered as a remedy to the non-robust weighted sample mean at the price of a considerable loss of efficiency. Let $\{y_{(j)}, j = 1, \ldots, n\}$ denote the ith order statistics of the sample $\{y_i, i = 1, \ldots, n\}$ and $w_j = w_i$ if $y_{(j)} = y_i$ (i.e. the weights are sorted according to the ordering of the y_i). For the population U, the pth quantile, Q_p, is obtained solving the estimating equation (EE) $\sum_{i=1}^{N}[\mathbb{1}\{y_i \leq Q_p\} - p] = 0$. For the sample s, a Horvitz-Thompson-type quantile estimator is given by solving the sample EE (cf. BINDER and PATAK, 1994)

$$\sum_{i \in s} w_i[\mathbb{1}\{y_i \leq \hat{Q}_p\} - p] = 0. \tag{2.2}$$

The sample median is obtained with $p = 0.5$. For a discussion of the limiting distribution of sample quantiles and (approximations to) their variance estimators, see FRANCISCO and FULLER (1991); and SHAO (1994) (who uses less restrictive assumptions on the sample design and the asymptotic setting).

For a simple random sample, the sample median is robust in the classical sense; see e.g., HAMPEL et al. (1986). Therefore, its design variance is essentially unaffected by the presence of an outlier in the finite population, no matter how large is that outlier. However, BEAUMONT and ALAVI (2004) show that the sampling error and the design bias of the sample median, when used as an estimator of the finite population mean, take an arbitrarily large value when one or more population unit takes an arbitrarily large value. This is explained by the fact that the finite population mean itself takes an arbitrarily large value in such a case. In fact, it is not the sample median which is not robust, but the estimand, i.e. the population mean. Unlike the sample median, the sample mean is design unbiased but it is not robust in the classical sense. The sampling error and the design variance of the sample mean can thus be heavily affected by the presence of an outlier in the finite population. This illustrates why outlier-robustness for finite populations is often viewed as a trade-off between bias and variance and why outliers must usually have an influence, at least to some extent, on estimators (HULLIGER, 1991, 1995).

If weights are needed in (2.2), the (weighted) sample median may be much less robust: The breakdown point of the median no longer depends on the number of contaminated observations but on the sum of their weights. If 5%, say, of the observations, account for 50% of the total weight then the breakdown point of the median is 5%!

2.2 Trimmed and Winsorized Weighted Sample Mean

It will prove useful to express the trimmed mean as a smooth location L-functional. We introduce these location functionals for the classical, infinite-sample context and adopt

them in a second step to the finite population sampling paradigm. Therefore, let $Y \sim F \in \mathcal{F}$ with $\mathcal{F}$ the set of all real-valued distribution functions with support $\mathcal{Y} \in \mathbb{R}$. The smooth location L-functional $T_T : \mathbb{R} \mapsto \mathcal{Y}$ is defined as $T_T(F) = \int xJ[F(x)]\mathrm{d}F(x)$ where J is a weight function $J : [0,1] \mapsto \mathbb{R}$ (depending on the characteristics of J, we may require additional restrictions on F for $T_T(F)$ to be behave well; see Serfling (1980) for details). The location L-functional is asymptotically equivalent to the estimator $T_T(F_n) = 1/N \sum_{j=1}^{N} J(j/N)Y_{(j)}$ where $Y_{(j)}$ is the jth order statistics.

For the α-trimmed sample mean, the weight function is defined as $J(t) = 1/(1-2\alpha)\mathbb{1}\{\alpha < t \leq 1-\alpha\}$ with $0 < \alpha < 0.5$. In fact, the α-trimmed sample mean discards all observations below the α and above the $(1-\alpha)$ quantile. Similarly, we may define a one-sided α-trimmed sample mean with $J_o(t) = 1/(1-\alpha)\mathbb{1}\{t \leq 1-\alpha\}$. This estimator is particularly useful for the positive-valued skewed income distribution.

Let $y_{(j)}$ denote the jth order statistics. The finite-population-sampling analogue of $T_T(F)$ can be obtained from

$$T_T(F_n) = \frac{1}{N}\sum_{j=1}^{N} c_j y_{(j)} \qquad \text{with } c_j = w_j J\left(\frac{\sum_{k=1}^{j} w_k}{N}\right), \tag{2.3}$$

where J is a weight function; see Shao (1994) for details. If the population size N is not known, it may be estimated by $\hat{N} = \sum_{i\in s} w_i$.

The α-trimmed and the one-sided α-trimmed sample mean are implemented in the R-package RHT (2011); see `tsvymean` therein.

Besides the trimmed sample mean, the winsorized sample mean has received a lot of attention (see e.g., Searls, 1966; Hulliger, 1991; Fuller, 1991; Rivest, 1994; Hulliger, 1999; Beaumont and Rivest, 2009). The winsorized sample mean cannot be represented as smooth location L-functional. Though, an adaptation of the winsorized mean to the finite population sampling context follows immediately. Let $\alpha \in [0, 0.5)$ be the wisorization-tuning constant. Define $F_n(t) = \sum_{j=1}^{t} w_j / \sum_{j=1}^{n} w_j$ with $w_j = w_i$ if $y_{(j)} = y_i$ (i.e., the weights ordered accoring to y). Find the numbers

$$a_l = \min\{j : F_n(j) \geq \alpha\}, \tag{2.4}$$
$$a_u = \max(\{j : F_n(j) < 1-\alpha\}, a_l). \tag{2.5}$$

These are the inner α and $1-\alpha$ quantiles. The weighted α-winsorized sample mean is given by (Hulliger, 1999)

$$T_W = \frac{1}{\sum_s w_i}\left(\sum_{j=a_l}^{a_u} w_j y_{(j)} + \sum_{j=1}^{a_l-1} w_j y_{(a_l)} + \sum_{j=a_u+1}^{n} w_j y_{(a_u)}\right). \tag{2.6}$$

The α-winsorized sample mean is implemented in the R-package RHT (2011); see `tsvymean` therein.

2.3 Robust Horvitz-Thompson and Ratio M-Estimators

HULLIGER (1995) introduced a robustification of the Horvitz-Thompson estimator based on M-estimators. The key to the robustification is to make the assisting model of the Horvitz-Thompson estimator explicit. The Horvitz-Thompson estimator is designed for the situation where the response variable of interest is linearly related to a design variable $z_i \propto \pi_i$: $y_i = \beta z_i + E_i$, where the error has variance $\sigma_E^2 = z_i\sigma^2$. It then turns out that the Horvitz-Thompson estimator is defined as $\bar{z}_U\hat{\beta}$ where $\hat{\beta}$ is defined through the following estimating equation:

$$\sum_{i=1}^{n} w_i \left(\frac{y_i - \beta z_i}{\sqrt{z_i}}\right)\sqrt{z_i} = 0, \tag{2.7}$$

where $z_i = \sum_{i=1}^{n} w_i/(nw_i)$. The Horvitz-Thompson estimator for the mean is $T_{HT} = \bar{z}_U\hat{\beta}$. Note that here we derive the measure of size z_i from the weight w_i and therefore the Hajek estimator for $\bar{z}_U = \sum_{i=1}^{n} w_i z_i / \sum_{i=1}^{n} w_i = 1$.

The robustification of the Horvitz-Thompson estimator proposed in HULLIGER (1995) assumes that the weights are not outlying and thus only the residuals $(y_i - \beta z_i)/\sqrt{z_i}$ must be robustified. This leads to a new estimating equation

$$\sum_{i=1}^{n} w_i\, \psi_c \left(\frac{y_i - \beta z_i}{\sqrt{z_i}}\right)\sqrt{z_i} = 0, \tag{2.8}$$

where ψ_c, as a default, is the Huber psi-function. Of course also robustification of the w_i, resp. z_i, are possible. The solution of (2.8) is the robustified Horvitz-Thompson estimator. This estimator can be expressed as weighted estimator, so that the solution can be obtained by an IRLS algorithm. For the IRLS algorithm a starting value $\beta^{(0)} = \text{med}_i(y_i, w_i)/\text{med}_i(z_i, w_i)$ is used, where $\text{med}_i(y_i, w_i)$ is the weighted median of y_i. As a next step a robust standard deviation of the standardized residuals $r_i = (y_i - \beta^{(0)} z_i)/\sqrt{z_i}$ is calculated. The median absolute deviation (mad) is used:

$$\hat{\sigma}_E = \text{mad}_i(r_i, w_i) \tag{2.9}$$

Then, one obtains a robustness weight u_i for each observation from

$$u_i = \frac{\psi_c(r_i/\hat{\sigma}_E)}{|r_i/\hat{\sigma}_E|}. \tag{2.10}$$

The robustness weight is 1 for observations which are not downweighted, i.e. which are considered good observations. For outliers, the robustness weight is lower than 1 and it

may become 0 or nearly so for extreme outliers. Therefore, an estimate of the robustified Horvitz-Thompson estimator on the $(t+1)$th iteration ($t \in \mathbb{N}_0^+$) can be expressed as a weighted estimator

$$\beta^{(t+1)} = \frac{\sum_{i=1}^n w_i u_i y_i}{\sum_{i=1}^n w_i u_i z_i}, \tag{2.11}$$

where u_i depends on $\beta^{(t)}$ (the functional dependence is suppressed for ease of display). The iterative estimation process is repeated until some convergence criterion is met (e.g., $|\beta^{(s+1)} - \beta^{(s)}| \leq acc$ with $s \in \mathbb{N}_0^+$ and $acc = 0.00001$, say). Using as a covariate $z_i = 1$ instead, gives an M-estimator that does not take into account a possible correlation between the weights and the response variable y_i.

Replacing in (2.8) z_i by a covariable x_i of which the population mean $\bar{x}_U$ is known and which is correlated to y_i we obtain a robust ratio estimator. The robust Horvitz-Thompson M-estimator and the robust, weighted M-estimator are implemented in the R-package RHT (2011); see `msvymean` therein.

The robust ratio estimator was introduced by Gwet and Rivest (1992) and Hulliger (1995). The robust ratio M-estimator is implemented in the R-package RHT (2011); see `rsvymean` therein.

2.4 Choosing the tuning constant

The difficult issue when using univariate robust estimators for asymmetric distributions is that, inevitably, the robust estimators have a bias. Fortunately the lower variance compensates at least partially the bias. Thus the question of choosing the tuning constant has two aspects:

1. How many outliers are in the sample even in the best case? We would like to be protected against such a minimal number of outliers. This sets an upper bound to the tuning constant. In fact, if this minimal proportion of outliers we have to expect is α then there is a tuning constant which declares a proportion α of the observations as outliers. Of course we would have to translate this into proportion of weight of the outliers when weighting is involved. In order to be protected against such a proportion of outliers we already accept a certain bias. This is a price to pay for the security of being protected against a small number of outliers at least.

2. If there are even more outliers than in the best case we would like to be protected but we know that the price may become larger in terms of bias. However, we hope that in terms of variance a slight gain is possible due to the downweighting of extremes. Thus the question is, where is the best trade-off between bias and variance.

The way to proceed is always to calculate the robust estimator not just for one tuning constant but for a whole series. Then the change from practically downweighting no

outlier to a robust estimator with a particular tuning constant can be compared and also corresponding variance estimates are available. Often the variance gain is already considerable for slight downweighting while the bias is small. Finding the point where the bias begins to grow fast when downweighting more and more is the difficult task.

When searching for the good tuning constant it is always helpful to observe the mean of the robustness weights (weighted or unweighted): $\bar{u}_S = \sum_{i=1}^{n} u_i/n$ or $\sum_{i=1}^{n} w_i u_i / \sum_{i=1}^{n} w_i$. The mean of the robustness weights measures how much weight is left in the effective sample. In many situations $\bar{u}$ should not drop below 0.99 or 0.98 to avoid a too large bias.

Bibliography

Beaumont, J.-F. and **Alavi, A.** (**2004**): *Robust Generalized Regression Estimation.* Survey Methodology, 30 (2), pp. 195–208.

Beaumont, J.-F. and **Rivest, L.-P.** (**2009**): *Dealing with outliers in survey data.* **Pfeffermann, D.** and **Rao, C.** (editors) Sample Surveys: Theory, Methods and Inference, *Handbook of Statistics*, vol. 29A, chapter 11, pp. 247–280, Amsterdam: Elsevier.

Binder, D. A. and **Patak, Z.** (**1994**): *Use of Estimating Functions for Estimation from Complex Surveys.* Journal of the American Statistical Association, 89 (427), pp. 1035–1043.

Chambers, R. L. (**1986**): *Outlier Robust Finite Population Estimation.* Journal of the American Statistical Association, 81 (396), pp. 1063–1069.

Francisco, C. A. and **Fuller, W. A.** (**1991**): *Quantile estimation with a complex survey design.* The Annals of Statistics, 19 (1), pp. 454–469.

Fuller, W. A. (**1991**): *Simple estimators for the mean of skewed populations.* Statistica Sinica, 1, pp. 137–158.

Gwet, J.-P. and **Rivest, L.-P.** (**1992**): *Outlier Resistant Alternatives to the Ratio Estimator.* Journal of the American Statistical Association, 87, pp. 1174–1182.

Hampel, F. R., **Ronchetti, E. M.**, **Rousseeuw, P. J.** and **Stahel, W. A.** (**1986**): Robust Statistics: The Approach Based on Influence Functions. New York: John Wiley & Sons.

Hulliger, B. (**1991**): Nonparametric M-estimation of a population mean. Ph.D. thesis, ETH Zurich, Nr. 9443.

Hulliger, B. (**1995**): *Outlier Robust Horvitz-Thompson Estimators.* Survey Methodology, 21 (1), pp. 79–87.

Hulliger, B. (**1999**): *Simple and robust estimators for sampling.* Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 54–63, American Statistical Association.

RHT (**2011**): *RHT: Robust Location Estimators for Complex Survey Samples.* R-Package; Beat Hulliger and Tobias Schoch, version 1.3.

Rivest, L.-P. (**1994**): *Statistical properties of Winsorized means for skewed distributions*. Biometrika, 81, pp. 373–383.

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 (316), pp. 1200–1204.

Serfling, R. J. (**1980**): Approximation theorems of mathematical statistics. New York: Wiley.

Shao, J. (**1994**): *L-Statistics in complex survey problems*. The Annals of Statistics, 22 (2), pp. 946–967.

Chapter 3

Semi-Parametric Robust Estimation

In this chapter, robust semiparametric estimation of social exclusion indicators and their application with the R package **laeken** is discussed. Special emphasis is thereby given to income inequality indicators, as the standard estimates for these indicators are highly influenced by outliers in the upper tail of the income distribution. This influence can be reduced by modeling the upper tail with a Pareto distribution in a robust manner. The focus of the paper is on both, to demonstrate the functionality of **laeken** beyond the standard estimation techniques and to give a brief mathematical description of the implemented procedures.

3.1 Introduction

From a robustness point of view, the standard estimators for some of the social exclusion indicators defined by Eurostat (2004, 2009) are problematic. In particular the income inequality indicators *quintile share ratio* (QSR) and *Gini coefficient* suffer from a lack of robustness. Consider, e.g., the QSR, which is estimated as the ratio of estimated totals or means (see Section 3.2.1 for an exact definition). It is well known that the classical estimates for totals or means have a breakdown point of 0, meaning that even a single outlier can distort the results to an arbitrary extent. In fact, the influence of a single observation in the upper tail of the income distribution on the estimation of the QSR is linear and therefore unbounded. For practical purposes, the standard QSR estimator thus cannot be recommended in many situations (cf. Hulliger and Schoch, 2009). It is also important to note that the behavior of the Gini coefficient is similar to the behavior of the QSR.

The data basis for the estimation of the social exclusion indicators according to Eurostat (2004, 2009) is the *European Union Statistics on Income and Living Conditions* (EU-SILC), which is an annual panel survey conducted in EU member states and other European countries. On the one hand, EU-SILC data typically contain a considerable amount of *representative* outliers in the upper tail of the income distribution, i.e., correct observations that behave differently from the main part of the data, but that are not unique in the population and hence need to be considered for computing estimates of the indicators. On the other hand, EU-SILC data frequently contain some even more

extreme *nonrepresentative* outliers, i.e., observations that are either incorrect or can be considered unique in the population. Consequently, such nonrepresentative outliers need to be excluded from the estimation process or downweighted.

As a remedy, the upper tail of the income distribution may be modeled with a *Pareto distribution* in order to recalibrate the sample weights or use fitted income values for observations in the upper tail when estimating the indicators (see Section 3.6). Nevertheless, classical estimators for the parameters of the Pareto distribution are highly influenced by the nonrepresentative outliers themselves. Using robust methods reduces the influence on fitting the Pareto distribution to the representative outliers and therefore on the estimation of the indicators.

Rather than evaluating these methods, the paper concentrates on showing how they can be applied in the statistical environment R (R DEVELOPMENT CORE TEAM, 2011) with the add-on package **laeken** (ALFONS et al., 2011a). The basic design of the package, as well as standard estimation of the social exclusion indicators is discussed in detail in vignette `laeken-standard` (TEMPL and ALFONS, 2011a). Furthermore, the general framework for variance estimation is illustrated in vignette `laeken-variance` (TEMPL and ALFONS, 2011b). Those documents can be viewed from within R with the following commands:

```
R> vignette("laeken-standard")
R> vignette("laeken-variance")
```

. Throughout the paper, the example data from package **laeken** is used. The data set is called `eusilc` and consists of 14 827 observations from 6 000 households. In addition, it was synthetically generated from Austrian EU-SILC survey data from 2006 using the data simulation methodology proposed by ALFONS et al. (2011b) and implemented in the R package **simPopulation** (ALFONS and KRAFT, 2010). More information on the example data can be found in vignette `laeken-standard` or in the corresponding R help page.

```
R> library("laeken")
R> data("eusilc")
```

The rest of the paper is organized as follows. Section 3.2 gives a mathematical description of the Eurostat definitions of the social exclusion indicators QSR and Gini coefficient. In Section 3.3, the Pareto distribution is briefly discussed. Section 3.4 discusses a rule of thumb for estimating the threshold for the upper tail of the distribution, and illustrates graphical methods for exploring the data in order to find the threshold. Classical and robust estimators for the shape parameter of the Pareto distribution are described in Section 3.5. How to use Pareto tail modeling to estimate the social exclusion indicators is then shown in Section 3.6. Finally, Section 3.7 concludes.

3.2 Social exclusion indicators

This paper is focused on the inequality indicators *quintile share ratio* (QSR) and *Gini coefficient*, which are both highly influenced by outliers in the upper tail of the distribution. Note that for the estimation of the social exclusion indicators, each person in a household is assigned the same *equivalized disposable income*. See vignette `laeken-standard` (Templ and Alfons, 2011a) for the computation of the equivalized disposable income with the R package **laeken**.

For the following definitions, let $\boldsymbol{x} := (x_1, \ldots, x_n)'$ be the equivalized disposable income with $x_1 \leq \ldots \leq x_n$ and let $\boldsymbol{w} := (w_i, \ldots, w_n)'$ be the corresponding personal sample weights, where n denotes the number of observations.

3.2.1 Quintile share ratio (QSR)

The income *quintile share ratio* (QSR) is defined as the ratio of the sum of the equivalized disposable income received by the 20% of the population with the highest equivalized disposable income to that received by the 20% of the population with the lowest equivalized disposable income (Eurostat, 2004, 2009).

For the estimation of the quintile share ratio from a sample, let $\hat{q}_{0.2}$ and $\hat{q}_{0.8}$ denote the weighted 20% and 80% quantiles, respectively. With $0 \leq p \leq 1$, these weighted quantiles are given by

$$\hat{q}_p = \hat{q}_p(\boldsymbol{x}, \boldsymbol{w}) := \begin{cases} \frac{1}{2}(x_j + x_{j+1}), & \text{if } \sum_{i=1}^{j} w_i = p \sum_{i=1}^{n} w_i, \\ x_{j+1}, & \text{if } \sum_{i=1}^{j} w_i < p \sum_{i=1}^{n} w_i < \sum_{i=1}^{j+1} w_i. \end{cases} \tag{3.1}$$

Using index sets $I_{\leq \hat{q}_{0.2}} := \{i \in \{1, \ldots, n\} : x_i \leq \hat{q}_{0.2}\}$ and $I_{>\hat{q}_{0.8}} := \{i \in \{1, \ldots, n\} : x_i > \hat{q}_{0.8}\}$, the quintile share ratio is estimated by

$$\widehat{QSR} := \frac{\sum_{i \in I_{>\hat{q}_{0.8}}} w_i x_i}{\sum_{i \in I_{\leq \hat{q}_{0.2}}} w_i x_i}. \tag{3.2}$$

With package **laeken**, the quintile share ratio can be estimated using the function `qsr()`. Sample weights can thereby be supplied via the `weights` argument.

```
R> qsr("eqIncome", weights = "rb050", data = eusilc)
```

```
Value:
[1] 3.971415
```

3.2.2 Gini coefficient

The *Gini coefficient* is defined as the relationship of cumulative shares of the population arranged according to the level of equivalized disposable income, to the cumulative share of the equivalized total disposable income received by them (Eurostat, 2004, 2009).

For the estimation of the Gini coefficient from a sample, the sample weights need to be taken into account. In mathematical terms, the Gini coefficient is estimated by

$$\widehat{Gini} := 100 \left[\frac{2 \sum_{i=1}^{n} \left(w_i x_i \sum_{j=1}^{i} w_j \right) - \sum_{i=1}^{n} {w_i}^2 x_i}{\left(\sum_{i=1}^{n} w_i \right) \sum_{i=1}^{n} \left(w_i x_i \right)} - 1 \right]. \tag{3.3}$$

The function `gini()` is available in **laeken** to estimate the Gini coefficient. As before, sample weights can be specified with the `weights` argument.

```
R> gini("eqIncome", weights = "rb050", data = eusilc)

Value:
[1] 26.48962
```

3.3 The Pareto distribution

The *Pareto distribution* is well studied in the literature and is defined in terms of its cumulative distribution function

$$F_\theta(x) = 1 - \left(\frac{x}{x_0} \right)^{-\theta}, \qquad x \geq x_0, \tag{3.4}$$

where $x_0 > 0$ is the scale parameter and $\theta > 0$ is the shape parameter (KLEIBER and KOTZ, 2003). Furthermore, its density function is given by

$$f_\theta(x) = \frac{\theta x_0^\theta}{x^{\theta+1}}, \qquad x \geq x_0. \tag{3.5}$$

Figure 3.1 visualizes the Pareto probability density function with scale parameter $x_0 = 1$ and different values of the shape parameter θ. Clearly, the Pareto distribution is a highly right-skewed distribution with a heavy tail. It is therefore reasonable to assume that a random variable following a Pareto distribution contains extreme values. The effect of changing the shape parameter θ is visible in the probability mass at the scale parameter x_0: the higher θ, the higher the probability mass at x_0.

In Pareto tail modeling, the cumulative distribution function on the whole range of x is modeled as

$$F(x) = \begin{cases} G(x), & \text{if } x \leq x_0, \\ G(x_0) + (1 - G(x_0)) F_\theta(x), & \text{if } x > x_0, \end{cases} \tag{3.6}$$

where G is an unknown distribution function (DUPUIS and VICTORIA-FESER, 2006).

Let n be the number of observations and let $\boldsymbol{x} = (x_1, \ldots, x_n)'$ denote the observed values with $x_1 \leq \ldots \leq x_n$. In addition, let k be the number of observations to be used for tail modeling. In this scenario, the threshold x_0 is estimated by

$$\hat{x}_0 := x_{n-k}. \tag{3.7}$$

[ht]

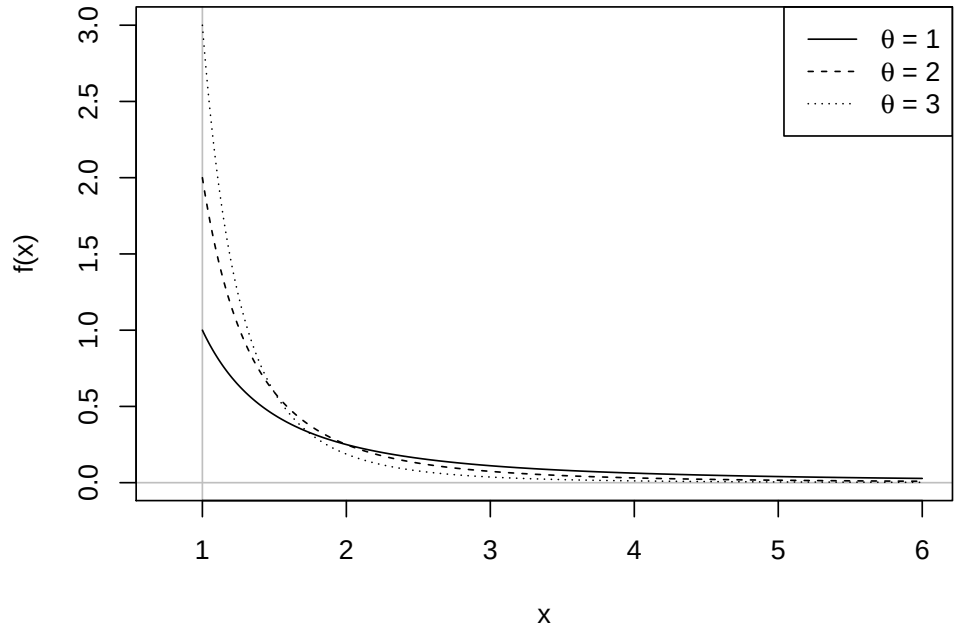


Figure 3.1: Pareto probability density functions with parameters $x_0 = 1$ and $\theta = 1, 2, 3$.

If an estimate $\hat{x}_0$ for the scale parameter of the Pareto distribution has been obtained, k is given by the number of observations larger than $\hat{x}_0$. Thus estimating x_0 and k directly corresponds with each other.

In the remainder of this chapter, the equivalized disposable income of the EU-SILC example data is of main interest. Consequently, the Pareto distribution will be modeled at the household level rather than the individual level. Moreover, the focus of the chapter would be on robust estimation of the social exclusion indicators. Hence the equivalized disposable income of the household with the largest income is replaced by a large outlier.

```
R> hID <- eusilc$db030[which.max(eusilc$eqIncome)]
R> eusilc[eusilc$db030 == hID, "eqIncome"] <- 1e+07
```

Since the aim is to model a Pareto distribution at the household level, the following command creates a data set that contains only the equivalized disposable income and the sample weights on the household level. This data set will be used in Sections 3.4 and 3.5 to estimate the parameters of the Pareto distribution.

```
R> eusilcH <- eusilc[!duplicated(eusilc$db030), c("eqIncome", "db090")]
```

3.4 Finding the threshold

The aim of the methods presented in this sections is to find the threshold x_0 for modeling the Pareto distribution. Several methods for the estimation of the threshold x_0 or the

number of observations k in the tail have been proposed in the literature, but those proposals typically do not consider sample weights.

BEIRLANT et al. (1996b,a) developed a procedure that analytically determines the optimal choice of k for the Hill estimator of the shape parameter (HILL, 1975, see also Section 3.5.1 of this paper) by minimizing the asymptotic mean squared error (AMSE). In package **laeken**, this approach is implemented in the function `minAMSE()`. However, the procedure is designed for the non-robust Hill estimator and is therefore not further discussed in this paper. Furthermore, DANIELSSON et al. (2001) proposed a bootstrap method to find the optimal k for the Hill estimator with respect to the AMSE, which has less analytical requirements than the approach by BEIRLANT et al. (1996b,a). Please note that this method is not robust either and that it is currently not available in package **laeken**. A robust prediction error criterion for choosing the number of observations k in the tail and estimating the shape parameter θ was developed by DUPUIS and VICTORIA-FESER (2006). Nevertheless, our implementation of this robust criterion was unstable and is therefore not included in **laeken**.

In any case, HOLZER (2009) concludes that graphical methods for finding the threshold outperform those analytical approaches in the case of EU-SILC data. While this section is thus focused on graphical methods, a simple rule of thumb designed specifically for the equivalized disposable income in EU-SILC data is described in the following as well.

3.4.1 Van Kerm's rule of thumb

VAN KERM (2007) presented a formula that is more of a rule of thumb for the threshold of the equivalized disposable income in EU-SILC data. Is is given by

$$\hat{x}_0 := \min(\max(2.5\bar{x}, q_{0.98}), q_{0.97}), \tag{3.8}$$

where $\bar{x}$ is the weighted mean, and $q_{0.98}$ and $q_{0.97}$ are weighted quantiles as defined in Equation (3.1).

In package **laeken**, the function `paretoScale()` provides functionality for computing the threshold with van Kerm's rule of thumb. The argument `w` is available to supply sample weights.

```
R> ts <- paretoScale(eusilcH$eqIncome, w = eusilcH$db090)
R> ts
```

```
Threshold: 48459.43
Number of observations in the tail: 119
```

It should be noted that the function returns an object of class `paretoScale`, which consists of a component `x0` for the threshold (scale parameter) and a component `k` for the number of observations in the tail of the distribution, i.e., that are larger than the threshold.

3.4.2 Pareto quantile plot

The *Pareto quantile plot* is a graphical method for inspecting the parameters of a Pareto distribution. For the case without sample weights, it is described in detail in Beirlant et al. (1996b).

If the Pareto model holds, there exists a linear relationship between the logarithms of the observed values and the quantiles of the standard exponential distribution, since the logarithm of a Pareto distributed random variable follows an exponential distribution. Hence the logarithms of the observed values, $\log(x_i)$, $i = 1, \ldots, n$, are plotted against the theoretical quantiles.

In the case without sample weights, the theoretical quantiles of the standard exponential distribution are given by

$$-\log\left(1 - \frac{i}{n+1}\right), \qquad i = 1, \ldots, n, \tag{3.9}$$

i.e., by dividing the range into $n + 1$ equally sized subsets and using the resulting n inner gridpoints as probabilities for the quantiles. If the data contain sample weights, the range of the exponential distribution needs to be divided according to the weights of the n observations. The Pareto quantile plot is thus generalized by using the theoretical quantiles

$$-\log\left(1 - \frac{\sum_{j=1}^{i} w_j}{\sum_{j=1}^{n} w_j} \frac{n}{n+1}\right), \qquad i = 1, \ldots, n, \tag{3.10}$$

where the correction factor $\frac{n}{n+1}$ ensures that the quantiles reduce to (3.9) if all sample weights are equal.

If the tail of the data follows a Pareto distribution, those observations form almost a straight line. The leftmost point of a fitted line can thus be used as an estimate of the threshold x_0, the scale parameter. All values starting from the point after the threshold may be modeled by a Pareto distribution, but this point cannot be determined exactly. Furthermore, the slope of the fitted line is in turn an estimate of $\frac{1}{\theta}$, the reciprocal of the shape parameter.

Figure 3.2 displays the Pareto quantile plot for the example data `eusilc` on the household level with the largest observation replaced by an outlier. The plot is generated using the function `paretoQPlot()`, which allows to supply sample weights via the argument `w`. In addition, the threshold can be selected interactively by clicking on a data point. Information on the selected threshold is then printed on the R console. When the interactive selection is terminated, which is typically done by a secondary mouse click, the selected threshold is returned as an object of class `paretoScale`.

Another advantage of the Pareto quantile plot is also illustrated in Figure 3.2. Nonrepresentative outliers such as the large income introduced into the example data in Section 3.3, i.e., extreme observations in the upper tail that deviate from the Pareto model, are clearly visible.

```
R> paretoQPlot(eusilcH$eqIncome, w = eusilcH$db090)
```

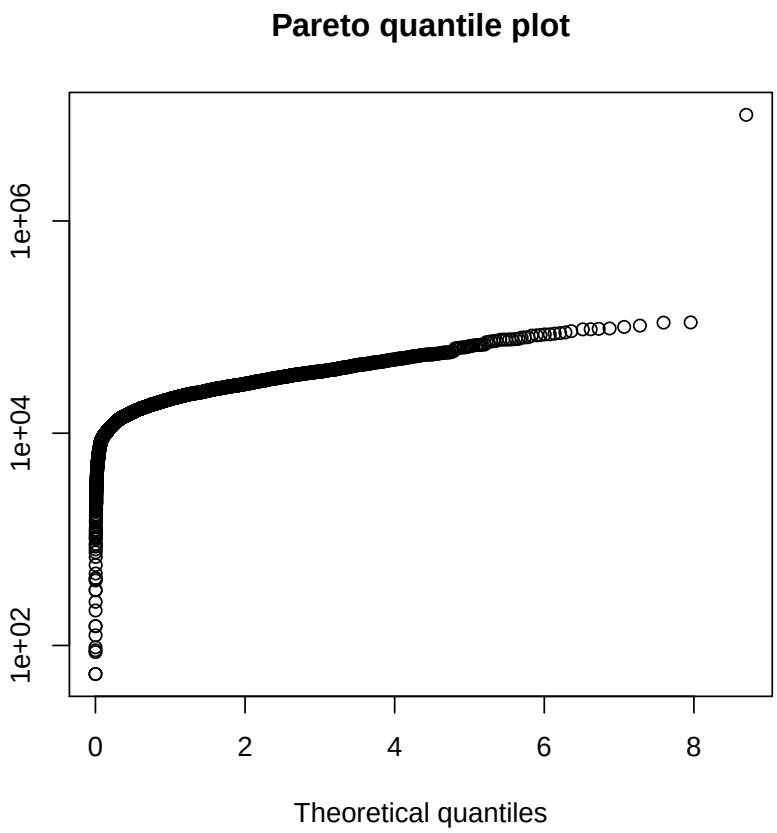


Figure 3.2: Pareto Quantile plot for the example data `eusilc` on the household level with the largest observation replaced by an outlier.

3.4.3 Mean excess plot

The *mean excess plot* is another graphical method for inspecting the threshold for Pareto tail modeling, but it does not provide information on the shape parameter. It is based on the excess function

$$e(x_0) := \mathbb{E}(x - x_0 | x > x_0), \qquad x_0 \geq 0. \tag{3.11}$$

A detailed description for the case without sample weights can be found in BORKOVEC and KLÜPPELBERG (2000).

For the following definition of the mean excess plot, keep in mind that the observations are sorted such that $x_1 \leq \ldots \leq x_n$. For each observation x_i, $i = 1, \ldots, \lfloor n - \sqrt{n} \rfloor$, the empirical excess function e_n is computed. In the case without sample weights, the expectation in Equation (3.11) is replaced by the arithmetic mean, and the empirical excess function is

given by

$$e_n(x_i) := \frac{1}{n-i} \sum_{j=i+1}^{n} (x_j - x_i), \qquad i = 1, \ldots, \lfloor n - \sqrt{n} \rfloor. \tag{3.12}$$

The values of the empirical excess function $e_n(x_i)$ are then plotted against the corresponding x_i, $i = 1, \ldots, \lfloor n - \sqrt{n} \rfloor$. If sample weights are available in the data, the mean excess plot is simply generalized by using the weighted mean for the empirical excess function:

$$e_n(x_i) := \frac{1}{\sum_{j=i+1}^{n} w_j} \sum_{j=i+1}^{n} w_j (x_j - x_i), \qquad i = 1, \ldots, \lfloor n - \sqrt{n} \rfloor. \tag{3.13}$$

If the tail of the data follows a Pareto distribution, those observations show a positive linear trend. The leftmost point of a fitted line can thus be used as an estimate of the threshold x_0, the scale parameter. As for the Pareto quantile plot, a disadvantage of the mean excess plot is that the threshold cannot be determined exactly.

```
R> meanExcessPlot(eusilcH$eqIncome, w = eusilcH$db090)
```

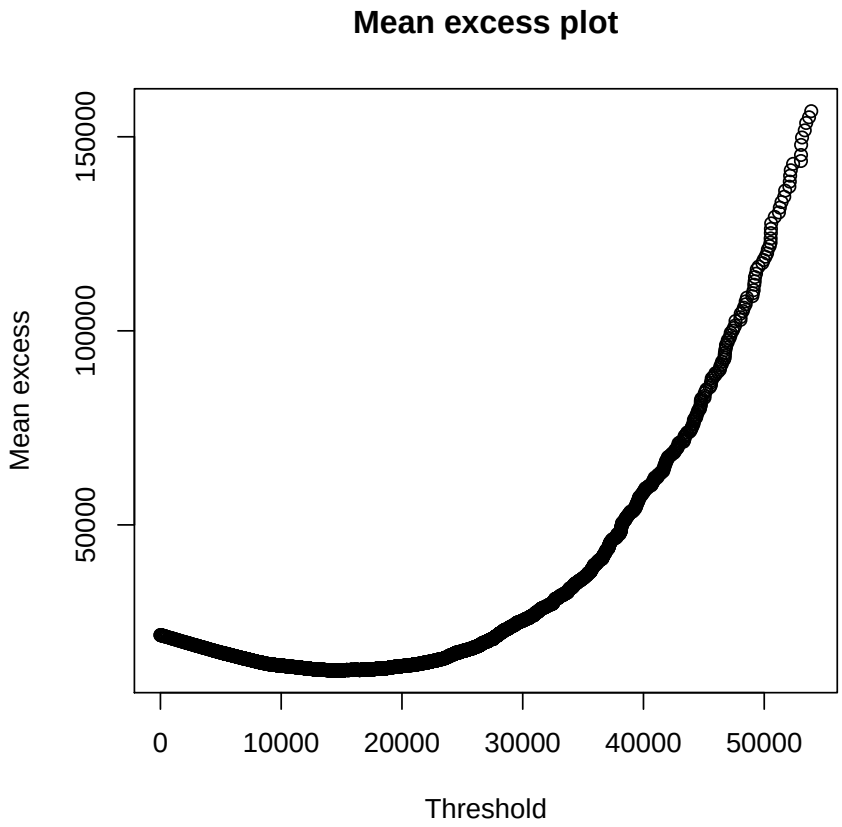


Figure 3.3: Mean excess plot for the example data `eusilc` on the household level with the largest observation replaced by an outlier.

Figure 3.3 shows the mean excess plot for the example data `eusilc` on the household level with the largest observation replaced by an outlier. The function `meanExcessPlot()` is thereby used to produce the plot. Sample weights can be supplied via the argument `w`. Interactive selection of the threshold works just like for the Pareto quantile plot. Again, the selected threshold is returned as an object of class `paretoScale`.

3.5 Estimation of the shape parameter

This section is focused on methods for estimating the shape parameter θ once the threshold x_0 is fixed. It should be noted that none of the original proposals takes sample weights into account. Most estimators presented in the following were therefore adjusted for the case of sample weights.

3.5.1 Hill estimator

The maximum likelihood estimator for the shape parameter of the Pareto distribution was introduced by HILL (1975) and is referred to as the *Hill* estimator. If the data do not contain sample weights, it is given by

$$\hat{\theta}_{\text{Hill}} = \frac{k}{\sum_{i=1}^{k} \log x_{n-k+i} - k \log x_{n-k}}. \tag{3.14}$$

In the case of sample weights, the *weighted Hill* (wHill) estimator is given by generalizing Equation (3.14) to

$$\hat{\theta}_{\text{wHill}} = \frac{\sum_{i=1}^{k} w_{n-k+i}}{\sum_{i=1}^{k} w_{n-k+i} \left(\log x_{n-k+i} - k \log x_{n-k}\right)}. \tag{3.15}$$

Package **laeken** provides the function `thetaHill()` to compute the Hill estimator. It requires to specify either the number of observations in the tail via the argument `k`, or the threshold via the argument `x0`. Furthermore, the argument `w` can be used to supply sample weights. In the following example, the shape parameter is estimated using the largest observations (first command) and the threshold (second command) as computed with van Kerm's rule of thumb in Section 3.4.1.

```
R> thetaHill(eusilcH$eqIncome, k = ts$k, w = eusilcH$db090)
```

```
[1] 3.437979
```

```
R> thetaHill(eusilcH$eqIncome, x0 = ts$x0, w = eusilcH$db090)
```

```
[1] 3.437979
```

3.5.2 Weighted maximum likelihood estimator

The *weighted maximum likelihood* (WML) estimator (Dupuis and Morgenthaler, 2002; Dupuis and Victoria-Feser, 2006) falls into the class of M-estimators and is given by the solution $\hat{\theta}$ of

$$\sum_{i=1}^{k} \Psi(x_{n-k+i}, \theta) = 0 \tag{3.16}$$

with

$$\Psi(x, \theta) := u(x, \theta) \frac{\partial}{\partial \theta} \log f(x, \theta) = u(x, \theta) \left(\frac{1}{\theta} - \log \frac{x}{x_0} \right), \tag{3.17}$$

where $u(x, \theta)$ is a weight function with values in $[0, 1]$ and $f(x, \theta)$ is a density function with unknown population parameter θ. In the implementation in package **laeken**, a Huber type weight function is used by default, as proposed by Dupuis and Victoria-Feser (2006). Let the logarithms of the relative excesses be denoted by

$$z_i := \log \left(\frac{x_{n-k+i}}{x_{n-k}} \right), \qquad i = 1, \ldots, k. \tag{3.18}$$

In the Pareto model, these can be predicted by

$$\hat{z}_i := -\frac{1}{\theta} \log \left(\frac{k+1-i}{k+1} \right), \qquad i = 1, \ldots, k. \tag{3.19}$$

The variance of z_i is given by

$${\sigma_i}^2 := \sum_{j=1}^{i} \frac{1}{\theta^2 (k-i+j)^2}, \qquad i = 1, \ldots, k. \tag{3.20}$$

Using the standardized residuals

$$r_i := \frac{z_i - \hat{z}_i}{\sigma_i}, \tag{3.21}$$

the Huber type weight function with tuning constant c is defined as

$$u(x_{n-k+i}, \theta) := \begin{cases} 1, & \text{if } |r_i| \leq c, \\ \frac{c}{|r_i|}, & \text{if } |r_i| > c. \end{cases} \tag{3.22}$$

For this choice of weight function, the bias of $\hat{\theta}$ is approximated by

$$\hat{B}(\hat{\theta}) = -\frac{\sum_{i=1}^{k} \left(u_i \frac{\partial}{\partial \theta} \log f_i \right) |_{\hat{\theta}} \left(F_{\hat{\theta}}(x_{n-k+i}) - F_{\hat{\theta}}(x_{n-k+i-1}) \right)}{\sum_{i=1}^{k} \left(\frac{\partial}{\partial \theta} u_i \frac{\partial}{\partial \theta} \log f_i + u_i \frac{\partial^2}{\partial \theta^2} \log f_i \right) |_{\hat{\theta}} \left(F_{\hat{\theta}}(x_{n-k+i}) - F_{\hat{\theta}}(x_{n-k+i-1}) \right)}, \tag{3.23}$$

where $u_i := u(x_{n-k+i}, \theta)$ and $f_i := f(x_{n-k+i}, \theta)$. This term is used to obtain a bias-corrected estimator

$$\tilde{\theta} := \hat{\theta} - \hat{B}(\hat{\theta}). \tag{3.24}$$

For details and proofs of the above statements, as well as for information on a probability-based weight function $u(x, \theta)$, the reader is referred to DUPUIS and MORGENTHALER (2002) and DUPUIS and VICTORIA-FESER (2006). However, note the WML estimator does not consider sample weights. An adjustment of the estimator to take sample weights into account is currently not available due to its complexity. For sampling designs that lead to equal sample weights, the WML estimator may still be useful, though.

The function `thetaWML()` is available in **laeken** to compute the WML estimator. Again, either the argument `k` or `x0` needs to be used to specify the number of observations in the tail or the threshold. Since the sample weights in the example data are not equal, the following example is only included to demonstrate the use of the function.

```
R> thetaWML(eusilcH$eqIncome, k = ts$k)

[1] 4.226204

R> thetaWML(eusilcH$eqIncome, x0 = ts$x0)

[1] 4.226204
```

3.5.3 Integrated squared error estimator

For the *integrated squared error* (ISE) estimator (VANDEWALLE et al., 2007), the Pareto distribution is modeled in terms of the relative excesses

$$y_i := \frac{x_{n-k+i}}{x_{n-k}}, \qquad i = 1, \ldots, k. \tag{3.25}$$

The density function of the Pareto distribution for the relative excesses is approximated by

$$f_\theta(y) = \theta y^{-(1+\theta)}. \tag{3.26}$$

The ISE estimator is then given by minimizing the integrated squared error criterion (TERRELL, 1990):

$$\hat{\theta} = \arg\min_\theta \left[\int f_\theta^2(y) dy - 2\mathbb{E}(f_\theta(Y)) \right]. \tag{3.27}$$

If there are no sample weights in the data, the mean is used as an unbiased estimator of $\mathbb{E}(f_\theta(Y))$ in order to obtain the ISE estimate

$$\hat{\theta}_{\text{ISE}} = \arg\min_\theta \left[\int f_\theta^2(y) dy - \frac{2}{k} \sum_{i=1}^{k} f_\theta(y_i) \right]. \tag{3.28}$$

See Vandewalle et al. (2007) for more information on the ISE estimator for the case without sample weights.

If sample weights are available in the data, the mean in Equation (3.28) is simply replaced by a weighted mean to obtain the *weighted integrated squared error* (wISE) estimator:

$$\hat{\theta}_{\text{wISE}} = \arg\min_{\theta} \left[\int f_\theta^2(y)dy - \frac{2}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_\theta(y_i) \right]. \tag{3.29}$$

With package **laeken**, the ISE estimator can be computed using the function `thetaISE()`. The arguments `k` and `x0` are available to specify either the number of observations in the tail or the threshold, and sample weights can be supplied via the argument `w`.

```
R> thetaISE(eusilcH$eqIncome, k = ts$k, w = eusilcH$db090)
```

```
[1] 3.993801
```

```
R> thetaISE(eusilcH$eqIncome, x0 = ts$x0, w = eusilcH$db090)
```

```
[1] 3.993801
```

3.5.4 Partial density component estimator

For the *partial density component* (PDC) estimator Vandewalle et al. (2007) minimizes the integrated squared error criterion using an incomplete density mixture model uf_θ. If the data do not contain sample weights, the PDC estimator in is thus given by

$$\hat{\theta}_{\text{PDC}} = \arg\min_{\theta} \left[u^2 \int f_\theta^2(y)dy - \frac{2u}{k} \sum_{i=1}^{k} f_\theta(y_i) \right]. \tag{3.30}$$

The parameter u can be interpreted as a measure of the uncontaminated part of the sample and is estimated by

$$\hat{u} = \frac{\frac{1}{k}\sum_{i=1}^{k} f_{\hat{\theta}}(y_i)}{\int f_{\hat{\theta}}^2(y)dy}. \tag{3.31}$$

See Vandewalle et al. (2007) and references therein for more information on the PDC estimator for the case without sample weights.

Taking sample weights into account, the *weighted partial density component* (wPDC) estimator is obtained by generalizing Equations (3.30) and (3.31) to

$$\hat{\theta}_{\text{wPDC}} = \arg\min_{\theta} \left[u^2 \int f_\theta^2(y)dy - \frac{2u}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_\theta(y_i) \right], \tag{3.32}$$

$$\hat{u} = \frac{\frac{1}{\sum_{i=1}^{k} w_{n-k+i}} \sum_{i=1}^{k} w_{n-k+i} f_{\hat{\theta}}(y_i)}{\int f_{\hat{\theta}}^2(y) dy}. \tag{3.33}$$

The function `thetaPDC()` is implemented in package **laeken** to compute the PDC estimator. As for the other estimators, it is necessary to specify either the number of observations in the tail via the argument `k`, or the threshold via the argument `x0`. Sample weights can be supplied using the argument `w`.

```
R> thetaPDC(eusilcH$eqIncome, k = ts$k, w = eusilcH$db090)
```

```
[1] 4.132596
```

```
R> thetaPDC(eusilcH$eqIncome, x0 = ts$x0, w = eusilcH$db090)
```

```
[1] 4.132596
```

3.6 Estimation of the indicators using Pareto tail modeling

Three approaches based on Pareto tail modeling for reducing the influence of outliers on the social exclusion indicators are implemented in the R package **laeken**:

Calibration for nonrepresentative outliers (CN): Values larger than a certain quantile of the fitted distribution are declared as nonrepresentative outliers. Since these are considered to be unique to the population data, the sample weights of the corresponding observations are set to 1 and the weights of the remaining observations are adjusted accordingly by calibration.

Replacement of nonrepresentative outliers (RN): Values larger than a certain quantile of the fitted distribution are declared as nonrepresentative outliers. Only these nonrepresentative outliers are replaced by values drawn from the fitted distribution, thereby preserving the order of the original values.

Replacement of the tail (RT): All values above the threshold are replaced by values drawn from the fitted distribution. The order of the original values is preserved.

An evaluation of the RT approach by means of a simulation study can be found in ALFONS et al. (2010).

Keep in mind that the largest observation in the example data `eusilc` was replaced by a large outlier in Section 3.3. With the following command, the Gini coefficient is estimated according to the Eurostat definition to show that even a single outlier can completely distort the results for the standard estimation (see Section 3.2.2 for the original value).

```
R> gini("eqIncome", weights = "rb050", data = eusilc)
```

```
Value:
[1] 29.24333
```

For Pareto tail modeling, the function `paretoTail()` is implemented in **laeken**. It returns an object of class `paretoTail`, which contains all the necessary information for further analysis using the three approaches described above. Note that the household IDs are supplied via the argument `groups` such that the Pareto distribution is fitted on the household level rather than the individual level. In addition, the PDC is used by default to estimate the shape parameter. Other estimators can be specified via the `method` argument.

```
R> fit <- paretoTail(eusilc$eqIncome, k = ts$k, w = eusilc$db090,
+     groups = eusilc$db030)
```

The function `reweightOut()` is available for semiparametric estimation with the CN approach. It returns a vector of the recalibrated weights. In this example, regional information is used as auxiliary variables for calibration. The function `calibVars()` thereby transforms a factor into a matrix of binary variables, as required by the calibration function `calibWeights()`, which is called internally. These recalibrated weights are then simply used to estimate the Gini coefficient with function `gini()`.

```
R> w <- reweightOut(fit, calibVars(eusilc$db040))
R> gini(eusilc$eqIncome, w)
```

```
Value:
[1] 26.45973
```

For the RN approach, the function `replaceOut()` is implemented. Since values are drawn from the fitted distribution to replace the observations flagged as outliers, the seed of the random number generator is set first for reproducibility of the results. The returned vector of incomes is then supplied to `gini()` to estimate the Gini coefficient.

```
R> set.seed(1234)
R> eqIncome <- replaceOut(fit)
R> gini(eqIncome, weights = eusilc$rb050)
```

```
Value:
[1] 26.46924
```

Similarly, the function `replaceTail()` is available for the RT approach. Again, the seed of the random number generator is set beforehand.

```
R> set.seed(1234)
R> eqIncome <- replaceTail(fit)
R> gini(eqIncome, weights = eusilc$rb050)

Value:
[1] 26.64921
```

It should be noted that `replaceTail()` can also be used for the RN approach by setting the argument `all` to `FALSE`. In fact, `replaceOut(x, ...)` is a simple wrapper for `replaceTail(x, all = FALSE, ...)`.

In any case, the estimates for the semiparametric approaches based on Pareto tail modeling are very close to the original value before the outlier has been introduced (see Section 3.2.2), whereas the standard estimation is corrupted by the outlier. Furthermore, the estimation of other indicators such as the quintile share ratio (see Section 3.2.1) using the semiparametric approaches is straightforward and hence not shown here.

