

Robustification of the Quintile Share Ratio

Beat Hulliger¹, Tobias Schoch²

¹University of Applied Sciences Northwestern Switzerland, School of Business,
e-mail: beat.hulliger@fhnw.ch

²University of Applied Sciences Northwestern Switzerland, School of Business,
e-mail: tobias.schoch@fhnw.ch

Abstract

The Quintile Share Ratio (QSR) – a primary European Laeken indicator for the inequality of the income distribution – is extremely sensitive to the presence of large outliers. This is clear from the influence function of the QSR. Outliers also increase the variability of the quintile share ratio considerably. Robust statistics has developed the tools to control the impact of outliers. However, straight-forward robustification, like trimming the extreme observations, entails a bias under asymmetric distributions. This paper presents first ideas and results which show the potential of a bias-compensated, robust Quintile Share Ratio. We propose three different non-parametric robust estimators for the QSR that rely on balanced and compensated trimming and asymmetric M-estimation. The estimators are simple structured to meet the practitioners needs. The most promising approaches will be studied in the AMELI¹ project by extensive simulation with SILC data.

Keywords: Robustness, Laeken indicator, inequality indicator

1 Introduction

The Quintile Share Ratio (QSR) of disposable income is one of the leading inequality indicators of the European Statistics on Income and Living Conditions (EU-SILC). It is defined as the ratio of total (equivalized disposable)² income received by the top 20% of a country's population with highest income to that received by the 20% of the country's population with the lowest income. It is therefore an inequality indicator, meaning that not the level of the income is of interest – like for poverty indicators – but the distribution of the income. The QSR is closely related to the well known but more complex Gini indicator (cf. Cowell, 2003). The QSR is part of a set of indicators which measure social cohesion across European countries and which have been denominated the Laeken indicators due to the place where the European Union heads of governments decided to use these indicators to monitor progress of

¹ AMELI stands for "Advanced Methodology for European Laeken Indicators". The project is funded under the European Commission's 7th Framework Programme. EC-Project Reference: 217322, Research area: SSH-2007-6.2-01 Improved ways of measuring both the potential for and impact of policies. Visit: <http://www.ameli.surveystatistics.net>.

² Equivalized income is defined as the household's total disposable income divided by its "equivalized size" (modified OECD-scale), to take account of the size and composition of the household, and is attributed to each household member including children.

Member States in their efforts against poverty and social exclusion (European Commission, 2003; Atkinson et al., 2002).

The QSR of the population is estimated based on the SILC Survey, a coordinated household survey in the countries of the European Economic Area and in Switzerland.³ These surveys have complex sample designs often with rotating panels. In addition complex estimation procedures are used to reduce possible bias due to non-response. QSR-estimators, therefore, use survey weights to ensure proper estimation of means.

The starting point for the estimation of the QSR is a sample of incomes y_1, \dots, y_n with corresponding sample weights w_1, \dots, w_n . The cumulative distribution function (cdf) of incomes in the population is $F_U(t) = \sum_{i \in U} w_i \mathbb{1}\{y_i \leq t\} / \sum_{i \in U} w_i$, where $\mathbb{1}\{A\}$ is an indicator function for a set A . The cdf $F_U(t)$ is estimated by the weighted empirical distribution function $F_S(t) = \sum_{i=1}^n w_i \mathbb{1}\{y_i \leq t\} / \sum_{i=1}^n w_i$. Hence, the target parameter QSR of the population is defined as

$$\eta = \frac{\sum_{i \in U} y_i \mathbb{1}\{F_U^{-1}(0.8) < y_i\}}{\sum_{i \in U} y_i \mathbb{1}\{y_i \leq F_U^{-1}(0.2)\}}. \quad (1)$$

In the following we use the notation $Q_F(\alpha) = F^{-1}(\alpha)$ ($0 \leq \alpha \leq 1$) for the quantiles of a distribution F . The calculation of the QSR of the population is a three-step procedure:

1. Estimation of the 20% and the 80% quantile of the population. That is: $Q_{F_S}(0.2) = F_S^{-1}(0.2)$ and $Q_{F_S}(0.8) = F_S^{-1}(0.8)$.⁴
2. Estimation of the mean or total of the lowest and highest quintile, i.e. estimation of the mean (or total) of incomes between the minimal income and the 20% quantile and of the mean (or total) of incomes between the 80% quantile and the maximum of incomes. The classical estimator for the (weighted) mean is $m_1 = \sum_{i \in S} w_i y_i \mathbb{1}\{y_i \leq Q_{F_S}(0.2)\} / \sum_{i \in S} w_i \mathbb{1}\{y_i \leq Q_{F_S}(0.2)\}$ and $m_5 = \sum_{i \in S} w_i y_i \mathbb{1}\{Q_{F_S}(0.8) < y_i\} / \sum_{i \in S} w_i \mathbb{1}\{Q_{F_S}(0.8) < y_i\}$.
3. The classical QSR-estimator is the ratio of these two means⁵, i.e.