3.7 Conclusions

This chapter introduces robust semi-parametric modelling which incooperates with sampling weights. It also shows the functionality of package **laeken** for robust semiparametric estimation of social exclusion indicators based on Pareto tail modeling. Most notably, it demonstrates that the functions are easy to use and that the implementation follows an object-oriented design.

Furthermore, it is shown that the standard estimation of the inequality indicators can be corrupted by a single outlier, thus underlining the need for robust alternatives. Three approaches for robust semiparametric estimation based on Pareto tail modeling are thereby implemented such that the corresponding functions share a common interface for ease of use.

Extensive simulations were carried out to gain insight into the properties of the methods. They are fully documented in Deliverable 7.1.

Clearly, the RT approach introduces to much additional uncertainty and is not recommendable, while the CN approach is favourable (see the outcome of the simulation study in Deliverable 7.1 where detailed recommondations were made).

Bibliography

Alfons, A., **Holzer, J.** and **Templ, M.** (**2011**a): **laeken**: Laeken indicators for measuring social cohesion. R package version 0.2.1.
URL http://CRAN.R-project.org/package=laeken

Alfons, A. and **Kraft, S.** (**2010**): **simPopulation**: Simulation of synthetic populations for surveys based on sample data. R package version 0.2.1.
URL http://CRAN.R-project.org/package=simPopulation

Alfons, A., **Kraft, S.**, **Templ, M.** and **Filzmoser, P.** (**2011**b): *Simulation of close-to-reality population data for household surveys with application to EU-SILC*. Statistical Methods & Applications, accepted for publication.

Alfons, A., **Templ, M.**, **Filzmoser, P.** and **Holzer, J.** (**2010**): *A comparison of robust methods for Pareto tail modeling in the case of Laeken indicators*. **Borgelt, C.**, **González-Rodríguez, G.**, **Trutschnig, W.**, **Lubiano, M.**, **Gil, M.**, **Grzegorzewski, P.** and **Hryniewicz, O.** (editors) Combining Soft Computing and Statistical Methods in Data Analysis, *Advances in Intelligent and Soft Computing*, vol. 77, pp. 17–24, Heidelberg: Springer, ISBN 978-3-642-14745-6.

Beirlant, J., **Vynckier, P.** and **Teugels, J.** (**1996**a): *Excess functions and estimation of the extreme-value index*. Bernoulli, 2 (4), pp. 293–318.

Beirlant, J., **Vynckier, P.** and **Teugels, J.** (**1996**b): *Tail index estimation, Pareto quantile plots, and regression diagnostics*. Journal of the American Statistical Association, 31 (436), pp. 1659–1667.

Borkovec, M. and **Klüppelberg, C.** (**2000**): *Extremwerttheorie für Finanzzeitreihen – ein unverzichtbares Werkzeug im Risikomanagement*. **Johanning, L.** and **Rudolph, B.** (editors) Handbuch Risikomanagement, pp. 219–241, Bad Soden: Uhlenbruch, ISBN 3933207150.

Danielsson, J., **de Haan, L.**, **Peng, L.** and **de Vries, C.** (**2001**): *Using a bootstrap method to choose the sample fraction in tail index estimation*. Journal of Multivariate Analysis, 76 (2), pp. 226–248.

Dupuis, D. and **Morgenthaler, S.** (**2002**): *Robust weighted likelihood estimators with an application to bivariate extreme value problems*. The Canadian Journal of Statistics, 30 (1), pp. 17–36.

Dupuis, D. and **Victoria-Feser, M.-P.** (**2006**): *A robust prediction error criterion for Pareto modelling of upper tails*. The Canadian Journal of Statistics, 34 (4), pp. 639–658.

Eurostat (**2004**): *Common cross-sectional EU indicators based on EU-SILC; the gender pay gap*. EU-SILC 131-rev/04, Unit D-2: Living conditions and social protection, Directorate D: Single Market, Employment and Social statistics, Eurostat, Luxembourg.

Eurostat (**2009**): *Algorithms to compute social inclusion indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC)*. Doc. LC-ILC/39/09/EN-rev.1, Unit F-3: Living conditions and social protection, Directorate F: Social and information society statistics, Eurostat, Luxembourg.

Hill, B. (**1975**): *A simple general approach to inference about the tail of a distribution*. The Annals of Statistics, 3 (5), pp. 1163–1174.

Holzer, J. (**2009**): Robust methods for the estimation of selected Laeken indicators. Diploma thesis, Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria.

Hulliger, B. and **Schoch, T.** (**2009**): *Robustification of the quintile share ratio*. New Techniques and Technologies for Statistics, Brussels.

Kleiber, C. and **Kotz, S.** (**2003**): Statistical Size Distributions in Economics and Actuarial Sciences. Hoboken: John Wiley & Sons, ISBN 0-471-15064-9.

R Development Core Team (**2011**): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

Templ, M. and **Alfons, A.** (**2011**a): *Standard Methods for Point Estimation of Social Inclusion Indicators using the R Package **laeken***. Research Report CS-2011-1, Department of Statistics and Probability Theory, Vienna University of Technology.
URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2011-1complete.pdf

Templ, M. and **Alfons, A.** (**2011**b): *Variance Estimation of Social Inclusion Indicators using the R Package **laeken***. Research Report CS-2011-3, Department of Statistics and Probability Theory, Vienna University of Technology.
URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2011-3complete.pdf

Terrell, G. (**1990**): *Linear density estimates*. Proceedings of the Statistical Computing Section, pp. 297–302, American Statistical Association.

Van Kerm, P. (**2007**): *Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC*. IRISS Working Paper Series 2007-01, CEPS/INSTEAD.

Vandewalle, B., **Beirlant, J.**, **Christmann, A.** and **Hubert, M.** (**2007**): *A robust estimator for the tail index of Pareto-type distributions*. Computational Statistics & Data Analysis, 51 (12), pp. 6252–6268.

Chapter 4

Robust Non-Parametric Quintile Share Ratio Estimator

4.1 Introduction

Lorenz curve, Gini coefficient, and the income share ratios – whereof the Quintile Share Ratio is the most prominent and represents the primary income inequality measure in the European Union's set of Laeken indicators (European Commission, 2003) – are central to the analysis of income distributions, embodying intuition about the income inequality and social cohesion. The formal welfare propositions can only be satisfactorily invoked if sample data can be taken as a reasonable representation of the income distribution under consideration. In particular, data from income distributions may be contaminated by recording errors, measurement errors and the like and, if the data cannot be purged of these, welfare conclusions drawn from the data can be seriously misleading (Cowell and Flachaire, 2007). In addition, extreme observations, which are not necessarily errors or some form of contamination, can exert strong influence on the estimate of an inequality measure (cf. Chambers, 1986). There is large bulk of literature discussing robustness issues in income inequality measurement (see e.g., Victoria-Feser and Ronchetti, 1994; Cowell and Victoria-Feser, 1996, 2002, 2003, 2006) – but not on income share ratios. In addition to potential outliers, appropriate statistical inference procedures are needed to take sampling variability and the complex nature of the sampling design into account in drawing conclusions. That is, estimators must cope with the heavily non-iid data structure, which is typical for household-survey samples (e.g., European Statistics on Income and Living Conditions (EU-SILC), Panel Study of Income Dynamics (PSID), or the Current Population Survey (CPS)). This pertains primarly to variance estimation (Hulliger and Münnich, 2006). Using the variance-covariance formulae and estimation procedures applicable for a simple random sample will likely lead to biased variance estimates for a complex random sample, which may ultimately lead to erroneous inference (see e.g., Nygård and Sandström, 1989; Zheng, 2002; Pfeffermann, 1993).

The problem of extreme values is accentuated in complex survey samples. Unlike the case of iid-data, it seems reasonable to define outliers not with respect to the sample because these may have been induced or masked by the sampling design. Insofar, the sampling design, or precisely the weighting scheme constitutes another, but indirect, channel by

means of which contamination may affect the inverse-probability-weightes (e.g., Horvitz-Thompson) estimators, aside from the contaminated observations. Indeed, extreme observations together with small inclusion probabilities have a particularly large influence on these estimators (cf. SMITH, 1987; HULLIGER, 1995; BEAUMONT and RIVEST, 2009).

The purpose of this paper is to provide robust income share ratio estimators under complex random sampling schemes, which accord well with pragmatic procedures that are adopted by applied researchers in this field. We show that the robust quantile share ratio estimators obtain a good trade-off between bias and efficiency. The estimators are asymptotically normal under weak conditions on the sampling design and their linearization variance estimators can be estimated by a inverse-probability weighted, design-unbiased estimator for a general class of stratified multi-stage cluster sampling designs, which are typical for household surveys (DEATON, 2000).

The remainder of the paper is organized as follows. In Section 4.2, we introduce some elementary functionals that will be used as building blocks. Section 4.3 is concerned with robust estimators of the quantile share ratio. In Section 4.4, we discuss estimating the quantile share ratios in stratified multi-stage cluster samples. Section 4.5 investigates variance estimation and in Section 4.6, we provide strategies to adaptively choose the trimming proportions.

4.2 Preliminaries

There is large number of approaches in the literature to the definition of an inequality measure. We confine ourselves to the discussion of the class of income share ratios. Typically, the income shares are defined in terms of the income distribution itself, that is quantiles. Income quantile share ratios (QSR) offer an intuitive interpretation of the income inequality (in contrast to e.g., the Gini coefficient) in terms of the ratio of total income received by richest, say, for instance 20% of a country's population (top quintile) to that received by the 20 % of the population with the lowest income (bottom quintile); or any similar choice of quantiles. The quintile share ratio is the most important and marks the primary income inequality measure in the European Union's set of Laeken indicators (denoted S80/S20; aka Indicators on Social Exclusion an Poverty), which play a central role in monitoring the performance of the EU member states in promoting social inclusion (cf. ATKINSON et al., 2002).

First, we introduce three elementary, statistical functionals (for infinite populations) as building blocks. Let $y_1, \ldots, y_N$ be independent and identically distributed (iid) realizations of a univariate (non-negative) random variable Y from a parametric model $F \in \mathcal{F}$, where $\mathcal{F}$ is the set of all absolutely continuous cumulative distribution functions (cdf) with support $\mathcal{Y} \subseteq \mathbb{R}$. Let $\underline{\mathcal{Y}}$ ($\overline{\mathcal{Y}}$) denote the lower (upper) limit of the support, $\mathcal{Y}$, (i.e., $\underline{\mathcal{Y}}$=0, $\overline{\mathcal{Y}} = \infty$). Second, let $\mathbb{Q} := [0, 1]$ denote the set of population proportions. We shall write any statistic as statistical functional $T(F)$ defined on the space of probability distributions. In particular we define the quantile functional as

Definition 1 (Quantile functional)**.** *For any $\beta \in \mathbb{Q}$, the βth quantile is the functional*

$\xi : \mathcal{F} \times \mathbb{Q} \mapsto \mathcal{Y}$ *such that*

$$\xi(F;\beta) = \inf\{y : F(y) \geq \beta\}. \tag{4.1}$$

The finite population analogue of $\xi(F;\beta)$ is obtained by replacing F with the empirical distribution F_U; this is a distribution consisting of N point-masses $1/N$, one at each unit $U = \{1,\ldots,i,\ldots,N\}$. For ease of simplicity, let $\beta \in \mathbb{Q}_U$ with $\mathbb{Q}_U := \{q|q = i/N,$ where $i \in U\}$. Thus, the βth quantile, $\xi(F_U;\beta)$, is obtained as the solution to the equation $F_U(\xi(\,;\beta)) = \beta$.

Given a predetermined quantile $\xi(F;\beta)$, we define an (income) quantile share mean, as the mean income received by all units (i.e., households or individuals) up to the particular quantile $\xi(F;\beta)$.

Definition 2 (Quantile share mean)**.** *For any $\beta \in \mathbb{Q}$, the βth quantile share mean is the functional $Q : \mathcal{F} \times \mathbb{Q} \mapsto \mathcal{Y}$ such that*

$$Q(F;\beta) = \frac{1}{\beta} \int_{\underline{\mathcal{Y}}}^{\xi(F;\beta)} y \mathrm{d}F(y). \tag{4.2}$$

Note that the quantile share mean functional is similar to the cumulative income functional in Cowell and Victoria-Feser (2002, Definition 3), but uses a different normalization. Again, the finite population analogue is obtained by plugging in the empirical distribution F_U instead of F and an estimate of the quantile $\xi(F;\beta)$. If $\beta \in \mathbb{Q}_U$, we can express the estimator as $Q(F_U;\beta) = (\beta N)^{-1} \sum_{i=1}^{\beta N} y_{(i)}$, where $y_{(i)}$ denotes the ith order statistics. Alternatively, we may represent the quantile share mean functional as a location L-estimator (i.e., linear function of order statistics) functional, which allows to invoke well-known result of Serfling (1980), which again will be helpful in proving the asymptotic properties of the respective estimators. In the following proposition we state the $Q(F;\beta)$ as an L-functional (adopting the convention that in absence of explicit limits of integration the range is to be taken from $\underline{\mathcal{Y}}$ to $\overline{\mathcal{Y}}$).

Proposition 3. *Let $X \sim F \in \mathcal{F}$ have finite variance then for any $\beta \in \mathbb{Q}$, the βth quantile share mean is the functional $Q : \mathcal{F} \times \mathbb{Q} \mapsto \mathcal{Y}$ and admits the following L-functional representation*

$$Q(F;\beta) = \int x J_q[F(x)] \mathrm{d}F(x), \quad F \in \mathcal{F}, \tag{4.3}$$

where $J_q(t)$ is a weight generating function on $[0,1]$ defined as

$$J_q(t) = \beta^{-1} \mathbb{1}\{t \leq \beta\}. \tag{4.4}$$

Note that the weighting generating function $J_q(t)$ is bounded, continuous (i.e., the set $D := \{x : J_q \text{ is discontinuous at } F(x)\}$ has Lebesgue measure zero), and $J_q(t) = 0$ when $t > \beta$.

For a proof in the case of iid-data see SERFLING (1980, p.279–285). We will prove this representation in a stratified multi-stage cluster sampling scheme in Section 4.5. The Assumption that D has Lebesgue measure zero is not needed to ascertain asymptotic normality of the estimators if we restrict attention to smooth L-statistics with weight generating functions sufficiently smoothly trimmed (STIGLER, 1973). This property is of considerable empirical relevance when sampling is from finite populations with grouped data (e.g., highly clustered populations; see Section 4.4) or when outliers are present in a proportion close to the trimming proportion. Therefore, we may use, for instance, $J_s(u) = h$ if $0 \leq u \leq \beta$, $J_s(u) = h[2(a-\beta)]^{-1}(a-u)$ if $\beta \leq u \leq a$, and zero otherwise, where $h = [\beta + (1/4)(a-\beta)]^{-1}$ (adopting the notation in STIGLER (1973)). The constant a is chosen such that $\beta \leq a \leq 1$.

The functional representation of L-statistics not only helps us to see what an L-estimator is actually estimating, but also brings into action the useful heuristic tool of the influence function (Section 4.3). For finite populations, an estimate is obtained by plugging in the empirical distribution F_U instead of F. Thus, we write

$$Q(F;\beta) = \int x J_q\left(F_U((x))\right) \mathrm{d}F_U\left(x\right) = \frac{1}{N}\sum_{i=1}^{N} J_q\left(\frac{i}{N}\right) X_{(i)}, \tag{4.5}$$

where $J_q(t)$ (or J_s) is defined as before, and $X_{(i)}$ denotes the ith order statistic.

The trimmed sample mean can be expressed as a linear combination of the quantile share means. This is shown in the following definition.

Definition 4 (Trimmed mean)**.** *For any $\alpha, \beta \in \mathbb{Q}$ and $\alpha \neq \beta$, an asymptotically equivalent formulation of the trimmed sample mean as a L-functional $T : \mathcal{F} \times \mathbb{Q} \times \mathbb{Q} \mapsto \mathcal{Y}$ is defined as*

$$T(F;\alpha,\beta) = \int x J_t[F(x)]\mathrm{d}F(x), \qquad F \in \mathcal{F}, \tag{4.6}$$

where $J_t(t)$ is a weight generating function on $[0,1]$ defined as $J_t(t) = (\beta-\alpha)^{-1}\mathbb{1}\{\alpha \leq t \leq \beta\}$.

Similarly, we may use J_s instead of J_t. Moreover, note that the quantile share means and the trimmed sample mean are related accordingly $T(F;\alpha,\beta) = [\beta Q(F;\beta) - \alpha Q(F;\alpha)][\beta-\alpha]^{-1}$.

Let $p, r \in \mathbb{Q}$ be associated with the quantiles ξ_p and ξ_r, respectively, adopting the notation that ξ_p and ξ_r refer to the income quantile share of the $\mathcal{P}$oor (bottom quantile) and the $\mathcal{R}$ich (top quantile). (One usually requires that $p = 1 - r$ where $p, r \in \mathbb{Q}$). Thus, we define the quantile share ratio functional.

Definition 5 (Quantile share ratio)**.** *The quantile share ratio is the functional* $QSR : \mathcal{F} \times \mathbb{Q} \times \mathbb{Q} \mapsto \mathcal{Y}$ *such that*

$$QSR(F; p, r) = \frac{(1-r) \cdot T(F; r, 1)}{p \cdot Q(F; p)} \tag{4.7}$$

Note that in the case of the quintile share ratio we choose $p = 0.2$ and $r = 0.8$. Strictly speaking, the normalization by p and $(1-r)$ is not necessary unless $p \neq 1-r$, but the general ratio formula will be useful in subsequent sections. For the finite population with cdf F_U and $p, r \in \mathbb{Q}_U$, an estimator is given replacing F by F_U.

Estimation for such ratios is, however, not simply a matter of applying standard methods for ratio estimation, since the quantiles must be estimated before estimating the respective shares. This is particularly relevant for robust estimation.

4.3 Robustness properties and data contamination

Estimates of income inequality indicators are known to be sensitive to outlying observations from the tails of the income distribution (Cowell and Victoria-Feser, 1996). The presence of only a few extreme observations can seriously distort the estimate of a statistic. Cowell and Flachaire (2007) showed that informal discussions of the empirical performance do not provide a reliable guide to the way in which the estimators respond to outliers and extreme values.

The principal tool for evaluating the influence of data contamination on estimates is the influence function IF in the theory of robust statistics (Hampel et al., 1986). First, let $y_1, \ldots, y_n$ be realizations of a parametric model F (according to the definition in Section 1) indexed by a parameter vector θ. Second, suppose that there is a small (but not directly observable) contamination at point z in the income distribution. Consider as contaminating distribution the elementary (degenerate) cdf $G(y) := \mathbb{1}\{y \geq z\}$, which has a unit point mass at z and zero elsewhere. As a result, the actually observed (and contaminated) distribution is the mixture distribution $F_\varepsilon(y) := (1-\varepsilon)\, F(y) + \varepsilon G(y)$, where ε captures the importance of the contamination relative to the true distribution. The influence function describes the effect of an infinitesimal contamination, εG, at the point z on the estimator T of θ standardized by the mass of the contamination ε, given the model $F(y)$. It is defined by

$$IF(z, T, F) := \lim_{\varepsilon \to 0} \frac{T\left((1-\varepsilon)\, F + \varepsilon G\right) - T(F)}{\varepsilon}. \tag{4.8}$$

When T is differentiable at F in direction of G (i.e., has a differential in the sense of Gâteaux) we may write $IF(z, T, F) := (\partial/\partial\varepsilon) T(F_\varepsilon)|_{\varepsilon=0}$. In other words, the linear approximation $\varepsilon \cdot IF(z, T, F)$ measures the asymptotic bias of the estimator T caused by a contamination of the relative weight ε at z. If the IF is unbounded for some income

value z it means that the estimate of the inequality index may be catastrophically affected by data contamination at z or a value close to z. Although we introduced the influence function approach to robustness in a strictly parametric setting, it is also useful as a heuristic tool (HAMPEL et al., 1986, 83) in the finite population sampling context HULLIGER (1995, 82).

Next we derive the influence function of the quantile share ratio. It becomes apparent that the estimation procedure constitutes several channels by means of which data contamination may bias the quantile share ratio estimates.

Lemma 6 (Influence function of the quantile share mean)**.** *Let $\beta \in \mathbb{Q}$ and β is not a discontinuity point of F^{-1}. The influence function of the quantile share mean functional $Q(F;\beta)$ (with an exogenously determined β, and weight generating function $J_q(t) = \beta^{-1}\mathbb{1}\{t \leq \beta\}$), is given by*

$$IF(z, Q(\cdot;\beta), F) = \xi(F;\beta) - Q(F;\beta) + \frac{1}{\beta}\mathbb{1}\{z \leq \xi(F;\beta)\}\left[z - \xi(F;\beta)\right]. \tag{4.9}$$

Proof. See Section 4.7. □

The IF is linear increasing and bounded above by $\xi(F;\beta) - Q(F;\beta)$. Note that the influence function for $\beta = 1$ becomes $z - Q(F;1)$, the IF of the arithmetic mean. The influence function of the trimmed mean functional follows directly from Lemma 6.

Corollary 7 (Influence function of the trimmed mean)**.** *Let $\alpha < \beta$, for $\alpha, \beta \in \mathbb{Q}$ (where α and β are not limit points of F^{-1}) be associated with the trimmed mean functional $T(F;\alpha,\beta)$ (with weight generating function J_t), then the influence function is*

$$\begin{aligned} IF(z, T(\cdot;\alpha,\beta), F) \;=\; & (\beta-\alpha)^{-1}[\beta\xi(F;\beta) - \beta Q(F;\beta) + \mathbb{1}\{z \leq \xi(F;\beta)\}\left(z - \xi(F;\beta)\right) - \\ & -\alpha\xi(F;\alpha) + \alpha Q(F;\alpha) - \mathbb{1}\{z \leq \xi(F;\alpha)\}\left(z - \xi(F;\alpha)\right)] \end{aligned} \tag{4.10}$$

Proof. The proof follows from $T(F;\alpha,\beta) = [\beta Q(F;\beta) - \alpha Q(F;\alpha)]/[\beta - \alpha]$ and the influence function of the quantile share mean functional. □

Note that $IF(z, T(\cdot;\alpha,\beta), F)$ for constants $0 < \alpha < \beta < 1$ is bounded, since we can show that $|IF(z, T(\cdot;\alpha,\beta); F)| \leq [\xi(F;\beta) - \xi(F;\alpha)][\beta - \alpha]^{-1}$. On the other hand, the IF becomes unbounded when $\beta = 1$ (given $\alpha < \beta$, $\underline{\mathcal{Y}} = 0$). That is, the influence function becomes

$$\begin{aligned} IF(z, T(\cdot;\alpha,1), F) \;=\; & (1-\alpha)^{-1}[z - \mu(F) - \alpha\xi(F;\alpha) + \alpha Q(F;\alpha) - \\ & -\mathbb{1}\{z \leq \xi(F;\alpha)\}\left(z - \xi(F;\alpha)\right)], \end{aligned} \tag{4.11}$$

where $\mu(F)$ denotes the mean functional. Next, the influence function of the quantile share ratio, $QSR(F;p,r)$, follows immediately.

Lemma 8 (Influence function of the quantile share ratio)**.** *Under the conditions of Lemma 6, the influence function of the QSR functional is*

$$\begin{aligned} IF(z, QSR(\cdot,p,r), F) \;=\; & \kappa[z - \mu(F) - r\xi(F;r) + rQ(F;r) \;- \\ & - \mathbb{1}\{z \le \xi(F;r)\}\,(z - \xi(F;r))] \;- \\ & - \upsilon[p\xi(F;p) - pQ(F;p) + \mathbb{1}\{z \le \xi(F;p)\}] \;\times \\ & \times\; (z - \xi(F;p))\,. \end{aligned} \tag{4.12}$$

where $\kappa = [pQ(F;p)]^{-1}$ *and* $\upsilon = [(1-r)T(F;r,1)][pQ(F;p)]^{-2}$ *are constants.*

Proof. Since the quantile share ratio can be expressed as a simple map in terms of the quantile share mean functionals, the influence function of the QSR follows immediately applying the chain rule. □

From the mathematical display in Lemma 8, we recognize that the influence function of the quantile share ratio is clearly unbounded. That is to say, that the gross error sensitivity of the QSR at F, $\sup_z |IF(z, QSR(\cdot;p,r), F)|$ (the supremum being taken over all z where the IF exists (Hampel et al., 1986, p.87)), indicates an arbitrarily large asymptotic bias when the infinitesimal contamination at z takes an arbitrarily large value. This formally means that a single observation, provided it is sufficiently large, can drive the estimate of the QSR arbitrarily large. The Quintile Share Ratio is as non-robust as the mean and thus a very unreliable estimator. Thus, as for the mean, the bias of the QSR described by the IF refers to non-representative outliers in the sense of Chambers (1986). That is to say, once an extreme value in a sample has been nominated as outlier, the question is whether it is correct or not and whether we should use it in inference on the population or not. Either an outlier may be a correct (but influential) observation from the target population (called representative outlier by Chambers (1986)) or it may be an incorrect observation, for instance, due to coding errors or from an element outside the target population. Discarding a correct observation leads to biased estimates. Keeping it, however, with full weight makes the estimator highly variable because typically the outlier would show up only in a few of the possible samples. Thus there is a trade-off between bias and variance in this case, which is particularly accentuated under asymmetric, heavy-tailed distributions (cf. Fuller, 1991). Obviously, if a representative outlier would tend to infinity then the QSR must tend to infinity, too. Therefore, any estimator of the QSR with a unbounded influence function will finally have an infinitely large bias when an outlier, in fact, is representative. On the other hand, if the outlier is an incorrect observation then keeping it with full weight may entail a large bias in addition to high variability. As a result, discarding or at least downweighting incorrect outliers reduces both bias and variance. Thus, we derive outlier robust estimators of the QSR reflecting the trade-off.

4.3.1 Robust income quantile share ratio estimators

If a known property of the data is presumed to be contaminated trimming these extreme observations (and doing inference conditional on the trimming) is a well-established operation and straight-forward robustification of economic inequality measures (cf. COWELL and VICTORIA-FESER, 2006). The following proposition defines the trimmed quantile share ratio estimator.

Proposition 9 (TQSR). *Let α_u and α_l denote two exogenously determined trimming proportions in $\mathbb{Q}$ such that $0 < \alpha_l \leq p$ and $r < 1 - \alpha_u < 1$, where p and r are the proportions used in the definition of the QSR such that $p = 1 - r$. The estimator of the trimmed quantile share ratio (TQSR) writes*

$$TQSR(\hat{F}; p, r, \alpha_l, \alpha_u) = \frac{(1 - \alpha_u - r) \cdot T(\hat{F}; r, 1 - \alpha_u)}{(p - \alpha_l) \cdot T(\hat{F}; \alpha_l, p)}. \tag{4.13}$$

Note that $TQSR$ trims the proportion α_l (α_u) of the observations at the bottom (top) of the distribution. Thus, the income share means of the poor (denominator) and the rich (numerator) are trimmed accordingly. The influence function of the $TQSR$ is bounded since α_u is chosen according to $r < 1 - \alpha_u < 1$. In addition, this setup guards the QSR to be affected by contamination at both tails (also by arbitrary negative values) of the underlying distribution.

However, this straight-forward robustification by trimming the extreme observations, entails a downward bias (HULLIGER and SCHOCH, 2009) insofar that trimming decreases the numerators, while the denominator is increased. This is obvious since the numerator decreases and the denominator increases under trimming. Therefore, we propose two bias-compensated robust estimators that are based on trimming and a particular correction term.

Proposition 10 (BQSR). *Let α_l and α_u denote two exogenously determined trimming proportions in $\mathbb{Q}$, such that $r < 1 - \alpha_u < 1$, and $0 < \alpha_l \leq p$, where p and r are the proportions used in the definition of the QSR. The bias-compensated, trimmed quantile share ratio (BQSR) writes*

$$BQSR(\hat{F}; p, r, \alpha_l, \alpha_u) = \frac{(1 - \alpha_u - r) \cdot T(\hat{F}; r, 1 - \alpha_u)}{(p - \alpha_l) \cdot Q(\hat{F}; p - \alpha_l)}. \tag{4.14}$$

Note that trimming of both the numerator and denominator compensates at least partially. For a given (small) upper trimming proportion α_u, there is an α_l such that BQSR is unbiased, that is, the identity $BQSR(\hat{F}; p, r, \alpha_l(\alpha_u), \alpha_u) = QSR$ holds. However, because the distribution is unknown and thus α_l cannot be derived from α_u the analyst must choose two predetermined parameters α_u and α_l for BQSR. In Section 4.6 we discuss methods and strategies for an appropriate choice.

From a practitioners perspective, the BQSR estimator features the disadvantage of choosing two tuning constants separately, instead of a single one. A promising approach that circumvents choosing two (or more) tuning constants, is based on following heuristic argument. Observe that trimming and compensation operate in general at different locations of the income distribution. The density estimates at (or in the neighborhood of) these locations differ strongly. Recall that the income share mean (QSM) of the poor writes $Q(F,p)$ (with $0<p<r$). Let $\Delta_p^{p-a}=Q(F;p)-Q(F,p-a)$ denote the difference in QSM when the range of the poor is reduced by a with $0<a<p$. Likewise we denote by Δ_r^{r-a} the difference in QSM that results when the range of the rich is reduced (from above) by a. In this respect, it is trivial to see that for a given a, Δ_p^{p-a} and Δ_r^{r-a} are not equal because of the skewed overall income distribution. By how far they differ, depends on the characteristics of the particular distribution. Thus, a reasonable BQSR-compensation strategy should try to balance Δ_p^{p-a} and Δ_r^{r-a} that results from a given choice a (trough trimming/compensation), at least partially. A promising approach is to let the compensation depend on the ratio of skewness of the lower versus the skewness upper tail of the distribution. We shall consider the simple and robust estimator of the skewness ratio, S, given by

$$S(F)=\frac{\xi(F;0.95)-\xi(F;0.9)}{\xi(F;0.95)-\xi(F;0.85)}\times\frac{\xi(F;0.15)-\xi(F;0.05)}{\xi(F;0.15)-\xi(F;0.1)}, \tag{4.15}$$

where $\xi(F,p)$ denotes the pth quantile. One may choose different quantiles; however, the reported choice of quantiles gives good result for moderate outlier scenarios. The following Proposition summarizes the skewness-balanced QSR estimator (SQSR).

Proposition 11 (SQSR)**.** *Let α_u denote an exogenously determined trimming proportion in $\mathbb{Q}$, such that $r<1-\alpha_u<1$ where r is the proportions used in the definition of the QSR. The amount of compensation is defined as*

$$\alpha_l=S(\hat{F})\cdot\alpha_u, \tag{4.16}$$

where $S(\hat{F})$ is the estimated skewness-ratio. The skewness compensated quantile share ratio estimator (SQSR) writes

$$SQSR(\hat{F};\alpha_l,p,r,\alpha_u)=TQSR(\hat{F};p,r,\alpha_l,\alpha_u) \tag{4.17}$$

Observe that the SQSR estimator is essentially a TQSR-type estimator, except that α_l is computed implicitly (i.e., data-based) instead of being user-supplied. Moreover, the amount of compensation is regulated by the skewness ratio and the chosen upper trimming constant, α_u. It is important to note (also for reasons of consistency) that SQSR coincides with QSR for the choice $\alpha_u=0$ (i.e., no trimming).

4.4 Estimation with Complex Survey Data

We emphasize that although the parameters of interest have been motivated by infinite population concepts, we are only concerned with the design-based sampling distribution of the estimators. For the purpose of finite sampling inference, the observations are to be taken as fixed and we are interested in estimating the finite population characteristic $Q(F_U)$. The connection to the concepts in Section 4.2 is that the N units of the finite population $U = \{1, \ldots, i, \ldots, N\}$ are thought as distinct realizations y_i of the characteristic Y with cdf F_U. In other words, we are interested in estimating, for instance, $QSR(F_U; p, r)$ for an unknown F.

4.4.1 Sampling Design

We consider in this paper a general stratified multi-stage cluster sampling design, which has been applied in most EU-SILC-countries.[1] Let the population U under consideration be stratified into L strata ($h = \{1, \ldots, L\}$) and the hth stratum contains N_h primary sampling units (PSU; or clusters $i = \{1, \ldots, N_h\}$). Furthermore, we assume that the design involves a relatively large number of strata with comparatively few PSU's. In the majority of country-specific EU-SILC sampling designs, the PSU's are counties or census areas that have been stratified by the degree of urbanization, some socio-economic criteria or geographical variables, and thus meet this assumption fairly well.[2] The (h, i)th PSU contains N_{hi} secondary sampling units (SSU) or clusters where $j = \{1, \ldots, N_{hi}\}$. Typically, households embody the SSU's. Associated with the kth ultimate unit (e.g., household member) in the (h, i, j)th SSU of the (h, i)th PSU of stratum h is a characteristic Y_{hijk} (with $h = 1, \ldots, L; i = 1, \ldots, N_h; j = 1, \ldots, N_{hi}; k = 1, \ldots, N_{hij}$). Here N_{hij} denotes the number of ultimate units in the (h, i, j)th SSU (e.g., number of persons in the household j). The finite population cdf $F_U(x)$ is

$$F_U(x) = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{N_{hi}} \sum_{k=1}^{N_{hij}} \mathbb{1}\{Y_{hijk} \leq x\}, \quad \forall x \in \mathbb{R}, \tag{4.18}$$

where $N = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j}^{N_{hi}} N_{hij}$.

Following standard survey sampling theory, we assume that estimation is based on a random sample s of size n with $0 < n < N$ from the finite population U. According to SÄRNDAL et al. (1992), let $p_r(s)$ be the sampling design for random sampling at the rth stage, where the sampling design may be any of the conventional designs. Each sampling

[1] In the 2004/05 SILC exercise, the EU member states Belgium, Czech Republic, France, Greece, Hungary, Italy, Ireland, Latvia, Poland, Portugal, Spain, and the United Kingdom applied a multi-stage random sample. The countries Denmark, Finland, Iceland, Norway, Slovenia, Sweden, and The Netherlands used population registers with income information to draw the samples. In the remaining 8 EU member states, the design consists of an indirect sampling of addresses (EUROSTAT, 2008).

[2] The SILC sampling design of the United Kingdom may serve as a showcase since stratification has been done on (24 regions) × (4 socio-economic levels of the head of the household) × (car ownership indicator).

design $p_r(s)$ for $r = 1, 2, \ldots$ is a function defined on the space of samples $\mathcal{S}$ that induces a probability distribution on the set of samples under the sampling scheme in use; see Münnich et al. (2011) for more details. In particular, suppose that a particular first-stage design, $p_1(s)$, has been fixed. Thus the first-order sample-inclusion probability that, say, element a is included in the first-stage sample s_1 is obtained from $p_1(s_1)$ as follows: $\pi_{1a} = \Pr(a \in s_1) = \sum_{a:a\in s_1} p_1(s_1)$ (see e.g. Särndal et al., 1992, p30-32).

In the matter at hand, we suppose that in the first-stage sampling $n_h \geq 2$ PSU's are selected (independently across strata) from stratum h with probability $p_{hi} > 0$, $i = 1, \ldots, N_h$; $h = 1, \ldots, L$ such that $\sum_{i=1}^{N_h} p_{hi} = 1$ (cf. Krewski and Rao (1981), Rao and Wu (1985), and Shao (1994) for a similar setting). For equal probability sampling within the strata, selection can be either with or without replacement (whereas variance estimation can be extremely heavy in designs without replacement because of the calculation of the second-order inclusion probabilities). In the case of unequal probability sampling, we assume that the clusters are selected independently across the strata by means of a with-replacement scheme (PSU's sampled more than once are independently sub-sampled as many times as they occur). The proposed stratified multi-stage (cluster) sampling scheme is very general and comprehends a series of simpler designs.

In particular, random sampling at the first stage is carried out as stratified random sampling of the PSU's (without replacement) according to the design $p_1(s)$. Since we assume selection of PSU's to be independent across strata, the sample inclusion probability of the (h, i)th PSU is defined as $\pi_{hi} = \pi_h \pi_{i|h}$, where $\pi_{i|h}$ denotes the sample inclusion probability of unit i in the hth stratum. For the without-replacement probability sample of SSU's within PSU's according to sampling design $p_2(s)$, we define similarly $\pi_{j|hi}$ as the conditional probability that SSU j is selected given (i.e., within) PSU i in stratum h. While assuming invariance and independence of the sampling stages subsequent to the first stage, the sample inclusion probability of the (h, i, j, k)th ultimate unit is $\pi_{hijk} = \pi_h \pi_{i|h} \pi_{j|hi} \pi_{k|hij}$ (Särndal et al., 1992, 144–146). Consequently, a design-based Horvitz-Thompson (HT) type estimator of the sample distribution function writes

$$\hat{F}(x) = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} \sum_{k=1}^{N_{hij}} w_{hijk} \mathbb{1}\{y_{hijk} \leq x\}, \quad \forall x \in \mathbb{R}, \tag{4.19}$$

where $w_{hijk} = 1/\pi_{hijk}$ denotes the inverse-probability weight associated with the (h, i, j, k)th ultimate unit in the sample. This choice of weights ensures that $\mathbb{E}[\hat{F}(x)] = F(x)$ for any x (where expectation is w.r.t. the design). The inverse-probability weights (typically adapted according to a calibration or raking procedure (cf. Deville and Särndal, 1992)) are routinely included in survey sampling data files released to analysts.

Note that the last sum in Eq. (4.19) is over all N_{hij} elements, since subsampling of the households is assumed to be exhaustive. If the population size N is unknown it can be replaced by a design-unbiased estimate, $\hat{N} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} \sum_{k=1}^{N_{hij}} w_{hijk}$, in Eq. (4.19), which yields a Hájek-type estimator. In the subsequent definitions we suppose the population size N to be known. This assumption does not affect the generalization of the proposed estimators, but simplifies notation considerably. Moreover, $\hat{F}(x)$ may not be a distribution function because $\hat{F}(\infty)$ is not necessarily equal to 1. Thus, one may

use the normalization $\hat{F}(x) = \hat{F}(x)/\hat{F}(\infty)$. Next, we give two illustrative (rather simple) sampling design examples of the well-established SILC exercise.

Example (2004 SILC Ireland)
*The sampling design of the Irish SILC in 2004 consists of a stratified simple random sample of blocks (geographical areas; PSU's) in the first stage, whereas stratification is according to $L = 8$ population density stratum groups (cities, suburbs, ..., rural areas). The second-stage involves a simple random sample of households (*CENTRAL STATISTICS OFFICE IRELAND*, 2005; EUROSTAT, 2008). Let s_1 denote the first-stage sample of number of blocks n_h, selected among all N_h blocks. Similarly, let n_{hi} be the number of household in the second-stage sample s_2. The HT estimator of the distribution function is*

$$\hat{F}(x) = \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{s_1} \frac{N_{hi}}{n_{hi}} \sum_{s_2} \frac{N_{hij}}{n_{hij}} \sum_{k=1}^{N_{hij}} \mathbb{1}\{y_{hijk} \leq x\}, \quad \forall x, \tag{4.20}$$

where $N = \sum_{h=1}^{L} N_h/n_h \sum_{s_1} N_{hi}/n_{hi} \sum_{s_2} n_{hijk}$. Alternatively, we may write the weights

$$w_{hijk} = \frac{1}{\pi_h \pi_{i|h} \pi_{j|hi} \pi_{k|hij}} = \frac{N_h N_{hi} N_{hij}}{n_h n_{hi} n_{hij}}, \tag{4.21}$$

and apply the HT estimator in Eq. (4.19). Note that usually $n_{hij} = N_{hij}$ because all eligible household members are selected, but the explicit notation with n_{hij} establishes a customary first non-response adjustment in the case not all household members could be contacted.

Another example is the EU-SILC sampling design in Switzerland in 2007. The Swiss design is considerably simpler than the general SILC multi-stage design.

Example (2007 SILC Switzerland)
The Swiss SILC sampling design in 2007 consists of a stratification along geographical regions (NUTS2 level; $L = 7$ strata), where the stratum size is proportional to the number of households within the respective stratum. The HT estimator of the distribution function writes

$$\hat{F}(x) = \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_{hi}} \frac{N_{hi}}{n_{hi}} \sum_{k=1}^{N_{hij}} \mathbb{1}\{y_{hij} \leq x\}, \quad \forall x, \tag{4.22}$$

where $N = \sum_{h=1}^{L} (N_h/n_h) \sum_{i=1}^{n_{hi}} (N_{hi}/n_{hi}) N_{hij}$.

4.4.2 Asymptotic framework

In order to make the derivation of distribution-free, asymptotic properties of the quantile share ratio estimates feasible, we impose regularity conditions on the sampling design (e.g., rule out zero variance in case of cluster sampling). The general framework for

the development of the asymptotic theory is provided by the concept of a sequence of finite populations $\{U\}_{L=1}^{\infty}$ with L strata. Consequently, the distribution function from Eq. (4.18) would actually be denoted by $F_L(x)$ and it's estimator $\hat{F}_L(x)$ for any L. For ease of notation, the population index L will be suppressed in what follows. In order to avoid unnecessary repetition, all limiting processes will be understood to be as $L \to \infty$. We always assume that $n \to \infty$ as $L \to \infty$ (where $n = \sum_{h=1}^{L} n_h$ is the number of sampled PSU's). In other words, the asymptotic framework is based on large numbers of strata and modest rates of sampling of PSU's within strata. An alternative asymptotic setup requires that the strata be fixed and all limiting results shall be obtained as the number of sampled PSU's within each stratum and the stratum sizes tend to ∞ (cf. Wolter, 2007). In order to make the asymptotic treatment feasible, we impose meaningful and workable regularity conditions on the weights w_{hijk}, and the sample sizes n at the different stages. The regularity conditions are based on Krewski and Rao (1981, p.1014), with modifications similarly to those in Rao and Wu (1985, p.621).

Assumption 1
Let the stratified multi-stage sampling design be such that the following conditions hold (all limiting processes are understood to be as $L \to \infty$):

$$\textit{each } \max_{h \le L}(n_h),\ \max_{i \le N_h}(n_{hi}) \textit{ and } \max_{j \le N_{hi}}(n_{hij}) \textit{ is } O(1), \tag{A1}$$

$$\max_{h \le L; i \le N_h; j \le N_{hi}} w_{hijk} \textit{ is } O(L^{-1}). \tag{A2}$$

In the case of the Irish SILC sampling design, one may substitute assumption (A1) and (A2) for the more general (combined) assumption (A1'): $\max_{h \le L; i \le N_h; j \le N_{hi}} w_{hijk} N_h N_{hi} N_{hij} / (n_h n_{hi} n_{hij})$ is $O(L^{-1})$; see also Shao (1994) for similar conditions. In general, assumption (A1) reflects the intention to focus on surveys with large numbers of strata and relatively few PSU's selected within each stratum, and requires additionally the allocation of strata, PSU's and SSU's to be bounded. Assumption (A2) means that no single stratum has disproportionate size. Note that, for instance, in the case of the Swiss SILC sampling design (see Example 2), a slightly more general setup is obtained while relaxing (A1) and requiring $max_h(nw_h)/n_h = O(1)$ instead. This condition allows for a trade-off between restrictions on L and n, and roughly says that the allocation of sample across strata should not be disproportionately small relative to the stratum weights. Under these conditions an approriate law of large numbers and a central limit theorem for the weighted mean exist (cf. Krewski and Rao, 1981, 1013-1015). By the assumptions and using

$$\begin{aligned} \mathrm{Var}(\hat{F}(x)) &\le \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} \mathbb{E} \left(\sum_{k=1}^{N_{hij}} w_{hijk} \mathbb{1}\{y_{hijk} \le x\} \right)^2 \\ &\le \frac{1}{N} \mathbb{E} \left(\sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} N_{hij} \sum_{k=1}^{N_{hij}} w_{hijk}^2 \right) \le \max_{1 \le h \le L} w_{hijk} \end{aligned} \tag{4.23}$$

we obtain (by standard randomization inference arguments and application of Jensen's and the Cauchy-Schwarz inequality) that, for any x, $\hat{F}(x) - F(x) \to_p 0$ (note that the index L is suppressed).

4.4.3 Finite population estimates

In this section we derive design-based estimators of the three functionals for the stratified multi-stage cluster sampling design. By means of the finite population cdf $\hat{F}$, a HT of the βth quantile is obtained as the solution, $\hat{\xi}(\hat{F};\beta)$, to the sample estimating function $\hat{u}(F_s, \xi(\cdot;\beta))$ (cf. BINDER and PATAK, 1994), such that

$$\hat{u}(\hat{F}, \xi(\cdot;\beta)) = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{k=1}^{N_{hij}} w_{hijk}\mathbb{1}\{y_{hijk} \leq \xi_\beta\} - \beta = 0. \tag{4.24}$$

Next, recall that the quantile share mean has been defined as a smooth L-functional. Therefore, we denote $\{y_{(l)},\ l = 1,\ldots,n\}$ the lth order statistic of the sample $\{y_{hijk},\ h = 1,\ldots,L; i = 1,\ldots,n_h; j = 1,\ldots,n_{hi}, k = 1,\ldots,n_{hij}\}$ and $w_l = w_{hijk}$ if $y_{(l)} = y_{hijk}$. Then

$$Q(\hat{F},\beta) = \int x J_q(\hat{F}(x))\mathrm{d}\hat{F}(x) = \frac{1}{N}\sum_{l=1}^{n} c_l y_{(l)}, \quad \text{with } c_l = w_l J_q\left(\frac{\sum_{t=1}^{l} w_t}{N}\right), \tag{4.25}$$

is an L-estimator (cf. SHAO, 1994, p.949), with the weight generating function $J_q(t)$. Similarly, an estimate of $T(F;\alpha,\beta)$ is obtained using J_t instead.

Bringing altogether, a Horvitz-Thompson type design-based quantile share ratio estimator writes

$$QSR(\hat{F};p,r) = \frac{(1-r)T(\hat{F};r,1)}{p\cdot Q(\hat{F};p)}. \tag{4.26}$$

4.5 Variance estimation

In this section we show that the estimators, QSR, BQSR, TQSR, and SQSR, are asymptotically normal and their linearization variance can be estimated by substituting the unknown quantities in the formula of the asymptotic variance by a design-unbiased HT-type estimator.

The relevance of the IF to the present analysis has been to study the robustness properties. On the other hand, it may be invoked to derive the limiting distribution of the estimators. First, let $T : \mathcal{F}\times[0,1] \mapsto \mathcal{Y}$ be some vector-valued statistical functional of the c.d.f $F \in \mathcal{F}$.

Let the distribution G be *near* (in some topological sense) F, then the first-order von Mises expansion of the particular functional T of G around F (assuming some differentiability properties of the functional) is given by

$$T(G) = T(F) + \int IF(z, T(\cdot), F)\mathrm{d}[G - F](z) + R(F, G), \tag{4.27}$$

where $R(F, G)$ denotes a remainder term. For sufficiently large n, we may replace G by F_U and obtain $\sqrt{n}(T(F_U) - T(F)) = n^{-1/2}\sum_{i=1}^{n} IF(z_i, T(\cdot), F) + \sqrt{n}R(F, F_U)$. When the remainder term, $\sqrt{n}R_n$, becomes negligible as $n \to \infty$, then $\sqrt{n}(T(F_U) - T(F))$ is asymptotically normal with asymptotic covariance matrix $\int [IF(z, T(\cdot), F)]'[IF(z, T(\cdot), F)]\mathrm{d}F(z)$. For the above treatment to be valid, the remainder term needs to be asymptotically negligible in an appropriate sense, which follows from the mode of stochastic differentiability of T at F, and the convergence of $F_U \to F$ (see e.g., Serfling, 1980, 219–221). In order to make the above treatment rigorous, we proof that the remainder $R(F, F_U)$ is negligible in an appropriate asymptotic setting for stratified multi-stage cluster sampling designs (cf. Shao, 1994).

Next, we shall derive asymptotic expressions of the functionals. The derivation of these expressions make much use of the following Lemma, where we establish the asymptotic normality of the quantile share means for general stratified multi-stage cluster sampling designs.

Lemma 12 (Asymptotic normality of the quantile share mean functional)**.** *Let* $\underline{\mathcal{Y}} = 0$ *(i.e., lower bound of the support of* Y*). Suppose the* p*-dimensional random vector*

$$\hat{\boldsymbol{\theta}} = \left[Q(\hat{F}; \beta_1),\ Q(\hat{F}; \beta_2), \ldots, Q(\hat{F}; \beta_p)\right]'$$

where $Q(\hat{F}; \beta_t)$ *are quantile share mean functionals* $(\forall \beta_t, t = 1, \ldots, p : 0 < \beta_t < 1$ *and* β_t *is not a limit point of* $F^{-1})$*. Suppose that assumptions (A1) and (A2) hold. Assume in addition that the* $Q(F; \beta_t)$ *functionals have strictly positive variance:*

$$\lim_{L\to\infty} \inf_{L} n\sigma^2(Q(\cdot; \beta_t), F) > 0, \quad \forall \beta_t, t = 1, \ldots, p, \tag{A5}$$

where

$$\sigma^2(Q(\cdot; \beta_t), F) = \mathbb{V}\left(\frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{k=1}^{N_{hij}} w_{hijk} IF(y, Q(\cdot; \beta_t), \hat{F})\right). \tag{4.28}$$

Then $\hat{\boldsymbol{\theta}}$ *is asymptotically normal in that* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ *has a limiting* p*-variate normal distribution with mean zero (i.e.,* $(p \times 1)$ *vector of zeros) and covariance matrix* $\boldsymbol{\Omega}$*, whose* i, j*th element (for* $i, j = \{1, \ldots, p\}$*) is*

$$\omega_{\beta_i,\beta_j} = \frac{1}{\beta_i\beta_j}S(\beta_i,F)+\xi_{\beta_i}Q(\beta_i,F)+\xi_{\beta_j}Q(\beta_j,F)-\frac{1}{\beta_j}Q(\beta_i,F)\left(\xi_{\beta_j}+\xi_{\beta_i}\right)+ \\ -Q(\beta_i,F)Q(\beta_j,F)+\xi_{\beta_i}\xi_{\beta_j}\left(\frac{1}{\beta_j}-1\right), \qquad \textit{for } i\leq j. \tag{4.29}$$

where ω_{β_i,β_j} *is a short-hand notation for* $\omega_{Q(F;\beta_i),Q(F;\beta_j)}$*, and* $S(\beta_i,F):=\int^{\xi(F;\beta_i)}y^2\mathrm{d}F(y)$ *denotes the variance of* Y *conditional on* $Y\leq\xi(F;\beta_i)$*.*

Proof. See Section 4.7. □

Remark.

i) The results of Lemma 12 remain valid under more general conditions (i.e., in the case of untrimmed smooth L-statistics), with either

(a) $0<\beta\leq 1$ *when* $\sup_L\int x\mathrm{d}F(x)<\infty$ *or*

(b) $0<\beta\leq 1$ *and* $\underline{\mathcal{Y}}=-\infty$ *when* $\sup_L\int|x|\mathrm{d}F(x)<\infty$*,*

given that the Liapounov-type moment condition holds: there is a $\delta>0$ *such that* $[n]^{1+\delta}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\mathbb{E}|u_{hij}-\mathbb{E}u_{hij}|^{2(1+\delta)}$ *is* $O(1)$*, where*

$$u_{hij}=(1/N)\sum_{k=1}^{N_{hij}}w_{hijk}IF(y_{hijk},Q(\cdot;\beta),\hat{F})$$

see also SHAO *(1994).*

ii) Note that in the case of the smoothly trimmed weight generating function, J_s*, Lemma 12 remains valid, but the asymptotic variance may have no closed form.*

iii) The linearization variance can be estimated by substituting the unknown quantities in the formula of the asymptotic variance (see Eq. (4.29)) by some design-unbiased HT type estimator for a given sampling design.

Next, we introduce a lemma on the limiting distribution of the trimmed mean functional. This lemma will prove useful to derive the asymptotic variance of TQSR, BQSR, and SQSR estimators. In particular, we approach the the main results of this paper in step-by-step manner.

Lemma 13 (Asymptotic normality of the trimmed mean functional)**.** *Let* $\alpha,\beta,\alpha',\beta'\in\mathbb{Q}$ *such that* $0\leq\alpha<\beta<\alpha'<\beta'<1$ *be associated with the trimmed mean functionals* $T(F;\alpha,\beta)$ *and* $T(F;\alpha',\beta')$ *where* α*,* β*,* α'*, and* β' *are not limit points of* F^{-1}*. Suppose that the assumptions of Lemma 12 hold, then the covariance of* $\sqrt{n}T(F;\alpha,\beta)$ *and* $\sqrt{n}T(F;\alpha',\beta')$ *is*

$$\Psi_{T(F;\alpha,\beta),T(F;\alpha',\beta')}=[(\beta-\alpha)(\beta'-\alpha')]^{-1}(\beta\beta'\omega_{\beta,\beta'}-\beta\alpha'\omega_{\beta,\alpha'}-\alpha\beta'\omega_{\alpha,\beta'}+\alpha\alpha'\omega_{\alpha,\alpha'}) \tag{4.30}$$

where $\omega_{i,j}$ is the covariance of $\sqrt{n}Q(F;i)$ and $\sqrt{n}Q(F;j)$ for $0 < i \leq j < 1$, as defined in Lemma 12. Moreover, $\sqrt{n}(T(\hat{F};\alpha,\beta) - T(F;\alpha,\beta))$ has a limiting normal distribution with mean zero and variance

$$\Phi_{T(F;\alpha,\beta)} = (\beta-\alpha)^{-2}\left(\beta^2\omega_{\beta,\beta} - 2\alpha\beta\omega_{\alpha,\beta} + \alpha^2\omega_{\alpha,\alpha}\right), \tag{4.31}$$

where $\omega_{i,j}$ is the covariance of $\sqrt{n}Q(F;i)$ and $\sqrt{n}Q(F;j)$ for $0 < i \leq j < 1$.

Proof. Note that $T(F;\alpha,\beta)$ can be expressed as a linear combination of quantile share functionals, i.e., $T(F;\alpha,\beta) = [\beta Q(F;\beta) - \alpha Q(F;\alpha)][\beta-\alpha]^{-1}$; and analogously for $T(F;\alpha',\beta')$. The covariance is given by $\int IF(z,T(\cdot;\alpha,\beta))IF(z,T(\cdot;\alpha',\beta'))\mathrm{d}F(z)$. Expanding the integral while using the fact that the influence function $IF(z,T(\cdot))$ is also a linear combination of the influence functions of $Q(\cdot)$, the first assertion follows by application of Lemma 12. The second assertion follows immediately, setting $\alpha' = \alpha$ and $\beta' = \beta$. □

By means of Lemma 12 and standard asymptotic arguments, we establish the asymptotic normality of TQSR.

Theorem 14 (Asymptotic normality of the TQSR)**.** *Let $\alpha_l, \alpha_u, r, p \in \mathbb{Q}$ such that $0 \leq \alpha_l < p < r \leq 1-\alpha_u$ be associated with the trimmed mean functionals $T(F;\alpha_l,p)$ and $T(F;r,1-\alpha_u)$, where α_l, α_u, r, and p are not limiting points of F^{-1}. Suppose that the assumptions of Lemma 12 hold, then $\sqrt{n}(TQSR(\hat{F};\alpha_l,p,r,\alpha_u) - TQSR(F;\alpha_l,p,r,\alpha_u))$ has a limiting normal distribution with mean zero and variance*

$$\sigma^2(TQSR(\cdot;\alpha_l,p,r,\alpha_u),F) = \kappa^{-4}[\kappa^2\Phi_{T(F;r,1-\alpha_u)} - 2\eta\kappa\Psi_{T(F;\alpha_l,p),T(F;r,1-\alpha_u)} + \eta^2\Phi_{T(F;\alpha_l,p)}] \tag{4.32}$$

where $\kappa = (1-\alpha_u-r)(p-\alpha_l)^{-1}T(F;\alpha_l,p)$ and $\eta = (1-\alpha_u-r)(p-\alpha_l)^{-1}T(F;r,1-\alpha_l)$ are constants, $\Psi_{T(F;\alpha_l,p),T(F;r,1-\alpha_u)}$ denotes the covariance of $\sqrt{n}T(F;\alpha_l,p)$ and $\sqrt{n}T(F;r,1-\alpha_u)$ according to Lemma 13. $\Phi_{T(F;\alpha_l,p)}$ and $\Phi_{T(F;r,1-\alpha_u)}$ denote the limiting variance of $\sqrt{n}T(F;\alpha_l,p)$ and $\sqrt{n}T(F;e,\alpha_u)$ from Lemma 13, respectively.

Proof. The assertion follows by application of the (functional) delta theorem and Lemma 13. □

By means of Theorem 14 the asymptotic normality of the bias-compensated trimmed quantile share ratio (BQSR) follows immediately.

Theorem 15 (Asymptotic normality of the BQSR functional)**.** *Let $\alpha_l, \alpha_u, r, p \in \mathbb{Q}$ such that $0 \leq \alpha_l < p < r \leq 1-\alpha_u$ be associated with the functionals $Q(F;p-\alpha_l)$ and $T(F;r,1-\alpha_u)$, where α_l, α_u, r, and p are not limiting points of F^{-1}. Suppose that the*

assumptions of Theorem 14 hold, then $\sqrt{n}(BQSR(\hat{F};\alpha_l,p,r,\alpha_u)-BQSR(F;\alpha_l,p,r,\alpha_u))$ *has a limiting normal distribution with mean zero and variance*

$$\sigma^2(SQSR(\cdot;\alpha_l,p,r,\alpha_u),F)=\kappa^{-4}[\kappa^2\Phi_{T(F;r,1-\alpha_u)}-2\eta\kappa\Psi_{T(F;0,p-\alpha_l),T(F;r,1-\alpha_u)}+\eta^2\Phi_{T(F;0,p-\alpha_l)}], \tag{4.33}$$

where $\kappa=(1-\alpha_u-r)(p-\alpha_l)^{-1}T(F;0,p-\alpha_l)$ *and* $\eta=(1-\alpha_u-r)(p-\alpha_l)^{-1}T(F;r,1-\alpha_l)$, *and* $\Phi_{T(F;r,1-\alpha_u)}$, $\Phi_{T(F;0,p-\alpha_l)}$, *and* $\Psi_{T(F;0,p-\alpha_l),T(F;r,1-\alpha_u)}$ *are defined in Lemma 13.*

Proof. The assertion follows from Theorem 14, since we can express BQSR as a variant of TQSR, i.e., $BQSR(F;\alpha_l,p,r,\alpha_u)\equiv TQSR(F;0,p-\alpha_l,r,\alpha_u)$. □

Corollary 16 (Asymptotic normality of the SQSR)**.** *Suppose* $S(F)$ *is a predetermined (or known) estimate of the skewness ratio. Denote by* $\alpha_l=\alpha_u S(F)$ *the compensation proportion, where* $\alpha_u,r,p\in\mathbb{Q}$ *are defined such that* $0\leq\alpha_l<p<r\leq 1-\alpha_u$ *holds. Associated with* α_l, p, r, α_u *are functionals* $Q(F;p-\alpha_l)$ *and* $T(F;r,1-\alpha_u)$, *assuming that neither* α_l, α_u, r, *or* p *is a limiting point of* F^{-1}. *Suppose that the assumptions of Theorem 14 hold, then* $\sqrt{n}(SQSR(\hat{F};\alpha_l,p,r,\alpha_u)-SQSR(F;\alpha_l,p,r,\alpha_u))$ *has the same limiting normal distribution as BQSR in Theorem 15.*

Proof. The assertion follows from Theorem 14, using the same arguments as in Theorem 15. □

Note that the limiting normal distribution of SQSR (Corollary 16) is derived under the assumption that the skewness ratio, $S(\hat{F})$, is known (or predetermined). As a result, the variance estimate is an underestimate of the true variance when $S(F)$ is indeed unknown. One may explicitly consider the contribution of estimating $S(F)$ on the variance of SQSR, however, this may make the overall variance estimate unstable because the variance contribution of $S(F)$ depends on density estimates in both tails of the distribution. Furthermore, and because the contribution to the overall variance is small, we ignore it.

The computation of the Horvitz-Thompson variance estimates in the case of general stratified multi-stage cluster sampling design, with a unequal probability without-replacement sampling scheme at the first stage, may become cumbersome or may be even infeasible because the second-order sample inclusion probabilities are not routinely included in the data files released to analysts. There exist numerous approximations for the second-order inclusion probabilities; see MÜNNICH et al. (2011) for details. However, the two illustrative SILC sampling designs may serve as showcase because the variance estimates are obtained immediately.

Example (Swiss SILC, ctd.)
The linearization variance estimate for the quantile share mean functional $Q(\hat{F},\beta)$ *with*

infuence function $z_{hi} = IF(y_{hi}, Q(\cdot;\beta), F)$ *writes in the case of the Swiss SILC sampling design (assuming the conditions of Lemma 12 hold)*

$$\hat{\mathbb{V}}(Q(\cdot;\beta),\hat{F}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2}{n_h(n_h-1)}\left(1-\frac{n_h}{N_h}\right)\sum_{i=1}^{n_{hi}}\left(z_{hi}-\hat{\phi}_h\right)^2, \tag{4.34}$$

where $\hat{\phi}_h = (1/n_h)\sum_{i=1}^{n_h} z_{hi}$*=0. (supressing the subscript for the household membership because all eglible household members have been contacted).*

For the sampling design of the 2004 SILC exercise in Ireland, we obtain a similar variance estimator. Note that as a consequence of the invariance and independence conditions of the sampling stages subsequent to the first stage the additional term (in comparison to the Swiss design) is additively related, which simplifies matters.

Example (Irish SILC, ctd.)
For the Irish SILC 2004 sampling design, we may use the following variance estimate of $Q(\hat{F};\beta)$ *for sufficiently large* n *(under the conditions of Lemma 12),*

$$\begin{aligned}\hat{\mathbb{V}}(Q(\cdot;\beta),\hat{F}) \;=\; & \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2}{n_h(n_h-1)}\left(1-\frac{n_h}{N_h}\right)\sum_{i=1}^{n_h}\left(z_{hi}-\hat{\phi}_h\right)^2 + \\ & + \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{j=i}^{n_h}\left(1-\frac{n_{hi}}{N_{hi}}\right)N_{hi}^2\left(\frac{1}{n_{hi}(n_{hi}-1)}\right)\sum_{j=1}^{n_{hi}}\left(z_{hij}-\hat{\phi}_{hi}\right)^2,\end{aligned} \tag{4.35}$$

where $\hat{\phi}_i = \hat{\phi}_{hi} = 0$.

On the other hand, if the design consists of sampling without replacement and unequal selection probabilites (e.g., 2004 SILC exercise in Belgium), we examine a variance estimation strategy appropriate for sampling with replacement (Hansen-Hurwitz strategy), which is the only feasible technique in the case of the released SILC data, but somewhat biased (cf. Särndal et al., 1992; Wolter, 2007). That is, we compute the variance estimator as if sampling had been done with replacement, whereas in actual fact it was without replacement. Generally, these variance estimates entail an upward bias (i.e., conservative confidence intervals); see also Münnich et al. (2011, Section 2.2).

4.6 Adaptive estimation

In empirical data analysis, we hardly face situations where the amount of outlyingness is known beforehand, such that the choice of the tuning constants for robustification is obvious. However, the trimming proportion can be chosen adaptively. According to Jaeckel (1971), the trimming proportion can be selected as follows. Consider a characteristic $S_U(F;\alpha,\beta)$ of the distribution of $T(\hat{F};\alpha,\beta)$, usually a measure of spread (e.g., asymptotic variance). For convenience, we assume that the trimming proportion α is known

beforehand (i.e., nuisance). The idea consist of estimating $S_U(F;\alpha,\beta)$ for all values β in a given set and choosing $\beta_U(\hat{F}(x))$ corresponding to the smallest estimate. Therefore, $\beta_U(\hat{F}) = \arg\min_{\beta\in B} S_U(\hat{F};\alpha,\beta)$, where B is some set of trimming proportions. The adaptive trimmed-mean estimate of location is $\hat{T}(\hat{F};\alpha,\hat{\beta}(\hat{F}))$. In the matter at hand, the underlying distribution is not symmetric and, thus, the location is no longer unambiguously defined (LÉGER and ROMANO, 1990). As a result, not only does the adaptive trimmed mean choose the trimming proportion, but in so doing, it also chooses the functional being estimated. Instead of e.g., the asymptotic variance, we propose to adapt the estimate according to the minimum estimated risk (MER) (cf. HULLIGER, 1991, 1995). Typically, the estimated mean squared error (MSE) can be used. Thus, the MER idea consists of estimating the MSE

$$r(\hat{F},\beta) = \max(\hat{\mathbb{V}}(\hat{T},\beta),0) + [\hat{T}(\hat{F};r,1) - \hat{T}(\hat{F};r,\beta)]^2 \tag{4.36}$$

of the functionals $T(\hat{F},\cdot)$ for all values β in the given set $B \subset [\alpha,1]$ (note that including 1 in B – i.e., admitting the non-trimmed and therefore non-robust mean as a candidate – ensures consistency of the HT). $\hat{\mathbb{V}}(\hat{T},\beta)$ denotes the variance estimator, where $\max(\cdot)$ safeguards the variance estimate from being negative. This estimate of the means squared error is inherently non-robust because it involves the non-robust arithmetic mean in its second summand, which estimates the squared bias. Thus, we propose the MER estimate of the trimmed mean.

Proposition 17. *Suppose $r(\hat{F},\cdot)$ has a global minimum at $\hat{\beta}(\hat{F}) = \arg\min_{\beta\in B} \hat{r}(\hat{F};\beta)$. The MER-estimator of the trimmed population mean is $M(\hat{F},\beta) = T(\hat{F},\alpha,\hat{\beta}(\hat{F}))$.*

The MER-estimator is not robust because it is consistent. However the efficiency gain compared with an unbiased estimator can be considerable. By downweighting the bias term in an appropriate way, a robust version of the MER results at the cost of an additional complexity. To estimate the asymptotic variance of the adaptive estimator, one usually has to estimate the variance of the nonadaptive estimator with its tuning constant equal to the adaptively chosen value. It is clear that such an estimate does not take into account the adaptiveness of the estimator. Therefore, the problem is caused by the fact that the estimator of variance pretends that the tuning constant was chosen *a priori* rather than adaptively. This in turn may likely result in a downward bias for the estimate of variance (see e.g., LÉGER and ROMANO, 1990).

4.7 Proofs

Proof. [Lemma 6] Let $F \in \mathcal{F}$, where $\mathcal{F}$ is the set of absolute continuous cdf, and denoted by $F_\varepsilon(y)$ the mixture distribution $F_\varepsilon(y) = (1-\varepsilon)F(y) + \varepsilon G(y)$, where $G(y) = \mathbb{1}\{y \geq z\}$ is an elementary (degenerate) cdf. For $\beta \in \mathbb{Q}$ such that β is not a limit point of F^{-1}, the influence function writes $IF(z, QSM(\cdot;\beta), F) = \partial/\partial\varepsilon \left[\beta^{-1} \int^{\xi_\beta(F_\varepsilon)} y \mathrm{d}F_\varepsilon(y)\right]$

for $\varepsilon = 0$. Thus, differentiating w.r.t. ε (by means of the Leibniz integration rule) and taking $\varepsilon \downarrow 0$, yields $IF(z, QSM(\cdot), F) = \beta^{-1} \int^{\xi_\beta(F)} y \mathrm{d}G(y) - \beta^{-1} \int^{\xi_\beta(F)} y \mathrm{d}F(y) + \beta^{-1}\xi_\beta(F) f(\xi_\beta(F)) \left[\mathrm{d}/\mathrm{d}\varepsilon \xi_\beta(F_\varepsilon)\right]_{\varepsilon=0}$. Note that $\mathrm{d}/\mathrm{d}\varepsilon[\xi_\beta(F_\varepsilon)]_{\varepsilon=0}$ is the influence function of the βth quantile functional (see e.g., Huber, 1981, 56-57) and defined as $IF(z, \xi_\beta(\cdot), F) = [\beta - \mathbb{1}\{\xi_\beta(F) \geq z\}][f(\xi_\beta(F))]^{-1}$. Bringing altogether, f cancels out and we get the influence function which completes the proof. □

Proof. [Lemma 12] Suppose $\underline{\mathcal{Y}} = 0$, and $\beta_t \in \mathbb{Q}$ be associated with $Q(F; \beta_t)$, where $\forall \beta_t, t = 1, \ldots, p : 0 < \beta_t < 1$; β_t is not a limit point of F^{-1}. The $Q(F; \beta_t)$ functional admits a first-order von Mises expansion at F around G, which is given by $Q(G; \beta_t) = Q(F; \beta_t) + \int IF(y, Q(\cdot; \beta_t), F)\mathrm{d}(G - F)(y) + R(G, F)$, with IF according to Lemma 6. For ease of notation, write $z_{hijk} = IF(y_{hijk}, Q(\cdot; \beta_t), F)$ and $Z_{hijk} = IF(Y_{hijk}, Q(\cdot; \beta_t), F)$ (adopting the convention that capital letters denote random variables). Under the assumption $0 < \beta_t < 1$ and for n sufficiently large, there exist constants c_t such that $\inf_L F(c_t) > \beta_t, \forall t$ (where $L \to \infty$ according to the asymptotic framework; and the fact that $\hat{F}_L(c_t) - F_L(c_t) \to_p 0$), then $\{z_{hijk}\}$ is bounded. Moreover, and under the regularity conditions on the sampling design, i.e., Assumptions A1 and A2, Liapounov's condition hold and we obtain for the weighted average

$$\int IF(y, Q(\cdot; \beta), F)\mathrm{d}\hat{F}(y) = 1/N \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} \sum_{k=1}^{N_{hij}} w_{hijk} z_{hijk} = \overline{z}, \tag{4.37}$$

and $\mathbb{E}\overline{z} = 1/N \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} \sum_{k=1}^{N_{hij}} Z_{hijk} = 0$ (cf. Shao, 1994, Theorem 1). Thus, by Krewski and Rao (1981, Theorem 3.1) $\overline{z}/\sigma(Q(\cdot; \beta_t), F) \to_d N(0, 1)$ (since $\mathbb{E}\overline{z} = 0$). For n sufficiently large, we may write $\hat{Q}(\hat{F}; \beta_t) = Q(F; \beta_t) + \overline{z} + R(\hat{F}, F)$. Finally, by Shao (1994, Theorem 1) $\sqrt{n}R(\hat{F}, F) \to_p 0$, and thus $[\hat{Q}(\hat{F}; \beta_t) - Q(F; \beta_t)]/\sigma(Q(\cdot; \beta_t), F) \to_d N(0, 1)$.

In particular, the asymptotic covariance of $\sqrt{n}Q(F; \beta_i)$ and $\sqrt{n}Q(F; \beta_j)$ using the result of Lemma 6 is given by

$$\omega_{\beta_i, \beta_j} = \int IF(z, Q(\cdot; \beta_i), F) IF(z, Q(\cdot; \beta_j), F)\mathrm{d}F(z) \tag{4.38}$$

Given $\beta_i \leq \beta_j$ and that $\mathbb{1}\{x \leq \xi(F; \beta_j)\} = 1$ whenever $\mathbb{1}\{x \leq \xi(F; \beta_i)\} = 1$ the right-hand side of (4.38) becomes

$$\begin{aligned} &[\xi_{\beta_i} - T_{\beta_i}(F)] \left[\xi_{\beta_j} - T_{\beta_j}(F)\right] + \int^{\xi_{\beta_j}} [\xi_{\beta_i} - T_{\beta_i}(F)] \frac{1}{\beta_j} \left[x - \xi_{\beta_j}\right] \mathrm{d}F(x) + \\ &+ \int^{\xi_{\beta_i}} \left[\frac{1}{\beta_j}\left(x - \xi_{\beta_j}\right) + \xi_{\beta_j} - T_{\beta_j}(F)\right] \frac{1}{\beta_i} \left[x - \xi_{\beta_i}\right] \mathrm{d}F(x) \end{aligned} \tag{4.39}$$

On simplifying Eq. (4.39), we obtain (in close relation to the "cumulative income functional" in COWELL and VICTORIA-FESER (2003, Appendix A.1))

$$\begin{aligned}\omega_{\beta_i,\beta_j} &= \frac{1}{\beta_i\beta_j}S(\beta_i,F)+\xi_{\beta_i}Q(\beta_i,F)+\xi_{\beta_j}Q(\beta_j,F)-\frac{1}{\beta_j}Q(\beta_i,F)\left(\xi_{\beta_j}+\xi_{\beta_i}\right)+\\ &\quad -Q(\beta_i,F)Q(\beta_j,F)+\xi_{\beta_i}\xi_{\beta_j}\left(\frac{1}{\beta_j}-1\right), \qquad \text{for } i\le j, \end{aligned} \tag{4.40}$$

where ω_{β_i,β_j} is a short-hand notation for $\omega_{Q(F;\beta_i),Q(F;\beta_j)}$, and $S(\beta_i,F) := \int^{\xi(F;\beta_i)} y^2 \mathrm{d}F(y)$ denotes the variance of Y conditional on $Y \le \xi(F;\beta_i)$.

Thus, for each $Q(F;\beta_i)$ with $0<\beta_i<1$ (and if β_i is not a limit point of F^{-1}), $i=1,\ldots,p$, we have $\sqrt{n}(\hat{Q}(\hat{F};\beta_i)-Q(F;\beta_i)) \to_d N(0,\omega_{\beta_i,\beta_i})$. The vector $(\hat{Q}(\hat{F};\beta_1),\ldots,\hat{Q}(\hat{F};\beta_p))^T$ can be shown (by a Cramer-Wold device; see e.g., SERFLING (1980, p.18)) to have a p-variate limiting normal distribution with covariance matrix $\mathbf{\Omega}$ whose i,jth element is equal to ω_{β_i,β_j} for $i \le j$. This completes the proof. □

Bibliography

Atkinson, T., **Cantillon, B.**, **Marlier, E.** and **Nolan, B.** (**2002**): Social Indicators: The EU and Social Inclusion. Oxford: Oxford University Press.

Beaumont, J.-F. and **Rivest, L.-P.** (**2009**): *Dealing with outliers in survey data.* **Pfeffermann, D.** and **Rao, C.** (editors) Sample Surveys: Theory, Methods and Inference, *Handbook of Statistics*, vol. 29A, chapter 11, pp. 247–280, Amsterdam: Elsevier.

Binder, D. A. and **Patak, Z.** (**1994**): *Use of Estimating Functions for Estimation from Complex Surveys.* Journal of the American Statistical Association, 89 (427), pp. 1035–1043.

Central Statistics Office Ireland (**2005**): *EU Survey on Income and Living Condition 2004.* Technical report, Central Statistics Office Ireland, Cork.

Chambers, R. L. (**1986**): *Outlier Robust Finite Population Estimation.* Journal of the American Statistical Association, 81 (396), pp. 1063–1069.

Cowell, F. A. and **Flachaire, E.** (**2007**): *Income distribution and inequality measurement: The problem of extreme values.* Journal of Econometrics, 141, pp. 1044–1072.

Cowell, F. A. and **Victoria-Feser, M.-P.** (**1996**): *Robustness properties of inequality measures.* Econometrica, 64 (1), pp. 77–101.

Cowell, F. A. and **Victoria-Feser, M.-P.** (**2002**): *Welfare rankings in the presence of contaminated data.* Econometrica, 70 (3), pp. 1221–1233.

Cowell, F. A. and **Victoria-Feser, M.-P.** (**2003**): *Distribution-free inference for welfare indices under complete and incomplete information.* Journal of Economic Inequality, 1, pp. 191–219.

Cowell, F. A. and **Victoria-Feser, M.-P.** (**2006**): *Distributional Dominance With Trimmed Data*. Journal of Business & Economic Statistics, 24 (3), pp. 291–300.

Deaton, A. (**2000**): The Analysis of Household Surveys: A Microeconomic Approach to Development Policy. World Bank Publications, Baltimore: The Johns Hopkins University Press, 3rd ed.

Deville, J.-C. and **Särndal, C.-E.** (**1992**): *Calibration Estimators in Survey Sampling*. Journal of the American Statistical Association, 87 (418), pp. 376–382.

European Commission (**2003**): *Laeken Indicators. Detailed calculation methodology*. Technical report, EUROSTAT working group statistics on income, poverty and social exclusion, Luxembourg. DOC. E2/IPSE/2003.

EUROSTAT (**2008**): *EU-SILC. Comparative Final EU Quality Report 2005*. Technical report, EUROSTAT, Luxembourg.

Fuller, W. A. (**1991**): *Simple estimators for the mean of skewed populations*. Statistica Sinica, 1, pp. 137–158.

Hampel, F. R., **Ronchetti, E. M.**, **Rousseeuw, P. J.** and **Stahel, W. A.** (**1986**): Robust Statistics: The Approach Based on Influence Functions. New York: John Wiley & Sons.

Huber, P. J. (**1981**): Robust Statistics. New York: John Wiley & Sons.

Hulliger, B. (**1991**): Nonparametric M-estimation of a population mean. Ph.D. thesis, ETH Zurich, Nr. 9443.

Hulliger, B. (**1995**): *Outlier Robust Horvitz-Thompson Estimators*. Survey Methodology, 21 (1), pp. 79–87.

Hulliger, B. and **Münnich, R.** (**2006**): *Variance Estimation for Complex Surveys in the Presence of Outliers*. ASA Proceedings of the Section on Survey Research Methods, American Statistical Association.

Hulliger, B. and **Schoch, T.** (**2009**): *Robustification of the quintile share ratio*. Proceedings of the NTTS Conference - New Techniques and Technologies for Statistics, Brussels: Eurostat.

Jaeckel, L. (**1971**): *Robust Estimates of Location: Symmetry and Asymmetric Contamination*. Annals of Mathematical Statistics, 42, pp. 1020–1034.

Krewski, D. and **Rao, J.** (**1981**): *Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication method*. The Annals of Statistics, 9 (5), pp. 1010–1019.

Léger, C. and **Romano, J. P.** (**1990**): *Bootstrap adaptive estimation. The trimmed mean example*. The Canadian Journal of Statistics, 18 (4), pp. 297–314.

Münnich, R., **Bruch, C.** and **Zins, S.** (**2011**): *Variance Estimation for Complex Surveys*. Technical report, AMELI deliverable D3.1.
URL http://ameli.surveystatistics.net/

Nygård, F. and **Sandström, A.** (**1989**): *Income inequality measures based on sample surveys*. Journal of Econometrics, 42, pp. 81–95.

Pfeffermann, D. (**1993**): *The role of sampling weights when modelling survey data*. International Statistical Review, 61 (2), pp. 317–337.

Rao, J. N. K. and **Wu, C. F. J.** (**1985**): *Inference From Stratified Samples: Second-Order Analysis of Three Methods for Nonlinear Statistics*. Journal of the American Statistical Association, 80 (391), pp. 620–630.

Särndal, C.-E., **Swensson, B.** and **Wretman, J.** (**1992**): Model Assisted Survey Sampling. New York: Springer, 2 ed.

Serfling, R. J. (**1980**): Approximation theorems of mathematical statistics. New York: Wiley.

Shao, J. (**1994**): *L-Statistics in complex survey problems*. The Annals of Statistics, 22 (2), pp. 946–967.

Smith, T. (**1987**): *Influential Observations in Survey Sampling*. Journal of Applied Statistics, 14 (2), pp. 143–152.

Stigler, S. M. (**1973**): *The Asymptotic Distribution of the Trimmed Mean*. Annals of Statistics, 1 (3), pp. 472–477.

Victoria-Feser, M.-P. and **Ronchetti, E. M.** (**1994**): *Robust Methods for Personal-Income Distribution Models*. The Canadian Journal of Statistics, 22 (2), pp. 247–258.

Wolter, K. M. (**2007**): Introduction to variance estimation. New York: Springer, 2nd ed.

Zheng, B. (**2002**): *Testing Lorenz curves with non-simple random samples*. Econometrica, 70 (3), pp. 1235–1243.

Chapter 5

Robust Basic Unit-Level Small-Area-Estimation Model

5.1 Introduction

Although the models involved in small area estimation have multivariate explanatory variables, the characteristic of interest is univariate. We therefore treat the robustification of small area estimation in the part of D4.2 dedicated to univariate estimators.

Small are estimation (SAE) has become of great importance due to the growing demand for reliable small-area statistics. In the basic setup, small area means (and totals) can be expressed as linear combinations of fixed and random effects, which are obtained by best linear unbiased prediction (BLUP) estimators, appealing to well-known results on BLUP estimation. These estimators minimize the MSE among the class of linear unbiased estimators without assuming normality of the random effects. Although the classical EBLUP method is useful for estimating the small area means efficiently under the normality assumptions, it can be highly influenced by the presence of outliers or departures from the assumed distribution. Therefore, Sinha and Rao (2009) proposed a robustification of the unit- and area-level models.

We discuss a related, but slightly different robustification. The main contribution is a fast algorithm that avoids inversion of large matrices and minimizes the number of matrix multiplication which in turn results in a tremendous speed-up in computing time and permits the user to apply the method to large datasets (e.g., datasets with $n = 2,000,000$ observations). Insofar the proposed method serves the needs in official statistics.

The remainder of the paper is organized as follows. In Section 5.2.1, we introduce the basic unit-level model and study its maximum likelihood estimators. Section 5.3 is concerned with robust, bounded-influence estimating equations (BIEE). In Section 5.4, we derive Newton-Raphson updating equations, introduce bounded-influence predicting equation (BIPE), and discuss the choice of starting values. Section 5.5 draws together the main findings.

5.2 Small Area Estimation

5.2.1 Unit-Level Models

A large class of unit-level small-area estimation models can be regarded as a special case of the mixed linear model (MLM) of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{t=1}^{c-1} \mathbf{Z}_t \mathbf{v}_t + \mathbf{e}, \tag{5.1}$$

where $\mathbf{y}$ is a $(n \times 1)$ vector of observations; $\mathbf{X}$ and $\mathbf{Z}_t$ are, respectively, known $(n \times q)$ and $(n \times p_t)$ matrices (of full rank); $\boldsymbol{\beta}$ is a q-vector of unknown fixed effects; the $\mathbf{v}_t$ are $(p_t \times 1)$ vectors of unobserved random effects, $1 \leq t \leq c-1$; and $\mathbf{e}$ is a $(n \times 1)$ vector of unobserved errors. The p_t levels of each random effect $\mathbf{v}_t$ are assumed to be independent with mean zero and variance σ_t^2; the random error $\mathbf{e}$ is assumed to be independent with mean zero and variance σ_c^2; and $\mathbf{v}_1, \ldots, \mathbf{v}_{c-1}$ and $\mathbf{e}$ are assumed to be independent. It follows that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ and $\mathbb{V}[\mathbf{y}|\mathbf{X}] =: \mathbf{V} = \boldsymbol{\theta}_c \mathbf{I}_n + \sum_{t=1}^{c-1} \boldsymbol{\theta}_t \mathbf{Z}_t \mathbf{Z}_t^T$, where $\boldsymbol{\theta} = (\sigma_1^2, \ldots, \sigma_c^2)^T$ and $\mathbf{I}_n$ denotes the $(n \times n)$ identity matrix. Moreover, it is assumed that we have adopted a parametrization in which all the $r = q + c$ unknown parameters $\boldsymbol{\tau} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ are identifiable.

The basic unit-level model (BULM), aka basic nested-error regression model (BATTESE et al., 1988), and several extensions are regarded as MLM with block-diagonal covariance structure; see RAO (2003, chap. 6.3). In particular, we shall define the BULM.

Definition 18. *The basic unit-level model, $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$, with Gaussian area-specific random effects, $i = 1, \ldots, g$, is defined as*

$$\mathbf{y}_i \sim F \in \mathcal{F} := \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i(\boldsymbol{\theta})), \tag{5.2}$$

with

$$\mathbf{V}_i(\boldsymbol{\theta}) = \sum_{l=1}^{2} \sigma_l^2 \mathbf{Z}_{li} \mathbf{Z}_{li}^T, \quad i = 1, \ldots, g, \tag{5.3}$$

where $\forall i = 1, \ldots, g$, $\boldsymbol{\beta}$ is an $(q \times 1)$ vector of fixed effects; $\boldsymbol{\theta}$ is defined as $\boldsymbol{\theta} = (\sigma_c^2, \sigma_1^2)^T$ with σ_c^2 and σ_1^2 the model-error- and random-effect variance, respectively; $\mathbf{Z}_{1i} = \mathbf{I}_i$ is the $(n_i \times n_i)$ identity matrix; $\mathbf{Z}_{2i} = \mathbf{J}_i$ is the $(n_i \times n_i)$ matrix of ones; $\mathbf{X}_i$ is the $(n_i \times q)$ design matrix of known co-variates. In addition we assume that the area-specific sample size satisfies $n_i \geq (q+2)$, $i = 1, \ldots, g$.

It follows that $\mathbb{E}[\mathbf{y}_i] = \mathbf{X}_i\boldsymbol{\beta}$ and $\mathbb{V}[\mathbf{y}_i|\mathbf{X}_i] = \mathbf{V}_i(\boldsymbol{\theta}) = \sigma_c^2 \mathbf{I}_i + \sigma_1^2 \mathbf{J}_i$, $i = 1, \ldots, g$. In what follows, we shall suppress the functional dependence of $\mathbf{V}_i(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ and write $\mathbf{V}_i$ for clarity of display, whenever no confusion can arise. Moreover, we shall define the parameter space of model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$.

Definition 19. *Suppose the basic unit-level model* $\mathcal{M}(\boldsymbol{\tau})$*, with* $\boldsymbol{\tau} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$*. The parameter space for* $\boldsymbol{\tau}$ *is assumed to be*

$$\Omega(\boldsymbol{\tau}) = \Omega_{\boldsymbol{\beta}} \times \Omega_{\boldsymbol{\theta}} \tag{5.4}$$

where

$$\Omega_{\boldsymbol{\beta}} = \{\beta_k \in \mathbb{R}, k = 1, \ldots, q\} \quad \textit{and} \quad \Omega_{\boldsymbol{\theta}} = \{(\sigma_c^2, \sigma_1^2) \in \mathbb{R}^2;\ \sigma_c^2 > 0,\ \sigma_1^2 \geq 0\} \tag{5.5}$$

Whenever no contamination is supposed to be present, estimates of the parameter vector $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{\theta})^T$ of $\mathcal{M}(\boldsymbol{\tau})$ shall be obtained by means of maximum likelihood (ML) estimators. Let $l(\hat{\boldsymbol{\tau}}_{ML})$ denote the (non-robust) log-likelihood of the core model. The ML estimator $\hat{\boldsymbol{\tau}}_{ML}$ of $\boldsymbol{\tau}$ is defined by $l(\hat{\boldsymbol{\tau}}_{ML}) = \sup_{\boldsymbol{\tau} \in \Omega} l(\boldsymbol{\tau})$, provided $\boldsymbol{\tau}$ is an interior point of Ω from Definition 19.

Upon having obtained the ML estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$, one then considers predicting the area-specific random effects $\mathbf{v}_i$. From the theory of best linear unbiased predictors (BLUP), where best is in the sense of minimal mean square error prediction, we note that for known $\tilde{\boldsymbol{\theta}}$ the BLUP (see e.g., Searle et al., 1992, chap. 7.4) is

$$\mathbb{E}[\mathbf{v}_i|\mathbf{y}_i] = \mathbf{G}_i(\tilde{\boldsymbol{\theta}})\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\tilde{\boldsymbol{\theta}})[\mathbf{y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}] =: \tilde{\mathbf{v}}_i(\boldsymbol{\theta}), \quad i = 1, \ldots, g, \tag{5.6}$$

where $\tilde{\boldsymbol{\beta}} = \text{BLUE}(\boldsymbol{\beta})$ and $\mathbf{G}_i$ follows from the decomposition $\mathbf{V}_i = \mathbf{R}_i + \mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i^T$, a property of models with a block-diagonal covariance matrix. In the case of the basic unit-level model, we have $\mathbf{R}_i = \sigma_c^2\mathbf{I}_i$ and $\mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i^T = \sigma_1^2\mathbb{1}_i\mathbb{1}_i^T$. Therefore, replacing $\tilde{\boldsymbol{\theta}}$ by the ML-estimate, $\hat{\boldsymbol{\theta}}$, and $\tilde{\boldsymbol{\beta}}$ by the ML-estimate $\hat{\boldsymbol{\beta}}$ leads to the empirical BLUP (EBLUP). Let $\bar{\mathbf{x}}_i$ denote the vector of known means for area i. The estimates $\hat{\mathbf{v}}_i$ (suppressing the dependence on $\hat{\boldsymbol{\theta}}$) are then used to obtain the EBLUP of μ_i given by $\hat{\mu}_i = \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}} + \hat{\mathbf{v}}_i$. The EBLUP of the mean, $\bar{Y}_i$, on the other hand, is obtained from

$$t_i(\hat{\boldsymbol{\theta}}, \mathbf{y}) = N_i^{-1}\Big(\sum_{j \in s_j} y_{ij} + \sum_{j \in \bar{s}_j} \hat{y}_{ij}\Big), \tag{5.7}$$

where s_i and $\bar{s}_i$ represent the set of sampled and non-sampled units in area i, respectively, and $\hat{y}_{ij} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \hat{v}_i$; see Rao (2003).

5.3 Bounded Influence-Equation Approach

Although the classical EBLUP method is useful for estimating the small area means efficiently under normality assumptions, it can be highly influenced by the presence of outliers in the data or departures from the assumed normal distribution of the random

effects. Furthermore, mixed linear models have, unlike location-scale or regression models, no nice invariance structure. Notably, this means that the *parameters cannot be estimated consistently in the presence of contamination*; there is an unavoidable asymptotic bias. Thus in the presence of contamination, any method estimates the parameter at the core model plus an unknown bias. In the case of ML estimates, the bias can be arbitrarily large and renders these estimators extremely inefficient when the model does not hold.

We consider as estimators any robust estimator (i.e., with bounded influence function) that is *Fisher-consistent* at the Gaussian core model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ when the model holds. For these estimators, the potential bias is bounded, the efficiency is reasonable if the model holds, and the estimators are much more efficient than e.g., ML estimators, if it does not (cf. WELSH and RICHARDSON, 1997).

For the basic unit-level model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$, SINHA and RAO (2009) propose to replace the ML estimating equations (1.10) and (1.11) by bounded-influence estimating equations (BIEE) in order to handle outliers in the response variable. Moreover, they indicate the usage of a Mallows/Schweppe-type weighting scheme. In contrast to SINHA and RAO (2009), we define the BIEE associated with the model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ (suppressing the Schweppe-type weighting) according to the proposal of RICHARDSON and WELSH (1995, p.1432) (called RML II therein; see also WELSH and RICHARDSON (1997, pp.361-362)). Thus, let $\mathbf{y}_i$, $\mathbf{X}_i$, and $\mathbf{V}_i^{-1}$, $i = 1, \ldots, g$, be specified as in the Definitions 18 and 19. Let $\boldsymbol{\psi}_k(\mathbf{r}_i) = (\psi_k(r_1), \cdots, \psi_k(r_{n_i}))$ denote the $(n_i \times 1)$ vector of winsorized residuals, $\mathbf{r}_i = \mathbf{V}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$, $i = 1, \ldots, g$, where $\psi_k(\cdot)$ is Huber's ψ-function indexed by the robustness-tuning constant k (or any other ψ-function). The system of BIEE writes

$$\sum_{i=1}^{g} \left\{ \mathbf{X}_i^T \mathbf{V}_i^{-1/2} \boldsymbol{\psi}_k(\mathbf{r}_i) \right\} = \mathbf{0}, \tag{5.8}$$

$$S := (1/2) \sum_{i=1}^{g} \left\{ \boldsymbol{\psi}_k^T(\mathbf{r}_i) \mathbf{V}_i^{-1} \boldsymbol{\psi}_k(\mathbf{r}_i) - \operatorname{tr}\left(\boldsymbol{\kappa}_i \mathbf{V}_i^{-1} \mathbf{I}_i\right) \right\} = 0, \tag{5.9}$$

$$M := (1/2) \sum_{i=1}^{g} \left\{ \boldsymbol{\psi}_k^T(\mathbf{r}_i) \mathbf{V}_i^{-1/2} \mathbf{J}_i \mathbf{V}_i^{-1/2} \boldsymbol{\psi}_k(\mathbf{r}_i) - \operatorname{tr}\left(\boldsymbol{\kappa}_i \mathbf{V}_i^{-1} \mathbf{J}_i\right) \right\} = 0, \tag{5.10}$$

where $\boldsymbol{\kappa}_i = c\mathbf{I}_i$ are consistency correction matrices with $c = \mathbb{E}[\psi(z)^2]$ and $z \sim N(0,1)$. For model $\mathcal{M}(\cdot)$ and the Huber ψ-function, we have $c = 2[k^2(1 - \Phi(k)) + \Phi(k) - 0.5 - k\phi(k)]$, where Φ and ϕ denote the cdf and pdf of the standard normal, respectively; see e.g., MARONNA et al. (2006, p.27).

Note that the system of BIEE proposed by SINHA and RAO (2009) (which uses a simpler normalization of the residuals) can be obtained by changing $\mathbf{V}_i^{-1/2}\boldsymbol{\psi}_k(\mathbf{r}_i)$ to $\mathbf{V}_i^{-1}\mathbf{U}_i^{1/2}\boldsymbol{\psi}_k(\tilde{\mathbf{r}}_i)$, where $\mathbf{U}_i = \operatorname{diag}(\mathbf{V}_i)$ and $\tilde{\mathbf{r}}_i = \mathbf{U}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$.

5.4 Proposed Method

5.4.1 Preparations

The crucial point is that the variance-covariance matrix of the model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$, $\mathbf{V}_i = \sigma_c^2\mathbf{I}_i + \sigma_1^2\mathbf{J}_i$, admits a closed-form expression of its inverse

$$\mathbf{V}_i^{-1} = \delta\mathbf{I}_i - \omega_i\mathbf{J}_i, \quad i = 1, \ldots, g, \tag{5.11}$$

with $\delta = 1/\sigma_c^2$ and $\omega_i = \sigma_1^2/(\sigma_c^2(\sigma_c^2 + n_i\sigma_1^2))$; see e.g., Searle et al. (1992, p.308).

We introduce a lemma on orthogonal factorizable variance-covariance matrices that will be useful to define the function of a matrix that correspond to the function of a scalar.

Lemma 20. *Suppose model* $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ *with* $\boldsymbol{\theta} = (\sigma_c^2, \sigma_1^2)^T$, $i = 1, \ldots, g$. *The* $(n_i \times n_i)$ *variance-covariance matrices* $\mathbf{V}_i$, $i = 1, \ldots, g$, *exhibit the orthogonal factorization*

$$\mathbf{V}_i = \mathbf{L}_i \text{diag}\,(\lambda_1 | \, \lambda_2, \lambda_2, \ldots, \lambda_2)\mathbf{L}_i^T, \tag{5.12}$$

where $\mathbf{L}_i$ *is the* $(n_i \times n_i)$ *matrix whose columns correspond to the eigenvectors of* $\mathbf{V}_i$; $\lambda_1 = \sigma_c^2 + n_i\sigma_1^2$ *and* $\lambda_2 = \sigma_c^2$ *denote the eigenvalues with multiplicities 1 and* $n - 1$, *respectively,* $i = 1, \ldots, g$. *(for clarity, the bar in* diag *indicates the partition)*

Proof. Since the factorization for all $\mathbf{V}_i$ $(i = 1, \ldots, g)$ analogous, we consider here the decomposition of $\mathbf{V}$ (suppressing the subscript). The characteristic equation of $\mathbf{V}$ writes $\det[(\sigma_c^2 - \lambda)\mathbf{I} + \sigma_1^2\mathbf{J}] = 0$, where λ is an eigenvalue. By the Matrix Determinant Lemma (see e.g., Gentle, 2007), we obtain $\det[(\sigma_c^2 - \lambda)\mathbf{I} + \sigma_1^2\mathbb{1}\mathbb{1}^T] = (1 + \mathbb{1}^T((\sigma_c^2 - \lambda)\mathbf{I})^{-1}\mathbb{1})\det[(\sigma_c^2 - \lambda)\mathbf{I}]$. Since $\det[(\sigma_c^2 - \lambda)\mathbf{I}] = (\sigma_c^2 - \lambda)^n\det[\mathbf{I}] = (\sigma_c^2 - \lambda)^n$ and $((\sigma_c^2 - \lambda)\mathbf{I})^{-1} = (\sigma_c^2 - \lambda)^{-1}\mathbf{I}$, the characteristic equation writes $(\sigma_c^2 - \lambda + \sigma_1^2 n)(\sigma_c^2 - \lambda)^{n-1} = 0$. From this result, we can read off the two solutions, i.e., two distinct eigenvalues, $\lambda_1 = \sigma_c^2 + n\sigma_1^2$ with algebraic multiplicity 1 and $\lambda_2 = \sigma_c^2$ with multicplicity $n - 1$. □

In passing we note that the assertions of the above Lemma could be extended to variance-covariance matrices with a more general structure. Searle and Henderson (1979, Theorem 3.1) give a procedure (involving a series of tedious operations) to obtain eigenvalues for very general matrices. However, there method is limited because it works only for variance matrices of models with balanced data. Alternatively one may derive expressions of the eigenvalues (and eigenvectors) by means of the matrix-inversion method proposed LaMotte (1972) that also applies for unbalanced data, but is restricted to nested designs (without cross-classification).

We introduce the following definition of matrix functions that correspond to functions of a scalar.

Definition 21 (Matrix function)**.** *Let $f : \mathbb{R} \mapsto \mathbb{R}$ and denote by $\mathbf{U} = \mathbf{L}\mathbf{D}\mathbf{L}^T$ a symmetric, orthogonal factorizable $(n \times n)$ matrix where $\mathbf{L}$ is the $(n \times n)$ matrix whose columns correspond to the eigenvectors of $\mathbf{U}$, and $\mathbf{D}$ is a diagonal matrix with the eigenvalues, $\lambda_1, \ldots, \lambda_n$, (incl. multiplicities) as diagonal elements. We define the function of $\mathbf{U}$ that corresponds to a function of a scalar as $f(\mathbf{U}) = \mathbf{L}diag\big[f(\lambda_1), \ldots, f(\lambda_n)\big]\mathbf{L}^T$ whenever the function for λ_i, $i = 1, \ldots, n$, exists.*

In effect, we could directly define functions of $\mathbf{V}$ (e.g., $\mathbf{V}^{1/2}$). However, this definition requires computing the matrices $\mathbf{L}_i$ $(i = 1, \ldots, g)$ whose columns correspond to the eigenvectors of $\mathbf{V}_i$, which is usually very involved (at least for large data sets). In the subsequent proposition, we show that a simple closed-form expression exists.

Proposition 22. *Suppose model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ and a function $f : \mathbb{R} \mapsto \mathbb{R}$. By Lemma 20, each area-specific covariance matrix, $\mathbf{V}_i$, $i = 1, \ldots, g$, posses an orthogonal factorization $\mathbf{V}_i = \mathbf{L}_i\mathbf{D}_i\mathbf{L}_i^T$, so that the matrix function f that corresponds to a function of a scalar from Definition 21, yields*

$$f(\mathbf{V}_i) = \frac{1}{n_i}\big[(f(\lambda_{1,i}) - f(\lambda_{2,i}))\mathbf{J}_{n_i} + n_i f(\lambda_{2,i})\mathbf{I}_{n_i}\big]. \tag{5.13}$$

Proof. By Lemma 20, we have obtained the first and second eigenvalues, $\lambda_{1,i}, \lambda_{2,i}$, of the matrix $\mathbf{V}_i$, for all $i = 1, \ldots, g$ (with the corresponding multiplicities). In what follows, we will suppress the subscript i for ease of simplicity. Given λ_1 and λ_2 we derive the matrix $\mathbf{L}$, the columns of which correspond to the eigenvectors of $\mathbf{V}$. By the symmetry of $\mathbf{V}$, the eigenvectors corresponding to one eigenvalue with multiplicity greater than one, are orthogonal (GENTLE, 2007). The characteristic polynomial associated with the first eigenvalue is $(\mathbf{V} - \lambda_1\mathbf{I})\mathbf{v} = \mathbf{0}$. On simplifying we get $\sum_{k=1}^{n} v_k = nv_k$, $k = 1, \ldots, n$. Similarly, for the eigenvectors (which form an $(n-1)$-dimensional eigen subspace) associated with the second eigenvalue, we have $\sum_{k=1}^{n} v_k = 0$. Suppose the following partitioning

$$\mathbf{L} = \begin{pmatrix} v_1 & \mathbf{v}_{-1}^T \\ \mathbf{v}_{-1} & \mathbf{Q} \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} \lambda_1 & \mathbf{z}_{-1}^T \\ \mathbf{z}_{-1} & \lambda_2\mathbf{I}_{n-1} \end{pmatrix}, \tag{5.14}$$

where $\mathbf{Q}$ is an $((n-1) \times (n-1))$ matrix whose columns correspond to the eigenvectors associated with the second eigenvalue; $\mathbf{v}_{-1} = (v_i)_{i=2,\ldots,n}$; $\mathbf{z}_{-1} = 0 \cdot \mathbb{1}_{n-1}$. Thus,

$$\begin{aligned} \mathbf{L}\mathbf{D}\mathbf{L}^T &= \begin{pmatrix} v_1 & \mathbf{v}_{-1}^T \\ \mathbf{v}_{-1} & \mathbf{Q} \end{pmatrix} \begin{pmatrix} \lambda_1 & \mathbf{z}_{-1}^T \\ \mathbf{z}_{-1} & \lambda_2\mathbf{I}_{n-1} \end{pmatrix} \begin{pmatrix} v_1 & \mathbf{v}_{-1}^T \\ \mathbf{v}_{-1} & \mathbf{Q}^T \end{pmatrix} \\ &= \begin{pmatrix} v_1^2\lambda_1 + \lambda_2\mathbf{v}_{-1}^T\mathbf{v}_{-1} & \lambda_1 v_1\mathbf{v}_{-1}^T + \lambda_2\mathbf{v}_{-1}^T\mathbf{Q}^T \\ \lambda_1 v_1\mathbf{v}_{-1} + \lambda_2\mathbf{Q}\mathbf{v}_{-1} & \lambda_1\mathbf{v}_{-1}\mathbf{v}_{-1}^T + \lambda_2\mathbf{Q}\mathbf{Q}^T \end{pmatrix} \end{aligned} \tag{5.15}$$

In order to simplify (5.15), we need to study $\mathbf{L}\mathbf{L}^T = \mathbf{I}_n$. Therefrom, we can retrieve several building blocks that will emerge to be useful. On simplifying $\mathbf{L}\mathbf{L}^T = \mathbf{I}_n$, we have

$$\begin{pmatrix} v_1 & \mathbf{v}_{-1}^T \\ \mathbf{v}_{-1} & \mathbf{Q} \end{pmatrix} \begin{pmatrix} v_1 & \mathbf{v}_{-1}^T \\ \mathbf{v}_{-1} & \mathbf{Q}^T \end{pmatrix} = \begin{pmatrix} v_1^2 + (n-1)v_1 & v_i^2\mathbf{v}_{-1}^T + \mathbf{v}_{-1}^T\mathbf{Q}^T \\ v_i\mathbf{v}_{-1} + \mathbf{Q}\mathbf{v}_{-1} & v_i^2\mathbf{J}_{n-1} + \mathbf{Q}\mathbf{Q}^T \end{pmatrix} = \mathbf{I}_n, \tag{5.16}$$

which in turn give rise to the following identities (i.e., equating block-wise), i) $v_i = 1/\sqrt{n}$, $i = 1, \ldots, n$; by application of i) we have, ii) $\mathbf{v}_{-1}^T\mathbf{Q}^T = -\frac{1}{n}\mathbb{1}_{n-1}^T$ (and analogously, $\mathbf{Q}\mathbf{v}_{-1} = -\frac{1}{n}\mathbb{1}_{n-1}$); and iii) $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}_{n-1} - v_i^2\mathbf{J}_{n-1}$. Thus, (5.15) can be expressed as

$$\begin{aligned} \mathbf{L}\mathbf{D}\mathbf{L}^T &= \frac{1}{n}\begin{pmatrix} \lambda_1 + (n-1)\lambda_2 & (\lambda_1 - \lambda_2)\mathbb{1}_{n-1}^T \\ (\lambda_1 - \lambda_2)\mathbb{1}_{n-1} & \lambda_2 n\mathbf{I}_{n-1} + (\lambda_1 - \lambda_2)\mathbf{J}_{n-1} \end{pmatrix} \\ &= \frac{1}{n}\big[(\lambda_1 - \lambda_2)\mathbf{J}_n + n\lambda_2\mathbf{I}_n\big]. \end{aligned} \tag{5.17}$$

This completes the proof. □

By application of Lemma 20, we can define $\mathbf{V}_i^{1/2}$ according to Definition 21 as

$$\mathbf{V}_i^{1/2} = \frac{1}{n_i}\big[(\sigma_c^2 + n_i\sigma_1^2)^{1/2} - (\sigma_c^2)^{1/2}\big]\mathbf{J}_{n_i} + (\sigma_c^2)^{1/2}\mathbf{I}_{n_i}, \quad i = 1, \ldots, g. \tag{5.18}$$

Similarly, we obtain a closed-form expression for $\mathbf{V}_i^{-1/2}$. Moreover, this definition of the matrix square root is unique (in contrast to the definitions of e.g., Welsh and Richardson (1997) or Huggins (1993) and admits simple expressions for $\partial\mathbf{V}_i^{-1/2}(\boldsymbol{\theta})/\partial\theta_1$ and the like.