$$\hat{\eta} = \frac{m_5}{m_1} = \frac{\sum_{i \in S} w_i y_i \mathbb{1}\{Q_{F_S}(0.8) \leq y_i\}}{\sum_{i \in S} w_i y_i \mathbb{1}\{y_i \leq Q_{F_S}(0.2)\}}. \quad (2)$$

The problems involved in the estimation of the QSR are

1. Development of an estimation procedure for population means, i.e. usually the development of sampling weights. Note that sampling weights for population means may not be optimal for the QSR.

³ See European Commission (2003) for a detailed description of the estimation methods and editing rules.

⁴ EUROSTAT additionally recommends to choose the quintiles such that persons from the same household belong to the same quintile (cf. European Commission, 2003). For the sake of simplicity of our notation we omit this distinction.

⁵ Note, we assumed in eq. (2) that the means over the particular income quintiles comprise the same total weight, for the sake of simplicity. Otherwise, an adjustment term must be added.

2. Estimation of the empirical cdf $F_S(t)$ or, more modestly of the corresponding quintiles. Note that there are various definitions to overcome the discreteness of these estimators and to cope with the definition of disposable income (the same disposable income is attributed to the persons of a household).
3. Estimation of the means m_1 and m_5
4. Calculation of the ratio $\hat{\eta} = m_5/m_1$.

From a robustness point of view the classical QSR-estimator is problematic. The means in step 3) have a breakdown point of 0, meaning that any outlier income may distort it to an arbitrary amount. Accordingly the influence of a single observation in the tails of the distribution is linear and therefore unbounded. Figure 1(a) shows the sensitivity curve of the classical QSR-estimator for the public use data set from AT-SILC 2004 (see Statistics Austria (2007a) also for explanations on AT-SILC 2004).⁶ The sensitivity curve is an empirical approximation to the influence function (cf. Hampel et al., 1986), see also Figure 1(b). Obviously large negative observations increase the QSR due to their influence on m_1 in the denominator of $\hat{\eta}$. One may object that the minimal possible income is bounded below by 0. In practice this is not true but admittedly negative income is a special problem which will not be considered here. Large incomes have a high impact on m_5 resulting in a linear and unbounded increase of the sensitivity curve (see Figure 1(a)).

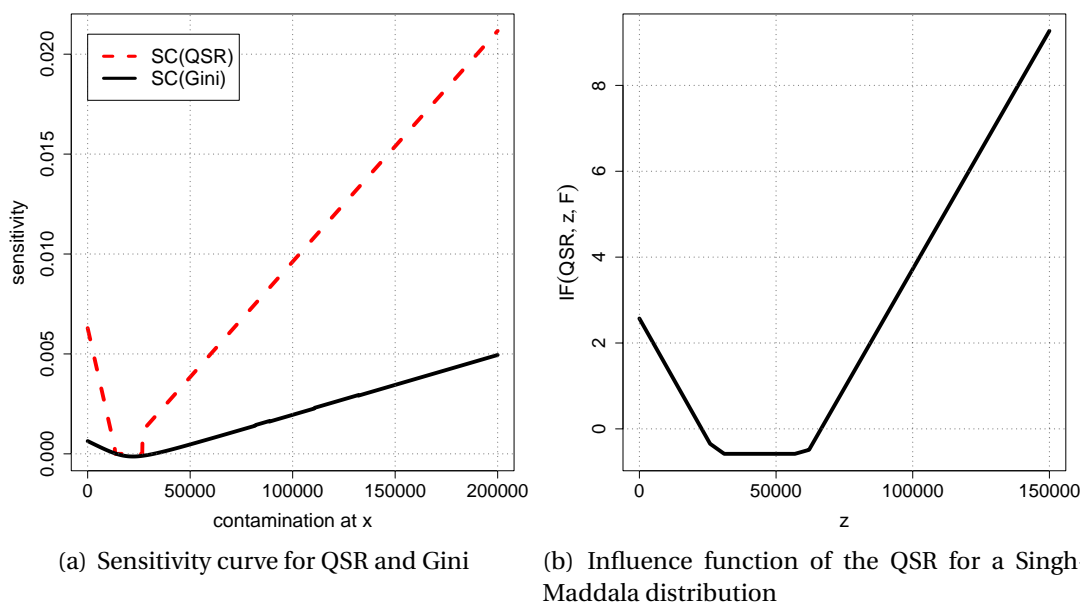


Figure 1: Sensitivity curve for the QSR and the Gini coefficient (using data from the public use sample AT-SILC 2004) and the influence curve for the QSR at a theoretical distribution (Singh-Maddala).