The following lemma states some useful identities that play an important role in the derivation of the Taylor series expansions subject to the Newton-Raphson updating equations.

Lemma 23 (Blocks)**.** *Suppose model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$. For the variance-covariance matrix $\mathbf{V}_i(\boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\sigma_c^2, \sigma_1^2)^T$, we note that $\forall i = 1, \ldots, g$*

i) $\partial\mathbf{V}_i^{-1}/\partial\sigma_c^2 = -\mathbf{V}_i^{-2}$ *ii)* $\partial\mathbf{V}_i^{-1}/\partial\sigma_1^2 = -\mathbf{V}_i^{-1}\mathbf{J}_i\mathbf{V}_i^{-1}$

iii) $\partial\mathbf{V}_i^{-1/2}/\partial\sigma_c^2 = -(1/2)\mathbf{V}_i^{-3/2}$ *iv)* $\partial\mathbf{V}_i^{-1/2}/\partial\sigma_1^2 = -(1/2)\mathbf{V}_i^{-1/2}\mathbf{J}_i\mathbf{V}_i^{-1}$

Proof. Assertions i) and ii) follow immediately from the chain rule of matrix derivatives, $\partial\mathbf{A}(\boldsymbol{\theta})^{-1}/\partial\theta_k = -\mathbf{A}^{-1}(\boldsymbol{\theta})[\partial\mathbf{A}(\boldsymbol{\theta})/\partial\theta_k]\mathbf{A}^{-T}$ (see e.g., Magnus and Neudecker, 1999, 96), and Lemma 20. For assertions iii) and iv) put $\partial\mathbf{x}^{1/2} = (1/2)\mathbf{x}^{-1/2}$ and evaluate it at $\mathbf{x} = \mathbf{V}^{-1}$, thus $\partial\mathbf{V}(\boldsymbol{\theta})^{-1/2}/\partial\theta_k = -(\partial\mathbf{V}_i^{-1}/\partial\boldsymbol{\theta})(\partial\mathbf{x}^{1/2}/\partial\mathbf{x})|_{\mathbf{x}=\mathbf{V}^{-1}}$. On simplifying, we are done (see e.g., Richardson and Welsh, 1995, 1439). □

Definition 24. *We shall make much use of the following definitions:*

i) *Let* $\mathbf{V}_i$ *denote the variance-covariance matrix associated with model* $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$, $i = 1, \ldots, g$. *According to Lemma 20, define* $\mathbf{V}_i^{-1/2} = \xi_i \mathbf{I}_i + \eta_i \mathbf{J}_i$ *with* $\xi_i = n_i(\sigma_c^2)^{-1/2}$ *and* $\eta_i = (\sigma_c^2 + n_i\sigma_1^2)^{-1/2} - (\sigma_c^2)^{-1/2}$, $i = 1, \ldots, g$,

ii) *let* $r_{i,m} = \left\{\mathbf{V}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}_{\{m:m=1,\ldots,n_i\}}$ *denote the residual,* $i = 1 \ldots, g$,

iii) $\psi_k(r_{i,m})$ *denotes the Huber* ψ*-function with tuning-constant* k, $m = 1, \ldots, n_i$, $i = 1, \ldots, g$,

iv) *let* $\psi_k'(r_{i,m}) = 1$ *if* $|r_{i,m}| < k$ *and* 0 *otherwise,* $m = 1, \ldots, n_i$, $i = 1, \ldots, g$.

We write $\{u\}^{(s)}$ with $s \in \mathbb{N}_0^+$ to denote the value of u at the sth iteration. In addition, we will suppress the functional dependence on the iteration-step-specific variance components (e.g., δ instead of $\delta(\{\boldsymbol{\theta}\}^{(s)})$) when no confusion can arise.

5.4.2 Updating Equations

Equations (5.8) to (5.10) can be solved iteratively to obtain robust estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We write $\{u\}^{(s)}$ with $s \in \mathbb{N}_0^+$ to denote the value of u at the sth iteration.

We adopt a combined Newton-Raphson (NR) and Fisher-scoring (FS) algorithm. For uniformity of display, we shall introduce some notation.

Theorem 25. *Suppose model* $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ *and Definition 24. Then, for* $\{\boldsymbol{\beta}\}^{(t+1)}$ *sufficiently close to* $\{\boldsymbol{\beta}\}^{(t)}$, *the updating equation for* $\boldsymbol{\beta}$ *is*

$$\begin{aligned}\{\boldsymbol{\beta}\}^{(t+1)} &= \{\boldsymbol{\beta}\}^{(t)} + \left[\sum_{i=1}^{g} \delta\big(\mathbf{S}_{x_i x_i} - \tilde{\mathbf{S}}_{x_i x_i}(\{\boldsymbol{\beta}\}^{(t)})\big) - \omega_i \mathbf{t}_{x_i}\tilde{\mathbf{t}}_{x_i}(\{\boldsymbol{\beta}\}^{(t)})\right]^{-1} \times \\ &\quad \times \left[\sum_{i=1}^{g} \xi_i \mathbf{X}_i^T \boldsymbol{\psi}_k(\mathbf{r}_i) + \eta_i \mathbf{t}_{x_i} t_{\psi_i}\right],\end{aligned}$$

where $\mathbf{t}_{x_i} = (t_{x_i,\cdot 1}, \ldots, t_{x_i,\cdot q})^T$, *with* $t_{x_i,\cdot j} = \sum_{m=1}^{n_i} x_{i,mj}$, *and* $\tilde{\mathbf{t}}_{x_i} = (\tilde{t}_{x_i,\cdot 1}, \ldots, \tilde{t}_{x_i,\cdot q})^T$, *with* $\tilde{t}_{x_i,\cdot j} = \sum_{m=1}^{n_i} x_{i,mj}\mathbb{1}\{\psi(r_{i,m}) \geq k\}$, $j = 1, \ldots, q$, $i = 1, \ldots, g$; $\mathbf{S}_{x_i x_i} = \mathbf{X}_i^T\mathbf{X}_i$, *and* $\tilde{\mathbf{S}}_{x_i x_i}(\boldsymbol{\beta}) = \tilde{\mathbf{X}}_i(\boldsymbol{\beta})^T\tilde{\mathbf{X}}_i(\boldsymbol{\beta})$, *where* $\tilde{\mathbf{X}}_i(\boldsymbol{\beta}) = \big(x_{i,mj}\big)_{m\in\mathcal{K}_i}$ *with* $\mathcal{K}_i = \{m : m = 1, \ldots, n_i; \psi_k(r_{i,m}) \geq k\}$, *which denotes the matrix* $\mathbf{X}_i$ *having removed the non-outlying observations,* $i = 1, \ldots, g$; $t_{\psi_i} = \sum_{m=1}^{n_i} \psi_k(r_{i,m})$, $i = 1, \ldots, g$.

Proof. The assertion follows on simplifying the first-order Taylor expansion of the BIEE around $\{\boldsymbol{\beta}\}^{(t)}$ (which follows from the Hessian in WELSH and RICHARDSON (1997, p.363)); see SCHOCH (2011) for details. □

Note that one has to compute $\mathbf{S}_{x_i x_i}$ and $\mathbf{t}_{x_i}$ only once; they remain the same for each iteration step. In contrast, $\tilde{\mathbf{t}}_{x_i}$ and $\tilde{\mathbf{S}}_{x_i x_i}$ may change during iteration. However, since they operate exclusively on outlying observations, which is a usually very small number, computation is very fast.

Similarly, we shall derive updating equations for the variance components. To this end, recall (5.9) and (5.10), the BIEE of σ_c^2 and σ_1^2, respectively. In the following theorem, we derive the NR updating equations based on $S' = \partial S/\partial\sigma_c^2$ and $M' = \partial M/\partial\sigma_1^2$. In addition, we obtain $C = \partial M'/\partial\sigma_c^2 \equiv \partial S'/\partial\sigma_1^2$.

Theorem 26. *Suppose model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ and Definition 24. Let M and S denote the BIEE of σ_1^2 and σ_c^2, respectively, and put $\gamma_i = \delta - n_i\omega_i$. The updating equations write*

$$\{\boldsymbol{\theta}\}^{(t+1)} = \{\boldsymbol{\theta}\}^{(t)} - \mathbf{J}\big(\boldsymbol{\theta}^{(t)}\big)^{-1}\mathbf{g}\big(\boldsymbol{\theta}^{(t)}\big) \tag{5.19}$$

with

$$\mathbf{g}\big(\boldsymbol{\theta}^{(t)}\big) = (S, M)^T, \quad \mathbf{J}\big(\boldsymbol{\theta}^{(t)}\big) = \begin{pmatrix} S' & C \\ C & M' \end{pmatrix}, \tag{5.20}$$

where

$$\begin{aligned} M &= (1/2)\sum_{i=1}^{g}\left\{\gamma_i\left[\left(\sum_{m=1}^{n_i}\psi_k(r_{i,m})\right)^2 - \kappa n_i\right]\right\}, \\ M' &= (1/4)\sum_{i=1}^{g}\left\{-2n_i\gamma_i^2\left(\sum_{m=1}^{n_i}\psi_k(r_{i,m})\right)^2 - \right. \\ &\quad \left. -2\gamma_i^2\left(\sum_{m=1}^{n_i} r_{i,m}\right)\left(\sum_{m=1}^{n_i}\psi_k(r_{i,m})\right)\left(\sum_{m=1}^{n_i}\psi_k'(r_{i,m})\right) + \kappa n_i^2\gamma_i^2\right\}, \end{aligned}$$

and

$$S = (1/2)\sum_{i=1}^{g}\left\{\delta\sum_{m=1}^{n_i}\psi_k(r_{i,m})^2 - \omega_i\left(\sum_{m=1}^{n_i}\psi_k(r_{i,m})\right)^2 - \kappa n_i(\delta - \omega_i)\right\},$$

$$\begin{aligned} S' &= (1/4)\sum_{i=1}^{g}\left\{-2\delta^2\sum_{m=1}^{n_i}\psi_k(r_{i,m})^2 - 2\omega_i(n_i\omega_i - 2\delta)\left(\sum_{m=1}^{n_i}\psi(r_{i,m})\right)^2 - \right. \\ &\quad -2\delta^2\sum_{m=1}^{n_i} r_{i,m}\psi_k(r_{i,m})\psi_k'(r_{i,m}) + 2\delta\omega_i\left(\sum_{m=1}^{n_i}\psi_k(r_{i,m})\right)\left(\sum_{m=1}^{n_i} r_{i,m}\psi_k'(r_{i,m})\right) + \\ &\quad +2\delta\omega_i\left(\sum_{m=1}^{n_i} r_{i,m}\right)\left(\sum_{m=1}^{n_i}\psi_k(r_{i,m})\psi_k'(r_{i,m})\right) - 2\omega_i^2\left(\sum_{m=1}^{n_i} r_{i,m}\right)\left(\sum_{m=1}^{n_i}\psi_k(r_{i,m})\right) \times \\ &\quad \left. \times\left(\sum_{m=1}^{n_i}\psi_k'(r_{i,m})\right) + n_i\kappa\big[\delta^2 + \omega_i(n_i\omega_i - 2\delta)\big]\right\}, \end{aligned}$$

and

$$C = (1/4)\sum_{i=1}^{g}\Bigg\{-2\gamma_i^2\Bigg(\sum_{m=1}^{n_i}\psi_k(r_{i,m})\Bigg)^2 - 2\gamma_i\Bigg(\delta\sum_{m=1}^{n_i}\psi_k(r_{i,m})\psi_k'(r_{i,m}) - \\ -\omega_i\sum_{m=1}^{n_i}\psi_k(r_{i,m})\sum_{m=1}^{n_i}\psi_k'(r_{i,m})\Bigg)\Bigg(\sum_{m=1}^{n_1}r_{i,m}\Bigg) + \kappa n_i\gamma_i^2\Bigg\},$$

Proof. The assertions follow on simplifying the first-order Taylor expansion of the BIEEs (which follow from the Hessian in WELSH and RICHARDSON (1997, p.363)); see SCHOCH (2011) for details. □

Note that from a computational perspective, certain (iteration step-specific) terms e.g., $\sum_{m=1}^{n_i}\psi_k(r_{i,m})$, have to be computed once, but can be used in the updating equations for both random-effect variances. In general, the computation of the iteration steps is not very involved, since it barely consists of evaluating sums.

5.4.3 Algorithm

The hitherto derived Newton-Raphson updating equations can be used to iteratively obtain robust estimates of the model parameters. However, it is usually helpful to start the algorithm with a couple of Fisher-scoring steps. When the estimates are sufficiently stable, the algorithm switches to Newton-Raphson. This procedure is recommended for two reasons. First, Fisher-scoring is more robust to poor starting values. Second, to avoid the computational burden (and associated potential numerical instability) of computing the second-derivative matrix the Hessian is replaced by the Fisher information matrix.

Denote by $\boldsymbol{\mathcal{I}}(\boldsymbol{\tau})$ the information matrix with (i,j)th element $\mathcal{I}_{ij}(\boldsymbol{\tau}) = -\mathbb{E}[H_{ij}(\boldsymbol{\tau})]$. The main advantage of the information over the Hessian matrix is that large sections of its elements are zero. In particular, we have

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\tau}) = \begin{bmatrix}\mathbf{B}(\boldsymbol{\beta}) & \mathbf{0}^T\\ \mathbf{0} & \mathbf{T}(\boldsymbol{\theta})\end{bmatrix}, \tag{5.21}$$

where $\mathbf{0}$ is an $(2\times q)$ matrix of zeros; $\mathbf{T}(\boldsymbol{\theta}) = -\mathbb{E}[\mathbf{J}(\boldsymbol{\theta})]$ with $\mathbf{J}(\boldsymbol{\theta})$ from (5.20); $\mathbf{B}(\boldsymbol{\beta})$ is the sub-information matrix of $\boldsymbol{\beta}$. For as $\boldsymbol{\mathcal{I}}$ is block-diagonal, the fixed-effects and the random-effect variances can be updated separately. After a few Fisher-scoring iterations the algorithm switches, as previously stated, to Newton-Raphson. However, we stick to updating the estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ separately for reasons of numerical stability, although the Hessian has no block-diagonal structure.

5.4.4 Initialization and choice of starting values

Up to now, we discussed the derivation of updating equations to iteratively solve the system of BIEE assuming that there is a set of starting values, $\{\boldsymbol{\beta}\}^{(0)}$ and $\{\boldsymbol{\theta}\}^{(0)}$. The

choice of initial values is crucial in terms of robustness of the fully iterated estimates and the number of iteration until convergence (if it converges). We propose to choose them by an educated guess. Alternatively one may choose a very robust initial estimate (but with low efficiency), such as LTS (cf. Maronna et al., 2006) for the fixed effects and a MAD for the random-effect variances; cf. Stahel and Welsh (1997).

Given reasonable candidates of starting values, one may directly solve the updating equations, or with very large data sets one may first subsample the data using a stratified design where stratification is according to the clusters.[1] In particular, this step consists of drawing samples of equal size from the strata and computing robust estimates of the parameters using the much simpler BIEE for balanced data (cf. Stahel and Welsh, 1997). In addition, the size of the subsamples may be enlarged during these burn-in steps. When the initial estimates have reached some numerical stability, the algorithm operates on the full sample.

5.4.5 Prediction

Based on the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$, one then considers predicting the area-specific random effects $\mathbf{v}_i$ according to the BLUP theory. For convenience, we repeat the BLUP predicting equation (5.6)

$$\mathbb{E}[\mathbf{v}_i|\mathbf{y}_i] = \mathbf{G}_i(\tilde{\boldsymbol{\theta}})\mathbf{Z}_i^T\mathbf{V}_i(\tilde{\boldsymbol{\theta}})^{-1}\big[\mathbf{y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}\big] =: \tilde{\mathbf{v}}_i, \quad i = 1, \ldots, g. \tag{5.6}$$

From this result, we notice that outyling observations in $\mathbf{y}_i$ may influence the predictions even if $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}}$ have been replaced by some robust estimates $\hat{\boldsymbol{\beta}}^R$ and $\hat{\boldsymbol{\theta}}^R$, respectively. Therefore, Sinha and Rao (2009) propose to solve Fellner's (1986) robust mixed-model equations for $\mathbf{v}_i$

$$\mathbf{J}_i\mathbf{R}_i^{-1/2}\boldsymbol{\psi}\big(\mathbf{R}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{J}_i\mathbf{v}_i)\big) - \mathbf{G}_i^{-1/2}\boldsymbol{\psi}\big(\mathbf{G}_i^{-1/2}\mathbf{v}_i\big) = \mathbf{0}, \quad i = 1, \ldots, g, \tag{5.22}$$

where $\mathbf{G}_i$ and $\mathbf{R}_i$ follow from $\mathbf{V}_i = \mathbf{R}_i + \mathbf{Z}_i^T\mathbf{G}_i\mathbf{Z}_i$. Sinha and Rao (2009) propose to solve (5.22) by a Newton-Raphson algorithm based on updating equations that are obtained by another first-order Taylor series expansion. Consequently, computation is very involved. However, one can obtain robust predictions far more easily; see Schoch (2011).

Proposition 27. *Suppose the robust estimates $\hat{\boldsymbol{\beta}}^R$ and $\hat{\boldsymbol{\theta}}^R$. Let $\psi_c(u)$ denote the Huber ψ-function indexed by the robustness-tuning constant c. Then, robust predictions of the*

[1] Note that the cluster with the smallest sample size determines the maximum size of the subsamples. For typical SAE problems, this constraint is no a serious limitation since the clusters' size is in general very large.

random effects are obtained as solutions of the bounded-influence predicting equations (BIPE)

$$\tilde{\mathbf{v}}_i^R = \kappa \mathbf{G}_i(\hat{\boldsymbol{\theta}}^R)\mathbf{Z}_i^T\mathbf{V}_i(\hat{\boldsymbol{\theta}}^R)^{-1/2}\boldsymbol{\psi}_c\big[\mathbf{V}_i(\hat{\boldsymbol{\theta}}^R)^{-1/2}[\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^R]\big], \quad i = 1, \ldots, g, \tag{5.23}$$

where

$$\kappa = \big[-2c\phi(c) + 2\Phi(c) - 1 + 2c^2\big(1 - \Phi(c)\big)\big]^{-1/2}, \tag{5.24}$$

with $\phi(u)$ and $\Phi(u)$ the pdf and cdf of the standard normal distribution, respectively. Note that κ is kind of a consistency correction term which has been chosen in order $\tilde{\mathbf{v}}_i^R$ to behave similarly to $\tilde{\mathbf{v}}_i$ at the core model. In particular, we impose the (implicit) moment conditions that $\mathbb{E}[\tilde{\mathbf{v}}_i^R] = 0$ and $\mathbb{V}[\hat{\mathbf{v}}_i^R] = \mathbb{V}[\tilde{\mathbf{v}}_i]$.

Proof. Let $\mathbf{z}_i = \mathbf{V}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$ denote the standardized residual. The condition $\mathbb{E}[\hat{\mathbf{v}}^R] = \mathbf{0}$ follows immediately noting that at the core model it holds that $\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{I}_i)$, $i = 1, \ldots, g$. Then, it easy to prove that $\mathbb{E}[\boldsymbol{\psi}_c(\mathbf{z}_i)] = \mathbf{0}$ $\forall i = 1, \ldots, g$ and thus $\mathbb{E}[\boldsymbol{\psi}_c(\mathbf{z})] = \mathbf{0}$. As a result of $\mathbb{E}[\boldsymbol{\psi}_c(\mathbf{z})] = \mathbf{0}$, the variance at the core model simplifies to $\mathbb{E}[\boldsymbol{\psi}_c(\mathbf{z})\boldsymbol{\psi}_c(\mathbf{z})^T]$. For the Huber ψ-function, we obtain $\kappa = \big[-2c\phi(c) + 2\Phi(c) - 1 + 2c^2\big(1 - \Phi(c)\big)\big]^{-1/2}$. □

Note that the derivation of the correction factor κ does not take into account that the estimation of $\boldsymbol{\beta}$ itself has a contribution to the variance of $\mathbf{v}_i$. In general, this leads to an overestimation of the variance. However, since the dominating term of the more sophisticated correction factor is p/n, most SAE applications render this term negligible since $p \ll n$ with p the number of fixed effects; see (SCHOCH, 2011).

All in all, the robust prediction of the area-specific random effects is computationally very simple.

5.5 Conclusion

We showed that robust estimates of the parameters of the basic unit-level model can be obtained in a computationally very efficient manner. The proposed algorithm usually converges in a couple of seconds on a standard personal computer, even for very large samples. Furthermore, we showed that the robust prediction of the random effects seems to be almost instantaneous, since no iteratively computation is required. Mean squared error estimation, on the other hand, remains an open problem (cf. SINHA and RAO, 2009).

Bibliography

Battese, G. E., **Harter, R. M.** and **Fuller, W. A.** (**1988**): *An error component model for prediction of county crop areas using.* Journal of the American Statistical Association, 83, pp. 28–36.

Fellner, W. (**1986**): *Robust estimation of variance components*. Technometrics, 28, pp. 51–60.

Gentle, J. E. (**2007**): Matrix Algebra. Theory, Computations, and Applications in Statistics. New York: Springer.

Huggins, R. (**1993**): *On the robust analysis of variance component models for pedigree data*. Australian Journal of Statistics, 35, pp. 43–58.

LaMotte, L. R. (**1972**): *Notes on the covariance matrix of a random, nested ANOVA model*. The Annals of Mathematical Statistics, 43, pp. 659–662.

Magnus, J. and **Neudecker, H.** (**1999**): Matrix Differential Calculus: With Applications in Statistics and Econometrics. Hoboken: John Wiley & Sons, 2nd and rev. ed.

Maronna, R. A., **Martin, D.** and **Yohai, V. J.** (**2006**): Robust Statistics: Theory and Methods. Chichester: John Wiley.

Rao, J. (**2003**): Small Area Estimation. (Wiley Series in Probability and Statistics), Hoboken: Wiley.

Richardson, A. M. and **Welsh, A. H.** (**1995**): *Robust Restricted Maximum Likelihood in Mixed Linear Models*. Biometrics, 51 (4), pp. 1429–1439.

Schoch, T. (**2011**): Robust Basic Unit-Level Small Area Model. Working Paper, University of Zurich.

Searle, S., **Casella, G.** and **McCulloch, C. E.** (**1992**): Variance Components. Hoboken: John Wiley & Sons.

Searle, S. R. and **Henderson, H. V.** (**1979**): *Dispersion Matrices for Variance Components Models*. Journal of the American Statistical Association, 74, pp. 465–470.

Sinha, S. K. and **Rao, J.** (**2009**): *Robust small area estimation*. The Canadian Journal of Statistics, 37 (3), pp. 381—399.

Stahel, W. A. and **Welsh, A. H.** (**1997**): *Approaches to robust estimation in the simplest variance components model*. Journal of Statistical Planning and Inference, 57, pp. 295–319.

Welsh, A. H. and **Richardson, A. M.** (**1997**): *Approaches to the Robust Estimation of Mixed Models*. **Maddala, G.** and **Rao, C.** (editors) Robust Inference, *Handbook of Statistics*, vol. 13, chapter 13, pp. 343–384, Amsterdam: Elsevier Science.

Part II

Robust Multivariate Methods for Incomplete Income Data

Chapter 6

Robust Multivariate Methods: An Overview

Univariate outlier-detection and robust estimation methods are well established. However, the univariate income variable is usually an aggregation of several distinct income sources or components (e.g., employee cash income, capital income, unemployment benefits, etc.). Notably outliers in the income components may be propagated or masked during the aggregation process, which renders outlier detection in the univariate income variable very difficult if not infeasible. Furthermore, multivariate outliers in the income components can seriously affect the estimates of the income inequality measures.

Therefore we propose to adopt multivariate outlier-detection methods directly on the joint distribution of the income components. The observations thus detected as outliers are subsequently imputed using a robustly fitted data model. Both outlier-detection- and imputation methods are adapted for the finite population sampling context and can cope with missing values.

The income components relate to the personal and the household level. Those components that are measured at the household level, are expressed as per-capita numbers and attached to the particular household members. Consequently, analysis can be conducted with individual-level data only. The multidimensional space of income components has the following characteristics:

- the marginal distribution of each component is very skewed and features a remarkable point mass at zero (zero-inflation),
- the joint distribution of the components is far from being elliptically contoured,
- an overwhelming majority of observations lies in subspaces, i.e., exhibits intrinsic zeros on certain dimension (e.g., individuals on working age with a positive employee-cash income do ordinarily neither receive old-age- nor unemployment benefits, and vice versa),
- within subspaces, the observations are clustered with respect to non-monetary, socio-economic characteristics.

All Multivariate Outlier Detection and Imputation (MODI) methods in this paper consist of, at least, three computation steps that depend upon another. The main steps are as follows. The first step is about data preparation such as transformations, treatment of zero, missing, and negative income values. This is followed by an outlier detection procedure. Finally, an imputation step replaces outlying and missing observations with realizations from a robustly estimated data model (or non-outlying donors).

The remainder of this part is organized as follows. In Chapter 7, we derive a typology of multivariate outlyingness and contamination mechanisms. The remaining chapters, 9 to 11, are devoted to the outlier detection and imputation methods. The MODI methods have been classified according to their requirements concerning the structure of the data.

6.1 Aggregation of the Income Components

The 2006 EU-SILC exercise collected data on 18 income components at household- and 14 at the individual level (only net incomes components considered). The resulting number of 32 variables is far too large to be fed directly into any outlier-detection method. In addition to the usual dimensional restrictions of the methods, zero-inflation in multiple dimensions and missing values evidently limit the number of feasible dimensions. The non-elliptically symmetric distribution, on the other hand, and potentially negative observation do not constitute another restriction to the dimensionality problem, since both phenomena can be treated by appropriate transformations.

With multivariate income data of this complexity, one has to rely on subject matter knowledge to reduce the dimensions while holding the loss of information minimal. It comes without saying that a large number of different dimension-reducing strategies may be applied. In Table 6.1, we report 16 income components from the synthetic population A-AT-SILC (TEMPL et al., 2011) and a proposal how to aggregate them. The motivation of this proposal is (1) to separate individual- from household-level variables, (2) to distinguish between income sources that refer to working life, purely capital income and social transfers, (3) to pool income sources of the same kind but with a large amount of zeros (e.g., variable survivor benefits, PY110n, features 99.34% zero observations), and (4) to keep the reduction-strategy sufficiently generalizable and applicable to a broad range of different populations. Nevertheless, we may possibly find better reduction strategies when optimizing for a single dataset of a single country (perhaps at risk of overfitting) , but discarding generalisability. Further, all household-level variables are expressed as per capita numbers. Thus, we study all components at the individual level.

Bibliography

Templ, M., **Alfons, A.**, **Filzmoser, P.**, **Kraft, S.**, **Hulliger, B.**, **Kolb, J.-P.** and **Münnich, R.** (**2011**): *Design of the simulation study and simulation environments.* Technical report, AMELI deliverable D6.1.
URL http://ameli.surveystatistics.net/

Table 6.1: Aggregation of the income components

workinc	=	PY010n *(employee cash or near cash income)*	+	PY050n *(cash benefit or losses from self-employment)*		
capinc	=	[HY040n *(income from rental of a property or land)*	+	HY090n]/hhsize *(interests, dividends, profit from capital investments in unincorporated business)*		
transh	=	[HY050n *(familiy/ children related allowances)*	+	HY110n *(income received by people aged under 16)*	+	HY070n + *(housing allowances)*
	+	HY080n *(inter-household cash transfers received)*	–	HY130n *(inter-household cash transfers paid)*	–	HY145n]/hhsize *(payments/receipts for tax adjustments)*
transp	=	PY090n *(unemployment benefits)*	+	PY110n *(survivor benefits)*	+	PY130n *(disability benefits)*
	+	PY100n *(old-age benefits)*	+	PY120n *(sickness benefits)*	+	PY140n *(education related allowances)*

Notes: variable names according to EUROSTAT definition. Data is from the synthetic A-AT-SILC population, see TEMPL et al. (2011).

Chapter 7

Multivariate Outliers

7.1 Introduction

Several of the Laeken indicators are based on equivalized income, which itself is a function of many income components, some of them may be negative, some of them are related to persons, others to the household. In addition most of these components have a semi-continuous distribution with a point mass at zero. Robustness of these income based indicators can be achieved through direct robustification of the indicators. For example, the Quintile Share Ratio of the income, an inequality measure, can be robustified (Hulliger and Schoch, 2009). However, a more adequate robustification may be achieved by a multivariate robust imputation of the income components. This avenue is explored in this article.

Detection of multivariate outliers in survey data with missing values has been treated in the literature and some experience with applications exist (Little and Smith, 1987; Maronna and Zamar, 2002; Chambers et al., 2004; Béguin and Hulliger, 2004; Ghosh-Dastidar and Schafer, 2006; Hulliger, 2006; Hulliger and Kilchmann, 2006). After the detection of outliers and influential observations, these suspicious observations may be revised interactively (Luzi et al., 2007) and/or an imputation considering their special status is carried out. Some robust imputation methods have been described in the literature (Charlton, 2003; Hulliger, 2007). This section discusses the conditions for successful robust multivariate imputation methods.

7.2 Outlier-, contamination- and missingness-mechanisms

Béguin and Hulliger (2008) set up a simulation where first an experiment decides whether an observation is an outlier and second if the value of the outlier is determined under the contaminating distribution. Also Ghosh-Dastidar and Schafer (2006) and Little and Rubin (2002) consider a model where an indicator on outlyingness and an outlier distribution is combined. We try to formalize these approaches and suggest that the outlier mechanism follows a two-step procedure. Alqallaf et al. (2009) discuss related contamination models.

7.2.1 Notation

The ingredients of the mechanisms we investigate are the following:

- U is the set of elements in the population of size N. We usually use the index i to indicate the elements of U.
- Y_i^* is the true, complete data. For the description of the mechanisms, the true Y_i^* are considered random variables which follow a superpopulation model. In the survey context we will fix them as one realisation of a superpopulation model. Any finite population characteristics would be a function of Y^*.
- Y_i is the observable data for unit i, Y_{ij} is its j-th component.
- R_{ij} a response indicator per observation i and variable j. Given the vector R_i we can split the observable data Y_i into an actually observed part Y_{io} and a missing part Y_{im}. We then write $Y_i = (Y_{io}, Y_{im})$.
- Y_{ci} is the contaminated data for unit i.
- O_i is an outlier indicator. For those observations with $O_i = 1$ the true data Y_i^* is replaced by the contaminated data Y_{ci}. Thus for $O_i = 1$ the observed data Y_i actually consists of Y_{ci} and for $O_i = 0$ we have $Y_i = Y_i^*$.
- S_i is the sample indicator.
- X_i are covariables which are fully observed.
- Z_i are unobserved covariables.

In Figure 7.1 a possible mechanism of contamination is shown. The outlier mechanism for representative outliers O_r is disinguished from the outlier mechanism for non-representative outliers O_n though in the following we do not use O_r and O_n is simply denoted O. In addition, the order of R, S and O_n is arbitrary and will be discussed later on. The parameter of interest is $\theta(Y^*)$, i.e. a characteristic of the data which contains already representative outliers. The outliers detection is denoted D and the result of detection is a prediction $\hat{O}$ based on the detection. Analogously the imputation yields a prediction $\hat{Y}$ of Y^* which is then used to make inference on the distribution of Y^*.

An important difference between R_i and O_i is that O_i is not directly observable. Thus the outlier indicator O_i is a latent variable. While R_i is a vector indicating for each variable j whether it is observed or not, O_i is a scalar. In other words, R_i indicates item response while O_i indicates unit outlyingness.

In the next section (7.2.2) we investigate outlyingness, contamination and missingness from the point of view of the mechanisms that create them. This step is necessary to develop an appropriate simulation setup. In the following section (7.2.3) we derive conditions that would allow proper inference for the interesting parameters of Y^*.

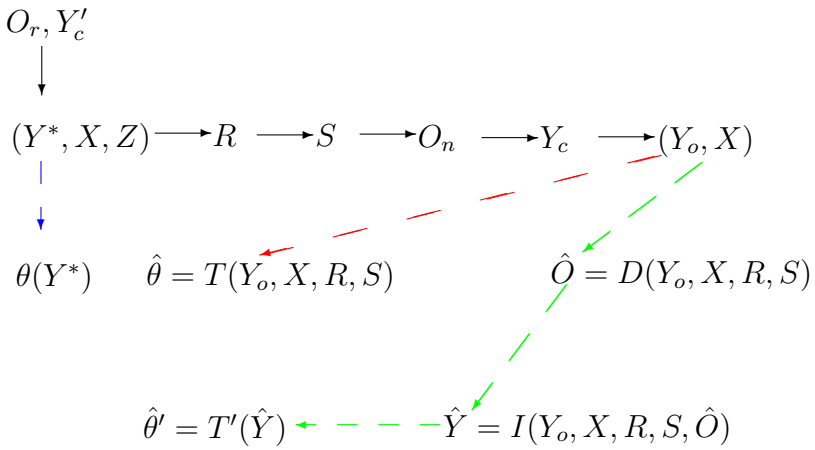


Figure 7.1: Outlier, contamination and missingness mechanisms which lead to the observed data

7.2.2 Mechanisms

From the start we assume that the sampling mechanism is ignorable and we omit it from the notation in this section. However, the models we are referring to always pertain to the population and the sample design should be taken into account when estimating and testing the model parameters.

We will not consider a process for representative outliers (Chambers, 1986). In other words we consider Y_i^* our starting point and assume therefore that representative outliers have been created already before and integrated into Y_i^*. Thus the outlier indicator O_i refers to non-representative outliers only.

We assume that no further error mechanism is disturbing the observable data than the outlier process, though we could add a general measurement error mechanism. Such mechanisms and the treatment of errors in a general data preparation system are discussed in Luzi et al. (2007).

The full density of all the above variables can be written as:

$$f(Y^*, Y_c, X, Z, R, S, O; \Xi), \tag{7.1}$$

where Ξ is the set of parameters. Note that for ease of notation and because we refer to random variables we usually leave away the subscripts i in Y_i etc. in the further development. There are many possible factorisations of this distribution. We will consider the following decomposition

$$\begin{aligned} &f(Y^*, Y_c, X, Z, R, S, O; \Xi) = \\ &\quad f(Z; \xi_Z) f(X|Z; \xi_X) f(Y^*|X, Z; \xi_*) \\ &\quad f(O|Y^*, X, Z; \xi_O) f(Y_c|Y^*, O, X, Z; \xi_c) \\ &\quad f(S|O, Y^*, Y_c, X, Z; \xi_S) f(R|S, O, Y^*, Y_c, X, Z; \xi_R), \end{aligned} \tag{7.2}$$

In this factorization we implicitely assume that the dependence of the truth Y^* on the outlier mechanism O is not important, while the dependence of the outlier mechanism on the truth is. In other words, the outlier mechanism is seen as noise and we are not interested in estimating its parameters ξ_O, while we are interested in ξ_*, the parameters of the distribution of the truth Y^*. Note that contrary to Figure 7.1 R and S occur after O and Y_c.

It seems unnecessary to let depend the contamination Y_c on the outlier indicator O in addition to its dependence on the true Y^* and thus we may assume $f(Y_c|Y^*, O, X, Z; \xi_c) = f(Y_c|Y^*, X, Z; \xi_c)$.

It may well be that the contamination depends on covariables X. A convenient assumption is that no other dependency exists, i.e.

$$f(Y_c|Y^*, O, X, Z; \xi_c) = f(Y_c|X; \xi_c'). \tag{7.3}$$

We call this situation **contaminated at random** (CAR) in the narrow sense. In Section 7.2.3 we will relax the condition somewhat by allowing a dependence on good observed data Y_{og}. There are many possibilites how X may affect the contamination. Simple cases are when the contamination is more heavy-tailed for self-employed or for retired persons than for employed. The simplest situation is, of course, when the contamination does not depend on any other variables, i.e.

$$f(Y_c|Y^*, O, X, Z; \xi_c) = f(Y_c; \xi_c''). \tag{7.4}$$

We call this situation **contaminated completely at random** (**CCAR**). Obviously this may not be a realistic assumption. However, most models in classical robust statistics assume a contamination which is CCAR.

The distribution of Y_c may depend on Y^*, of course. For example the well-known thousands-error, i.e. $Y_c = 1000 \cdot Y^*$ is such a case. This contamination may be non-ignorable if no observed proxy for Y^* helps to detect the contamination mechanism, e.g. when the outlier has missing values in the outlying components.

We think of the contamination Y_c as a potential contamination which is plugged into the data only when the outlier indicator $O = 1$.

What we actually would observe in the case of no missing values is a mixture of Y^* and Y_c, where the mixing is due to the outlier mechanism:

$$Y = (1 - O)Y^* + OY_c. \tag{7.5}$$

We call Y observable and not observed because we actually observe Y only if there is no missingness. The observable variable Y, conditional on O, X, Z, has a density

$$f(Y|O, X, Z; \xi_Y) = (1 - O)f(Y^*|X, Z; \xi_*) + Of(Y_c|Y^*, X, Z; \xi_c). \tag{7.6}$$

This is a mixture of the distribution of the true data and the distribution of the contamination as in the well known contamination model from classical robustness theory.

However, the conditioning on O hides a complication, i.e. that the outlier mechanism O may depend on Y^*.

We assume that the sample design is independent of all other variables except possibly X, i.e. $f(S|R, Y^*, Y_c, X, Z) = f(S|X)$. In other words the fully observed covariables X contain all design information. We assume further that the rest of the involved random variables is independent of S. In other words we assume S is ignorable for inference on Y^*.

We call a mechanism missing at random (MAR) in the narrow sense if $f(R|Y^*, Y_c, X, Z; \xi_R) = f(R|X, Z; \xi_R')$. This assumption is strong in what concerns the outlyingness. It may be more realistic to assume that R depends on representative outliers in Y^* or even on O. For example, a true outlier in the population like a self-employed person who has been exceptionally successful, may not participate in the survey and it may happen that this behaviour cannot be modelled through X. It may be more reasonable that the response R does not depend on the contamination because, in the case of non-representative outliers, the contamination is in fact a measurement error and may have no effect on the response. Note that even this assumption may not hold in reality because it may happen that a person filling in a paper questionnaire realises that he or she made a mistake and decides not to respond as a consequence. This latter scenario should, in principle, not happen for the SILC survey since the survey instrument is administered through personal or telephone interviews.

In analogy with MAR and similarly to R we assume that the outlier indicator O should not depend on unobserved observation:

$$f(O|R, S, Y^*, Y_c, X, Z; \xi_O') = f(O|X, Z; \xi_O{}'') . \tag{7.7}$$

We call this situation **outlying at random** (OAR) in a narrow sense (Béguin and Hulliger, 2008). Particularly in a case where $O = 1$, such that the observable $Y = Y_c$ it becomes obvious that this is a strong assumption. Even stronger, of course, is the assumption that the outlier mechanism does not depend on any other variable, i.e.

$$f(O|R, S, Y^*, Y_c, X, Z; \xi_O) = f(O; \xi_O''') . \tag{7.8}$$

This latter mechanism is called, in analogy to MCAR, **outlying completely at random** (OCAR) (Béguin and Hulliger, 2008).

For a simulation set-up we can now formulate mechanisms, which are able to capture interesting dependencies, though not all dependencies, of course. We may denote the situtations where no MAR missingness, or no OAR outlyingness or no CAR contamination holds as MNAR, ONAR, CNAR. And we may denote the situation where no missingness occurs at all as NoR and where no outlyingness, and therefore no contamination, at all occurs with NoC. Then the feasible and interesting simulations would use the following scheme of crossings:

$$(NoR, MCAR, MAR, MNAR) \tag{7.9}$$
$$\times$$

$$(NoC, (OCAR, OAR, ONAR) \times (CCAR, CAR, CNAR)). \tag{7.10}$$

For example the likelihood under MAR, OAR, CAR in the narrow sense and assuming no dependence on unobserved covariates Z would be

$$\begin{aligned} &f(Y^*, Y_c, X, R, S, O; \Xi) = \\ &\quad f(X|Z;\xi_X) f(Y^*|X, Z;\xi_*) \\ &\quad f(O|X;\xi_O) f(Y_c|X;\xi_c) \\ &\quad f(S|X;\xi_S) f(R|X;\xi_R). \end{aligned} \tag{7.11}$$

In other words the full likelihood factors in marginal densities which all are only conditional on observed covariates X. Of course this is an oversimplification of the full likelihood. However, the mechanisms implied are already much more complicated than the classical robust contamination model which does not consider missingness and assumes OCAR and CCAR.

7.2.3 Inference

In analogy to the development by SCHAFER (1997) we look for a possibility of inference on the parameter ξ_* governing the conditional distribution $Y^*|X$. We assume now, in addition, that no unobserved variables Z influence the rest of the variables.

For the moment we assume that the outlier indicator O is observed like the response indicator R. The observable data Y is then partitioned into $Y = (Y_{og}, Y_{oc}, Y_{mg}, Y_{mc})$, where Y_{og} is the observed good (not contaminated) data, Y_{oc} is the observed contaminated data, Y_{mg} is the missing good data and Y_{mc} denotes the part of the data which is missing and contaminated. The observed good data Y_{og} is the part of the data where $O_i = 0$ and $R_{ij} = 1$, or shorter where $(1 - O_i)R_{ij} = 0$. We concentrate on Y_{og} since we cannot rely on Y_{oc} the observed contaminated data for inference on ξ_*. Note that in very special situations we may use the contaminated data for inference. This happens for example when the true data and the contamination have a normal distribution with mean μ and the contamination distribution differs only by a larger variance from the distribution of the true data. We may then estimate μ from the contaminated distribution, too, though with a large variance. This is the assumption in the classical contamination model. The observed good data likelihood is obtained as follows

$$f(Y_{og}, X, Z, R, O; \xi) = \int f(Y, R, O, X; \theta) dY_m dY_{oc} \tag{7.12}$$

We see that the marginal distribution of Y_{og} cannot be derived from this density if O is latent. The condition that O is observable is needed to integrate over Y_{oc}, the contaminated observed data.

The likelihood of the observed good data Y_{og} together with the indicators R and O can be written as

$$f(Y_{og}, R, O|X; \xi'_*, \xi_R, \xi_O) = \int f(R, O|Y, X; \xi_R, \xi_O)\ f(Y|X; \xi''_*) dY_{oc} dY_m, \tag{7.13}$$

Note that $Y_{og} = Y^*_{og}$ and therefore

$$f(Y_{og}, R, O|X; \xi'_*, \xi_R, \xi_O) = f(Y^*_{og}, R, O|X; \xi'_*, \xi_R, \xi_O).$$

For $O = 1$ Y_{og} is empty. In other words we integrate over all the components of Y and what remains is $f(R|X, \xi_R)$, the distribution of the response only. Thus we cannot use the outliers for inference on Y^*. When all data is missingg, i.e. $R_{ij} = 0, \forall j$ it remains $f(O|X, \xi_O)$ which again does not involve Y^* and, of course, is not helpful because it relies on the assumption that O_i is not latent.

In other words we do not consider the contaminated observations or the completely missing observations for inference. We can separate the likelihood (7.13) into two parts

$$f(Y^*_{og}, R, O|X; \xi'_*, \xi_R, \xi_O) = f(R, O|Y^*_{og}, X; \xi_R, \xi_O, \xi_*) \int f(Y|X; \xi_*) dY_{oc} dY_m \tag{7.14}$$

if the following condition holds:

$$f(R, O|Y, X; \xi_R, \xi_O) = f(R, O|Y^*_{og}, X; \xi_R, \xi_O). \tag{7.15}$$

This condition is ät randomcondition for the joint distribution of R and O. It is not in the narrow sense as in Section 7.2.2 because it does condition at least on part of the observed data in addition to X, namely on Y_{og}. We call this condition **outlying and missing at random** (OMAR). In other words, if we can assume that the indicators R and O do not depend on unobserved data or on data which is contaminated, then we can factor out the part of the likelihood which contains information on the parameter ξ_*. We then can make inference on ξ_* using just Y_{og}, which is Y^*_{og}.

We may add a further assumption on the conditional independence of R and O, namely

$$f(R, O|Y, X; \xi_R, \xi_O) = f(R|Y^*_{og}, X; \xi_R) f(O|Y^*_{og}, X; \xi_O). \tag{7.16}$$

Then R and O are OMAR if they are separately MAR and OAR (not in the narrow sense). Note that the condition MAR (not in the narrow sense) in a situation where we exclude outlyingness becomes the original condition of Little and Rubin (2002). In a situation with outlyingness and missingness we need condition OMAR for inference.

We now have to come back to the assumption that O is observable, which is needed for the derivation of the above results. Outlyingness is a latent fact and O must be estimated

from the data. Estimating O means that we try to detect the outliers. In order that the above derivation is applicable we need a perfect detection of the outliers or equivalently an estimation $\hat{O}_i$ which coincides with O_i for all i with $O_i = 1$. In other words, we must be sure that no outlier remains in Y_{og}. It may be inefficient to obtain $\hat{O}_i = 1$ if $O_i = 0$, i.e. to declare an observation an outlier though it is good. Thus we would discard observations which contain information on ξ_*.

Under what circumstances can we detect all outliers with certainty, or perfectly? The outliers must be detected based on Y_o whether good or contaminated. Perfect detection can be seen as a discrimination problem, where the number of "correct falsemust be 0. This is possible if the support of the contamination-distribution and the support of the true data is disjoint.

The conditions for the perfect detection of outliers are:

1. The distribution $f(Y^*|X, Z; \xi_*)$ must have a support which does not overlap with the support of $f(Y_c|Y^*, X, Z; \xi_c)$. Let $G^* = \{Y^* : f(Y^*|X, Z; \xi_*) > 0\}$ and $G^c = \{Y_c : f(Y_c|Y^*, X, Z; \xi_c) > 0\}$. The first condition is then $G^* \cap G^c = \{\}$.

   We may relax this condition somewhat in the sense, that only $G^* \setminus G^c \neq \{\}$. This means that there may be true observations in the outlier support but there must a part of the support of the true data where no contamination occurs. We may call $G^* \setminus G^c$ a safe support.

   Of course the support G^c cannot span the whole space because this would exclude the possibility of a safe support.

2. The response R must not set all information from the safe support $G^* \setminus G^c$ to missing, i.e. $P[R_{ij} = 1|Y^* \in G^* \setminus G^c] > 0$.

If both of these conditions hold we call the contamination **separable**.

Note that even if the contamination is separable, the region $G^* \setminus G^c$ or a proper subset of it must still be determined from the data by an outlier detection rule. Since this region may be complex there is still no guarantee to detect all outliers. Of course we may use a conservative rule for outlier nomination and thus minimise the risk of not detecting an outlier. This may cause a loss of efficiency since the usable data Y_{og} is reduced.

If for an outlier contamination occurs exclusively in components of Y which are missing, then the outlier can only be detected through covariate information in X. This is possible under the CAR setting but not in a CCAR setting. In any case, because we base our inference only on the observed part of such an observation the un-observed contamination should not harm the inference on ξ^*.

Using $\hat{O}_i$ instead of O_i in (7.14) we obtain $f(Y^*_{o\hat{g}}|X; \xi_*)$ for inference on ξ^*. Note the $\hat{}$ on g in the index, indicating that now we rely on $\hat{O}$ to determine which observations we rely on to make inference on ξ^*.

Since we may have restricted inference to $G^* \setminus G^c$ we will have to take this restriction into account when estimating ξ_*. If we postulate OAR and CAR then the contamination does

not depend on Y^* conditionally on X and thus there is no selection bias in considering only $G^* \setminus G^c$ for inference on Y^* if the inference is conditionally on X.

To understand the role of a safe region $G^* \setminus G^c$ better we can look at the problem of inliers. Inliers occur for example in a contaminaton model $(1-\epsilon)N(\mu, \sigma^2) + \epsilon N(\nu, \tau^2)$, which is OCAR-CCAR, when an observation stems from $N(\nu, \tau^2)$ but is sufficiently close to μ to have a high probability to stem from $N(\mu, \sigma^2)$. Such inliers cannot exist if the contamination is separable. In other words, there must be a region where no contamination occurs but good observations from the bulk of the data do occur. Obviously in the above contamination model this is not the case because true data and outliers have the same support, the whole real line (or the whole space). Suppose now, for example, that the true data is distributed with $N(\mu, \sigma^2)$ and the contamination is an exponential distribution with support $[x_0, \infty]$. The safe region then would be $[-\infty, x_0]$. For inference on (μ, σ^2) we would have to find a number $g \leq x_0$ and then we would work with the likelihood truncated at g. Even if g depends on a covariate X we must take the truncation into account.

Separable contamination is an extreme situation where we understand how inference on ξ_* is possible. In such a situation we can impute for the missing values and for the outliers from the distribution $f(Y^*|X; \hat{\xi}_*)$. In practice we seldom have such a nice situation and usually contamination is not separable, i.e. inliers may occur and we cannot ensure that observations which are not nominated outliers, i.e. with $\hat{O}_i = 0$, stem from the true distribution.

In conclusion we would need OMAR outlyingness and missingness and separable contamination to obtain a good imputation. But hope is small that these conditions really hold. Nevertheless these conditions are helpful for the study of detection and imputation procedures or robust estimators in a complex multivariate setting when missing values and outliers occur.

Bibliography

Alqallaf, F., **Aelst, S. V.**, **Yohai, V. J.** and **Zamar, R. H.** (**2009**): *Propagation of outliers in multivariate data*. The Annals of Statistics, 37, pp. 311–331.

Béguin, C. and **Hulliger, B.** (**2004**): *Multivariate Outlier Detection in Incomplete Survey Data: the Epidemic Algorithm and Transformed Rank Correlations*. Journal of the Royal Statistical Society, Series A: Statistics in Society, 167 (2), pp. 275–294.

Béguin, C. and **Hulliger, B.** (**2008**): *The BACON-EEM Algorithm for Multivariate Outlier Detection in Incomplete Survey Data*. Survey Methodology, Vol. 34, No. 1, pp. 91–103.

Chambers, R., **Hentges, A.** and **Zhao, X.** (**2004**): *Robust Automatic Methods for Outlier and Error Detection*. Journal of the Royal Statistical Society, Series A: Statistics in Society, 167 (2), pp. 323–339.

Chambers, R. L. (**1986**): *Outlier Robust Finite Population Estimation*. Journal of the American Statistical Association, 81 (396), pp. 1063–1069.

Charlton, J. (editor) (**2003**): Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project, vol. 1 and 2. EUREDIT consortium (published as web reference only), http://www.cs.york.ac.uk/euredit/results/results.html.

Ghosh-Dastidar, B. and **Schafer, J.** (**2006**): *Outlier Detection and Editing Procedures for Continuous Multivariate Data*. Journal of Official Statistics, Vol. 22, No. 3„ pp. 487–506.

Hulliger, B. (**2006**): *On Two Aspects of Outlier Treatment: Univariate vs. Multivariate Approach and Transformation of Data*. Proceedings of Q2006, the European Conference on Quality in Survey Statistics, Office for National Statistics UK and Eurostat.

Hulliger, B. (**2007**): *Multivariate Outlier Detection and Treatment in Business Surveys*. Proceedings of the III International Conference on Establishment Surveys, Montréal, American Statistical Association.

Hulliger, B. and **Kilchmann, D.** (**2006**): *HANDLING OF OUTLIERS AT SFSO*. Work Session on Statistical Data Editing, p. WP 31, Bonn: UN/ECE,CONFERENCE OF EUROPEAN STATISTICIANS.

Hulliger, B. and **Schoch, T.** (**2009**): *The Robustification of the Quintile Share Ratio*. **Eurostat** (editor) NTTS-Conference on New Techniques and Technologies for Statistics, Brussels 18-20 February 2009, Publications Office of the European Union.

Little, R. and **Rubin, D.** (**2002**): Statistical Analysis With Missing Data (2ed.). Wiley.

Little, R. and **Smith, P.** (**1987**): *Editing and imputation for quantitative survey data*. Journal of the American Statistical Association, 82, pp. 58–68.

Luzi, O., **De Waal, T.**, **Hulliger, B.**, **Di Zio, M.**, **Pannekoek, J.**, **Kilchmann, D.**, **Guarnera, U.**, **Hoogland, J.**, **Manzari, A.** and **Tempelman, C.** (**2007**): Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys. Italian Statistical Institute ISTAT„ institutions: ISTAT, CBS, SFSO, Eurostat.

Maronna, R. and **Zamar, R.** (**2002**): *Robust Estimates of Location and Dispersion for High-Dimensional Datasets*. Technometrics, 44 (4), pp. 307–317.

Schafer, J. L. (**1997**): Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC.

Chapter 8

EM-based Regression Imputation Using Robust Methods

Abstract Imputation of missing values is one of the major tasks for data pre-processing in many areas. Whenever imputation of data from official statistics, such as within the EU-SILC data, comes into mind, several (additional) challenges almost always arise, like large data sets, data sets consisting of a mixture of different variable types, or data outliers.

We summarize an automatic algorithm called `IRMI` for iterative model-based imputation using robust methods which was developed (and implemented and available in R) during the AMELI project and which encounters for the mentioned challenges. The corresponding software can be freely downloaded at http://cran.r-project.org/package=VIM.

The proposed algorithm is compared to the algorithm `IVEWARE`, which is the "recommended software" for imputations in international and national statistical institutions. Using artificial data and the EU-SILC data, the advantages of `IRMI` over `IVEWARE` – especially with respect to robustness – are demonstrated.

Keywords: EM-based regression imputation, robustness, R

8.1 Introduction

The imputation of missing values is especially important in official statistics, because virtually all data sets from this area deal with the problem of missing information due to non-responses, or because erroneous values have been set to missing. This has especially consequences for statistical methods using the multivariate data information. The naive approach, namely omitting all observations that include at least one missing cell, is not attractive because a lot of valuable information might still be contained in these observations. On the other hand, omitting observations may only lead to non-biased estimates when the missing data are *missing completely at random* (MCAR) (see, e.g., Little and Rubin, 1987)

The estimation of the missing cells can even introduce additional bias depending on the method used. Valid estimates and inferences can mostly only be made if the missing data

are at least *missing at random* (MAR) (see, e.g., Little and Rubin, 1987). Even in this case there are further challenges, and these are very typical in data sets from official statistics:

Mixed type of variables in the data: Data from official statistics typically consist of variables that have different distributions, i.e. various variables consist of an alternative distribution (binary data), some variables might be categorical, and the distribution of some variables could be determined to be continuous. If missing values are present in all these variable types, the challenge is to estimate the missing values based on the whole multivariate information.

Semi-continuous variables: Another challenge is the presence of variables in the data set where the distribution of one part of the data is continuous and the other part includes a certain proportion of equal values (typically zeros). The distribution of such variables is often referred to as "semi-continuous" distribution (see, e.g., Schafer and Olson, 1999). Data consisting of semi-continuous variables are, for example, income components in the *European Union Statistics of Income and Living Condition (EU-SILC)* survey, or tax components in tax data, in which one part of such a variable origins from a continuous distribution, and the other part consists of (structural) zeros.

Large data sets: Since data collection is a requirement in many fields nowadays, the resulting data sets can become "large", and thus the computation time of imputation methods is an important issue. One might argue that many such data sets can be decomposed into subsets referring to sub-populations, which are for instance defined by the NACE-codes in *Structural Business Survey* (SBS) data. Still, these subsets can contain more that 50000 observations, which calls for fast methods for data imputation.

Far from normality: A common assumption used for multivariate imputation methods is usually that the data originate from a multivariate normal distribution, or that they can be transformed to approximate multivariate normal distribution. This is violated in presence of outlying observations in the data. In this case, standard methods can result in very biased estimates for the missing values. It is then more advisable to use robust methods, being less influenced by outlying observations (see, e.g., Beguin and Hulliger, 2008; Serneels and Verdonck, 2008; Hron et al., 2010).

Note that prior exclusion of outliers before imputation is not straightforward. For example, when regression imputation is applied, leverage points might only be detected when analyzing the residuals from robust regression but might not be reliably identified from a least-squares fit nor by other multivariate outlier detection methods.

Note that rather than using sampling weights in the imputation process we recommend to impute data within reasonable subsets of the data set and to include the variables used for computing the sampling weights in the imputation model (see, e.g., Lumley, 2010a).

8.1.1 Imputation methods

Many different methods for imputation have been developed over the last few decades. The techniques for imputation may be divided into univariate methods such as column-wise (conditional) mean imputation, and multivariate imputation. In the latter case there are basically three approaches: distance-based imputation methods such as k-nearest neighbor imputation, covariance-based methods such as the approaches by Verboven et al. (2007) or Serneels and Verdonck (2008), and model-based methods such as regression imputation.

If an imputation method is able to deal with the randomness inherent in the data, it can be used for multiple imputation, generating more than one candidate for a missing cell (Rubin, 1987). Multiple imputation is one way to reflect the sampling variability, but it should only be used with careful consideration of the underlying distributional assumptions and underlying models (see also Fay, 1996; Durrant, 2005). In addition, if assumptions for the distribution of the occurrence of non-response are made but violated, poor results might be obtained (see also Schafer and Olsen, 1998). The sampling variability can also be reflected by adding a certain noise to the imputed values, and valuable inference can also be obtained by applying bootstrap methods (Little and Rubin, 1987; Alfons et al., 2009). However, most of the existing methods assume that the data originate from a multivariate normal distribution (e.g. the Gibbs sampling methods of the imputation software MICE (van Buuren and Oudshoorn, 1999; Buuren and Groothuis-Oudshoorn, 2011), Amelia (Honaker et al., 2009), mi (Yu-Sung et al., 2009) or mitools (Lumley, 2010b)). This assumption becomes inappropriate as soon as there are outliers in the data, or in case of skewed or multimodal distributions. Since this is a very frequent situation with practical data sets, imputation methods based on robust estimates are gaining increasing importance.

The basic procedure behind most model-based imputation methods is the EM-algorithm (Dempster et al., 1977), which can be thought of as a guidance for the iterative application of estimation, adaption and re-estimation. For the estimation, usually regression methods are applied in an iterative manner, which is known under the names regression switching, chain equations, sequential regressions, or variable-by-variable Gibbs sampling (see, e.g., van Buuren and Oudshoorn, 1999; Muennich and Rässler, 2004).

8.1.2 Software for imputation

The R package **mix** by Schafer (2009, 1997) considers most of the challenges described above, but it cannot handle semi-continuous variables, although Schafer and Olson (1999) described the problems with semi-continuous variables. In addition also the R package **mice** by van Buuren and Oudshoorn (1999) not supports to deal with semi-continuous variables. The R package **mi** (Yu-Sung et al., 2009) is well suited for multiple imputation in general, but it has the same limitations related to semi-continuous variables. This problem is treated in `IVEWARE`, a set of C and Fortran functions for which also SAS Macros are available (Raghunathan et al., 2001). The algorithms in **mi** and `IVEWARE` are based on iterative regression imputation. **mi** starts the algorithm by a rough initialization (randomly chosen values). The same concept is used by **MICE** (Multiple Imputation by

Chain Equations) the package of VAN BUUREN and OUDSHOORN (1999); BUUREN and GROOTHUIS-OUDSHOORN (2011) and the Amelia package of HONAKER et al. (2009), where first bootstrap samples with the same dimensions as the original data are drawn, and used for EM-based imputation. For a detailed review of software, see also WHITE et al. (2010).

All these algorithms and procedures cannot adequately cope with data including outliers. The aim is to develop a procedure that is competitive with the above algorithms, but has the additional feature of being robust with respect to data outliers. Since `IVEWARE` takes care of all the mentioned problems except robustness, and because this software is also recommended by EUROSTAT (see, e.g., EUROSTAT, 2008), it is natural to use it as a basis for our task. `IVEWARE` is mentioned by Eurostat in internal task force reports and presentation slides, and it is routinely used in national statistical institutions as well as in various research organizations.

A drawback of `IVEWARE` is that the exact procedure of the algorithm is not well documented. Therefore, we analyzed the software and provide a mathematical description of the algorithm in Section 8.2. Section 8.3 introduces the robust counterpart to `IVEWARE` which we call `IRMI` (Iterative Robust Model-based Imputation). Also other improvements were included in `IRMI`, like a different strategy for the initialization of the missing values. Simple comparisons of the two algorithms on two-dimensional artificial data are made in Section 8.4, and more detailed comparisons based on simulations are in Section 8.5. Applications to real data sets are provided in Section 8.6. The final section concludes.

8.2 The algorithm `IVEWARE`

`IVEWARE` estimates the missing values by fitting a sequence of regression models and drawing values from the corresponding predictive distributions (RAGHUNATHAN et al., 2001). Unfortunately, a detailed description of the algorithm does not exist. RAGHUNATHAN et al. (2001) provide only a rather vague outline of the functionality of this algorithm. However, the algorithm is fully described in TEMPL et al. (2011b). In few words, the algorithm includes the following steps:

- Sort the variables according to the amount of missing values.
- Initialization loop: Initialize the missing values by using one variable with missing values as response and the variables which includes either no missing values and variables which are already initialized as predictors. Apply this procedure for all variables which includes missing values.
- Iteration: Use one variable as response and the others as predictors and update the missing values in the response by drawing from the predictive distribution. The link function of the generalized linear regression method has to be selected based on the distribution of the response. Do that for all variables. Repeat this procedure until convergency.

8.3 The algorithm IRMI

The algorithm called IRMI for iterative robust model-based imputation has been implemented in function `irmi()` in the R package **VIM**. Basically it mimics the functionality of IVEWARE (Raghunathan et al., 2001), but there are several improvements with respects to the stability of the initialized values, or the robustness of the imputed values. In each step of the iteration, one variable is used as a response variable and the remaining variables serve as the regressors. Thus the "whole" multivariate information will be used for imputation in the response variable. The proposed iterative algorithm can be summarized as follows:

1. Initialisation of the missing values.
2. Choose one variable as response and the others as predictors and update the former missing values in the response.
3. Go to the next variable and repeat the procedure.
4. Repeat the whole procedure starting from 2 until convergency.
5. Add noise to the final estimates in a proper way to allow multiple imputation.

Note that robust regression methods are used instead of classical ones. This implies to solve many problems concerning these methods. A discussion about these problems and the detailed mathematical description of the method can be found in Templ et al. (2011b).

Within our implementation it is possible to use least trimmed squares (LTS) regression (Rousseeuw and Van Driessen, 2002), MM-estimation (default) (Yohai, 1987) and M-estimation (Huber, 1981) whenever the response is continuous or semi-continuous. If the response variable is binary, a robust generalized linear model with family *binomial* is applied (Cantoni and Ronchetti, 2001). When the response variables is categorical, a multinomial model is chosen, which is based on neural networks (for details, see Venables and Ripley, 2002; Ripley, 1996).

Note that robust regression for continuous or semi-continuous responses also protects against poorly initialized missing values, because the estimation of the regression coefficients is based only on the majority of the observations.

The function `irmi()` also provides the option to add a random error term to the imputed values, creating the possibility for multiple imputation. The error term has mean 0 and a variance corresponding to the (robust) variance of the regression residuals from the observations from the observed response. To provide adequate variances of the imputed data the error term has to be multiplied by a factor $\sqrt{1 + \frac{1}{n}\#m^l}$, considering the amount of missing values ($\#m^l$) in the response (additionally, the level of noise can be controlled by a scale parameter, which is by default set to 1). Conceptionally, this is different from all other implementations of EM-based regression imputation methods. It's somehow a simplification because only expected values are used to update former missing values until convergence. However, it guarantees much faster convergence and full control of the

convergence of the algorithm. Note that to keeping track of convergence of sequential methods used in `IVEWARE`, `mice` or `mi` is rather difficult because in each step predictive values are used to update former missing values instead of expected values like in `IRMI`. Within `IRMI`, errors to provide correct (co-)variances are included in a final iteration in an adequate way. The chosen factor allows multiple imputation with proper coverage rates as well (see Section 8.3.1). Within practical application using real-world data this approach is preferable and the results shows correct variances and coverage rates, i.e. this change of paradigm has a lot of advantages when working with complex data sets from official statistics, for example.

8.3.1 Properties

The imputation method should be "proper", i.e. to incorporate the variability that affects the imputed value, in order to lead to consistent standard errors (see, e.g., RUBIN, 1987). Since a mathematical proof whether a complex robust imputation method is proper in Rubin's sense is virtually impossible, the problem can be addressed by Monte Carlo simulation studies. RAESSLER and MÜNNICH (2004) give a detailed description on how to use simulations to determine if a multiple imputation method is proper or at least approximately proper. They investigated in a simulation to estimate coverage rates of the imputation procedures. We investigated this problem by reproducing the simulation study given in RAESSLER and MÜNNICH (2004). Let

$$(AGE, INCOME) \sim N\left(\begin{pmatrix} 40 \\ 1500 \end{pmatrix}\right), \left(\begin{pmatrix} 10 & 44 \\ 44 & 300 \end{pmatrix}\right)$$

the universe from which samples of size 2000 are drawn, whereas variable AGE is recoded in 6 categories. We set 30% of the income values to missing values using MCAR, MAR and MNAR mechanisms. Note that under MCAR the missing values are generated completely at random, under MAR, income is missing with higher probability the higher the value of AGE, under MNAR the probability of missing in INCOME is higher the higher INCOME (see, e.g., LITTLE and RUBIN, 1987). Then the three data sets are imputed using `IVEWARE`, `IMI` and `IRMI`, whereas 10 multiple imputed data sets are generated. The results from multiple imputed data were combined by well known rules (RUBIN, 1987). The aim is to estimate the total variance of the arithmetic mean which is the sum of squared distances from the MI estimates to the mean of the MI estimates (between-imputaton variance) plus the variance of the MI estimates itself (within-imputation variance) including small sample correction (see also RAESSLER and MÜNNICH, 2004). The whole procedure is repeated 2000 times and the coverage rate is counted, which is defined as the ratio between the amount of how often the true mean is covered by the estimated confidence intervals and the number of replications (2000).

Table 8.1 shows the coverage rate from complete case analysis (CA), mean imputation, `IVEWARE`, `IMI` and `IRMI-MM` (using MM-regression for robust imputation of continuous and semi-continuous variables).

All sequential imputation methods lead to comparable results and even in a MAR situation the coverage rate is reasonable. However, no contamination is introduced in that

Table 8.1: Coverage rates (0.95% confidence interval) using complete case analysis (CA), mean imputation, IVEWARE, IMI (the non-robust version of IRMI) and IRMI and different missing values mechanisms.

Missing	mPop	CA	Mean	IVEWARE	IMI	IRMI-MM
none	0.96					
MCAR		0.954	0.818	0.916	0.914	0.902
MAR		0.709	0.527	0.904	0.894	0.882
MNAR		0.822	0.820	0.918	0.916	0.906

simulation, and it becomes later clear that then the methods (IRMI) based on robust estimation are preferable. The coverage rate from complete analysis (CA) and arithmetic mean imputation is quite low, especially in MAR situations.

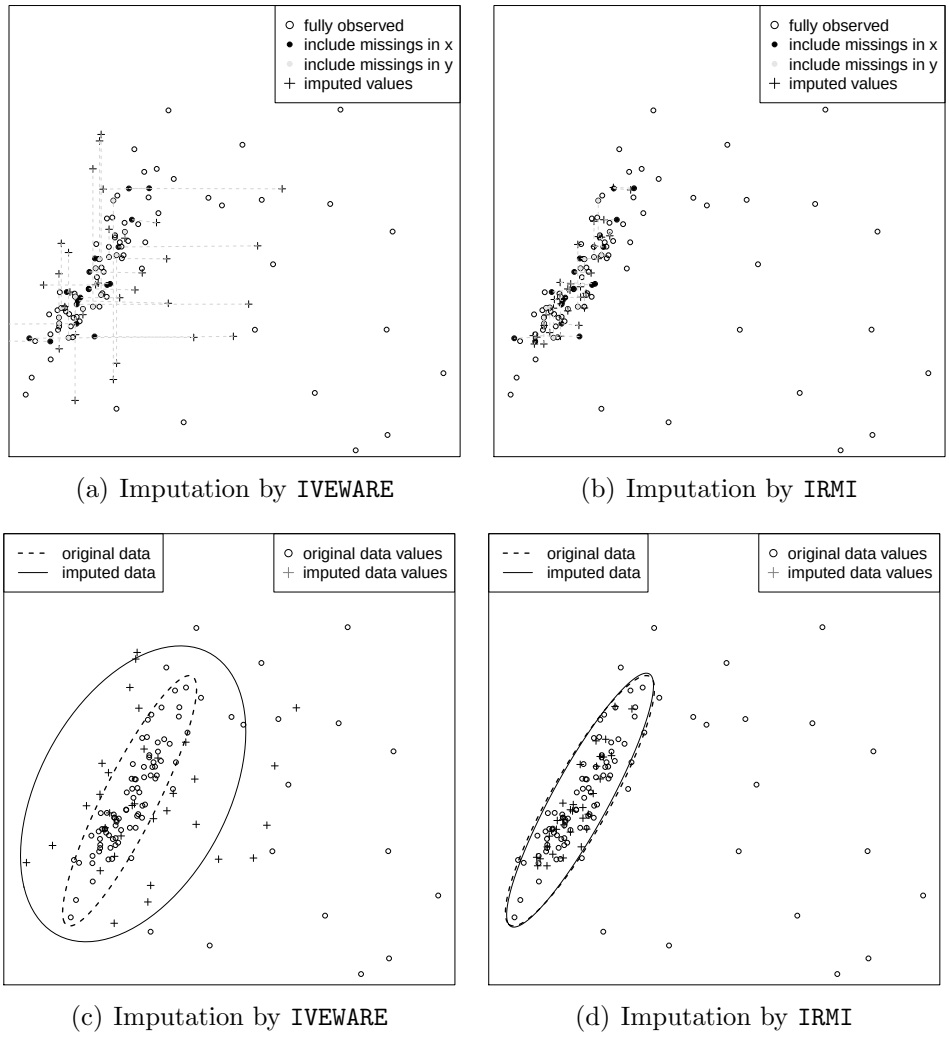
8.4 Comparison using exploratory examples including outliers

The different behaviour of IVEWARE and IRMI can be investigated by simple data configurations where the structure is clearly visible. Here we focus on the robustness aspect of IRMI, and thus on the effect of outlying observations in the data.

In Figure 8.1, two-dimensional data with both variables continuous distributed are shown. A complete data set including outliers is generated, and after that certain values of the non-outlying part are set to be missing. The aim is to impute those missing values and to evaluate if the covariance of the non-outlying part has changed after imputation. The dashed lines in the upper plots (Figure 8.1(a) and 8.1(b)) join the original values with their imputed ones. It is apparent that IVEWARE is highly influenced by outlying observations (Figure 8.1(a)) while IRMI leads to imputed values in the central part of the point cloud (Figure 8.1(b)). This becomes again visible when comparing the 95% tolerance ellipses, constructed by the non-outlying (and imputed) part of the data. A 95% tolerance ellipse covers (theoretically) 95% of the observations in case of two-dimensional normal distribution. The non-robust imputation by IVEWARE results in an inflated tolerance ellipse when compared to the tolerance ellipse using the original complete outlier-free data (Figure 8.1(c)). In contrast, the robust imputation by IRMI causes both ellipses to be almost indistinguishable, and thus IRMI generates practically the same bivariate data structure (Figure 8.1(d)).

Figure 8.2 already shows both the imputation results and the 95% tolerance ellipses from another two-dimensional data set, where the variable on the vertical axis originates from a semi-continuous distribution. This situation is very typical for data from official statistics. The figures also contain boxplots in the margins of the horizontal axes, providing information of both the distribution of the original (dark-grey colored boxes) and the imputed (light-grey colored boxes) constant data part. The 95% tolerance ellipses are based on the continuous and non-outlying part of the data. Although IVEWARE should be able to cope with semi-continuous variables. In IVEWARE we have used the data type `mixed` and the default values for all other parameters. Figure 8.2(a) shows that the imputed values are influenced by the outliers and the constant data part. In contrast, IRMI works as expected (see Figure 8.2(b)).

Similar experiments were made with other data configurations (for example, one binary distributed variable versus one variable where numbers are drawn from a normal distribution) and with other choices of the means and covariances for generating the data. The conclusions are analogous. A more detailed comparison of IVEWARE and IRMI will be provided by simulation studies in the next section.



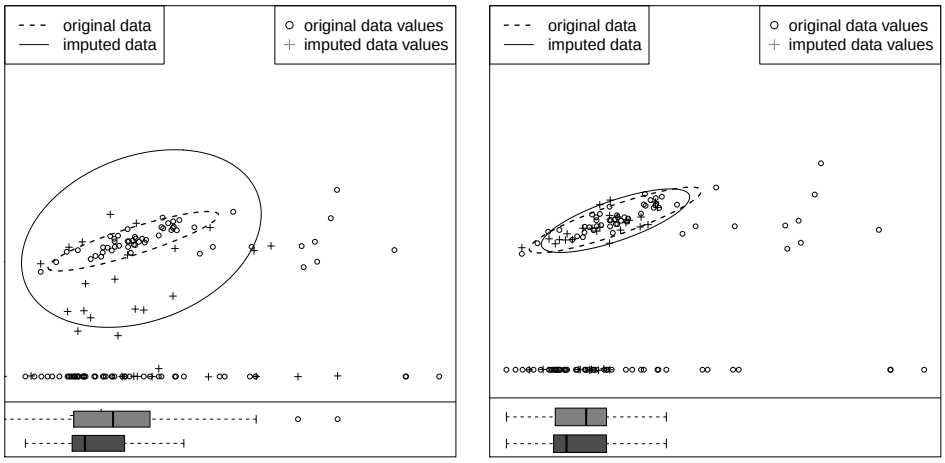
(a) Imputation by `IVEWARE`

(b) Imputation by `IRMI`

(c) Imputation by `IVEWARE`

(d) Imputation by `IRMI`

Figure 8.1: Imputation for continuous distributed two-dimensional data by `IVEWARE` and `IRMI`. Upper plots: The dashed lines join the original values with their imputations. Lower plots: 95% tolerance ellipses characterizing the multivariate data structure of the non-outlying part of the data. The imputation by `IVEWARE` is sensitive to the outliers, while `IRMI` succeeds to impute according to the original data structure.



(a) Imputation by `IVEWARE` (b) Imputation by `IRMI`

Figure 8.2: Imputation results by `IVEWARE` and `IRMI` for a two-dimensional data set consisting of a continuous variable and a semi-continuous variable. The covariance structure of the non-outlying continuous part of the data is visualized by 95% tolerance ellipses. The constant data part is summarized by boxplots for the original (lower boxplot) and the imputed (upper boxplot) data.

8.5 Simulation studies

For all simulations presented in this section we randomly generate data with $n = 500$ observations and p variables from a multivariate normal distribution. The population mean of (the non-outlier part of) each variable is fixed at 10. Based on the multivariate normal distribution, variables drawn from a binary and a semi-continuous distribution (as many other authors (see, e.g., Raghunathan et al., 2001) we do not consider the multinomial variables because within extensive simulation studies the computation time would then grow up a lot) are constructed by the following procedures:

Binary variable: A binary variable y with values $y_1, \ldots, y_n$ is created on the basis of a variable x with values $x_1, \ldots, x_n$ from the generated multivariate data by

$$y_i = \begin{cases} 0 & \text{with} \quad P(y_i = 0) = 1 - F_{N(\mu,\sigma^2)}(x_i) \\ 1 & \text{with} \quad P(y_i = 1) = 1 - P(y_i = 0) = F_{N(\mu,\sigma^2)}(x_i) \end{cases} ,$$

for $i = 1, \ldots, n$. $F_{N(\mu,\sigma^2)}$ denotes the distribution function of x, a normal distribution with mean μ and variance σ^2. Hence, if x_i is high (low) the probability that y_i becomes zero is high (low). Depending on the choice of μ the ratio of zero and ones differs (default 50% zeros on the average)

Semi-continuous variable: Without loss of generality, we set the constant part of the variable from a semi-continuous distribution to zero. We use two variables from the multivariate data. One variable is used to generate a binary variable y with values $y_1, \ldots, y_n$. This is done in the same way as above for binary distributed variables. A second variable $\tilde{x}$ with values $\tilde{x}_1, \ldots, \tilde{x}_n$ determines the non-constant part of the semi-continuous variable z with values $z_1, \ldots, z_n$ by

$$z_i = \begin{cases} 0 & \text{if} \quad y_i = 0 \\ \tilde{x}_i & \text{if} \quad y_i = 1 \end{cases} ,$$

for $i = 1, \ldots, n$.

These procedures allow that the correlation structure generated for the multivariate normally distributed data is also reflected by the variables with mixed distribution.

In order to avoid complicated notation, the resulting data values are denoted by x_{ij}^{orig}, with $i = 1, \ldots, n$ and $j = 1, \ldots, p$, and the imputed values by x_{ij}^{imp}.

8.5.1 Error measures

The use of variables with different distribution has also consequences for an error measure, providing information on the quality of the imputed data. A solution is to use different measures for categorical and binary variables, and for continuous and semi-continuous variables.

Note, that the specified error measures in the following are suitable to measure precision of the imputations. Other error measures which are designed to estimate the error in terms of variance preservation are not discussed in this contribution.

Error measure for categorical and binary variables: This error measure is defined as the proportion of imputed values taken from an incorrect category on all missing categorical or binary values:

$$err_c = \frac{1}{m_c} \sum_{j=1}^{p_c} \sum_{i=1}^{n} \mathbb{I}(x_{ij}^{orig} \neq x_{ij}^{imp}) \quad , \tag{8.1}$$

with $\mathbb{I}$ the indicator function, m_c the number of missing values in the p_c categorical variables, and n the number of observations.

Error measure for continuous and semi-continuous variables: The two different situations continuous and semi-continuous have to be distinguished. For the continuous parts we use the absolute relative error between the original and the imputed value. For the categorical (constant) part we count the number of incorrect categories, similar to Equation (8.1). Here we assume that the constant part of the semi-continuous variable is zero. Hence, the joint error measure is

$$\begin{aligned} err_s = \quad & \frac{1}{m_s} \sum_{j=1}^{p_s} \sum_{i=1}^{n} \left[\left| \frac{(x_{ij}^{orig} - x_{ij}^{imp})}{x_{ij}^{orig}} \right| \cdot \mathbb{I}(x_{ij}^{orig} \neq 0 \ \wedge \ x_{ij}^{imp} \neq 0) + \right. \\ & \left. \mathbb{I}((x_{ij}^{orig} = 0 \ \wedge \ x_{ij}^{imp} \neq 0) \ \vee \ (x_{ij}^{orig} \neq 0 \ \wedge \ x_{ij}^{imp} = 0)) \right] \quad , \end{aligned} \tag{8.2}$$

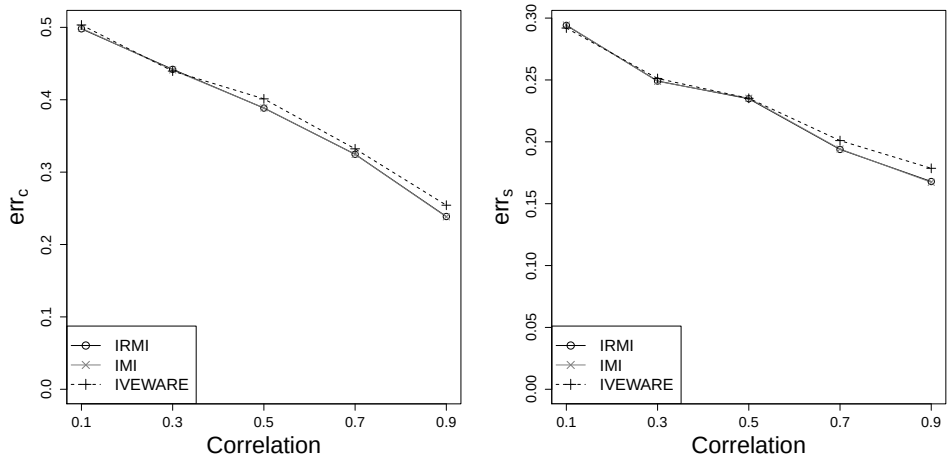
with m_s the number of missing values in the p_s continuous and semi-continuous variables. For continuous variables we assume that both the original value and the imputed value are different from zero, and thus the first part of Equation (8.2) measures the imputation error. In the other case, if either the original or the imputed value is zero, the second part of the equation is used.

8.5.2 First configuration: varying the correlation structure

In a first simulation setting we want to study the effect of the correlation structure between the variables. Therefore, the covariance matrix of the underlying multivariate normally distributed data is taken as a matrix with variances of one on the main diagonal, and otherwise constant values. These are chosen in 4 steps as 0.1, 0.3, 0.5, 0.7 and 0.9, respectively. The following type of variables are included - two continuous variables, one binary variable and one semi-continuous variable. As for all simulations in this section, the number of repetitions is 500, and the final error measure is the average of all 500 resulting error measures. The proportion of missing values is fixed at 5% in each variable.

The results of the algorithms `IVEWARE`, `IRMI`, and `IMI`, the non-robust version of `IRMI`, are presented in Figure 8.3. Generally, the error measure decreases with increasing correlation, because then the multivariate information is more and more useful for the estimation of the missing values. The error measure for the categorical variables (see Figure 8.3(a)) and the

continuous and semi-continuous variables (Figure 8.3(b)) is compariable among the robust method (`IRMI`) and the non-robust OLS-based regression method, because no outliers have been generated in the simulated data. The difference for the error measures between `IVEWARE` and our proposed methods gets more pronounced with increasing correlation (Figure 8.3(b)). Here, `IVEWARE` obviously may suffer from a less optimal strategy to initialize the missing values, which is then reflected in slightly poorer imputation quality.



(a) Imputation error for binary and categorical parts

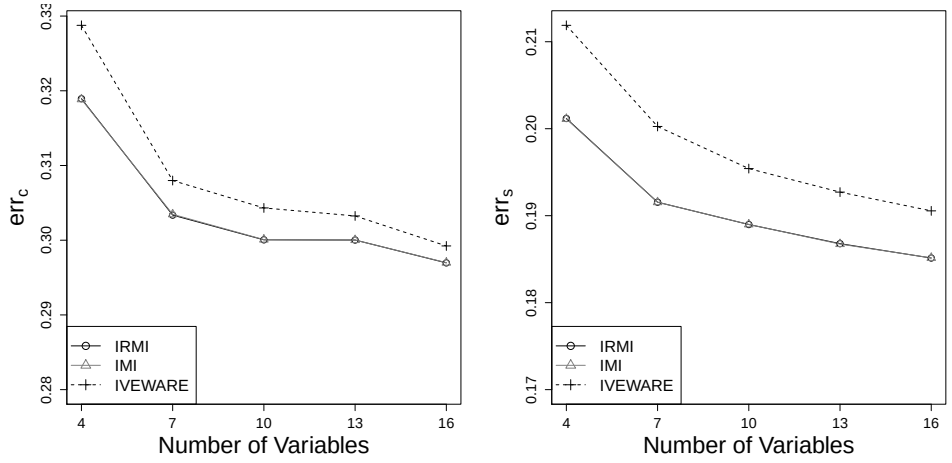
(b) Imputation error for continuous and semi-continuous parts

Figure 8.3: Comparison of the error measures resulting from the three algorithms by varying the correlation structure of the generated data.

8.5.3 Second configuration: varying the number of variables

With increasing dimensionality of the data, the gain in multivariate information can be used by model-based regression imputation as long as the additional variables are not uncorrelated to the variable where missing information needs to be estimated. This effect is demonstrated by a simulation setting which starts from 4 variables (two continuous, one binary and one semi-continuous), and increases in each step the dimensionality of the data by including one further variable of each type. Figure 8.4 presents the results of this study. Here the simulated data are based on multivariate normally distributed data with fixed covariances of 0.7 and variances of 1. Again the proportion of missing values is fixed with 5% in each variable.

Figure 8.4 shows that all three algorithms have a similar performance. `IMI` and `IRMI` have a slightly better precision than `IVEWARE` with respect to lower dimensionality of the data. While the errors from the categorical and binary variables remain almost constant when increasing the number of variables (see Figure 8.4(a)), the error from imputing the continuous and semi-continuous variables is first decreasing and then increasing (see Figure 8.4(b)).



(a) Imputation error for binary and categorical parts

(b) Imputation error for continuous and semi-continuous parts

Figure 8.4: Comparison of the error measures resulting from the three algorithms by varying the number of variables.

8.5.4 Third/fourth configuration: varying the amount of outliers using variables with high/low correlation

To illustrate the influence of outliers on the considered imputation algorithms, n_1 out of n observations will be replaced by outlying observations. The non-outlying part of the data is generated in the same way as described in the previous settings, with covariances of 0.9 (third configuration) and 0.4 (fourth configuration), respectively, including two continuous variables, two binary variables and one semi-continuous variable. The outlier part is generated with the mean vector

$$\mu_{out} = (5, 15, 10, 10, 10)^t$$

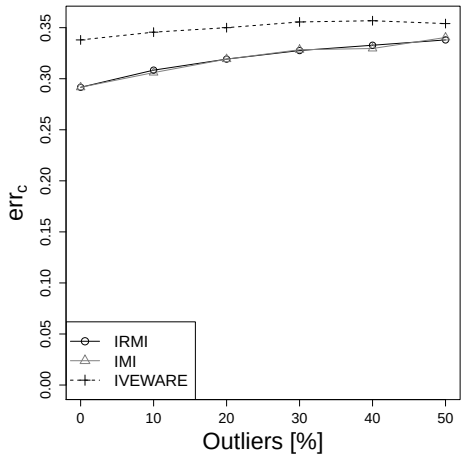
and covariances 0.5 (third configuration) and 0.4 (fourth configuration), the variances are 1. The generation of binary variables and semi-continuous variables is done as described in the first part of this chapter. The percentage of outliers is varied from 0 to 50. The proportion of missing values in the variables is $(0.1, 0.06, 0.05, 0.04)$, and they are only chosen in the non-outlying part. Accordingly, the error measures are only based on non-outlying observations. The results are shown in Figure 8.5.

With respect to the errors in the categorical parts (Figure 8.5(a) and 8.5(c)), the non-robust method `IMI` and its robust counterpart `IRMI` performs almost identical since non-robust regression is used by `IRMI` for imputing categorical responses per default, because robust methods tends to be very instable for categorical responses. Especially for the highly correlated data including outliers, both methods outperforms `IVEWARE`. The error in the continuous parts reveals a contrasting behavior (Figure 8.5(b) and 8.5(d)). Here, the robust version `IRMI` clearly dominates, and it remain stable until about 30% outliers. The non-robust version `IMI` perfoms slightly better than `IVEWARE`.

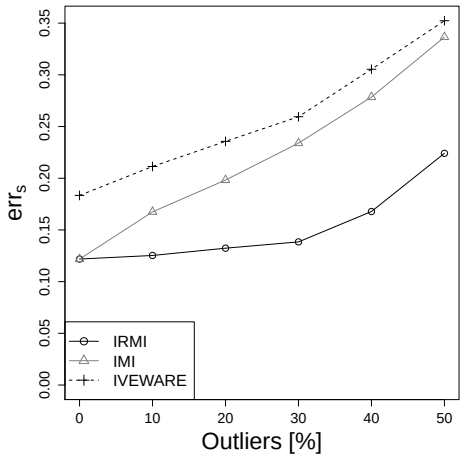
8.6 Application to EU-SILC

Originally, all the EU-SILC data set come with missing values and few outlying observations. We took the available complete observations, almost ignoring the possible dependencies of the missing values in the data. However, also within this simplification the results should reflect how the algorithm performs with the data. We set missing values randomly to the available information before imputing these artificial missing values. After imputation, the imputed values are compared with their "true" original values.

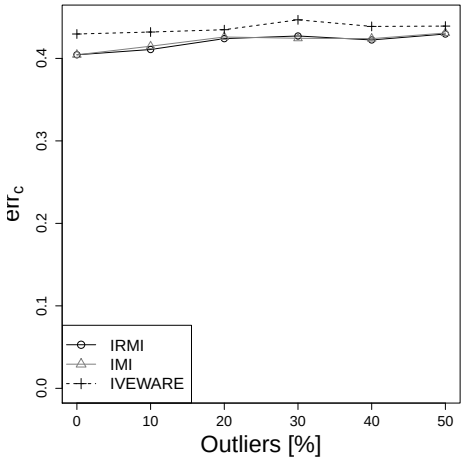
The EU-SILC data set includes a moderate amount of missing values in the semi-continuous part of the data. `IVEWARE` is used by various statistical agencies (e.g. by the Federal Statistical Office in Swiss) to impute the income components of the EU-SILC data, for example. Statistics Austria, for example, uses (non-iterative) least-squares regression imputation whereas a random error based on the variance of the residuals is added to the response (see, e.g., Rubin, 1987; Ghellini and Neri, 2004). Fisher (2006) imputed the income components of "similar" data (consumer expenditure data) separately, i.e. he uses one income component as the response variable and as predictors demographic characteristics of the consumer unit and a variable that equals the quarterly expenditure outlays



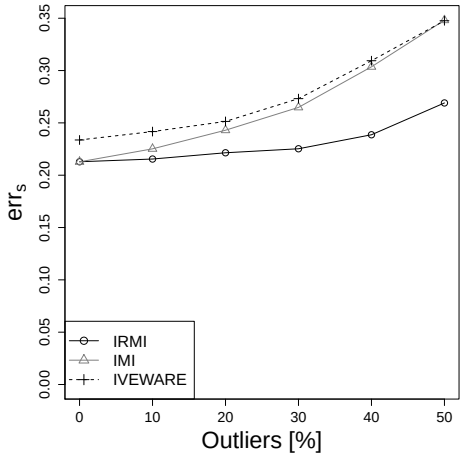
(a) Imputation error for binary and categorical parts; variables with high correlation



(b) Imputation error for continuous and semi-continuous parts; variables with high correlation



(c) Imputation error for binary and categorical parts; variables with low correlation



(d) Imputation error for continuous and semi-continuous parts; variables with low correlation

Figure 8.5: Comparison of the error measures resulting from the three algorithms by varying the percentage of outliers.

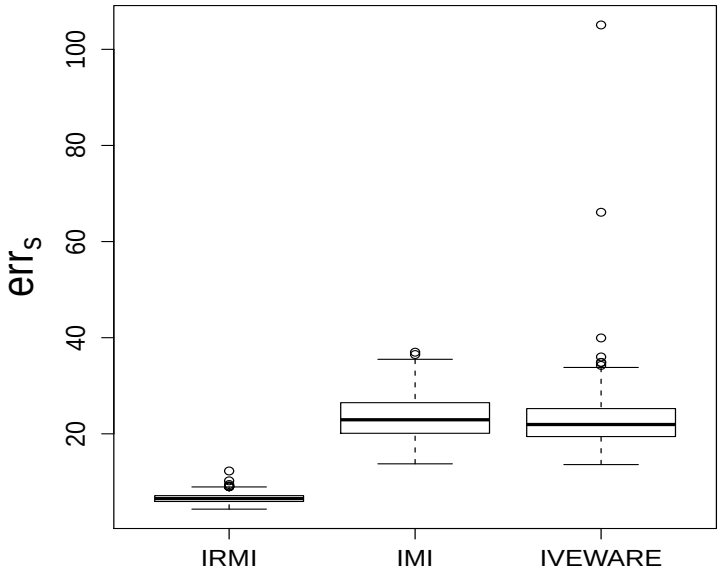


Figure 8.6: Results for the Austrian EU-SILC data comparing original data points with the imputed data points.

for the consumer unit. He performs a stepwise backward approach to select only the most important predictors.

For imputation we used the household income variables (semi-continuous variables) with the largest amount of missing values in the raw data set, namely *hy050n* (family/children related allowances), *hy060n* (social exclusion not elsewhere classified), *hy070n* (housing allowances), *hy090n* (interest, dividends, profit from capital investments in unincorporated business), and three categorical variables, namely *household size*, *region*, and *number of childrens in the household*. The available complete observations of this data set are used (3808 observations), and missing values are set completely at random in the income components respecting the rate of missing values in the income variables from the complete data. Therefore, 25 percent missing values are generated in variable `hy090n`, and 2 percent missing values are generated in the other income components. `IMI`, `IRMI` and `IVEWARE` are applied to impute the missing values.

The procedure is repeated 1000 times. Figure 8.6 displays the results of the simulation. Since missing values are only obtained in the semi-continuous income components, only the error measure for continuous and semi-continuous variables is reasonable. It is easy to see that `IRMI` leads to much better results than `IMI` and `IVEWARE`. Also the variance of the errors is much smaller.

An additional result within the AMELI project by imputing this data set with `IRMI` is obtained by Alfons et al. (2009). They used `IRMI` (without describing the algorithm) to estimate the additional uncertainty (with respect to missing values) of indicators via the bootstrap approach from Little and Rubin (1987). In fact they estimate the additional

uncertainty due the presence of imputations when estimating the GINI coefficient but also the weighted mean of the equivalised household income from the Austrian EU-SILC data. The additional uncertainty was evaluated for the point estimates but also for the variance estimates. Their simulations was not designed to show that `IRMI` is proper according to definitions in NIELSEN (2003) or RUBIN (1987), but it results in realistic estimates and consider small additional uncertainty due to point and variance estimates.

8.7 Conclusions

All real-world data sets we have seen so far, especially in official statistics, include outlying observations and they often include different types of distributions. We summarize an EM-based iterative robust model-based imputation procedure for automatic imputation of missing values, which can deal with the mentioned data problems. All simulation results show that our robust method shows either equal behaviour or outperforms the investigated non-robust methods. Additionally, the results from the imputation of the complex and popular EU-SILC data set which includes several semi-continuous variables showed that `IRMI` performs very well in a real-world settings.

Therefore, we would suggest to apply `IRMI` whenever an automatical approach for imputation is needed, and especially, when the variables are realizations of different types of distributions possible including all, binary, categorical, semi-continuous and continuous variables.

Furthermore, our methods are implemented in the package **VIM** version 1.4.4. (TEMPL et al., 2011a) written in R (R DEVELOPMENT CORE TEAM, 2009). The function `irmi()` can be used to impute missing values with our proposed method. It has sensible defaults and various user-friendly tools to automatically detect the distribution of variables (optinally), for example. The application is straightforward and explained in the manual of the package. As mentioned before, the package **VIM** can be freely downloaded from the comprehensive R archive network (see http://cran.r-project.org/package=VIM).

Bibliography

Alfons, A., **Templ, M.** and **Filzmoser, P.** (**2009**): *ON THE INFLUENCE OF IMPUTATION METHODS ON LAEKEN INDICATORS: SIMULATIONS AND RECOMMENDATIONS*. UNECE Work Session on Statistical Data Editing; Neuchatel, Switzerland, p. 10, to appear.

Beguin, C. and **Hulliger, B.** (**2008**): *The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data*. Survey Methodology, 34 (1), pp. 91–103.

Buuren, S. and **Groothuis-Oudshoorn, K.** (**2011**): *MICE: Multivariate Imputation by Chained Equations in R*. Journal of statistical software, in press.

Cantoni, E. and **Ronchetti, E.** (**2001**): *Robust Inference for Generalized Linear Models*. JASA, 96 (455), pp. 1022–1030.

Dempster, A., **Laird, N.** and **Rubin, D.** (**1977**): *Maximum likelihood for incomplete data via the EM algorithm (with discussions)*. Journal of the Royal Statistical Society, 39, pp. 1–38.

Durrant, G. (**2005**): *Imputation methods for handling item-nonresponse in the social sciences: a methodological review*. Ncrm methods review papers, Southampton Statistical Sciences Research Institute (S3RI), University of Southampton.

Eurostat (**2008**): *Survey sampling reference guidelines. Introduction to sample design and estimation techniques*. Methodologies and working papers, ISNN 1977-0375, European Commission.

Fay, R. (**1996**): *Alternative paradigms for the analysis of imputed survey data*. Journal of the American Statistical Association, 91 (434), pp. 490–498.

Fisher, J. (**2006**): *Income imputation and the analysis of consumer expenditure data*. Monthly Labor Review, 129 (11), pp. 11–19.

Ghellini, G. and **Neri, L.** (**2004**): *Proper Imputation of Missing Income Data for the Tuscany Living Condition Survey*. In Proceedings of the Atti della XLII Riunione Scientifica, Universitá di Bari, Italy, p. 4.

Honaker, J., **King, G.** and **Blackwell, M.** (**2009**): Amelia: Amelia II: A Program for Missing Data. R package version 1.2-2.
URL http://CRAN.R-project.org/package=Amelia

Hron, K., **Templ, M.** and **Filzmoser, P.** (**2010**): *Imputation of missing values for compositional data using classical and robust methods*. Computational Statistics & Data Analysis, 54 (12), pp. 3095–3107, ISSN 0167-9473, doi:DOI:10.1016/j.csda.2009.11.023.

Huber, P. (**1981**): Robust Statistics. Wiley.

Little, R. and **Rubin, D.** (**1987**): Statistical Analysis with Missing Data. New York: Wiley.

Lumley, T. (**2010**a): Complex Surveys: A Guide to Analysis Using R. Wiley.

Lumley, T. (**2010**b): mitools: Tools for multiple imputation of missing data. R package version 2.0.1.
URL http://CRAN.R-project.org/package=mitools

Muennich, R. and **Rässler, S.** (**2004**): *Variance Estimation under Multiple Imputation*. Proceedings of Q2004 European Conference on Quality in Survey Statistics, Mainz, p. 19.

Nielsen, S. (**2003**): *Proper and Improper Multiple Imputation*. Internat. Statist. Rev., 71 (3), pp. 593–607.

R Development Core Team (**2009**): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

Raessler, S. and **Münnich, R.** (**2004**): *The Impact of multiple imputation for DACSEIS*. Research report ist-2000-26057-dacseis, 5/2004, University of Tübingen.

Raghunathan, T., **Lepkowski, J.** and **Hoewyk, J.** (**2001**): *A multivariate technique for multiply imputing missing values using a sequence of regression models*. Survey Methodology, 27 (1), pp. 85–95.

Ripley, B. D. (**1996**): Pattern Recognition and Neural Networks. Cambridge University Press.

Rousseeuw, P. and **Van Driessen, K.** (**2002**): *Computing LTS regression for large data sets*. Estadistica, 54, pp. 163–190.

Rubin, D. (**1987**): Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Schafer, J. (**1997**): Analysis of Incomplete Multivariate Data. Chapman & Hall, chapter 9.

Schafer, J. (**2009**): mix: Estimation/multiple Imputation for Mixed Categorical and Continuous Data. R package version 1.0-7, see also his implementation in SAS and SPLUS.
URL http://CRAN.R-project.org/package=mix

Schafer, J. and **Olson, M.** (**1999**): *Modeling and Imputation of Semicontinuous Survey Variables*. Fcsm research conference papers, Federal Commitee on Statistical Methodology.
URL http://www.fcsm.gov/99papers/shaffcsm.pdf

Schafer, J. L. and **Olsen, M. K.** (**1998**): *Multiple imputation for multivariate missing-data problems: a data analyst's perspective*. Multivariate Behavioral Research, 33, pp. 545–571.

Serneels, S. and **Verdonck, T.** (**2008**): *Principal component analysis for data containing outliers and missing elements*. Computational Statistics & Data Analysis, 52 (3), pp. 1712–1727.

Templ, M., **Alfons, A.** and **Kowarik, A.** (**2011**a): VIM: Visualization and Imputation of Missing Values. R package version 1.4.4.
URL http://cran.r-project.org/package=VIM

Templ, M., **Kowarik, A.** and **Filzmoser, P.** (**2011**b): *Iterative stepwise regression imputation using standard and robust methods*. Computational Statistics and Data Analysis, resubmitted after minor revision.

van Buuren, S. and **Oudshoorn, C.** (**1999**): *Flexible multivariate imputation by MICE*. Tno/vgz/pg 99.054, Netherlands Organization for Applied Scientific Research (TNO).
URL http://web.inter.nl.net/users/S.van.Buuren/mi/docs/rapport99054.pdf

Venables, W. N. and **Ripley, B. D.** (**2002**): Modern Applied Statistics with S. Springer.