Thus for practical purposes the classical QSR-estimator is not to be recommended in many situations. It is easy to see that this non-robust behavior of the QSR-estimator

⁶ An extensive description of the Austrian SILC data can be found in Statistics Austria (2007b) and Statistics Austria (2006).

$\hat{\eta}$ is directly inherited from the QSR in the population η . The reason for choosing QSR as an estimand is exactly this behavior which yields a very sensitive indicator of inequality. Unfortunately the sensitivity of the QSR-estimator not only makes it unreliable when outliers occur but, in addition, it may entail the QSR a very inefficient estimator with a large variance. Nevertheless we accept this challenge to estimate a population characteristic which inherently cannot be estimated in a robust way. The challenge is similar to estimating a population mean: the only way to obtain a robust estimator is to accept some bias (cf. Hampel et al., 1986). As long as there is a gain in mean squared error due to a lower variance this small bias will still pay off even in terms of efficiency. Robustification will particularly be adopted in step 3) of the estimation. However, when considering the bias of the final estimator step 4) will be important, too.

In the pertinent literature robust parametric estimation and to a less extent semi-parametric approaches are commonly discussed, an overview is provided e.g. by Victoria-Feser (2000) and Marazzi and Ruffieux (1999), Cowell and Victoria-Feser (2007), respectively. On the other hand, these methods are seldom applied in practice, aside from the study of van Kerm (2007) on the robustness properties of Laeken indicators. The robust nonparametric estimation of the mean of asymmetric distributions is treated in various articles. Particularly interesting are Searls (1966), Fuller (1991), Hulliger (1995), Cowell and Victoria-Feser (2003) and (2006).

2 Methods

We consider several methods to obtain a robustified version of the QSR-estimator. They differ in two ways: Whether only the largest incomes are treated, i.e. whether both shares m_1 and m_5 are treated, and how the robustness tuning constants are chosen. We use cumulative means – (cf. Cowell and Victoria-Feser, 2003) – to describe the estimators

$$C_{F_S}(\alpha) = \frac{\sum_{i \in S} w_i y_i \mathbb{1}\{y_i \leq Q_F(\alpha)\}}{\sum_{i \in S} w_i \mathbb{1}\{y_i \leq Q_F(\alpha)\}}. \quad (3)$$

Note that the mean of the observations between the α_l and α_u quantiles is $C_F(\alpha_u) - C_F(\alpha_l)$. Thus $m_1 = C_{F_S}(0.2)$ and $m_5 = C_{F_S}(1) - C_{F_S}(0.8)$. Hence, the classical QSR-estimator can be written as

$$\hat{\eta}_{QSR} = \frac{C_{F_S}(1) - C_{F_S}(0.8)}{C_{F_S}(0.2)}. \quad (4)$$

We use symmetric trimming of extreme observations as a baseline comparison estimator. This has been extensively studied by Cowell and Victoria-Feser (2006). The basic idea is simple: the observations above $Q_{1-\alpha/2}$ and below $Q_{\alpha/2}$ are trimmed and the shares of the (original) quintiles are estimated with the remaining observations. We use the obvious adaptation to cope with sampling weights. The bias introduced by removing the lowest incomes for estimating the lowest quintile share, which is in the denominator of the QSR, and the largest observations, which act on the numerator of the QSR, pull the trimmed QSR in the same direction, i.e. decrease the estimate. Thus we can expect that this estimator rapidly builds up a large bias. We

call this estimator TQSR in the following, for Trimming Quintile Share Ratio. Thus we have

$$\hat{\eta}_{TQSR}(\alpha_l, \alpha_u) = \frac{C_{F_5}(1 - \alpha_u) - C_{F_5}(0.8)}{C_{F_5}(0.2) - C_{F_5}(\alpha_l)}, \quad (5)$$

where upper and lower trimming proportion are equal, i.e. $\alpha_l = \alpha_u$. An adaptation is to trim only the largest observations of the sample, i.e. to set $\alpha_l = 0$. This estimator is then called one-sided TQSR.