Verboven, S., **Branden, K.** and **Goos, P.** (**2007**): *Sequential imputation for missing values*. Computational Biology and Chemistry, 31, pp. 320–327.

White, I., **Daniel, R.** and **Royston, P.** (**2010**): *Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables*. Computational Statistics & Data Analysis, 54 (10), pp. 2267–2275, ISSN 0167-9473, doi:DOI:10.1016/j.csda.2010.04.005.

Yohai, V. (**1987**): *High breakdown-point and high efficiency estimates for regression*. The Annals of Statistics, 15, pp. 642–665.

Yu-Sung, S., **Gelman, A.**, **Hill, J.** and **Yajima, M.** (**2009**): *Multiple imputation with diagnostics (mi) in R: Opening windows into the black box*. Journal of Statistical Software, to appear.

Chapter 9

Robust Methods for Elliptically Contoured Data

All the methods in this chapter assume that the data come from a model (i.e., from an elliptically contoured distribution; see below). This assumption is less restrictive than it may seem at first sight since the income data can be transformed (by a series of data preparation steps) to fit into the model; at least nearly so. From the model perspective, a general framework for multivariate outlier identification in a p-dimensional data set $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ is to compute some measure of the distance of a particular data point from the center of the data and declare as outliers those points which are too far away from the center. Usually, as a measure of öutlyingnessfor a data point $\mathbf{x}_i$, $i = 1, \ldots, n$, a robust version of the (squared) Mahalanobis distance RD_i is used, computed relative to high breakdown point robust estimates of location $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of the data set $\mathbf{X}$

$$RD_i^2(\mathbf{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}). \tag{9.1}$$

The most common estimators of multivariate location and scatter are the sample mean, $\mathbf{m}$, and the sample covariance matrix, $\mathbf{C}$, i.e. the corresponding ML estimates (when the data follow a normal distribution). These estimates are optimal if the data come from a multivariate normal distribution but are extremely sensitive to the presence of even a few outliers in the data.

We restrict attention to the familiy of ellipitically contoured distributions (EC); a natural extension of the multivariate normal distribution. Notably, we assume that the data come from an EC. Denote by $F_{p,\mathbf{I}} \in \mathcal{F}_{p,\mathbf{I}}$ a distribution in the class of spherically symmetric distributions in $\mathbb{R}_p$. A $(p \times 1)$ random variable $\mathbf{x}$ is said to have a spherically symmetric distribution $S(\omega)$ if there exists a scalar function $\omega(\cdot)$ (characteristic generator) such that for its characteristic function $f(u)$ it holds that $f(u) = \omega(\mathbf{u}^T\mathbf{u})$, for ω in the family of all generators and $\mathbf{u} \in \mathbb{R}^p$; see Fang et al. (1990, p.28-29) for the details. Thus, an $(p \times 1)$ random variable $\mathbf{y}$ is said to have elliptically symmetric distribution with parameters $\boldsymbol{\mu}$ $(p \times 1)$ and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^T\boldsymbol{\Lambda}$ of dimension $(p \times p)$ and rank p if $\mathbf{y} =_d \boldsymbol{\mu} + \boldsymbol{\Lambda}^T\mathbf{x}$, with $\mathbf{x}$ a spherically symmetric distributed $(p \times 1)$ random variable (where $a =_d b$ denotes that a

and b have the same distribution). We shall write $\mathbf{y} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \omega)$ with $\text{rank}(\boldsymbol{\Sigma}) = p$. For $\omega(u) = \exp(-u/2)$, $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \omega)$ is the p-dimensional normal distribution.

In the presence of contamination, outlier identification procedures based on sample mean and sample covariance will suffer from the following two problems (ROUSSEEUW and LEROY, 1987):

- *Masking*: multiple outliers can distort the classical estimates of mean and covariance in such a way (attracting $\mathbf{m}$ and inflating $\mathbf{C}$) that they do not get necessarily large values of the Mahalanobis distance, and
- *Swamping*: multiple 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.

Consequently, sample mean and sample covariance matrix are not reliable candidates of location and scatter estimators for multivariate outlier detection. Suppose we have some robust estimators of multivariate location and scatter. The second issue is to determine how large the robust distances should be in order to declare a particular point an outlier. If the data $\mathbf{X}$ have a p-variate normal distribution, one may base outlier declaration on a cutoff value $d_c = \chi^2_p(c)$. Consequently, all observations with RD_i^2 larger than d_c would be declared outliers. This procedure will no more be valid if robust estimators are applied and/or if the data have other than multivariate normal distribution. Therefore, MARONNA and ZAMAR (2002) propose to use the following transformation

$$\tilde{d}_c = \frac{\chi^2_p(c)\text{median}[RD_1^2, \ldots, RD_n^2]}{\chi^2_p(0.5)}. \tag{9.2}$$

Once outliers have been declared by an outlier-detection method, one considers imputing (non-outlying) observations for the declared outliers in order to end up with a *clean* dataset. However, a drawback of all methods considered so far is that they work only with complete data. As for the income components from EU-SILC consist of a large number of missing observations, the outlier-detection methods in this chapter take into account missing values. Notably, the imputation methods impute for both the declared outliers and the missing observations.

The Multivariate Outlier Detection and Imputation (MODI) methods based on the EC-model consist of, at least, three computation steps that depend upon another. The main steps are as follows. The first step is about data preparation such as transformations, treatment of zero, missing, and negative income values. This is followed by an outlier detection procedure. Finally, an imputation step replaces outlying and missing observations with realizations from a robustly estimated data model (or non-outlying donors). In the sequel, we shall discuss the properties of MODI methods in more detail.

9.1 Data Preparation

In order to invoke EC-distribution-based MODI methods, the income data from the EU-SILC exercise needs to be preprocessed. In this respect, the data must be transformed such that they behave similarly to EC-data. The data prepration consists of the following tasks.

- *Dimension reduction:* The 2006 EU-SILC exercise collected data on 18 income components at household- and 14 at the individual level (only net incomes components considered). Having in mind the complex data structure, the resulting number of 32 variables is far too large to be fed directly into any outlier-detection method. Therefore, the income components have to aggregated; see Chapter 6 for more on that issue. Moreover, through aggregation information is accumulated.
- *Symmetrizing transformation:* The income components are transformed such that their marginal distribution is close to symmetric (e.g., coordinate-wise transform by the logarithm with base 10, $\tilde{y} = \log_{10}(y + 1)$). Note that the application of coordinate-wise transformations is sufficient since the aggregated data (see Chapter 6 for more details) are nearly orthogonal to each other. If this would not be the case, one may process the data by a truely multivariate transformation; see e.g., Andrews et al. (1971).
- *Negative income values:* For negative values, the transformation is applied to the absolute value. In addition, the information that a particular observation is negative is stored in order to reset the sign after back-transformation.
- *Structural zero observations:* All structural zeros are replaced by a missing value, but the information that a particular observation was zero is stored in order to reset it to zero after back-transformation. As a result, the method cannot detect false zeros, i.e., observations with a truely positive value that are falsely zero.

9.2 Outlier Detection

9.2.1 BACON-EEM

The BACON-EEM algorithm by Béguin and Hulliger (2008) is based on the BACON algorithm proposed by Billor et al. (2000) which in turn is an improvement of an earlier *forward search* based algorithm by one of the authors. The original BACON algorithm is limited to complete, multivariate normal iid data. The BACON-EEM, on the other hand, has been adapted to cope with missing values and the finite population sampling context. The BACON-EEM starts from a subset of the data which is supposed outlier free or nearly so. It uses an adjusted version of the EM-algorithm to estimate the population sufficient statistics. Next, it calculates the Mahalanobis distances according to these statistics for the whole sample and includes or excludes observations according to their Mahalanobis distance. These steps are iterated until the final set remains stable; see Béguin and Hulliger (2008) for details. The algorithm is supposed to be a balance between affine equivariance and robustness (Todorov et al., 2011).

9.2.2 GIMCD: Gaussian Imputation followed by MCD-detection

The GIMCD detection method imputes for the missing values on grounds of a multivariate normal model using data augmentation (Gaussian imputation; GI step). Since the underlying MV model has not been robustly fitted, the GI step may impute outliers. However, with the imputed – and thus complete – data one subsequently estimates the robust location vector and the scatter matrix by means of MCD; see MARONNA et al. (2006) for details on MCD. Accordingly, outliers are detected using a cutoff on the Mahalanobis distance (with the robustly estimated statistics) (cf. BÉGUIN and HULLIGER, 2004). It is noteworthy that in situations where the data feature an overwhelming majority of missing values (e.g., 60-70% in a single variable) in addition to extreme outliers, GI may impute too many outlying observations for MCD in order not to breakdown. The finite breakdown point (in the sense of Huber-Donoho) of MCD (assuming that the data are in general position) is equal to $(n + p)/2$ (MARONNA et al., 2006, p. 190).

9.2.3 TRC: Tranformed Rank Correlations

The Transformed Rank Correlation (TRC) is one of the algorithms proposed by BÉGUIN and HULLIGER (2004) and is based, similarly as the OGK algorithm of MARONNA and ZAMAR (2002), on the proposal of GNANADESIKAN and KETTENRING (1972) for pairwise construction of the covariance matrix. The initial matrix is calculated using bivariate Spearman Rank correlations $\rho(\mathbf{x}_k, \mathbf{x}_h)$, $0 \geq i,\ j \leq p$ which is symmetric but not necessarily positive definite. To ensure positive definiteness of the covariance matrix the data are transformed into the space of the eigenvectors of the initial matrix and univariate estimates of location and scatter are computed which are used then to reconstruct an approximate estimate of the covariance matrix in the original space. The resulting robust location and covariance matrix are used to compute robust distances for outlier identification. In case of incomplete data the Spearman rank correlations are computed only from observations which have values for both variables involved and thus the initial covariance matrix estimate can be computed using all available information. From this matrix the transformation matrix can be computed but in order to apply the transformation complete data matrix is needed which is the most difficult problem in this algorithm. To obtain complete data each missing item x_{ik} is imputed by a simple robust regression of $\mathbf{x}_k$ on $\mathbf{x}_h$ where $\mathbf{x}_h$ is the variable with highest rank correlation $\rho(\mathbf{x}_k, \mathbf{x}_h)$. In order to be able to impute x_{ik}, the item x_{ih} must be observed and the quality of the regression on a given variable is controlled by an overlap criterion, i.e. in order to choose a variable as a regressor, the number of observations in which both variables are observed must be greater than some γn where $0 < \gamma < 1$ is a tuning constant. After all missing items have been imputed (or some observations with too few observed items have been removed) the complete data matrix can be used to perform the transformation and compute the final robust location and covariance matrix as in the case of complete data.

9.3 Imputation

Once outliers have been declared by an outlier-detection method (e.g., BACON-EEM), one considers robust imputation for both the missing values and the declared outliers. We distinguish three types of imputation for outliers: (1) treating the outlier as if it were a completely missing observation, (2) winsorizing the outlier with the help of the outlying values and (3) nearest neighbour imputation.

Let $\mathbf{y}^* = (y^*_{i,1}, \ldots, y^*_{i,p})^T$ denote the $(p \times 1)$ vector of true, un-contaminated income components for the ith observation, $i = 1, \ldots, n$. Both outlier-detection methods produce a flag variable o_i indicating outliers. We shall write (for ease of notation) y_c if some or all components are outlying to distinguish contaminated from un-contaminated observations, $i = 1, \ldots, n$.

- *Treating outliers as missing (TOaM):* A simple solution to imputation for outliers is to impute them as if the outlying component(s) were completely missing. Insofar imputation does only take into account the information in the flag variable $\mathbf{o}_i$, that is, whether a particular component in observation i has been declared outlying or not. The procedure does not make use of the information contained in $\mathbf{y}^c_i$. This is often inefficient, and it is usually more appropriate to use the information in $\mathbf{y}^c_i$ (i.e., sign and magnitude). By way of example, suppose an outlier of the kind error-of-thousand in several components (e.g., $\mathbf{y}^c_i = 1000\ \mathbf{y}^*_i$). Winsorizing such an observation would at least preserve the direction of the correct value while setting the outlier to missing would imply that a mean value is imputed which may lie in a different direction. In other words winsorizing the outlier may preserve the correlation structure while imputing a mean does not.

- *Winsorization and Gaussian imputation (WGI):* The multivariate outlier-detection methods that are based on the robust $(p \times 1)$ location vector $\hat{\boldsymbol{\mu}}$ and the robust $(p \times p)$ scatter matrix $\hat{\boldsymbol{\Sigma}}$ (e.g., for BACON-EEM) lead to a direct model-based imputation. Let $\hat{d}_i = [RD^2_i(\mathbf{y}_i; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})]^{1/2}$ denote the Mahalanobis distance of observation $\mathbf{y}^c_i$ from the robust location $\hat{\boldsymbol{\mu}}$ w.r.t. $\hat{\boldsymbol{\Sigma}}$. Observations are declared outliers if their Mahalanobis distance is larger than a constant k. The imputation for an observation $\mathbf{y}^c_i$ with Mahalanobis distance $\hat{d}_i > k$ is

$$\hat{\mathbf{y}}_i = \hat{\boldsymbol{\mu}} + (\mathbf{y}^c_i - \hat{\boldsymbol{\mu}})u_i, \tag{9.3}$$

with

$$u_i = \min[k/\hat{d}_i,\ 1], \tag{9.4}$$

where k is a robustness tuning constant. The squared Mahalanobis distance of the winsorized value $\hat{\mathbf{y}}_i$ w.r.t. $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ is equal to k^2. Thus this imputation corresponds to winsorizing the Mahalanobis distance of the vector $\mathbf{y}^c_i - \hat{\boldsymbol{\mu}}$ to k while leaving its direction unchanged. In general, the tuning constant k may be larger than the

corresponding tuning constant in the foregone outlier detection in order to accommodate representative outliers. This is often necessary to avoid heavy bias when the underlying data is heavily asymmetric.

If there are missing values in the outliers then the observed variables of the outlier may be winsorized in the same way as above but with the mean and covariance are then only referring to the sub-space of observed variables. Note that when calculating the Mahalanobis distance with missing values, a factor p/q (where q is the number of present values) is applied to compensate at least partially for the number of missing dimensions.

Once outliers are winsorized, the rest of the data may be imputed under the model, conditioning on the observed variables. Since the covariance matrix used for this imputation is robust and there are no outliers left due to winsorization, imputing missing values should not create outliers (unless the model is wrong).

Bibliography

Andrews, D., **Gnanadesikan, R.** and **Warner, J.** (**1971**): *Transformations of Multivariate Data*. Biometrics, 27 (4), pp. 825–840.

Béguin, C. and **Hulliger, B.** (**2004**): *Multivariate Outlier Detection in Incomplete Survey Data: the Epidemic Algorithm and Transformed Rank Correlations*. Journal of the Royal Statistical Society, Series A, 167 (2), pp. 275–294.

Béguin, C. and **Hulliger, B.** (**2008**): *The BACON-EEM Algorithm for Multivariate Outlier Detection in Incomplete Survey Data*. Survey Methodology, Vol. 34, No. 1, pp. 91–103.

Billor, N., **Hadi, A. S.** and **Vellemann, P. F.** (**2000**): *BACON: Blocked Adaptative Computationally-efficient Outlier Nominators*. Computational Statistics and Data Analysis, 34, pp. 279–298.

Fang, K.-T., **Kotz, S.** and **Ng, K.-W.** (**1990**): Symmetric Multivariate and Related Distributions. Cha.

Gnanadesikan, R. and **Kettenring, J.** (**1972**): *Robust estimates, residuals, and outlier detection with multiresponse data*. Biometrics, 28, pp. 81–124.

Maronna, R. and **Zamar, R.** (**2002**): *Robust estimates of location and dispersion for high-dimensional datasets*. Technometrics, 44, pp. 307–317.

Maronna, R. A., **Martin, D.** and **Yohai, V. J.** (**2006**): Robust Statistics: Theory and Methods. Chichester: John Wiley.

Rousseeuw, P. J. and **Leroy, A. M.** (**1987**): Robust Regression and Outlier Detection. (Wiley Series in Probability and Statistics), Hoboken: John Wiley & Sons.

Todorov, V., **Templ, M.** and **Filzmoser, P.** (**2011**): *Detection of Multivariate Outliers in Business Survey Data with Incomplete Information*. Advances in Data Analysis and Classification (in press), accepted for publication.

Chapter 10

Robust Methods for Non-Elliptically Contoured Data

The multivariate outlier detection and imputation (MODI) methods in the previous chapters depend on strong distributional assumptions, notably, on p-variate elliptically contoured (EC) distributions. There we argued that subtable transformations can bring the data into shape, so that the distributional assumption holds (at least partially). However, the multivariate zero-inflation property of income data can not be directly addressed by transformation. Thus, for the MODI methods that are based on EC distributions (e.g., BACON-EEM), we could rely on those method's strong capability of sustaining a large amount of missing values (NA), and set all structural zeros to NA prior to detection. That is, we could cede the treatment of these NA observations to the methods and therefore to their strong underlying model. If the model holds (at least nearly so), this strategy should give reasonably good results. However, this strategy of leaving the task of dealing with the NA of this kind to the method, works fine up to a certain share of missing observations. In addition to the total amoung of missing values, the dimension, p, of the data under consideration matters, too. Thus, the larger p, the larger the possibility of observing at least one p-vector of income, the components of which are missing. As a result, we may end up with a rather large number of observations containing no information. Observe that *no information* means here that the methods cannot distinguish whether a particular NA resulted from setting a zero observation to NA or whether the observation is truely missing. In other words, such an information is uninformative. The problem of structural zeros is accentuated in the case of data on income data because several income components feature a large amount of truely zero observations. To a certain extent, one can avoid having variables with an overwhelming share of zeros by aggregating the data from similar income-component variables, such that resulting variable features more non-zero values.

Another shortcomming of setting all zero observations to NA before detection (and setting them back to zero afterwards) is that outlier-zero observations are masked. In doing so, we cannot detect wrong zeros, that is observations that were truly positive (more general, non-zero valued). Depending on the application, this may depict a serious limitation.

The Epidemic Algorithm (EA) can be considered a remedy to the above problems because it does not depend on any assumption on the distribution (Béguin and Hulliger, 2004). For EA to work, neither a transformation of the data nor setting the zero observations

to NA is needed. The Epidemic Algorithm has basically no assumption on the form of the distribution except that there is a bulk of relatively dense data while outliers are more or less separated from this bulk. But, this simplicity comes at the price of higher computational costs because EA has to compute a (upper-triangular) distance matrix consisting of $[n(n+1)]/2$ between-observation distances for a sample of size n.

10.1 Detection by the Epidemic Algorithm

The Epidemic Algorithm has been proposed by BÉGUIN and HULLIGER (2004). Here, we give only a short overview; for more details, see the original paper.

An epidemic is simulated starting at the centre of the data and spreading through it stepwise. The epidemic is, in fact, thought to be running in the population and the infection process in the population must be estimated from the sample. Missing values are accounted for in the distance underlying the infection probability. The probability that an infected point transmits the disease to an uninfected point in the next step decreases with the distance between both points. Eventually most or all of the points are infected. Typically outliers are infected late in this process or not at all. The infection times of the points are used to judge their outlyingness. To put this simple idea to work a distance is needed. The Euclidean distance is used here. To avoid unbalanced contributions of the different variables to the Euclidean distance, the variables are standardized with the median and the median absolute deviation (MAD) beforehand. Thus, let $d_{i,j}$ denote the standardized Euclidean distance between two points, x_i and x_j.

The epidemic is started from the sample spatial median (which is preferred to the coordinate-wise median because the former lies, unlike the latter, always in a dense region of the data). Assume that observation i is infected. The probability that point i infects point j depends on the distance $d_{i,j}$ through a transmission function, $h(d_{i,j})$: $P[i|j] = h(d_{i,j}) = P[j|i]$. The function h is monotone decreasing and obeys $h(0) = 1$ and $0 \leq h(d) \leq 1$. The function that is used here is the root function, $h(d) = max(1-(d_0/d)^{1/l}, 0)$, l denoting the stopping criterion of the algorithm. The reach d_0 is determined as the maximum of the distances to the nearest neighbor, $d_0 = max_i(min_{j,d_{i,j}>0}(d_{i,j}))$. In addition, we assume that at each step the infections are independent of each other.

Suppose the sample spatial median is the starting value at $t = 1$, i.e., the first infected observation. The algorithm then proceeds as follows.

1. increase the running time of the epidemic by 1 ($t := t + 1$),
2. calculate the total infection probability for all non-infected points
3. calculate the expected number of infections, ν, by the sum of the total probabilities of infection of the non-infected points; then infect those ν non-infected points with largest total probability of infection and set their infection time to t.
4. if the number of infected observations at time t, I_i, is equal to n or if $t - max(t_i : i \in I_t) > l$, then the algorithm stops. Otherwise go to step 1.

Thus the epidemic stops if all points are infected or if during l steps no infection occurred.

The main problem is the tuning of the infection process such that it has the right timing. It should not be too fast, in order to obtain a good differentiation of the infection times. And it should not be too slow in order to see accelerations and de-accelerations in the infection process which may indicate clustering or special sub-groups of the population.

10.2 Imputation by Reverse Epidemic Algorithm

Subsequent to outlier detection, one considers imputing *good* observations for those observations that have been declared outliers and the missing observations. Similarly to detecting, imputation cannot build on any model. Therefore, we start the Epidemic Algorithm from an outlying observation of the sample (Hulliger and Schoch, 2009). An imputation where the algorithm is started at an outlier and propagated until it touches one or several observations, which are not considered outliers is a natural extension of the concept for outlier detection. In order to obtain a successful imputation only complete non-outlying observations may be used as donors. The imputation is a nearest neighbour imputation since the infection time is a distance measure. The tuning of this infection running backwards is somewhat delicate: to avoid overusing certain donors the epidemic should reach several potential donors when started from an outlier. In order to impute missing values in non-outliers the Epidemic Algorithm may be started at the observation with missing values and propagated until one or several observations are infected which are able to donate the missing values. Restrictions on the amount of overlap of non-missing values between receiver and donor should be imposed in order to obtain a meaningful distance.

Bibliography

Béguin, C. and **Hulliger, B.** (**2004**): *Multivariate Outlier Detection in Incomplete Survey Data: the Epidemic Algorithm and Transformed Rank Correlations*. Journal of the Royal Statistical Society, Series A, 167 (2), pp. 275–294.

Hulliger, B. and **Schoch, T.** (**2009**): *Robustification of the quintile share ratio*. Proceedings of the NTTS Conference - New Techniques and Technologies for Statistics, Brussels: Eurostat.

Chapter 11

Robust Imputation for Compositional Data

Abstract The aim of this contribution is to present the advantages of imputation algorithms which consider the special nature of compositional data. These kind of data consist of multivariate observations that carry only relative information. The Euclidean geometry is no longer valid and standard statistical methods should only be applied after an isometric transformation to an unrestricted space. It is shown that our proposed algorithms are designed to deal with compositional data. Within a simulation study it is demonstrated that these algorithms perform better than other imputation algorithms. It is also shown that multiple imputation can be made, and we give an outline to possible problems when applying these methods to real-world data sets such as EU-SILC.

Note that an paper is already published (Hron et al., 2010) which detaily explains the algorithm for model-based imputation of compositional data. In this chapter we summarize the paper and concentrate on problems in official statistics.

11.1 Introduction

11.1.1 Imputation

Many different methods for imputation have been developed over the last few decades. The techniques for imputation can be subdivided into four categories: univariate methods such as column-wise (conditional) mean or median imputation, distance-based imputation methods such as k-nearest neighbor imputation, covariance-based methods such as the well-known expectation maximization imputation method, and model-based methods such as regression imputation. Most of these methods are able to deal with missing completely at random (MCAR) and missing at random (MAR) missing values mechanism (see, e.g., Little and Rubin, 1987). However, most of the existing methods assume that the data originate from a multivariate normal distribution. In addition to that, almost all methods are not designed to deal with compositional data.

11.1.2 Compositional Data

Compositional data occur frequently in official statistics. Examples are expenditure data, income components in tax data or in the EU-SILC data, wage components in the Earnings Structure Survey GRAF (2006), components of turnover of enterprises etc., and all data that carry all the relevant information in the ratios between the components (parts).

Compositions are also known as data which parts sum up to a certain constant, e.g. 100 in case of percentages.. The estimation of missing values in compositional data is a common problem not only in official statistics, but also in various other fields (see, e.g., FILZMOSER and HRON, 2008b). The estimation becomes even more complex if outliers are present in the data.

Advanced (robust) imputation methods have turned out to work well for data with a direct representation in the Euclidean space. However, this is not the case when dealing with compositional data (see, e.g., MARTÍN-FERNÁNDEZ et al., 2003; BOOGAART et al., 2006; HRON et al., 2008; ALBALADEJO and MARTÍN-FERNÁNDEZ, 2008). It can be shown that existing imputation methods are not appropriate for compositional data.

An observation $\boldsymbol{x} = (x_1, \ldots, x_D)$ is called a D-part composition if, and only if, all its components are strictly positive real numbers and all the relevant information is included in the ratios between them AITCHISON (1986). One can thus define the *simplex*, which is the sample space of D-part compositions, as

$$\mathcal{S}^D = \{\boldsymbol{x} = (x_1, \ldots, x_D),\ x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa\} \quad . \tag{11.1}$$

Note that the constant sum constraint κ implies that D-part compositions are only $D-1$ dimensional, so they are singular by definition. It is, however, possible that the constant κ is different for each observation without loss of information.

The application of standard statistical methods, like correlation analysis or principal component analysis, directly to compositional data can lead to biased and meaningless results (see, e.g., FILZMOSER and HRON, 2008b,a). This is also true for imputation methods.

11.2 One-to-One Transformations for Compositional Data and Their Properties Related to Imputation

A way out is to first transform the data with an appropriate transformation. Such transformations, preserving the specific geometry of compositional data on the simplex (also called Aitchison geometry), are represented by the family of logratio transformations: additive (alr), centered (clr) and isometric (ilr) logratio transformations AITCHISON (1986); EGOZCUE et al. (2003). Standard statistical methods can then be applied to the transformed data, and the results can be back-transformed to the original space.

By applying the alr transformation, all values are divided by the values of the j-th variable (compositional part),

$$\boldsymbol{x}^{(j)} = \left(x_1^{(j)}, \ldots, x_{D-1}^{(j)}\right) = \left(\ln \frac{x_1}{x_j}, \ldots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \ldots, \ln \frac{x_D}{x_j}\right) \quad . \tag{11.2}$$

The index $j \in \{1, \ldots, D\}$ refers to the "ratioing" part used.

There are two disadvantages of the alr transformation in the context of imputation: Firstly, the "ratioing" variable should not contain any missing values, and, secondly, in general the imputed values will differ when using another "ratioing" variable, i.e. this transformation is not isometric which, for example, plays a crutial role in detecting outliers.

The ilr transformation can be identified with representation of compositions in coordinates with respect to an orthonormal basis of D-1 compositions on the simplex, and it is also an isometric transformation.

However, for imputation purposes a special choice of the ilr transformation is necessary: We define the ilr transformed data as $ilr(\boldsymbol{x}) = \boldsymbol{z} = (z_1, \ldots, z_{D-1})$, where

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{\sqrt[D-j]{\prod_{l=j+1}^{D} x_l}}{x_j}, \qquad \text{for } j = 1, \ldots, D-1 \quad . \tag{11.3}$$

Equation (11.3) ensures that missing values in the first compositional part only affect the first ilr variable z_1, leading to a better stability of the transformed variables. It is thus useful to re-arrange the original variables in descending amounts of missing values.

The corresponding inverse transformation is $ilr^{-1}(\boldsymbol{z}) = \boldsymbol{x} = (x_1, \ldots, x_D)$, with

$$x_1 = \exp\left(-\frac{\sqrt{D-1}}{\sqrt{D}} z_1\right), \tag{11.4}$$

$$x_j = \exp\left(\sum_{l=1}^{j-1} \frac{1}{\sqrt{(D-l+1)(D-l)}} z_l - \frac{\sqrt{D-j}}{\sqrt{D-j+1}} z_j\right), \text{ for } j = 2, \ldots, D-1, \tag{11.5}$$

$$x_D = \exp\left(\sum_{l=1}^{D-1} \frac{1}{\sqrt{(D-l+1)(D-l)}} z_l\right). \tag{11.6}$$

11.3 Challenges

11.3.1 Outliers

It is well-known that outliers influence standard imputation methods and therefore robust methods should be preferred (see, e.g, Hron et al., 2008). This also holds for compositional data, where standard imputation methods will be influenced by outlying observations.

The effect of outliers in the geometry of compositional data can be visualized using 3-part compositions. In that case the data can be presented in a planar graph, called ternary diagram, which is an equilateral triangle $X_1X_2X_3$ such that a composition $\boldsymbol{x} = (x_1, x_2, x_3)$ is plotted at a distance x_1 from the opposite side of vertex X_1, at a distance x_2 from the opposite side of vertex X_2, and at a distance x_3 from the opposite side of vertex X_3 (see, e.g., AITCHISON, 1986).

We generate 90 observations with 3 parts, being normally distributed on the simplex (i.e. they are multivariate normally distributed in the 2-dimensional ilr space). The data points are shown in the ilr space in Figure 11.1 (right), and in a ternary diagram in Figure 11.1 (left). In addition to the normally distributed data, 5 outliers (*group 1*) are added, that are shown as green crosses in the plots. Furthermore, another group (*group 2*) of 5 outliers is added that is only affecting the Euclidean space (blue triangles). Both types of outliers are simulated to have a considerably higher sum of their parts, which is not visible in the ternary diagram AITCHISON (1986) in Figure 11.1 (left) where the parts are re-scaled to have sum 1. The points of group 2 originate from the same distribution as the clean data, whereas the points from group 1 are simulated from a different distribution.

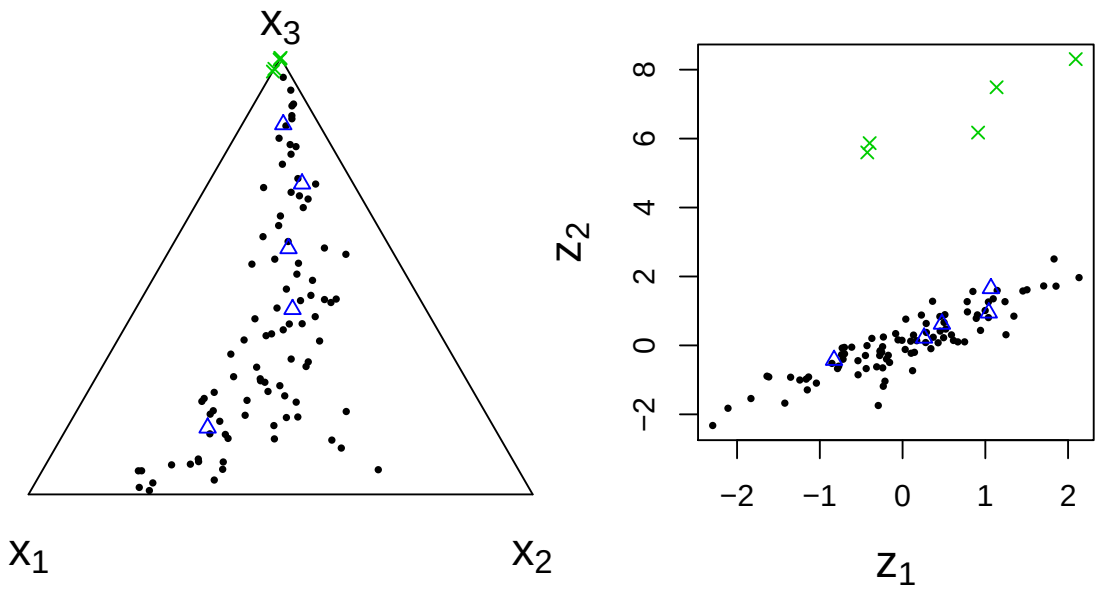


Figure 11.1: Simulated data set with 5 points from *outlier group 1* (symbol ×) and 5 points from *outlier group 2* (symbol △). Left plot: 3-part compositions shown in the ternary diagram; right plot: data after ilr transformation.

11.3.2 The Structure of Missing Values

Before imputation, especially before model-based imputation, one should be aware of the multivariate structure of the missing values. The multivariate structure could be explored by using the R-package `VIM` TEMPL and ALFONS (2008); TEMPL and FILZMOSER (2008). 13 different plot methods are implemented in this package which allow to interactively visualize data with missing values.

11.3.3 Zeros

Practical data sets like income components often have zero entries (e.g. zero income for a specific income component). This problem is in the context of compositions also known under the term structural zeros problem. Compositional data including zeros are problematic for logratio transformations because the logarithm of zero is infinity. As a way out, one could replace zeros by small values Aitchison and Kay (2003). However, by applying this procedure, the ratios between the parts might become very high or small. In other words, outliers will be introduced which might affect non-robust imputation algorithms.

Another way to deal with zeros is to combine variables so that no zeros remain in the data. This is straightforward because often some compositional parts have a very similar meaning, e.g., household income from *children allowances* are similar to income from *other family allowances*, and it may be reasonable to combine such variables. After eliminating the zeros, the imputation can be done in the usual way. Finally, the imputed values have to be split into the original income components. For this purpose, the information of the current sample can be used (see, e.g., Kraft, 2009) which, however, is not easy to perform in the Aitchison geometry.

Usually, those compositions which include zeros in a certain part are conceptionally different from compositions including no zeros in this part Bacon-Shone (2003). The algorithms proposed by Hron et al. (2008) and presented below could be adapted to deal with zeros. Firstly, it must be determined if a zero or a positive value should be imputed by using a certain model. Secondly, a subset of the variables can be used to impute the missing values for those missing values that are grouped in the "non-zero" part.

11.3.4 Measuring the Uncertainty of the Imputations

Little and Rubin (1987) suggest to estimate standard errors for estimators via bootstrapping, and they outline two approaches – a modified bootstrap approach and a modified jackknife procedure – to obtain consistent standard errors when data are imputed. Each bootstrap sample is drawn from the raw data and afterwards the missing values in the bootstrap samples are imputed and the bootstrap replicates are computed. Confidence intervals obtained from these bootstrap replicates are in general slightly larger as if the imputation were done before the bootstrap samples are drawn. A result within our proposed procedure is shown in Templ et al. (2009a). It reveals that mean imputation – a simple method still frequently applied – can lead to higher uncertainty, and that the results are biased, whereas our proposed model-based procedure provides reasonable results.

Another possibility to take care of the variability due to imputations is to use (proper) multiple imputation methods (e.g., Rubin, 1987). While a proof that a method is proper in Rubin's sense is almost impossible, methods can be evaluated by simulation if they are proper or at least approximately proper.

11.4 Imputation Algorithms for Compositional Data

In the following we briefly describe the imputation methods that have been implemented in the R-package `robCompositions`. The detailed description of the algorithms can be found in HRON et al. (2010).

11.4.1 k-Nearest Neighbor Imputation

k-nearest neighbor imputation is usually based on Euclidean distances. Since compositional data are represented only in the simplex sample space, a different distance measure needs to be used, like the Aitchison distance, being defined for two compositions $\boldsymbol{x} = (x_1, \ldots, x_D)$ and $\boldsymbol{y} = (y_1, \ldots, y_D)$ as

$$d_a(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2} \quad . \tag{11.7}$$

Thus, the Aitchison distance takes care of the property that compositional data include their information only in the ratios between the parts.

Once the k-nearest neighbors to an observation with missing parts have been identified, their information is used to estimate the missings. For reasons of robustness, the estimation can be based on using medians rather than means. If the compositional data do not sum up to a constant, it is important to use an adjustment according to the sum of all parts prior to imputation. For details, see HRON et al. (2010).

11.4.2 alr-EM algorithm

This approach was introduced by PALAREA-ALBALADEJA and MARTÍN-FERNÁNDEZ (2008). Here a kind of EM-algorithm DEMPSTER et al. (1977) for alr-transformed (see Equation (11.2)) compositional data was used for imputation. In fact, this algorithm is an iterative regression-based method: at every step of the algorithm one variable is chosen to be the response and a regression fit is taken with all other transformed variables. Missing values in the response are then updated using the obtained regression coefficients. The procedure is continued until "convergence". Although this algorithm was originally introduced for replacing rounded zeros of geochemical data under a certain detection limit, it can be easily adapted for estimating missing values without using the restrictions that the missings originate from such rounded zeros. This approach, however, is not robust against outlying observations, and the alr-transformation is not isometric. Thus, robustifying this method may not improve the imputation because of the lack of good geometrical properties. In the following, this algorithm will be denoted by alr-EM.

11.4.3 Iterative Robust Model-Based Imputation

In the second approach we initialize the missing values with the proposed k-nearest neighbor approach. Then the data are transformed to the $D-1$ dimensional real space using

the ilr transformation from Equation (11.3). The ilr transformation holds the so-called isometric property,

$$d_a(\boldsymbol{x}, \boldsymbol{y}) = d_e(ilr(\boldsymbol{x}), ilr(\boldsymbol{y})) \tag{11.8}$$

Egozcue and Pawlowsky-Glahn (2005). Here, d_e denotes the Euclidean distance. Consequently, one can use standard statistical methods like multiple linear regression, that work correctly in the Euclidean space. We use a special form of the ilr transformation and its inverse shown in Equations (11.3) and (11.4) to (11.6). Here, the compositional parts are rearranged such that x_1 includes the highest amount of missings, x_2 the second highest, and so on. Thus, when performing a regression of z_1 on $z_2, \ldots, z_{D-1}$, only z_1 will be influenced by the initialized missings in x_1, but not the remaining ilr variables.

The idea of the procedure is thus to iteratively improve the estimation of the missing values. After the regression of z_1 on $z_2, \ldots, z_{D-1}$, the results are back-transformed to the simplex, and the cells that were originally missing are updated. Next we consider the variable which originally has the second highest amount of missings, and the same regression procedure as before is applied in the ilr space. After each variable containing missings has been proceeded, one can start the whole process again until the estimated missings stabilize. The detailed description of this algorithm can be found in Hron et al. (2010).

As a regression method we propose to use robust regression, like LTS regression (Maronna et al., 2006), especially if outliers are included in the data.

Multiple imputation is provided by drawing values from their posterior predictive distribution. In the context of regression, normal noise with mean zero and variance corresponding to the residual variance is added to the point estimation of a specific value in direction of the response variable. It can easily be shown that this modification of the algorithm (denoted by LTS MI (ilr), see Figure 11.2 and 11.3) leads to a proper, or at least an approximative proper multiple imputation method. However, this can only hardly be shown in an analytic form. Nevertheless, if estimations are taken from a complete data set and if this estimations have similar properties as the average of multiple imputations of several data sets generated randomly from the complete data set, then the imputation method is proper or approximately proper in a frequentist's way of thinking.

11.4.4 Other Imputation Methods for Compositional Data

Mean imputation is still frequently applied, although it is well-known that such an imputation reduces the variance of the variables, and that the multivariate structure of the data is not respected. For compositional data one should replace (arithmetic) mean imputation by using the column-wise geometric mean, because the geometric mean accounts for the relative information of compositional data. Moreover, the geometric mean represents the best linear unbiased estimator of the center of the distribution with respect to the Aitchison geometry.

In Martín-Fernández et al. (2003) the estimation of missing values in compositional data was done in the sense of the Aitchison geometry, but with the constraint of constant sum of the parts. They impute missing compositional parts and adjust the non-missings

so that the row sums are constant. However, this concept is not meaningful whenever the compositions sum up to different constants, which is usually the case for expenditures data, income components, tax components, etc. In official statistics it may even not be acceptable to change non-missing values that fulfill given editing rules.

11.5 Results

11.5.1 Data Used

As an example, the 5-dimensional expenditure data set from AITCHISON (1986) is used. Missing values are generated in the first four parts with equal amount in order to keep things simple. Note that our proposed iterative method provides the best results if the amount of missing values is different for each variable (see, e.g., TEMPL et al., 2009a). While in the simulations performed in HRON et al. (2008) and TEMPL et al. (2009a) the percentage of the outliers is varied, we concentrate on varying the amount of missing values without introducing outliers. 1000 simulations were run for 3%, 5%, 10%, 15% and 20% of missing values in the first four parts, and the arithmetic means of the simulation results are presented in Figure 11.2. The two graphics on the left report the results of imputation methods that do not account for the compositional nature of the data, while the graphics on the right show the results from methods considering the special properties of compositional data. The method *iterative LTS (ilr)* refers to the algorithm proposed by HRON et al. (2008), and *iterative LS (ilr)* replaces robust regression by classical regressions. The counterparts denoted by *no transf.* are applied in the original space without using the ilr transformation.

Two different quality measures of the imputed values are used: The two upper graphics provide the results for the compositional error variance TEMPL et al. (2009a), measuring closeness of the imputed values in the Aitchison geometry. The two bottom graphics report the relative difference in covariance structure HRON et al. (2008); TEMPL et al. (2009a), expressing the influence of the imputation to the multivariate covariance structure.

The results in Figure 11.2 show that working in the inappropriate geometry (left two graphics) generally leads to much poorer quality of the imputed values. When working in the correct geometry, varying the percentage of missing values does not much effect the results. Moreover, the considered methods lead to comparable quality. The regression based methods are preferable in terms of compositional error variance. Since classical and robust regression only yield a marginal difference, there might not be severe outliers included in the expenditures data set.

In the second experiment we will use simulated data, where the data structure and the outliers are exactly known. The data structure of one single realization is presented in Figure 11.1. Additionally, 10 values (of the non-outliers) are set to zero. We keep the percentage of outliers fixed with 5%. The average amount of missing values vary from 1 percent to 20 percent whereas the first part includes twice as high missing values than the second part. The probabilty of missing of the second part depends on the third part (MAR situation). Again the results presented in Figure 11.3 show the same tendency as before. The quality of the imputation is generally improved if methods are used that

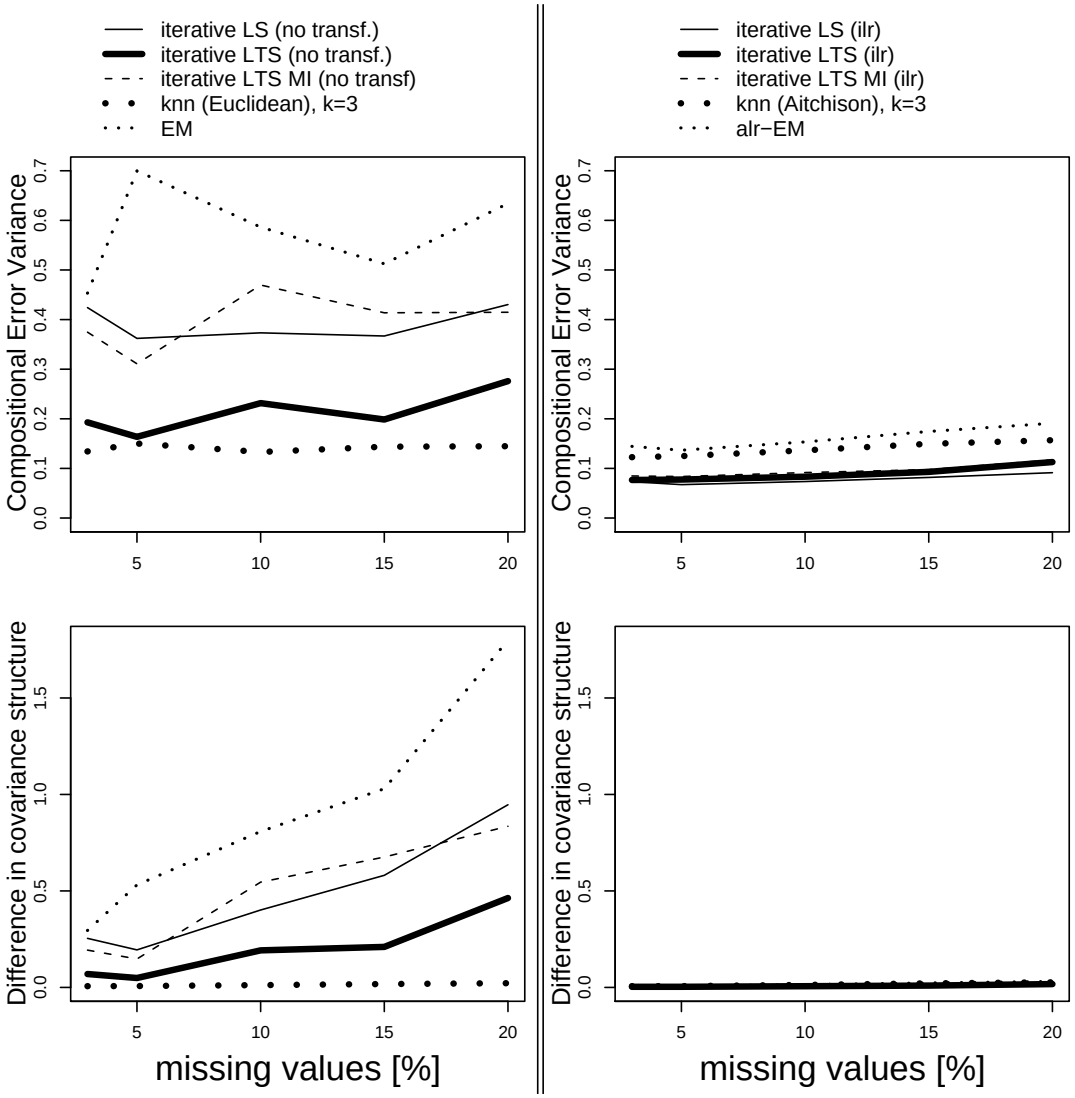


Figure 11.2: Simulation results with varying percentage of missing values for the expenditures data set: methods are applied in the original space (left column) and in the appropriate space (right column); the quality of the imputation is measured by the compositional error variance (upper row) and by the difference in covariance structure (bottom row).

consider the compositional nature of the data. The best results are obtained with our proposed robust iterative method.

We performed numerous simulations with different data dimension (up to 10 compositional parts), with different covariance structure, different outlier models, different missing values mechanisms, etc. We also tested more than 15 standard imputation methods, with the conclusion that they show poorer performance in the context of compositional data. All results supported that our iterative model based method based on a special ilr transformation performs best.

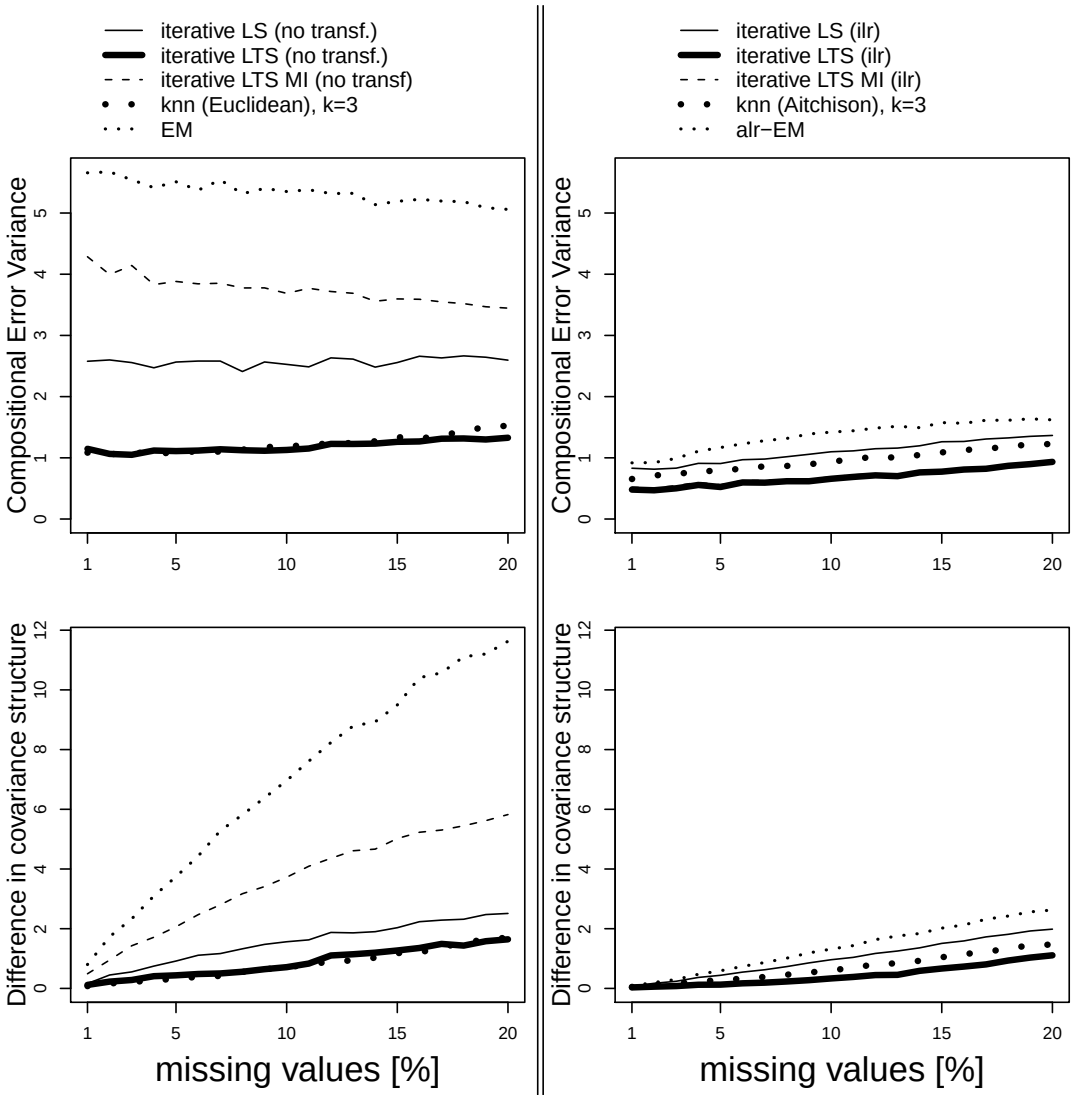


Figure 11.3: Simulation results with varying percentage of missing values for simulated data according to Figure 11.1. The results are presented in analogy to Figure 11.2 expect that robust methods performs better than non-robust.

11.6 Conclusions

This paper focuses on various aspects for the imputation of missing values in the context of compositional data. Different data transformations are proposed, and their advantages and shortcomings are discussed. Some of this shortcomings like a high amount of zeros in the income parts of EU-SILC are still under research focus and no solution have been presented there. The lack of the methods for compositional data is that a log-ratio transformation cannot deal with zeros. While there exists already some solutions for detection limit problems in geochemical data sets, the zero problem in official statistics is untouched untill now. Note that we was not be aware of this problem and so nothing is outlined in the work package description. However, we found the problem of interest and showed some possible solutions and give some simulation results on simplified situations without zeros. This is typically the situation when we want to impute only those subgroups with

having positive values in each income parts.

Simulation results show that our model-based robust imputation method performs best, and that standard imputation methods should not be applied to compositional data. The algorithm is described in detail in HRON et al. (2010), where also simulation results with different settings of outliers are presented, see also TEMPL et al. (2009a). This paper compares the quality of the imputation when the amount of missing values is varied and when the compositions include zeros. The algorithm is provided in the R-package `robCompositions` TEMPL et al. (2009b).

Bibliography

Aitchison, J. (**1986**): The Statistical Analysis of Compositional Data. Chapman & Hall, London.

Aitchison, J. and **Kay, J.** (**2003**): *Possible solutions of some essential zero problems in compositional data analysis*. In Proceedings of the CoDa Workshop (CoDa 2003), Spain, p. 6.

Albaladejo, P. and **Martín-Fernández, J.** (**2008**): *A modified EM alr-algorithm for replacing rounded zeros in compositional data sets*. Computer & Geosciences, 34 (8), pp. 902–917.

Bacon-Shone, J. (**2003**): *Modelling structural zeros in compositional data*. In Proceedings of the Compositional Data Analysis Workshop (CoDaWork 2003), Spain, p. 4.

Boogaart, K., **Tolosana-Delgado, R.** and **Bren, M.** (**2006**): *Concept for handling with zeros and missing values in compositional data*. **Pirard, E.** (editor) Proceedings of IAMG'06 - The XI annual conference of the International Association for Mathematical Geology, University of Liege, Belgium. CD-ROM., 4 pages.

Dempster, A., **Laird, N.** and **Rubin, D.** (**1977**): *Maximum Likelihood from Incomplete Data via the EM Algorithm*. J. Royal Stat. Soc., 39, pp. 1–38, doi:http://dx.doi.org/10.2307/2984875.
URL http://web.mit.edu/6.435/www/Dempster77.pdf

Egozcue, J. and **Pawlowsky-Glahn** (**2005**): *Groups of parts and their balances in compositional data analysis*. Mathematical Geology, 37 (7), pp. 795–828, doi:10.1007/s11004-005-7381-9.
URL http://www.springerlink.com/content/fx037244708561v5/

Egozcue, J., **Pawlowsky-Glahn, V.**, **Mateu-Figueras, G.** and **Barceló-Vidal, C.** (**2003**): *Isometric log-ratio transformations for compositional data analysis*. Mathematical Geology, 35 (3), pp. 279–300, doi:10.1023/A:1023818214614.
URL http://www.springerlink.com/content/wx1166n56n685v82/

Filzmoser, P. and **Hron, K.** (**2008**a): *Correlation Analysis for Compositional Data*. Research report sm-2008-2, Department of Statistics and Probability Theory, Vienna University of Technology.

URL http://www.statistik.tuwien.ac.at/forschung/SM/SM-2008-2complete.pdf

Filzmoser, P. and **Hron, K.** (**2008**b): *Outlier Detection for Compositional Data Using Robust Methods*. Mathematical Geosciences, 40 (3), pp. 233–248, doi:http://dx.doi.org/10.1007/s11004-007-9141-5.
URL http://www.springerlink.com/content/d662421553216861

Graf, M. (**2006**): *Swiss earnings structure survey. Compositional data in a stratified two-stage sample*. Metholology report, isbn: 3-303-00338-6, Swiss Federal Statistical Office.
URL http://www.bfs.admin.ch/bfs/portal/de/index/themen/00/07/blank/02.Document.77975.pdf

Hron, K., **Templ, M.** and **Filzmoser, P.** (**2008**): *Imputation of compositional data using robust methods*. Research report sm-2008-4, Department of Statistics and Probability Theory, Vienna University of Technology.
URL http://www.statistik.tuwien.ac.at/forschung/SM/SM-2008-4complete.pdf

Hron, K., **Templ, M.** and **Filzmoser, P.** (**2010**): *Imputation of missing values for compositional data using classical and robust methods*. Computational Statistics & Data Analysis, 54 (12), pp. 3095–3107, ISSN 0167-9473, doi:DOI:10.1016/j.csda.2009.11.023.

Kraft, S. (**2009**): *The generation of a population for the EU-SILC data*. Vienna University of Technology. Master Thesis, 2009, to appear.

Little, R. and **Rubin, D.** (**1987**): Statistical Analysis with Missing Data. Wiley, New York.

Maronna, R., **Martin, R.** and **Yohai, V.** (**2006**): Robust Statistics: Theory and Methods. John Wiley & Sons, New York.

Martín-Fernández, J., **Barceló-Vidal, C.** and **Pawlowsky-Glahn, V.** (**2003**): *Dealing with zeros and missing values in compositional data sets using nonparametric imputation*. Mathematical Geology, 35 (3), pp. 253–278.

Palarea-Albaladeja, J. and **Martín-Fernández, J.** (**2008**): *A modified EM alr-algorithm for replacing rounded zeros in compositional data sets*. Computer & Geosciences, 34 (8), pp. 902–917.

Rubin, D. (**1987**): Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Templ, M. and **Alfons, A.** (**2008**): VIM: Visualization and Imputation of Missing Values. R package version 1.2.4.
URL http://cran.r-project.org

Templ, M. and **Filzmoser, P.** (**2008**): *Visualization of missing values using the R-package VIM*. Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology.
URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2008-1complete.pdf

Templ, M., **Filzmoser, P.** and **Hron, K.** (**2009**a): *Compositional Data Using the R ,-Package robCompositions*. In Proceedings of the International Conference on New Techniques and Technologies in Statistics (NTTS 2009), Bruessels, p. 11.

Templ, M., **Hron, K.** and **Filzmoser, P.** (**2009**b): robCompositions: Robust Estimation for Compositional Data. R package version 1.2.2.

Chapter 12

Robust Methods for Semi-Continuous Data

This chapter is based on the master thesis of Meraner (2010) that was part of the AMELI project.

In general, data on income exhibit high percentage of zeros which, in turn, results in semi-continuous variables. This leads to a serious limitation of outlier-detection methods. One approach consists of treating the strctural zeros in the data as if they were missing values and subsequently impute for these *missings* with an appropriate (but not necessary robust) imputation method. Finally, one may apply conventional outlier-detection methods on the imputed data. A possible disadvantage of this approach is the strong dependence on the performance of the imputation method used. Therefore, we concentrate on estimates which omit observations with zeros. However, this causes a problem for multivariate methods due to the fact that excluding observations with zeros might render a data matrix far too small for drawing significant conclusions.

Hence, a pairwise approach of certain multivariate methods seems sensible because it is possible to make use of a considerable amount of observations from the actual data without having to resort to imputation. In this context, we adapted three robust estimators for the estimation of location and dispersion using the pairwise approach, namely the OGK estimator (Maronna and Zamar, 2002), the quadrant correlation estimate (Shevlyakov, 1997; Blomqvist, 1950; Mosteller, 1946) and an estimator based on robust PCA (Locantore et al., 1999).

These adaptations were implemented in the statistical environment R and compared to the original pairwise procedures as well as to two multivariate procedures, the MCD estimator (Rousseeuw, 1985) and the BACON-EEM (Béguin and Hulliger, 2008) algorithm.

12.1 Adaption of Robust Methods for Semi-continuous Variables

Semi-continuous variables frequently appear in survey data and are particularly difficult to handle because of the large proportion of zeros. Most of the effective methods for

multivariate outlier detection require a prior imputation since too many observations may be lost by merely excluding all the zeros. This however, makes the estimates strongly dependent on the type of imputation used. Consequently, it seems advantageous to use a pairwise approach for the robust estimation of location and scatter because in this case, it is possible to exclude the zeros in a pairwise manner which does not sacrifice as much information and increases the amount of äctual dataïncluded in the process of estimation.

12.1.1 The OGK Estimator for Semi-continuous Variables

The orthogonalized Gnanadesikan-Kettenring (OGK) covariance matrix estimator (MARONNA and ZAMAR, 2002) uses a specific pairwise approach which was first proposed by GNANADESIKAN and KETTENRING (1972) and is defined as

$$cov(X,Y) = \frac{1}{4}(\sigma(X+Y)^2 - \sigma(X-Y)^2) \quad , \tag{12.1}$$

where σ is a robust estimate of the standard deviation and X, Y is a pair of random variables. However, the resulting multivariate scatter matrix is neither affine equivariant nor necessarily positive semidefinite. Affine equivariance is, however, desirable but not mandatory, and it can be sacrificed for other properties such as computational speed (MARONNA et al., 2006).

OGK covariance-matrix estimator

MARONNA and ZAMAR (2002) subsequently propose a way to obtain a robust covariance matrix predicated on Equation (12.1), which is both positive semidefinite and approximately affine equivariant. Their procedure, also described e.g. in FILZMOSER et al. (2008) and in more detail in MARONNA et al. (2006), is implemented in R, e.g. in the R packages **robustbase** (MAECHLER et al., 2009) and **rrcov** (TODOROV, 2010). It is based on the idea that the eigenvalues of a covariance matrix are the variances along the directions given by the respective eigenvectors. If the variables in eigenvector space are orthogonal, the covariances are equal to zero. Hence it is sufficient to obtain robust variance estimates of the data projected onto each eigenvector direction and then replace the eigenvalues with these robust variances. Finally, the eigenvector transformation is applied in reverse in order to obtain a positive semidefinite robust covariance matrix.

Notably, MARONNA and ZAMAR (2002) take the identity

$$\sigma^2(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\mathbf{\Sigma}\mathbf{a}, \qquad \forall \mathbf{a} \in \mathbb{R}^p \tag{12.2}$$

as their starting point, where $\mathbf{\Sigma}$ is the covariance matrix of the p-dimensional random vector $\mathbf{x}$ and σ denotes the standard deviation. The lack of positive semi-definiteness is

then overcome by a modification that forces Equation (12.2) for a robust σ and a set of *principal directions*.

This whole procedure is expressed in more mathematical terms by Maronna and Zamar (2002) as follows. Let $\mathbf{X} = [x_{ij}]$ be an $n \times p$ matrix with rows $\mathbf{x}_i^T$, $i = 1, \ldots, n$, and columns $\mathbf{x}^j$, $j = 1, \ldots, p$. Let $\sigma(\cdot)$ and $\mu(\cdot)$ be robust univariate dispersion and location statistics, and let $cov(\cdot, \cdot)$ denote a robust estimate of the covariance of two variables. A robust scatter matrix $\mathbf{C}(\mathbf{X})$ and a robust location vector $\mathbf{t}(\mathbf{X})$ are obtained by the following computational steps:

1. Let $\mathbf{D} = \text{diag}(\sigma(\mathbf{x}^1), \ldots, \sigma(\mathbf{x}^p))$. To make the estimate scale equivariant, compute a normalized data matrix $\mathbf{Y}$ with columns

$$\mathbf{y}^j = \mathbf{x}^j / \sigma(\mathbf{x}^j), \qquad j = 1, \ldots, p \tag{12.3}$$

   and hence rows

$$\mathbf{y}_i = \mathbf{D}^{-1}\mathbf{x}_i, \qquad i = 1, \ldots, n \tag{12.4}$$

2. Compute the robust "correlation matrix" $\mathbf{U} = [U_{jk}]$ of $\mathbf{X}$ by applying $cov(\cdot, \cdot)$, i.e. (12.1), to the columns of $\mathbf{Y}$. Respectively,

$$U_{jj} = 1, \qquad \text{and} \qquad U_{jk} = \frac{1}{4}(\sigma(\mathbf{y}^j + \mathbf{y}^k)^2 - \sigma(\mathbf{y}^j - \mathbf{y}^k)^2), \qquad j \neq k. \tag{12.5}$$

3. Determine the eigenvalues λ_j and eigenvectors $\mathbf{e}_j$ of $\mathbf{U}$, $j = 1, \ldots, p$, and let $\mathbf{E}$ be the matrix whose columns are the $\mathbf{e}_j$'s. The principal components of the standardized variables are obtained from the eigenvectors of the correlation matrix $\mathbf{U}$ of the original variables, hence we get

$$\mathbf{U} \equiv \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T \qquad \text{where} \qquad \mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p). \tag{12.6}$$

   Here the λ_j's may be negative. (12.6) corresponds to the principal component decomposition of $\mathbf{Y}$.

4. Let

$$\mathbf{A} = \mathbf{D}\mathbf{E}, \qquad \text{and} \qquad \mathbf{z}_i = \mathbf{E}^T\mathbf{y}_i = \mathbf{E}^T\mathbf{D}^{-1}\mathbf{x}_i = \mathbf{A}^{-1}\mathbf{x}_i \tag{12.7}$$

   so that

$$\mathbf{x}_i = \mathbf{A}\mathbf{z}_i \quad , \tag{12.8}$$

with $\mathbf{z}^1, \ldots, \mathbf{z}^p$ being the principal components of $\mathbf{Y}$.

5. Compute $\sigma(\mathbf{z}^j)$ and $\mu(\mathbf{z}^j)$ for $j = 1, \ldots, p$ and set

$$\mathbf{\Gamma} = \text{diag}(\sigma^2(\mathbf{z}^1), \ldots, \sigma^2(\mathbf{z}^p)) \qquad \text{and} \qquad \nu = (\mu(\mathbf{z}^1), \ldots, \mu(\mathbf{z}^p))^T. \tag{12.9}$$

Since the $\mathbf{z}^1, \ldots, \mathbf{z}^p$ should be approximately uncorrelated, $\mathbf{\Gamma}$ corresponds to their diagonal covariance matrix.

6. Transform back to $\mathbf{X}$ using Equation (12.8) and finally define

$$\mathbf{C}(\mathbf{X}) = \mathbf{A}\mathbf{\Gamma}\mathbf{A}^T \tag{12.10}$$

and

$$\mathbf{t}(\mathbf{X}) = \mathbf{A}\nu \quad , \tag{12.11}$$

where the λ_j's are replaced by robust variances $\sigma^2(\mathbf{z}^j)$.

Motivation for Formula (12.11) is the fact that results are better when $\mathbf{t}(\mathbf{X})$ is computed by applying a coordinate-wise location estimate to the (approximately uncorrelated) $\mathbf{z}^j$'s and then transform back to the $\mathbf{X}$-coordinates rather than applying it directly to the $\mathbf{x}^j$'s. In general, Equations (12.10) and (12.11) can be justified by the argument that if ν and $\mathbf{\Gamma}$ were the mean and covariance matrix of $\mathbf{Z}$, the mean and covariance matrix of $\mathbf{X}$ would be given by said equations because of the identity $\mathbf{x}_i = \mathbf{A}\mathbf{z}_i$ from Equation (12.8).

The procedure can be iterated by computing $\mathbf{C}$ and $\mathbf{t}$ according to Equations (12.10) and (12.11) following steps 1-6, now, however, for the matrix $\mathbf{Z}$ with rows $\mathbf{z}_i = \mathbf{A}^{-1}\mathbf{x}_i$ from Equation (12.7), rendering $\mathbf{C}(\mathbf{Z})$ and $\mathbf{t}(\mathbf{Z})$, which are then expressed back in the original coordinate system by solving the following equations

$$\mathbf{C}_{(2)}(\mathbf{X}) = \mathbf{A}\mathbf{C}(\mathbf{Z})\mathbf{A}^T \tag{12.12}$$

and

$$\mathbf{t}_{(2)}(\mathbf{X}) = \mathbf{A}\mathbf{t}(\mathbf{Z}). \tag{12.13}$$

This second step makes use of the fact that the $\mathbf{z}^j$'s are expectedly less correlated than the original variables. Further iterations are possible (see MARONNA et al., 2006) but MARONNA and ZAMAR (2002) found that iterations beyond the second do not lead to

an improvement. A potential final step is a re-weighting procedure which is supposed to improve the estimators by increasing their efficiency and making them *more equivariant*. It is done by using a weight function W, with W being an indicator function

$$W(d) = I(d \leq d_0), \tag{12.14}$$

which is employed to obtain the weighted location estimate

$$\mathbf{t}_w = \frac{\sum_{i=1}^{n} w_i \mathbf{x}_i}{\sum_{i=1}^{n} w_i} \tag{12.15}$$

and the weighted scatter matrix

$$\mathbf{C}_w = \frac{\sum_{i=1}^{n} w_i (\mathbf{x}_i - \mathbf{t}_w)(\mathbf{x}_i - \mathbf{t}_w)^T}{\sum_{i=1}^{n} w_i} \quad , \tag{12.16}$$

where each $\mathbf{x}_i$ has weight $w_i = W(d_i)$ for the robust distances d_i. In other words, observations whose robust distances satisfy $d \leq d_0$ receive full weight while for the remaining observations with $d > d_0$ zero weight is assigned. The threshold d_0 is set to

$$d_0 = \frac{\text{med}(d_1, \ldots, d_n)\sqrt{\chi^2_p(\beta)}}{\sqrt{\chi^2_p(0.5)}} \quad , \tag{12.17}$$

where $\chi^2_p(\beta)$ is the β-quantile of the χ^2-distribution with p degrees of freedom. According to Maronna and Zamar (2002), $\beta = 0.9$ usually yields the best results, with $\beta = 0.95$ being almost as good.

For the computation of the robust distances d, no matrix inversion is needed since they can be calculated in the eigenvector space where the p components are orthogonal. Hence

$$d_i = \sum_{j=1}^{p} \left(\frac{z_{ij} - \mu(\mathbf{z}^j)}{\sigma(\mathbf{z}^j)} \right)^2, \qquad i = 1, \ldots, n \quad , \tag{12.18}$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ are robust univariate dispersion and location statistics as stated above. The entries z_{ij} of the transformed data matrix $\mathbf{Z}$ and the principal components $\mathbf{z}^j$ correspond to Equation (12.7).

Modifications of the OGK method for Semi-Continuous Data

Consider an $n \times p$ data matrix $\mathbf{X}$ with rows $\mathbf{x}_i$, $i = 1, \ldots, n$ and columns $\mathbf{x}^j$, $j = 1, \ldots, p$, where each column may contain a certain amount of zeros. A robust scatter matrix $\mathbf{C}(\mathbf{X})$ and a robust location vector $\mathbf{t}(\mathbf{X})$ are obtained according to the following computational steps.

1. Let $\sigma(\cdot)$ and $\mu(\cdot)$ be the robust univariate dispersion and location statistics and denote a column $\mathbf{x}^j$ deprived of its zeros by $\tilde{\mathbf{x}}^j$, $j = 1, \ldots, p$.

2. Let $\mathbf{D} = \text{diag}(\sigma(\tilde{\mathbf{x}}^1), \ldots, \sigma(\tilde{\mathbf{x}}^p))$ and compute a normalized data matrix $\mathbf{Y}$ with columns

$$\mathbf{y}^j = \mathbf{x}^j / \sigma(\tilde{\mathbf{x}}^j), \qquad j = 1, \ldots, p. \tag{12.19}$$

3. For each pair of columns $\{\mathbf{y}^j, \mathbf{y}^k\}$ exclude the rows that contain zeros in either one of the two variables, thus creating a new pair of (equal equal length) columns $\{\tilde{\mathbf{y}}^j_{(jk)}, \tilde{\mathbf{y}}^k_{(jk)}\}$ which form a matrix $\tilde{\mathbf{Y}}_{(jk)}$. Subsequently, use Equation(12.1) to compute the robust correlation matrix $\mathbf{U}$ according to

$$U_{jj} = 1, \qquad \text{and} \qquad U_{jk} = \frac{1}{4}(\sigma(\tilde{\mathbf{y}}^j_{(jk)} + \tilde{\mathbf{y}}^k_{(jk)})^2 - \sigma(\tilde{\mathbf{y}}^j_{(jk)} - \tilde{\mathbf{y}}^k_{(jk)})^2), \qquad j \neq k. \tag{12.20}$$

4. Find the eigenvalues λ_j and eigenvectors $\mathbf{e}_j$ of $\mathbf{U}$, $j = 1, \ldots, p$, so that $\mathbf{U} \equiv \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$, where $\mathbf{E}$ consists of the eigenvectors $\mathbf{e}_j$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$.

5. Replace the zeros in $\mathbf{Y}$ by imputed values stemming from an adequate imputation method and compute the principal components $\mathbf{z}^1, \ldots, \mathbf{z}^p$ of the imputed data set $\mathbf{Y}_{(imp)}$ with rows $\mathbf{y}_{i_{(imp)}}$, $i = 1, \ldots, n$, by

$$\mathbf{z}_i = \mathbf{E}^T \mathbf{y}_{i_{(imp)}} = \mathbf{A}^{-1} \mathbf{x}_{i_{(imp)}} \tag{12.21}$$

so that

$$\mathbf{x}_{i_{(imp)}} = \mathbf{A}\mathbf{z}_{i_{(imp)}}, \tag{12.22}$$

where

$$\mathbf{A} = \mathbf{D}\mathbf{E} \tag{12.23}$$

and $\mathbf{X}_{(imp)}$ is the original data set $\mathbf{X}$ containing imputed values instead of zeros.

6. Compute the robust univariate dispersion and location statistics $\sigma(\mathbf{z}^j)$ and $\mu(\mathbf{z}^j)$ for $j = 1, \ldots, p$ and set

$$\mathbf{\Gamma} = \text{diag}(\sigma^2(\mathbf{z}^1), \ldots, \sigma^2(\mathbf{z}^p)) \qquad \text{and} \qquad \nu = (\mu(\mathbf{z}^1), \ldots, \mu(\mathbf{z}^p))^T. \tag{12.24}$$

   $\mathbf{\Gamma}$ corresponds to the diagonal covariance matrix of the approximately uncorrelated $\mathbf{z}^j$'s.

7. Transform back to the $\mathbf{X}$-coordinates with the help of Equation (12.22) or perform an iteration step.

   In the first case, compute

$$\mathbf{C}(\mathbf{X}) = \mathbf{A}\mathbf{\Gamma}\mathbf{A}^T \tag{12.25}$$

   and

$$\mathbf{t}(\mathbf{X}) = \mathbf{A}\nu. \tag{12.26}$$

   In the second case, apply steps 1-6 on the matrix $\mathbf{Z}$ from Equation (12.21) whose columns are the principal components $\mathbf{z}^1, \ldots, \mathbf{z}^p$, rendering $\mathbf{C}(\mathbf{Z})$ and $\mathbf{t}(\mathbf{Z})$. Following this, transform back to the original coordinate system by solving

$$\mathbf{C}_{(2)}(\mathbf{X}) = \mathbf{A}\mathbf{C}(\mathbf{Z})\mathbf{A}^T \tag{12.27}$$

   and

$$\mathbf{t}_{(2)}(\mathbf{X}) = \mathbf{A}\mathbf{t}(\mathbf{Z}). \tag{12.28}$$

8. Since our approach is based on omitting zeros in a pairwise manner, the re-weighting step can not be applied as in the case of the standard OGK.

   To be more precise, a weighted location estimate can still be computed after slight modifications of Equation (12.15) but the weighted scatter matrix in Equation (12.16) can not be realized in a pairwise way without possibly losing important properties such as positive definiteness or affine equivariance. Consequently, one could go forward, and apply another pairwise estimate on only those observations with weight 1. This however did not improve our results. For said reason we only employed re-weighting on the location estimate, rendering the weighted location

estimate $\mathbf{t}_w(\tilde{\mathbf{X}}) = (t_w(\tilde{\mathbf{x}}^1), \ldots, t_w(\tilde{\mathbf{x}}^p))$ which is computed by

$$t_w(\tilde{\mathbf{x}}^j) = \frac{\sum_{i=1}^{n(\tilde{\mathbf{x}}^j)} w_i \tilde{x}_{ij}}{\sum_{i=1}^{n(\tilde{\mathbf{x}}^j)} w_i}, \qquad \forall j = 1, \ldots, p, \tag{12.29}$$

where $\tilde{\mathbf{X}}$ is created by excluding all zeros in $\mathbf{X}$ column by column and $n(\tilde{\mathbf{x}}^j)$ refers to the number of entries in $\tilde{\mathbf{x}}^j$.

The weights w_i, with

$$w_i = \begin{cases} 1 & \text{if} \quad d_i \leq d_0 \\ 0 & \text{otherwise} \end{cases}$$

are found according to Equations (12.17) and (12.18), where the distances d_i, $i = 1, \ldots, n$, are computed by applying Equation (12.18) on the matrix $\mathbf{Z}_{(2)}$ which is calculated during the computation of $\mathbf{C}(\mathbf{Z})$ and $\mathbf{t}(\mathbf{Z})$ in step 7.

The final estimates of location and scatter in this algorithm are hereby $\mathbf{t}_w(\mathbf{X})$ and $\mathbf{C}_{(2)}(\mathbf{X})$. We shall denote them by $\mathbf{t}_{OGK}(\mathbf{X})$ and $\mathbf{C}_{OGK}(\mathbf{X})$.

12.1.2 The sign1 Covariance Matrix for Semi-continuous Variables

The idea of computing robust principal components after robustly sphering and normalizing the data in order to use them for the computation of the covariance matrix was originally proposed by LOCANTORE et al. (1999) and then refined and implemented as the function `sign1` in the R package **mvoutlier** by FILZMOSER and GSCHWANDTNER (2009). For more detail, see also FILZMOSER et al. (2008). The sign1 covariance matrix is obtained by carrying out the following computational steps.

1. Robustly sphere the data set $\mathbf{X} = [x_{ij}]$ by computing

$$\mathbf{x}^{j^*} = \frac{\mathbf{x}^j - \text{med}(\mathbf{x}^j)}{\text{MAD}(\mathbf{x}^j)} \quad , \qquad j = 1, \ldots, p \tag{12.30}$$

with entries

$$x^*_{ij} = \frac{x_{ij} - \text{med}(x_{1j}, \ldots, x_{nj})}{\text{MAD}(x_{1j}, \ldots, x_{nj})} \quad , \tag{12.31}$$

using the coordinate-wise sample medians and the respective median absolute deviations (MAD), which are defined as

$$\mathrm{MAD}(\mathbf{x}) = \mathrm{MAD}(x_1, \ldots, x_n) = \mathrm{med}\{|\mathbf{x} - \mathbb{1} \cdot \mathrm{med}(\mathbf{x})|\}, \tag{12.32}$$

where $\mathbb{1}$ is a vector of ones. Then bring the resulting rows of the rescaled data matrix $\mathbf{X}^*$ to norm 1 with

$$y_{ij} = \frac{x^*_{ij}}{\sqrt{\sum_{l=1}^{p} x^{*2}_{il}}}, \qquad i = 1, \ldots, n; j = 1, \ldots, p \tag{12.33}$$

resulting in the robustly sphered and normalized data matrix $\mathbf{Y}$. Variables with MAD zero should be either omitted or associated with a different measure of scale.

2. Apply singular value decomposition (SVD) on the matrix $\mathbf{Y}$, i.e.

$$\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{12.34}$$

with $\mathbf{U}$ and $\mathbf{V}$ being orthogonal $n \times r$ and $r \times p$ matrices and $\mathbf{D}$ being a diagonal $r \times r$ matrix with the singular values in the diagonal. r is the rank of $\mathbf{Y}$ and without loss of generality we assume $r = p$. The column vectors of $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors of $\mathbf{Y}$.

The diagonal values σ_j, $j = 1, \ldots, p$, of $\mathbf{D}$ with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$ are the square roots of the eigenvalues of $\mathbf{Y}^T\mathbf{Y}$ and $\mathbf{Y}\mathbf{Y}^T$. Furthermore, $\mathbf{V}$ corresponds to the matrix $\mathbf{E}$ consisting of the eigenvectors of $\mathbf{Y}^T\mathbf{Y}$, i.e.

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{V}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}^T\mathbf{D}\mathbf{V}^T \tag{12.35}$$

and hence

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T = \mathbf{V}\mathbf{D}^T\mathbf{D}\mathbf{V}^T \tag{12.36}$$

where $\mathbf{\Lambda}$ is the diagonal matrix containing the eigenvalues λ_j, $j = 1, \ldots, p$, of $\mathbf{Y}^T\mathbf{Y}$ and $\mathbf{E}$ is the matrix whose columns are the corresponding eigenvectors $\mathbf{e}^j$ of $\mathbf{Y}^T\mathbf{Y}$. For more detail see, e.g., Gonnet and Scholl (2009).

There exists an immediate connection between singular value decomposition and principal component analysis (see e.g. Wall et al., 2003). To be more specific, since $\mathbf{Y}$ is a centered and scaled matrix with p variables and n observations, $\mathbf{Y}^T\mathbf{Y}$ is proportional to the sample covariance matrix of the matrix $\mathbf{Y}$.

As a consequence of Equation (12.36), the matrix $\mathbf{Z}$ of principal components can be computed analogously to Equations (12.6) and (12.7) by (FILZMOSER et al., 2008)

$$\mathbf{Z} = \mathbf{YV}, \qquad \text{with} \qquad \mathbf{V} = \mathbf{E}. \tag{12.37}$$

3. Let $\sigma(\cdot)$ denote the MAD. We replace the eigenvalues in $\mathbf{\Lambda}$ from Equation (12.36) with robust variances $\sigma^2(\mathbf{z}^j)$, $j = 1, \ldots, p$, assembling the diagonal covariance matrix of the principal components as in Equation (12.9):

$$\mathbf{\Gamma} = \text{diag}(\sigma^2(\mathbf{z}^1), \ldots, \sigma^2(\mathbf{z}^p)). \tag{12.38}$$

Since the principal components are defined by those directions that maximize the variance along each component, we can select the $p-1$ components with the largest variances (or accordingly, if $p > n$, the $n-1$ components), thereby excluding those components which usually represent useless noise and have an obscuring effect on the underlying data structure. Furthermore we avoid singularity problems for the case $p >> n$.

With regard to the new dimension $p^* = \min\{p-1, n-1\}$, we obtain the reduced matrix $\mathbf{E}^*$ of corresponding eigenvectors and a reduced matrix $\mathbf{\Gamma}^*$ containing only the first $1, \ldots, p^*$ robust variances of the principal components.

4. To facilitate the calculation of the Mahalanobis distances later on in the process of outlier detection, we can directly compute the inverted covariance matrix $\mathbf{C}^{-1}$ by

$$\mathbf{C}^{-1} = \mathbf{E\Gamma}^{-1}\mathbf{E}^T. \tag{12.39}$$

Since $\mathbf{E}$ is an orthogonal matrix with its inverse being equal to its transpose, i.e. $\mathbf{E}^T = \mathbf{E}^{-1}$ and hence, for the covariance matrix $\mathbf{C}$ as in (12.10), it holds that

$$\mathbf{C}^{-1} = (\mathbf{E\Gamma E}^T)^{-1} = \mathbf{E}^{T^{-1}}\mathbf{\Gamma}^{-1}\mathbf{E}^{-1} = \mathbf{E\Gamma}^{-1}\mathbf{E}^T. \tag{12.40}$$

Proceeding by approximating $\mathbf{C}^{-1}$ with only p^* dimensions, we get

$$\mathbf{C}_{sign1}^{-1} = \mathbf{E}^*\mathbf{\Gamma}^{*^{-1}}\mathbf{E}^{*^T} \tag{12.41}$$

By directly computing the inverse of the approximation matrix $\mathbf{C}_{sign1}$, we also avoid the singularity problems resulting from the dimension reduction.

Modifications of the sign1 Covariance Matrix Estimator for Semi-Continuous Data

For semi-continuous data, the sign1 algorithm can be adapted as follows.

1. Consider the $n \times p$ data set $\mathbf{X} = [x_{ij}]$ which may consist of a (high) proportion of zeros and compute the robust univariate location and scale estimates, i.e. the coordinate-wise sample medians and the respective median absolute deviations (see Equation (12.32)) after excluding all zeros column by column. We shall denote these estimates by $\text{med}(\tilde{\mathbf{x}}^j)$ and $\text{MAD}(\tilde{\mathbf{x}}^j)$, where $\tilde{\mathbf{x}}^j$, $j = 1, \ldots, p$ are the columns of $\mathbf{X}$ without zeros like in Section 12.1.1.

2. Robustly sphere the data set by computing

$$\mathbf{x}^{j^*} = \frac{\mathbf{x}^j - \text{med}(\tilde{\mathbf{x}}^j)}{\text{MAD}(\tilde{\mathbf{x}}^j)} \quad , \qquad j = 1, \ldots, p \tag{12.42}$$

3. Let each pair of columns $\{\mathbf{x}^j, \mathbf{x}^k\}$, $j, k = 1, \ldots, p$, form a matrix $\mathbf{X}_{(jk)}$. Subsequently, create a reduced matrix $\tilde{\mathbf{X}}_{(jk)}$ with columns $\{\tilde{\mathbf{x}}^j_{(jk)}, \tilde{\mathbf{x}}^k_{(jk)}\}$ for every matrix $\mathbf{X}_{(jk)}$ by excluding all rows containing zeros in either one of the two variables of $\mathbf{X}_{(jk)}$. In this context, the same rows are excluded from every sphered matrix $\mathbf{X}^*_{(jk)}$, resulting in the reduced robustly sphered matrices $\tilde{\mathbf{X}}^*_{(jk)}$ with entries $\tilde{x}^*_{ij_{(jk)}}$, $i = 1, \ldots, n(\tilde{\mathbf{x}}^j_{(jk)})$, where $n(\tilde{\mathbf{x}}^j_{(jk)})$ refers to the number of entries of column $\tilde{\mathbf{x}}^j_{(jk)}$.

4. Bring the rows of the reduced rescaled data matrices $\tilde{\mathbf{X}}^*_{(jk)}$, $j, k = 1, \ldots, p$, to norm 1 with

$$y_{ij_{(jk)}} = \frac{\tilde{x}^*_{ij_{(jk)}}}{\sqrt{\sum_{l=1}^{p} \tilde{x}^{*2}_{il_{(jk)}}}}, \qquad i = 1, \ldots, n; j = 1, \ldots, p \tag{12.43}$$

   resulting in the robustly sphered and normalized reduced data matrix $\mathbf{Y}_{(jk)}$. If a row sum in the denominator of Equation (12.43) should be zero, hence leading to no result, the respective cells in $\mathbf{Y}_{(jk)}$ are filled with zeros since this problem can only occur if all the values in the corresponding row of $\tilde{\mathbf{X}}^*_{(jk)}$ are equal to zero.

5. Apply singular value decomposition on the matrix $\mathbf{Y}_{(jk)}$, i.e.

$$\mathbf{Y}_{(jk)} = \mathbf{U}_{(jk)} \mathbf{D}_{(jk)} \mathbf{V}^T_{(jk)}, \tag{12.44}$$

where $\mathbf{U}_{(jk)}$ and $\mathbf{V}_{(jk)}$ contain the left and right singular vectors of $\mathbf{Y}_{(jk)}$ and the diagonal matrix $\mathbf{D}_{(jk)}$ its singular values. Moreover, $\mathbf{V}_{(jk)}$ also satisfies

$$\mathbf{Y}_{(jk)}^T\mathbf{Y}_{(jk)} = \mathbf{V}_{(jk)}\mathbf{D}_{(jk)}^T\mathbf{D}_{(jk)}\mathbf{V}_{(jk)}^T = \mathbf{E}_{(jk)}\mathbf{\Lambda}_{(jk)}\mathbf{E}_{(jk)}^T, \tag{12.45}$$

where $\mathbf{E}_{(jk)}$ and $\mathbf{\Lambda}_{(jk)}$ contain the eigenvectors and eigenvalues of $\mathbf{Y}_{(jk)}^T\mathbf{Y}_{(jk)}$ respectively.

Furthermore, $\mathbf{Y}_{(jk)}^T\mathbf{Y}_{(jk)}$ is proportional to the sample covariance matrix of $\mathbf{Y}_{(jk)}$ since $\mathbf{Y}_{(jk)}$ is a centered and scaled matrix with two variables and $n(\tilde{\mathbf{x}}_{(jk)}^j)$ observations.

It follows, that the matrix $\mathbf{Z}_{(jk)}$ of principal components can be computed by

$$\mathbf{Z}_{(jk)} = \mathbf{Y}_{(jk)}\mathbf{V}_{(jk)}, \qquad \text{with} \qquad \mathbf{V}_{(jk)} = \mathbf{E}_{(jk)}. \tag{12.46}$$

6. Let $\sigma(\cdot)$ denote the MAD and replace the eigenvalues in $\mathbf{\Lambda}_{(jk)}$ with the robust variances $\sigma^2(\mathbf{z}^j), \sigma^2(\mathbf{z}^k)$ by substituting $\mathbf{\Lambda}_{(jk)}$ in Equation (12.45) with the diagonal matrix $\mathbf{\Gamma}_{(jk)} = \text{diag}(\sigma^2(\mathbf{z}^j), \sigma^2(\mathbf{z}^k))$. In other words, we perform the inverse transformation

$$\mathbf{C}_{(jk)} = \mathbf{E}_{(jk)}\mathbf{\Gamma}_{(jk)}\mathbf{E}_{(jk)}^T. \tag{12.47}$$

7. Compute the preliminary covariance matrix $\mathbf{C}_{(1)} = [C_{(1)_{jk}}]$, $j, k = 1, \ldots, p$, by using the results of the previous steps according to

$$C_{(1)_{jj}} = \text{MAD}(\tilde{\mathbf{x}}^j), \qquad \text{and} \qquad C_{(1)_{jk}} = C_{(1)_{kj}} = C_{(jk)_{12}}, \qquad j \neq k \tag{12.48}$$

with the $\text{MAD}(\tilde{\mathbf{x}}^j)$'s from steps 1-2 and with the corresponding 2×2 matrices $\mathbf{C}_{(jk)} = [C_{(jk)_{ab}}]$, $a, b = 1, 2$, where of course $C_{(jk)_{12}} = C_{(jk)_{21}}$.

8. Finally, it is necessary to ensure the positive definiteness of $\mathbf{C}_{(1)}$, which is not guaranteed under given circumstances. In this context, find the eigenvalues and eigenvectors of $\mathbf{C}_{(1)}$ so that

$$\mathbf{C}_{(1)} = \mathbf{E}_{(1)}\mathbf{\Lambda}_{(1)}\mathbf{E}_{(1)}^T, \tag{12.49}$$

where $\mathbf{E}_{(1)}$ consists of the eigenvectors and $\mathbf{\Lambda}_{(1)} = \text{diag}(\lambda_{(1)_1}, \ldots, \lambda_{(1)_p})$ of the eigenvalues of $\mathbf{C}_{(1)}$. Subsequently, keep only the positive eigenvalues $\lambda_{(1)_j} \geq 0$, $j = 1, \ldots, p$ and the pertinent eigenvectors. Store them in the respective matrices

$\mathbf{\Lambda}^*_{(1)}$ and $\mathbf{E}^*_{(1)}$ and compute the inverted covariance matrix

$$\mathbf{C}^{-1}_{sign1} := \mathbf{C}^{-1}_{(2)} = \mathbf{E}^*_{(1)}\mathbf{\Lambda}^*_{(1)}\mathbf{E}^{*^T}_{(1)} \tag{12.50}$$

which is equal to $\mathbf{C}^{-1}_{(1)}$ if $\dim(\mathbf{\Lambda}^*_{(1)}) = \dim(\mathbf{\Lambda}_{(1)})$ and which approximates the inverse of $\mathbf{C}_{(1)}$ if $\dim(\mathbf{\Lambda}^*_{(1)}) < \dim(\mathbf{\Lambda}_{(1)})$. Direct computation of the inverse prevents singularity problems resulting from a possible dimension reduction.

12.1.3 Quadrant Correlation and Covariance for Semi-continuous Variables

First of all, we introduce quadrant correlation methods for non-semi-continuous data.

For a bivariate sample $\{(x_i, y_i)\}$, $i = 1, \ldots, n$, denoted by the vectors $\mathbf{x}$ and $\mathbf{y}$, the quadrant correlation is computed as (see Shevlyakov, 1997; Mosteller, 1946)

$$r_Q = \frac{1}{n}\sum_{i=1}^{n} \text{sign}\left\{(x_i - \text{med}(\mathbf{x}))(y_i - \text{med}(\mathbf{y}))\right\} \tag{12.51}$$

In other words, the two coordinate-wise sample medians med($\mathbf{x}$) and med($\mathbf{y}$) divide the plane into 4 quadrants so that r_Q is based on the number of observations in the first or third quadrant, minus the number of observations in the second or fourth quadrant.

According to Maronna et al. (2006), the estimator is not consistent under a given model but can be corrected by (see Blomqvist, 1950)

$$\hat{r}_Q = \sin\left(\frac{r_Q \pi}{2}\right) \tag{12.52}$$

to ensure consistency at the normal model. The robust quadrant covariance is then computed as usual by

$$cov_Q(\mathbf{x}, \mathbf{y}) = \sigma(\mathbf{x})\sigma(\mathbf{y})\hat{r}_Q \quad , \tag{12.53}$$

where the univariate scale estimate σ corresponds to the MAD (see Equation (12.32)). Based on the definitions (12.51), (12.52) and (12.53) for two variables, the quadrant covariance matrix $\mathbf{C}_Q = [C_{Q\,jk}]$ for the multivariate case is then computed by applying $cov_Q(\cdot, \cdot)$ to the columns $\mathbf{x}^j$, $j = 1, \ldots, p$, of $\mathbf{X}$, i.e.

$$C_{Q\,jj} = \sigma(\mathbf{x}^j), \qquad \text{and} \qquad C_{Q\,jk} = cov_Q(\mathbf{x}^j, \mathbf{x}^k), \qquad j \neq k. \tag{12.54}$$

Finally, to assure positive semidefiniteness, an eigenvalue decomposition is performed, subsequently keeping only the positive eigenvalues before transforming back to the covariance matrix. Since the eigenvector matrix is orthogonal, one can directly transform back to the inverse of the covariance matrix, see Equation (12.40), which saves us from the matrix inversion later on in the process of outlier detection, i.e. for computing the Mahalanobis distances.

Modification of Quadrant Correlations for Semi-Continuous Data

The quadrant correlation method can be adapted to semi-continuous data as follows.

1. Consider the $n \times p$ data set $\mathbf{X} = [x_{ij}]$ where each column $\mathbf{x}^j$, $j = 1, \ldots, p$, may contain a certain amount of zeros. As in Sections 12.1.1 and 12.1.2, the robust univariate location and scale estimates are denoted by $\text{med}(\tilde{\mathbf{x}}^j)$ and $\text{MAD}(\tilde{\mathbf{x}}^j)$, $j = 1, \ldots, p$, where the $\tilde{\mathbf{x}}^j$'s are the columns of $\mathbf{X}$ without zeros. In this context, we also use the notation from the previous sections for the reduced two-dimensional matrices deprived of all zeros in a pairwise manner. To be more specific, the matrices denoted by $\tilde{\mathbf{X}}_{(jk)}$ with columns $\{\tilde{\mathbf{x}}^j_{(jk)}, \tilde{\mathbf{x}}^k_{(jk)}\}$ of length $n(\tilde{\mathbf{x}}^j_{(jk)})$ are each created by removing the rows with zeros in the corresponding matrices $\mathbf{X}_{(jk)}$ which are composed of the pair of columns $\{\mathbf{x}^j, \mathbf{x}^k\}$, $j, k = 1, \ldots, p$.

   In the following, find the respective $2 \times n(\tilde{\mathbf{x}}^j_{(jk)})$ matrices $\tilde{\mathbf{X}}_{(jk)}$ for every pair $\{\mathbf{x}^j, \mathbf{x}^k\}$ and execute steps 2-4 for each of these matrices. In step 5, combine the bivariate results in a multivariate scatter estimate which is ensured to be positive definite in step 6.

2. Calculate the quadrant correlation according to Equation (12.51) by

$$r_{Q_{(jk)}} = \frac{1}{n(\tilde{\mathbf{x}}^j_{(jk)})} \sum_{i=1}^{n(\tilde{\mathbf{x}}^j_{(jk)})} \text{sign}\left\{\left(\tilde{x}_{ij_{(jk)}} - \text{med}(\tilde{\mathbf{x}}^j_{(jk)})\right)\left(\tilde{x}_{ik_{(jk)}} - \text{med}(\tilde{\mathbf{x}}^k_{(jk)})\right)\right\}. \tag{12.55}$$

3. Apply Equation (12.52) to the estimator $r_{Q_{(jk)}}$ to ensure consistency at the normal model, i.e.

$$\hat{r}_{Q_{(jk)}} = \sin\left(\frac{r_{Q_{(jk)}}\pi}{2}\right) \tag{12.56}$$

4. Compute the robust quadrant covariance according to Equation (12.53) by

$$cov_{(jk)}(\tilde{\mathbf{x}}^j_{(jk)}, \tilde{\mathbf{x}}^k_{(jk)}) = \hat{r}_{Q_{(jk)}} \sqrt{\text{MAD}(\tilde{\mathbf{x}}^j)\text{MAD}(\tilde{\mathbf{x}}^k)}, \tag{12.57}$$

   with the univariate scale estimates $\text{MAD}(\tilde{\mathbf{x}}^j)$, $j = 1, \ldots, p$, as mentioned in step 1.

5. Assemble a preliminary multivariate covariance matrix $\mathbf{C}_{(1)} = [C_{(1)_{jk}}]$, $j, k = 1, \ldots, p$, by using the results from the previous steps according to

$$C_{(1)_{jj}} = \text{MAD}(\tilde{\mathbf{x}}^j), \qquad \text{and} \qquad C_{(1)_{jk}} = C_{(1)_{kj}} = cov_{(jk)}(\tilde{\mathbf{x}}^j_{(jk)}, \tilde{\mathbf{x}}^k_{(jk)}), \qquad j \neq k. \tag{12.58}$$

6. Use step 8 from Section 12.1.2 to ensure the positive definiteness of $\mathbf{C}_{(1)}$ and denote the resulting inverted covariance matrix estimate $\mathbf{C}^{-1}_{(2)}$ by $\mathbf{C}^{-1}_{Q}$.

12.1.4 Outlier Detection for Semi-continuous Variables

The following distance-based method for detecting outliers is obviously applicable on any result from the previous sections. Hence, the location and covariance estimates in this current section will generally be denoted by $\mathbf{t} = \mathbf{t}(\mathbf{X})$ and $\mathbf{C} = \mathbf{C}(\mathbf{X})$.

Start by preparing an imputed data set $\mathbf{X}_{(imp)}$ with rows $\mathbf{x}_{i_{(imp)}}$, $i = 1, \ldots, n$, for computational purposes. Hereby, an adequate imputation method is used to replace the zeros in $\mathbf{X}$ by imputed values. Subsequently, compute the robust distances for the observations $\mathbf{x}_{i_{(imp)}}$ of the imputed data set according to

$$d_i = d(\mathbf{x}_{i_{(imp)}}) = \sqrt{(\mathbf{x}_{i_{(imp)}} - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x}_{i_{(imp)}} - \mathbf{t})} \quad . \tag{12.59}$$

Note that the sign1 and quadrant covariance algorithms in Sections 12.1.2 and 12.1.3 already deliver an inverted covariance estimate $\mathbf{C}^{-1}$.

Now transform the d_i's, $i = 1, \ldots, n$, by

$$d^*_i = d_i \cdot \frac{\sqrt{\chi^2_p(0.5)}}{\text{med}(d_1, \ldots, d_n)} \quad , \tag{12.60}$$

which should make the distribution of the robust distances better resemble the χ^2_p distribution and eventually use the cutoff-value $d_0 = \sqrt{\chi^2_p(\beta)}$ to flag all observations as outliers whose robust distances d^*_i, $i = 1, \ldots, n$, satisfy

$$d^*_i \geq d_0. \tag{12.61}$$

12.2 Simulations and Results

Simulations under different scenarios were run to compare the methods discussed in the previous sections. We compare the modified pairwise methods from Section 12.1 with each another and with the methods from the previous sections that were applied to data, which has been imputed beforehand (except of the BACON-EEM).

12.2.1 Data Generation

To evaluate and compare the performance of the different methods, contaminated data consisting of semi-continuous variables have to be generated synthetically. For this purpose, data sets of dimension $n \times p$ with $p << n$ were created, where the data follow a contaminated multivariate normal model with "mean-shift outliers" (see e.g. TODOROV et al., 2011), i.e. $(1-\varepsilon)n$ regular observations in each data set $\mathbf{X} = [x_{ij}]$, $i = 1, \ldots, n$, $j = 1, \ldots, p$ are generated from a multivariate normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ while the remaining εn outlying observations are sampled from $\mathcal{N}_p(\boldsymbol{\mu}_{(out)}, \boldsymbol{\Sigma}_{(out)})$. We used the following values for the parameters:

1. $\boldsymbol{\mu} = (10, \ldots, 10)^T$
2. $\boldsymbol{\Sigma} = [\Sigma_{jk}]$ with $\Sigma_{jj} = 1$ and $\Sigma_{jk} = 0.5$ for $j \neq k$
3. For $\boldsymbol{\mu}_{(out)}$, we considered the following cases, where $p_{(out)}$ denotes the number of values being different from the entry in $\boldsymbol{\mu}$, i.e. the number of variables that should contain (univariate) outliers:
   (a) $p_{(out)} = 2$:
      i. *Moderate Outliers:* Replace the first 2 entries of $\boldsymbol{\mu} = (10, \ldots, 10)^T$ by $(5, 15)^T$ to get $\boldsymbol{\mu}_{(out)}$. For $p = 5$, for example, this would render the mean vector $\boldsymbol{\mu}_{(out)} = (5, 15, 10, 10, 10)^T$.
      ii. *Extreme Outliers:* Replace the first 2 entries of $\boldsymbol{\mu} = (10, \ldots, 10)^T$ by $(-30, 50)^T$ to get $\boldsymbol{\mu}_{(out)}$.
   (b) $p_{(out)} > 2$: The distance between the (centers of the) two point clouds of observations with and without outliers should be the same for any number $p_{(out)}$ of variables with outliers. This is accomplished by computing a constant

$$c = \sqrt{\frac{\sum_{j=1}^{p(out)} (\mu_j - \mu_{(out)_j})^2}{p_{(out)}}} \tag{12.62}$$

   with the Euclidean distance between $\boldsymbol{\mu}_{(out)}$ from alternative (a), where only the first two columns of the generated data set contain outliers, and $\boldsymbol{\mu} = (10, \ldots, 10)^T$. The constant is then used to replace the first $p_{(out)}$ entries of $\boldsymbol{\mu}$ by $10 - c$ and $10 + c$ in equal parts to form the new $\boldsymbol{\mu}_{(out)}$. If $p_{(out)}$ is not an even number, $\frac{p_{(out)}}{2} + 1$ entries are replaced by $10 + c$.

4. $\mathbf{\Sigma}_{(out)} = b\mathbf{\Sigma}$, with $b = 1.1$.

To make the data set semi-continuous, zeros were inserted according to the following approach:

1. Let n_0 be the total number of rows containing zeros so that n_0/n is the proportion of rows with zeros. Then denote number and proportion of zeros in such rows by p_0 and p_0/p.

2. The observations without zeros will occupy the first $i = 1, \ldots, (n - n_0)$ rows of $\mathbf{X}$ and the observations with zeros the last $i = (n - n_0 + 1), \ldots, n$ rows.

3. Now let $\mathbf{X}_{(reg)}$ and $\mathbf{X}_{(out)}$ denote the regular and outlying parts of the continuous data set $\mathbf{X} = (\mathbf{X}_{(reg)}, \mathbf{X}_{(out)})$. The position of the observations belonging to $\mathbf{X}_{(out)}$ depends on the proportion n_0/n of rows with zeros:

   (a) If $\varepsilon \leq n_0/n$, half of the εn outlying observations will occupy the last rows of the observations without zeros while the remaining $\frac{\varepsilon n}{2}$ outlying observations will occupy the first rows of the observations with zeros.

   (b) If $\varepsilon > n_0/n$, $\frac{n_0}{2}$ outlying observations will occupy the first half of the rows with zeros while the remaining $\varepsilon n - \frac{n_0}{2}$ outlying observations occupy the last rows of the observations without zeros.

   For illustrative purposes, Figure 12.1 shows the principal structure of such a data set. $\mathbf{X}_0$ and $\mathbf{X}_{\neq 0}$ shall denote the parts with and without zeros of a data set $\mathbf{X} = (\mathbf{X}_{\neq 0}, \mathbf{X}_0)$ respectively.

4. After these structural arrangements, p_0 zeros are included randomly in every row of $\mathbf{X}_0$. However, if $p_0 \geq p_{(out)}$, there is a chance that all outlying entries in an observation belonging to $\mathbf{X}_0$ and $\mathbf{X}_{(out)}$ are replaced by zeros. This is avoided by creating a slightly restricted sample pool of possible indices $\{1, \ldots, p\}$ for the zeros in each row. To be more specific, $p_{(out)} - 1$ elements are randomly selected out of the index set $\{1, \ldots, p_{(out)}\} \subset \{1, \ldots, p\}$ and then combined with the elements of $\{(p_{(out)} + 1), \ldots, p\} = \{1, \ldots, p\} \setminus \{1, \ldots, p_{(out)}\}$. The indices of the p_0 zeros can then be randomly selected out of this new index set.

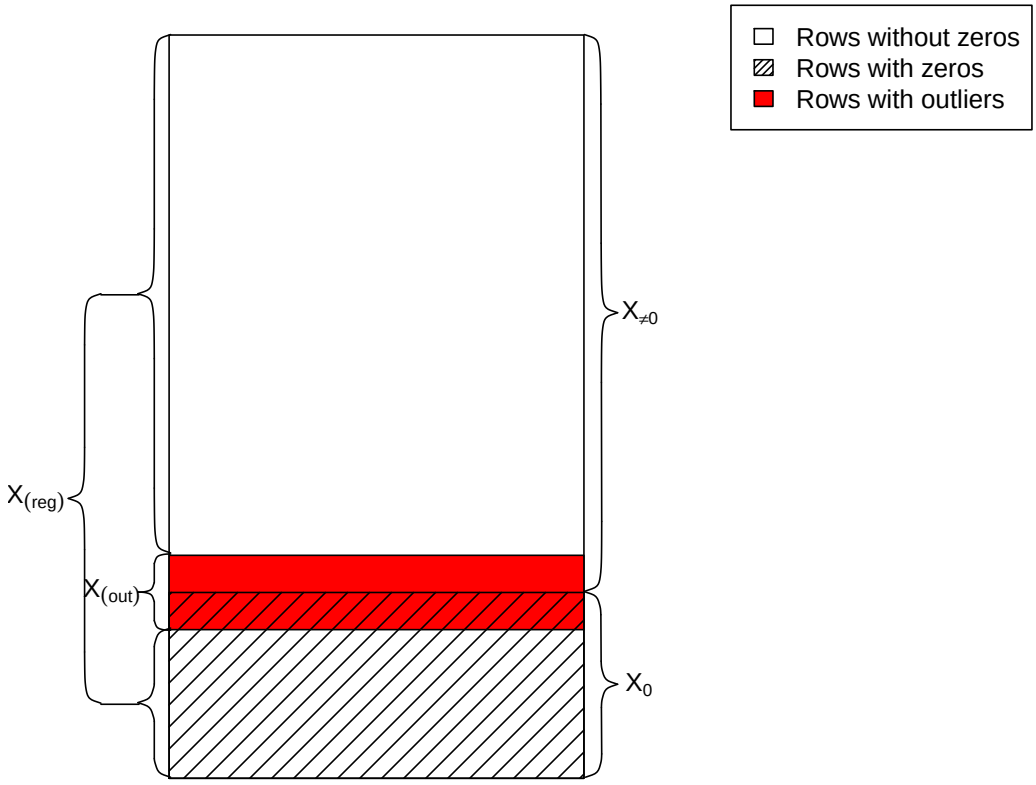


Figure 12.1: Structure of a data set with $\varepsilon = 0.1$ and $n_0/n = 0.25$

12.2.2 Simulation

The performance of the methods is evaluated on grounds of the following criteria.

1. *Mean proportion of false negatives:* the average percentage of false negatives (FN), i.e. outliers that were not identified

2. *Mean proportion of false positives:* the average percentage of false positives (FP), i.e. non-outliers that were classified as outliers

Whether an observation is considered an outlier depends strongly on the choice of the threshold d_0 and hence, in the case of semi-continuous data, on the imputation method used for the computation of the Mahalanobis distances in Equation (12.59). This is illustrated by way of example.

Consider the semi-continuous data set $\mathbf{X}$ with $\mathbf{X} = (\mathbf{X}_{\neq 0}, \mathbf{X}_0)$ and $\mathbf{X} = (\mathbf{X}_{(reg)}, \mathbf{X}_{(out)})$ as defined in the previous section. The parameters are set to $p = 5$, $n = 1000$, $p_0 = 3$, $n_0/n = 0.4$, $p_{(out)} = 2$ moderate, $\varepsilon = 0.1$. Applying the OGK algorithm on this data set

with $\beta = 0.975$ (see Section 12.1.4) and with varying types of imputation, renders the results depicted in Figure 12.2.

Figures 12.2(a)-12.2(c) show the Mahalanobis distances of observations with/without zeros and observations with/without outliers as well as the cutoff-value d_0. The scale of the y-axis is inverted for easier comparison with Figure 12.1 depicting the structure of such a data set. It can be seen, that the Mahalanobis distances belonging to $\mathbf{X}_0$, i.e. index $601 - 1000$, exhibit a different structure than those belonging to $\mathbf{X}_{\neq 0}$, i.e. index $1 - 600$, which can make the choice of a "good" threshold rather difficult. False positives can be identified as those observations above the threshold belonging to $\mathbf{X}_{(reg)}$, i.e. index $1 - 550$ and $651 - 1000$, and false negatives correspondingly as those observations below the threshold belonging to $\mathbf{X}_{(out)}$, i.e. index $551 - 650$.

In the following, simulations were performed for semi-continuous data sets of dimension 1000×5, 1000×10 and 100×5 as defined in Section 12.2.1, for which $p_{(out)} = 2$ when $p = 5$ and $p_{(out)} = 4$ when $p = 10$. For comparative purposes, a fixed value $\beta = 0.975$ was used for all simulated scenarios and the proportion of outliers was set to $\varepsilon = 0.1$.

Simulations were also carried out for $\varepsilon = 0.01$ since small amounts of outliers are often of interest, but only marginal differences to the findings in Sections 12.2.2 and 12.2.2 could be observed; see also Meraner (2010) for more details.

100 simulations were made for all scenarios where $n = 1000$ and 500 simulations were run when $n = 100$.

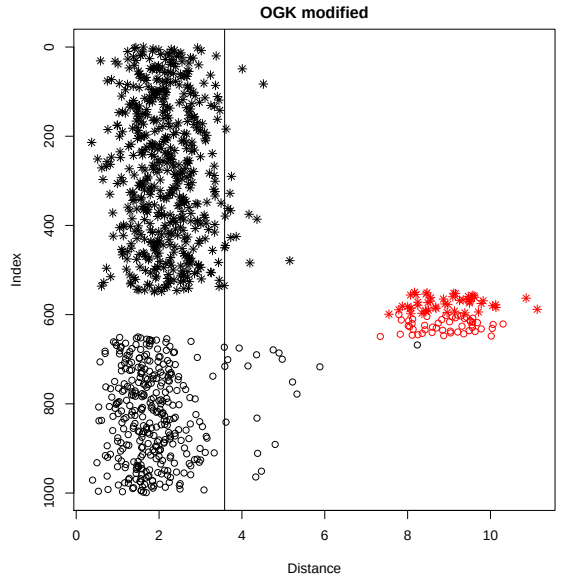
The resulting plots show the proportion n_0/n of rows with zeros on the x-axis and the average proportion of false negatives or false positives on the y-axis. There are two sets of plots for every parameter setting. One set compares modified and original pairwise methods for different types of imputation while the other set compares modified pairwise and multivariate methods, where the modified pairwise methods as well as the MCD algorithm use knn-imputation for outlier detection which seemed best when looking at the first sets of plots. The second set of plots also shows the proportion of false negatives and false positives corresponding to the parts $\mathbf{X}_0$ and $\mathbf{X}_{\neq 0}$ with and without zeros. To be more specific, for every proportion of rows with zeros n_0/n, the mean proportion of false negatives/false positives of the whole data set $\mathbf{X}$ is equal to the corresponding mean proportion of false negatives/positives of $\mathbf{X}_0$ plus the mean proportion of false negatives/positives of $\mathbf{X}_{\neq 0}$.

Note that the scale of the plots is different for every scenario and that the scale of the plots depicting false negatives within each scenario is different from the scale used to depict false positives.

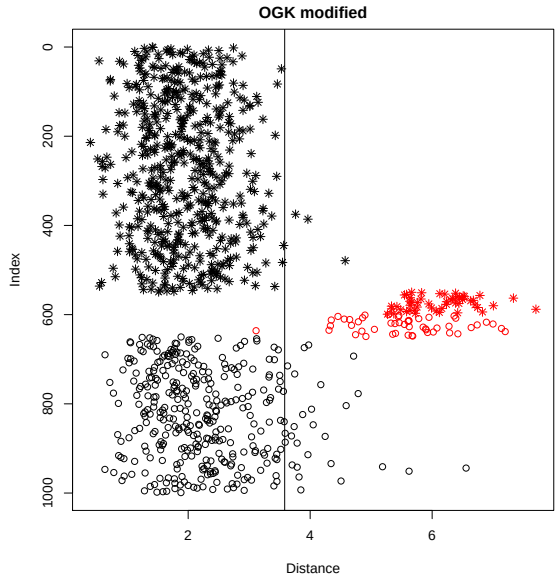
Moderate Outliers

In this section, we discuss the simulation results for $p = 5$ and $p = 10$ dimensional data and a scenario with moderate outliers. In Figure 12.3, we depict the general structure of such data sets without zeros.

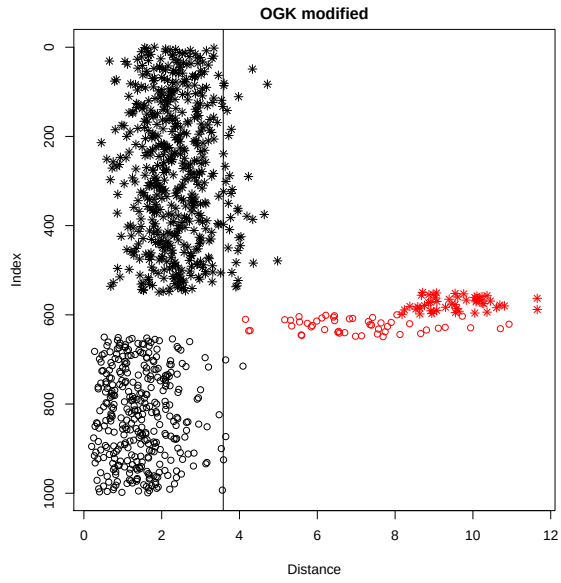
The simulation results are presented visually (Figures 12.4 – 12.11). It can be seen that the modified pairwise methods do – in general – not yield better results than the original

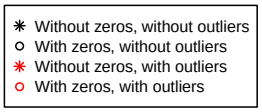


(a) knn



(b) norm





(c) IRMI

Figure 12.2: OGK for different types of imputation with $p = 5$, $n = 1000$, $p_{(out)} = 2$ moderate, $p_0 = 3$, $n_0/n = 0.4$ and $\varepsilon = 0.1$

pairwise methods do. It is also remarkable that BACON-EEM gives usually the best results. Furthermore, both modified and original methods strongly depend on the type of imputation used. The results indicate that the knn imputation performs best.

The results for $p = 5$, $n = 1000$ and $p_0 = 1$, show hardly any difference in the performance between the modified pairwise and the standard pariwise methods (Figure 12.4). It is also evident that the knn imputation algorithm is to be preferred (Figure 12.4). The BACON-EEM algorithm performs best up until a proportion of approximately $n_0/n = 0.65$ rows with zeros. For $n_0/n > 0.65$, the MCD algorithm for example is slightly better (Figure 12.5). The results remain essentially the same for the setup $p = 5$ and $p = 10$, $n = 1000$

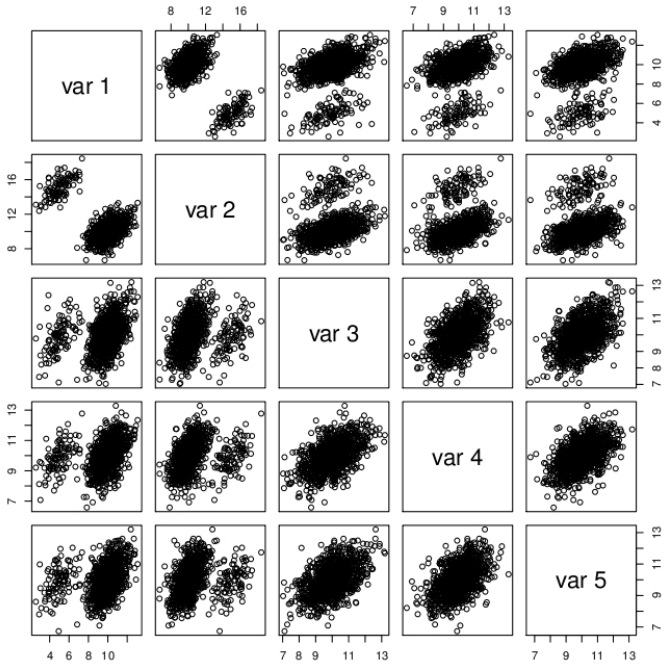


Figure 12.3: Five variables without zeros and moderate outliers with $\varepsilon = 0.1$

and $p_0 = 3$ and $p_0 = 6$ (Figures 12.6, 12.7, and 12.8). Again, BACON-EEM tends to be better than the other methods (Figure 12.9). However, for $n_0/n > 0.65$, the MCD and OGK algorithms are slightly better (Figure 12.11).

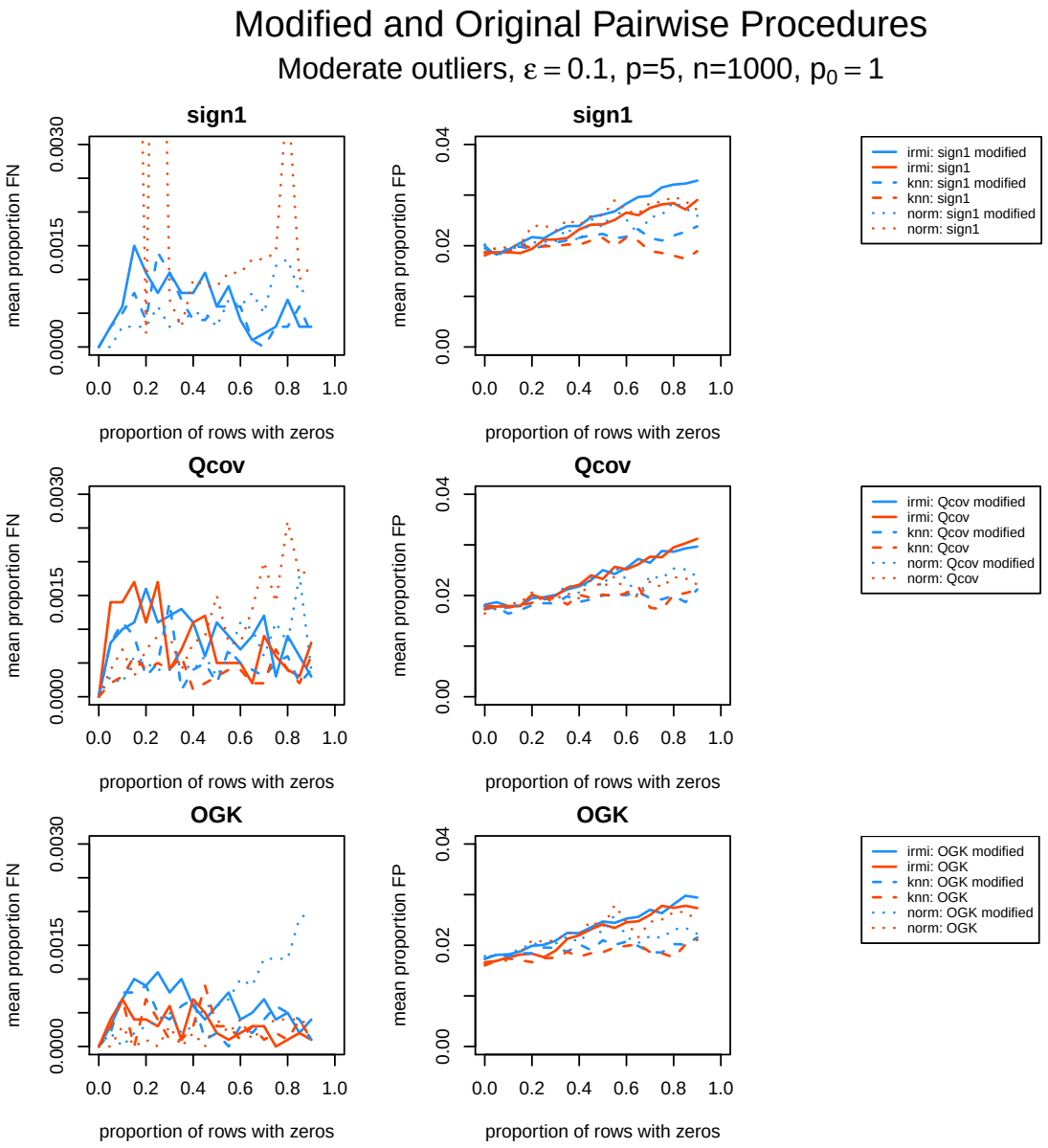


Figure 12.4: Average proportion of false negatives (left) and false positives (right) for a fixed proportion of outliers and varying percentage of rows with zeros. Comparison of pairwise outlier detection procedures with different types of imputation.

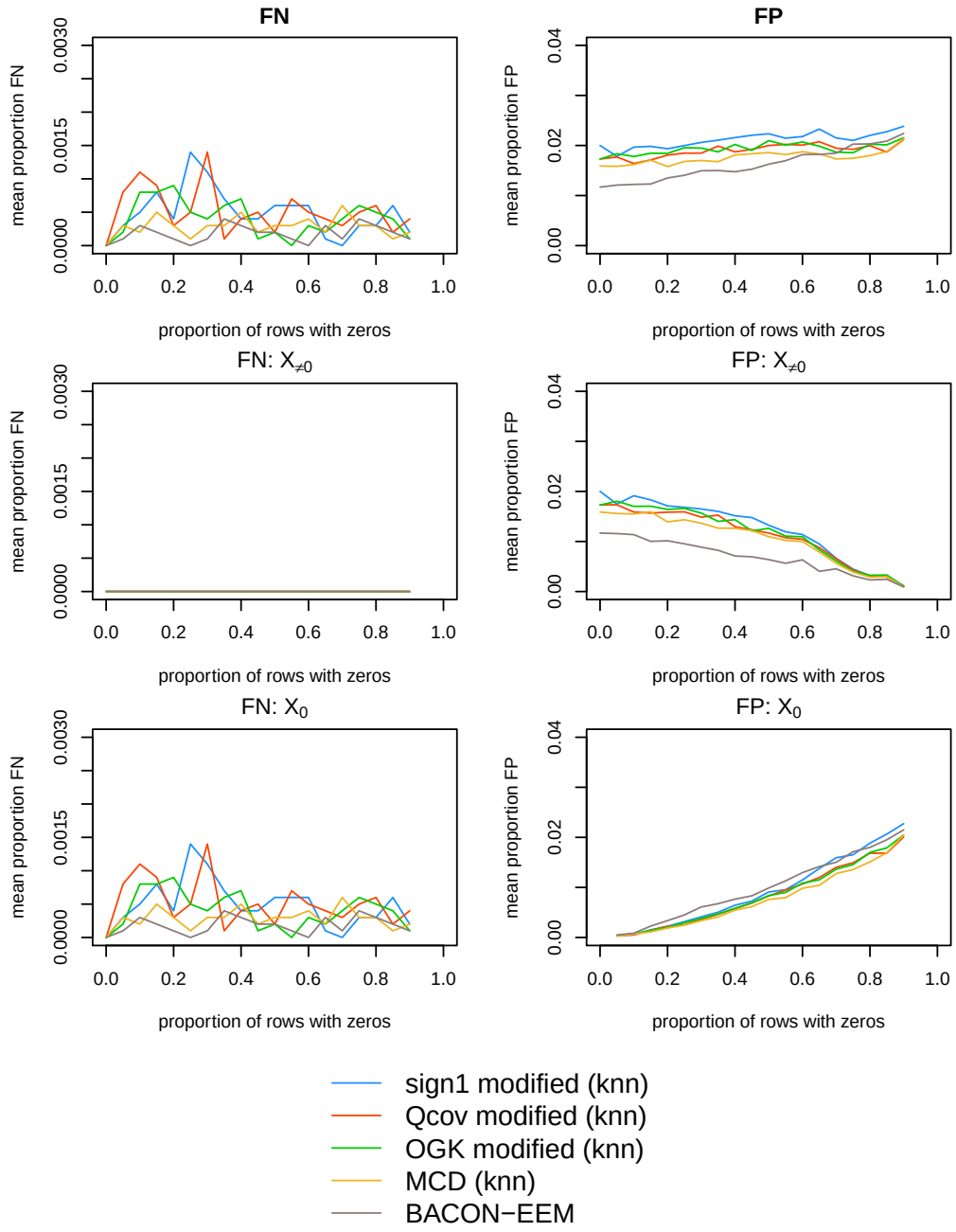


Figure 12.5: Average proportion of false negatives (left) and false positives (right) for a fixed proportion of outliers and varying percentage of rows with zeros (moderate outliers, $\varepsilon = 0.1$, $p = 5$, $n = 1000$, $p_0 = 1$). Comparison of $\mathbf{X}$, $\mathbf{X}_{\neq 0}$ and $\mathbf{X}_0$. *Legend:* [blue] sign1 modified (knn); [red] Qcov modified (knn); [green] OGK modified (knn); [yellow] MCD (knn); [grey] BACON-EEM.

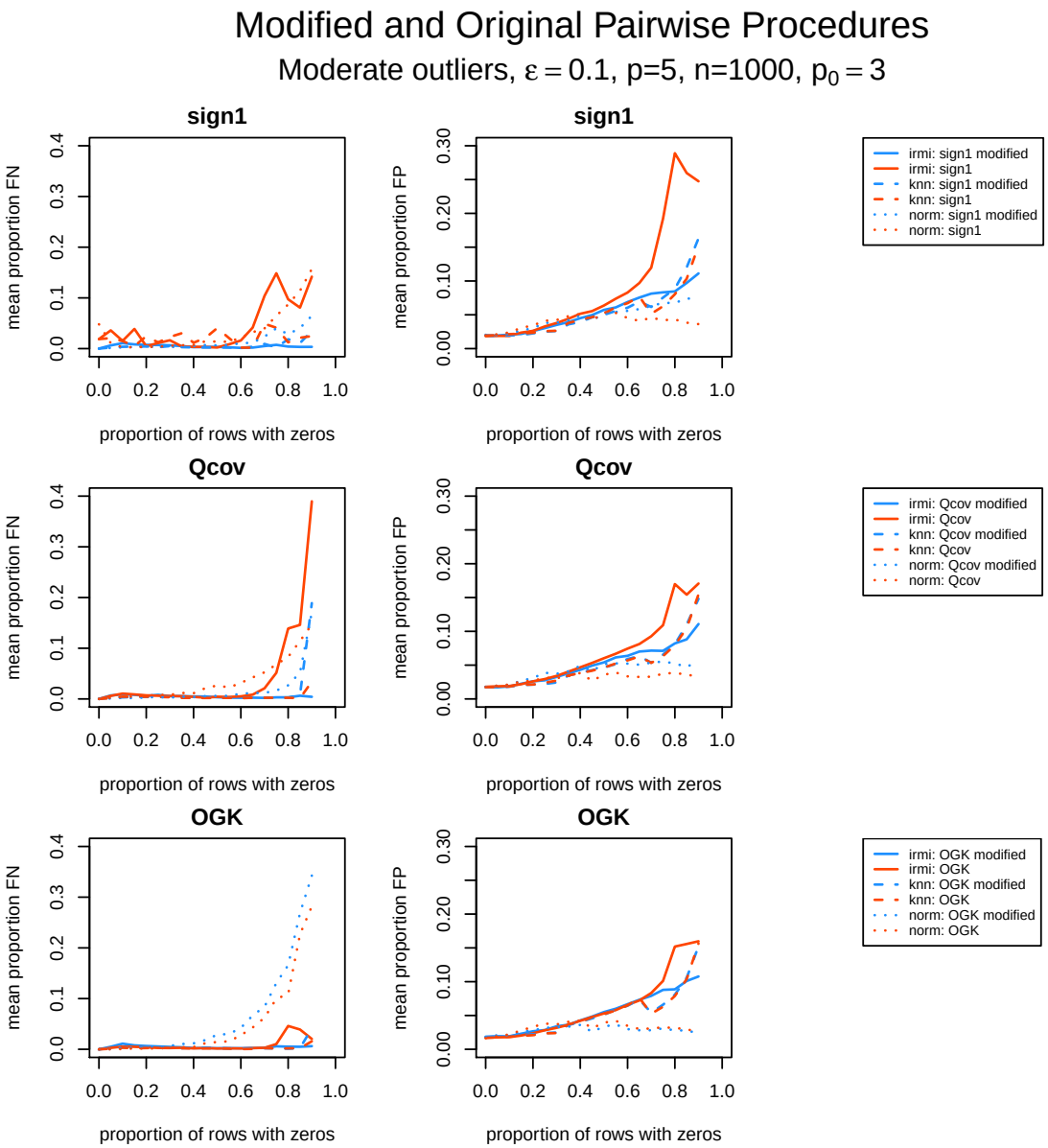


Figure 12.6: Average proportion of false negatives (left) and false positives (right) for a fixed proportion of outliers and varying percentage of rows with zeros. Comparison of pairwise outlier detection procedures with different types of imputation.

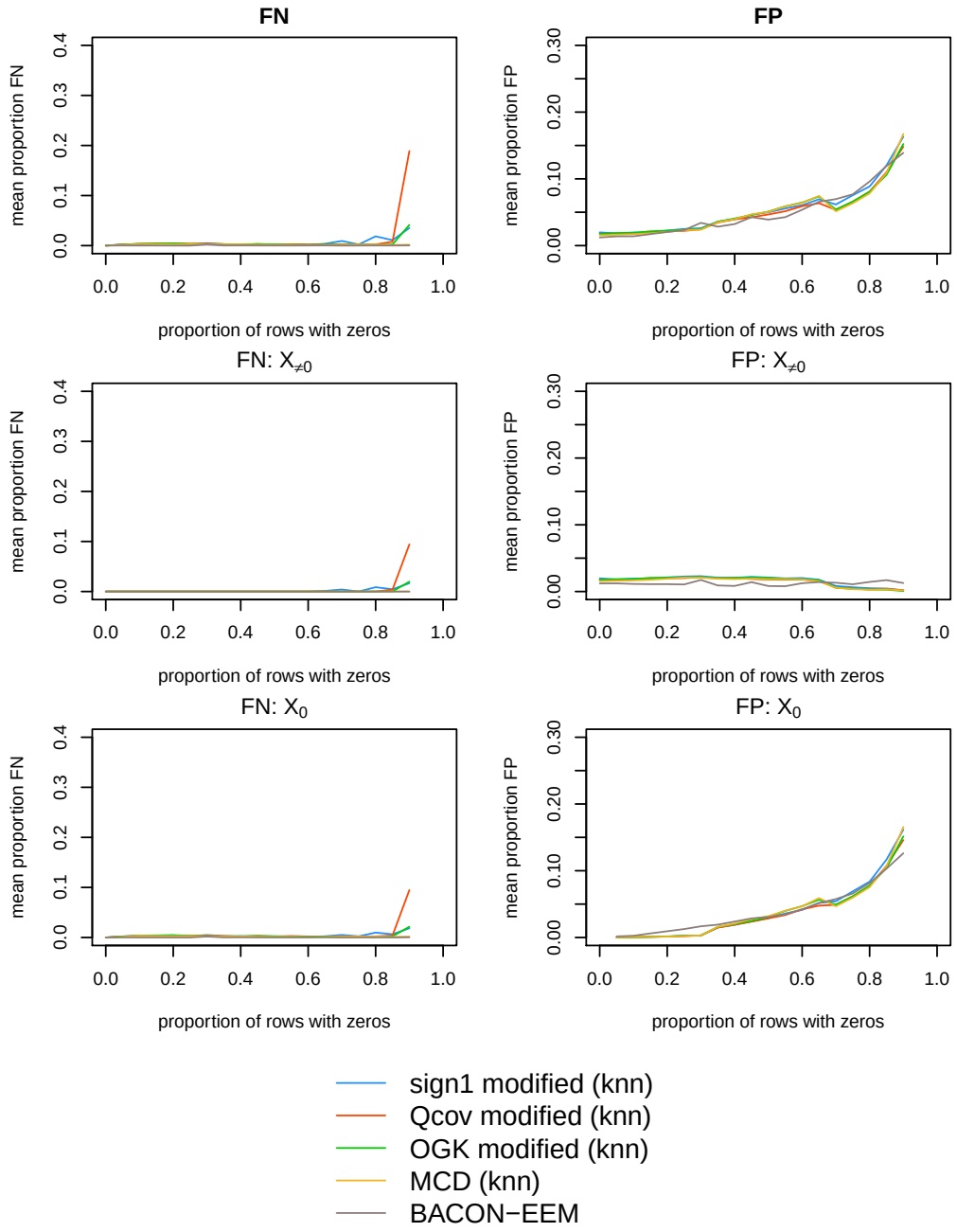


Figure 12.7: Average proportion of false negatives (left) and false positives (right) for a fixed proportion of outliers and varying percentage of rows with zeros (moderate outliers, $\varepsilon = 0.1$, $p = 5$, $n = 1000$, $p_0 = 3$). Comparison of $\mathbf{X}$, $\mathbf{X}_{\neq 0}$ and $\mathbf{X}_0$. *Legend:* [blue] sign1 modified (knn); [red] Qcov modified (knn); [green] OGK modified (knn); [yellow] MCD (knn); [grey] BACON-EEM.

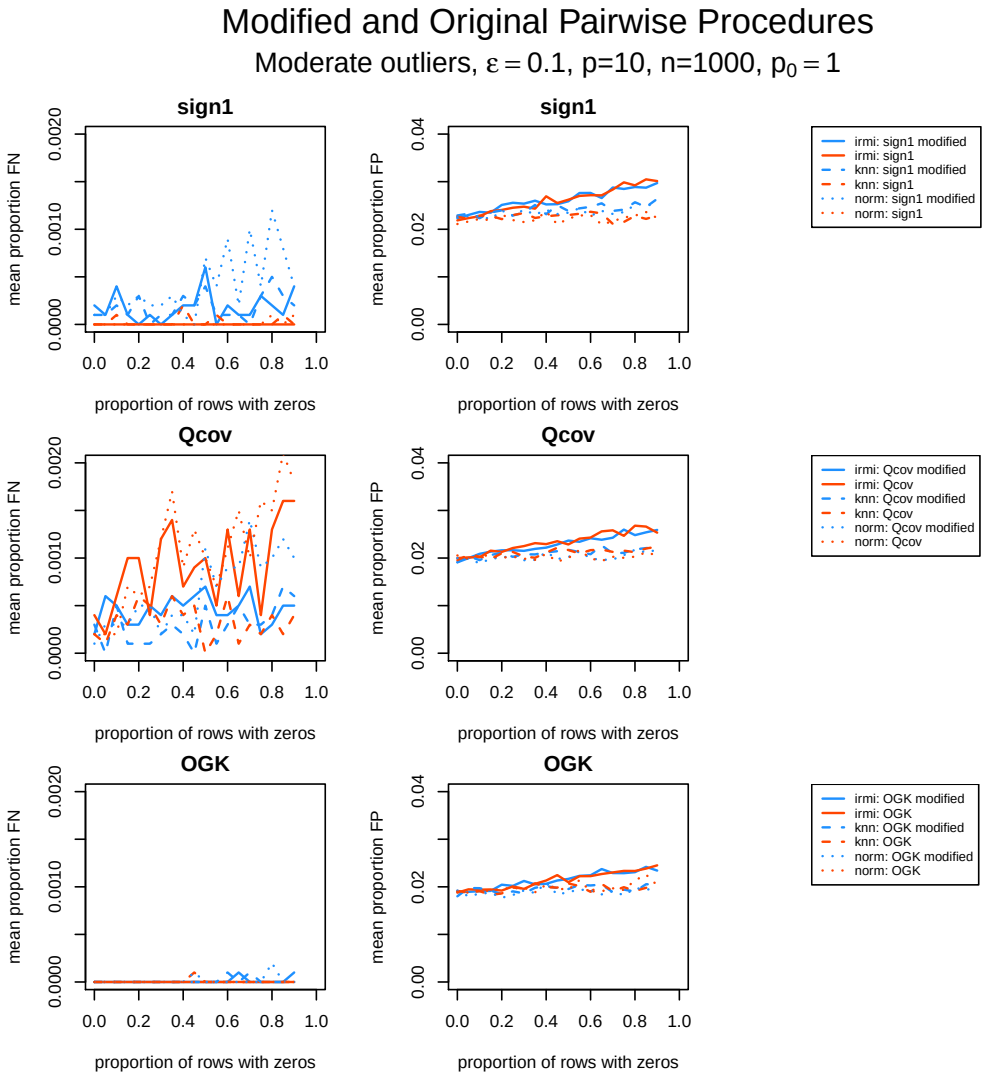


Figure 12.8: Average proportion of false negatives (left) and false positives (right) for a fixed proportion of outliers and varying percentage of rows with zeros. Comparison of pairwise outlier detection procedures with different types of imputation.

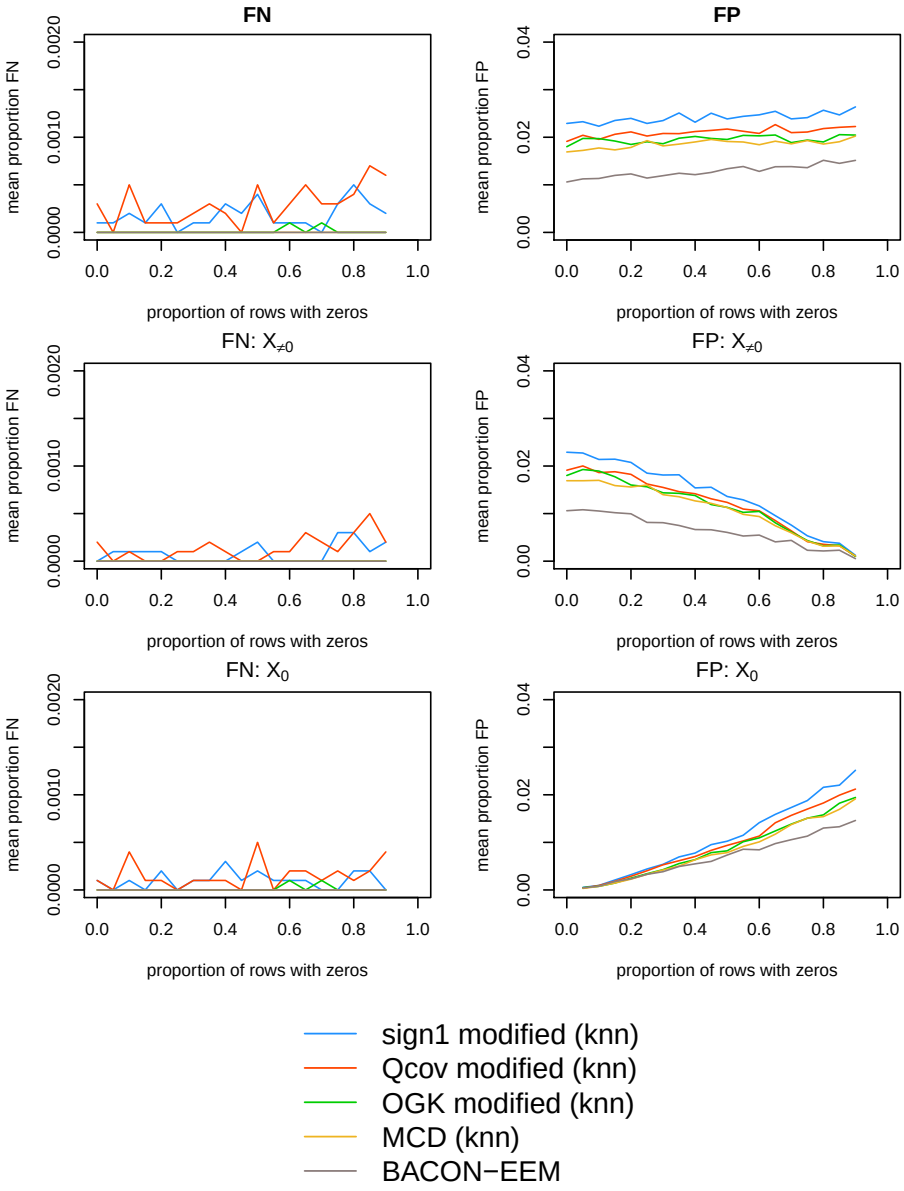


Figure 12.9: Average proportion of false negatives (left) and false positives (right) for a fixed proportion of outliers and varying percentage of rows with zeros (moderate outliers, $\varepsilon = 0.1$, $p = 10$, $n = 1000$, $p_0 = 1$). Comparison of $\mathbf{X}$, $\mathbf{X}_{\neq 0}$ and $\mathbf{X}_0$. *Legend:* [blue] sign1 modified (knn); [red] Qcov modified (knn); [green] OGK modified (knn); [yellow] MCD (knn); [grey] BACON-EEM.

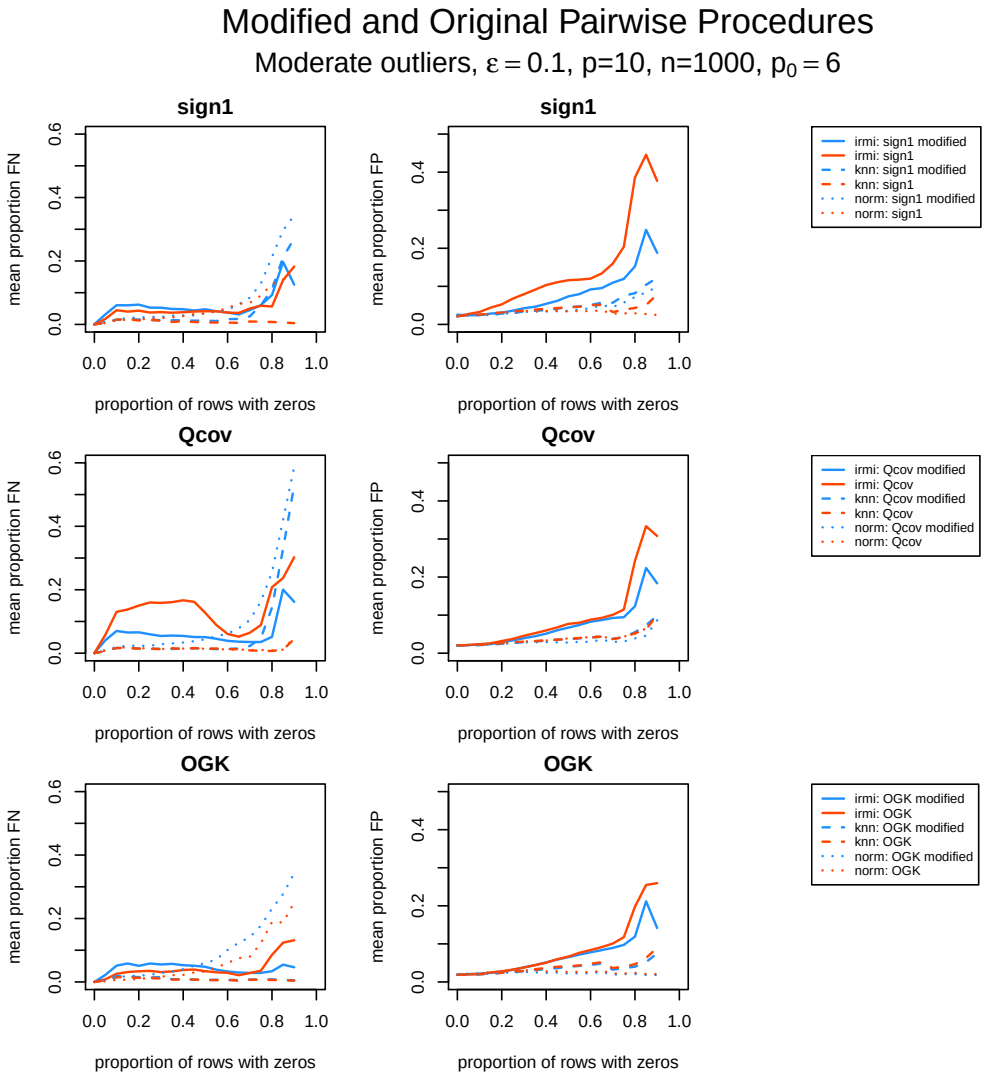


Figure 12.10: Average proportion of false negatives (left) and false positives (right) for a fixed proportion of outliers and varying percentage of rows with zeros. Comparison of pairwise outlier detection procedures with different types of imputation.

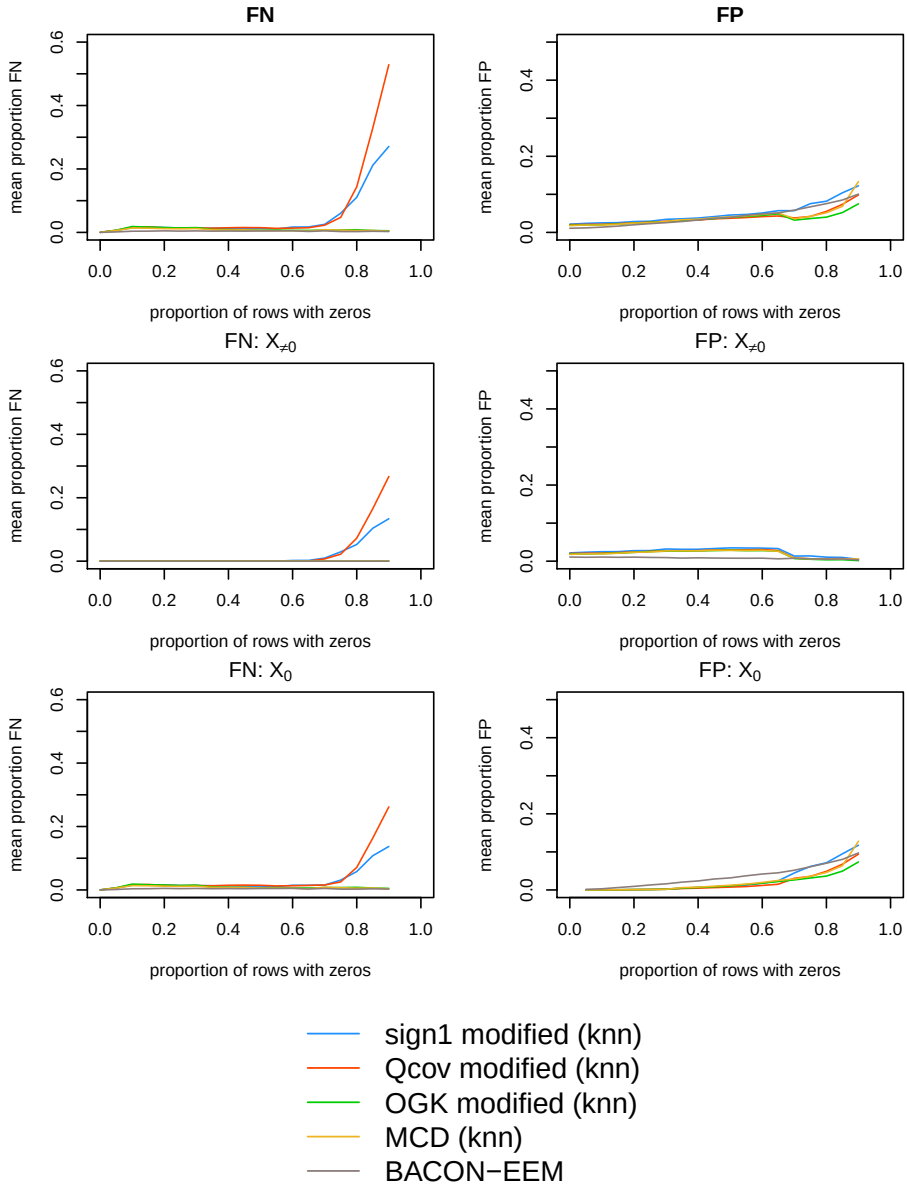


Figure 12.11: Average proportion of false negatives (left) and false positives (right) for a fixed proportion of outliers and varying percentage of rows with zeros (moderate outliers, $\varepsilon = 0.1$, $p = 10$, $n = 1000$, $p_0 = 6$). Comparison of $\mathbf{X}$, $\mathbf{X}_{\neq 0}$ and $\mathbf{X}_0$. *Legend:* [blue] sign1 modified (knn); [red] Qcov modified (knn); [green] OGK modified (knn); [yellow] MCD (knn); [grey] BACON-EEM.

Extreme Outliers

In MERANER (2010), a simulation setup with extreme outliers is studied. With a low proportion of structural zeros ($n_0/n < 0.65$) BACON-EEM performs best. For $n_0/n > 0.65$, on the other hand, the modified OGK is best.

Smaller Number of Observations

In MERANER (2010), a simulation setup with a small number of observations is studied.

Simulations were therefore also performed for data sets of dimension 100×5. In this section, the results for the extreme case of $p_0 = 3$, i.e. 3 zeros per row, are presented. MERANER (2010) shows that with moderate outliers shows no improvement between original and modified pairwise methods and indicates that the knn imputation algorithm is to be preferred in most cases. However, for a very high proportion n_0/n of rows with zeros, IRMI imputation tends to render better results in the case of the modified pairwise methods Qcov and OGK. Moreover, the BACON-EEM algorithm is best up to a proportion of approximately $n_0/n = 0.3$ rows with zeros and the OGK algorithm is best for a proportion of $n_0/n > 0.3$ rows with zeros.

12.3 Summary and Conclusions

This anaylsis was motivated by the question whether certain modifications of pairwise methods for robust estimation of location and scatter would render better results in connection with outlier detection than the original methods when applied to semi-continuous variables. These modifications were based on the idea that omitting the zeros in the data in a pairwise manner could prove effective for pairwise procedures when dealing with a high percentage of zeros.

The pairwise methods chosen for this purpose were the OGK estimator, the sign1 covariance matrix and the quadrant correlation. The modified OGK estimator now includes an imputation technique, as do the outlier detection procedures corresponding to the adapted methods. Hence, there is a strong dependency on a "good choice" of imputation which is analyzed in a series of plots comparing modified and original pairwise methods for different types of imputation, where the original pairwise methods are applied to previously imputed data sets.

The most appropriate type of imputation is then used for another series of plots comparing the modified pairwise methods to two multivariate procedures, the MCD estimator and the BACON-EEM algorithm. The MCD estimator is also applied to a previously imputed data set, which is not necessary for the BACON-EEM algorithm where an imputation procedure is already included.

The first series of plots (see Section 12.2) indicate that the modifications of the pairwise methods do not improve the results. They also show that knn is the best imputation technique for the chosen parameters as opposed to imputation with norm or the IRMI

algorithm. Subsequently, knn imputation was used for the second series of plots which generally identify the BACON-EEM procedure as being the best choice.

All of the above methods are implemented in the statistical environment R. The source code of the pairwise procedures that were modified in the course of this analysis can be found in MERANER (2010) and also the corresponding software.

Bibliography

Béguin, C. and **Hulliger, B.** (**2008**): *The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data*. Survey Methodology, 34 (1), pp. 91–103.

Blomqvist, N. (**1950**): *On a Measure of Dependence Between two Random Variables*. The Annals of Mathematical Statistics, 21 (4), pp. 593–600.

Filzmoser, P. and **Gschwandtner, M.** (**2009**): mvoutlier: Multivariate outlier detection based on robust methods. R package version 1.4.
URL http://cran.r-project.org/web/packages/mvoutlier/

Filzmoser, P., **Maronna, R.** and **Werner, M.** (**2008**): *Outlier identification in high dimensions*. Computational Statistics and Data Analysis, 52, pp. 1694–1711.

Gnanadesikan, R. and **Kettenring, J.** (**1972**): *Robust estimates, residuals, and outlier detection with multiresponse data*. Biometrics, 28, pp. 81–124.

Gonnet, G. and **Scholl, R.** (**2009**): Scientific Computation. Cambridge: Cambridge University Press, iSBN: 978-0-521-84989-0.

Locantore, N., **Marron, J.**, **Simpson, D.**, **Tripoli, N.**, **Zhang, J.** and **Cohen, K.** (**1999**): *Robust principal component analysis for functional data*. Test, 8 (1), pp. 1–73, sociedad de Estadística e Investigación Operativa.

Maechler, M., **Rousseeuw, P.**, **Croux, C.**, **Todorov, V.**, **Ruckstuhl, A.**, **Salibian-Barrera, M.** and **Verbeke, T.** (**2009**): robustbase: Basic Robust Statistics. R package version 0.5-0-1.
URL http://cran.r-project.org/web/packages/robustbase/

Maronna, R., **Martin, R.** and **Yohai, V.** (**2006**): Robust Statistics. Chichester: Wiley & Sons, iSBN: 0-470-01092-4.

Maronna, R. and **Zamar, R.** (**2002**): *Robust Estimates of Location and Dispersion for High-Dimensional Datasets*. Technometrics, 44 (4), pp. 307–317.

Meraner, A. (**2010**): Outlier Detection for Semi-continuous Variables. Diplomarbeit, Institut f. Statistik und Wahrscheinlichkeitstheorie, Technische Universität, Wien.

Mosteller, F. (**1946**): *On some useful ïnefficientßtatistics*. Annals of Mathematical Statistics, 17 (4), pp. 377–408.

Rousseeuw, P. (**1985**): *Multivariate estimation with high breakdown point.* **Grossmann, W.**, **Pflug, G.**, **Vincze, I.** and **Wertz, W.** (editors) Mathematical Statistics and Applications, pp. 283–297, Dordrecht: Reidel.

Shevlyakov, G. (**1997**): *On Robust estimation of a correlation coefficient.* Journal of Mathematical Sciences, 83 (3), pp. 434–438.

Todorov, V. (**2010**): rrcov: Scalable Robust Estimators with High Breakdown Point. R package version 1.0-00.
URL http://cran.r-project.org/web/packages/rrcov/

Todorov, V., **Templ, M.** and **Filzmoser, P.** (**2011**): *Detection of Multivariate Outliers in Business Survey Data with Incomplete Information.* Advances in Data Analysis and Classification, accepted for publication.

Wall, M., **Rechtsteiner, A.** and **Rocha, L.** (**2003**): *Singular Value Decomposition and Principal Component Analysis.* **Berrar, D.**, **Dubitzky, W.** and **Granzow, M.** (editors) A Practical Approach to Microarray Data Analysis, chapter 5, pp. 91–109, Kluwer Academic Publishers.