As a first attempt to reduce the bias we propose to trim the largest observations in the first quintile instead of the lowest observations as in TQSR. Trimming the largest observations of the lowest quintile will have the opposite effect on the bias as trimming the largest observations of the largest quintile. Thus there is hope that the bias of this version of QSR-estimator is reduced compared with a one-sided TQSR or the ordinary TQSR. The amount of trimming in the lowest quintile may be chosen such that the bias of the resulting QSR-estimator vanishes. We call this bias corrected trimming QSR-estimator (BQSR). To distinguish it from our next proposal we denominate it BQSR1. Thus BQSR1 is defined as

$$\hat{\eta}_{BQSR1}(\alpha_l, \alpha_u) = \frac{C_{F_5}(1 - \alpha_u) - C_{F_5}(0.8)}{C_{F_5}(0.2 - \alpha_l)}. \quad (6)$$

The choice of α_u is mostly guided by the assumption on how many outliers we have to expect in a sample. Usually it will be very low, say $\alpha_u = 0.005$. The choice of α_l is mostly guided by the wish to reduce the bias of the BQSR1 as much as possible. In fact, under theoretical distributions, it is possible to calculate α_l as a function $h(\alpha_u)$ such that the BQSR1 estimator has no bias, i.e.

$$\hat{\eta}_{BQSR}(h(\alpha_u), \alpha_u) = \hat{\eta}. \quad (7)$$

In practice choosing the right function $h(\alpha_u)$ is equivalent to estimate the tail mass above α_u which is exactly the task of optimal robustification and therefore we will use a rule of thumb instead to let α_l depend on α_u .

Our second proposal leaves the estimation of the share of the lowest quintile untouched and aims at a bias correction directly for the share of the highest quintile when a portion of the largest observations is trimmed. Thus we start by estimating m_5 by $C_{F_5}(1 - \alpha_u) - C_{F_5}(0.8)$. We propose to approximate the bias incurred when trimming α_u observations by observing the additional bias incurred when trimming a further proportion α_u of the observations. We then apply a bias correction term to the first estimate. The bias correction should reflect the form of $C_F(1 - \alpha_u)$ as a function of α_u and, therefore again, a perfect bias correction would come down to estimating the tail distribution perfectly right. Here we test an approach where we assume $C_F(1 - \alpha_u) - C_F(0.8) = C_F(1 - 2\alpha_u) \beta^{\alpha_u}$. In other words we assume that the trimmed mean increases exponentially when reducing the trimming proportion. Obviously other functional forms may be reasonable. This proposal leads to

$$\hat{\eta}_{BQSR2}(\alpha_u) = \frac{C_{F_5}(1 - \alpha_u) - C_{F_5}(0.8)}{C_{F_5}(0.2)} \frac{C_{F_5}(1 - \alpha_u) - C_{F_5}(0.8)}{C_{F_5}(1 - 2\alpha_u) - C_{F_5}(0.8)}. \quad (8)$$

The advantage of BQSR2 is that it depends on only one tuning constant α_u at the price of relying on a heavy assumption about the form of the bias.

Our third proposal is an adaptation of the Minimum Estimated Risk (MER) estimators proposed in Hulliger (1995). He treats the problem of the robust estimation of the mean of an asymmetric distribution. The basic idea is the following: We use an M-estimator $T(F_S, k)$ with asymmetric (Huber) ψ -function $\psi(x, k) = \min(x, k)$ with ($k > 0$) to estimate the mean of the highest quintile. It is defined as the solution of

$$\sum_{i \in S} w_i \mathbb{1}\{Q_{F_S}(0.8) < y_i\} \psi(y_i - T, k) = 0. \quad (9)$$

Note, that in practice a preliminary scale estimator is used to standardize $y_i - T$. This is an important help in choosing a tuning constant k . We can use this M-estimator directly if we can decide on the tuning constant k to be chosen. Alternatively we use the following approach to determine k : We estimate the bias of $T(F_S, k)$ by its difference to $m_5 = C_{F_S}(1) - C_{F_S}(0.8)$ and we estimate the sampling variance of $T(F_S, k)$ by $\hat{V}(T(F_S, k))$. Next we search for a minimum of the estimated mean squared errors

$$r(k) = \hat{V}(T(F_S, k)) + (T(F_S, k) - m_5)^2 \quad (10)$$

as a function of the tuning constant k . Then the M-estimator $T(F_S, k_0)$ with the minimizing tuning constant k_0 (MER-estimator) is chosen to estimate the share of the highest quintile. We call this variant of the QSR-estimator MQSR. This estimator does not need a tuning constant to be chosen. However, its breakdown point is 0 because the bias estimator involves the non-robust mean of the highest quintile as the classical QSR. Hulliger (1991) and (1995) showed that the resulting estimator is more efficient and more robust than the original estimator as long as the squared bias does not dominate the variance.

3 Data and preliminary results

We confine ourselves to the discussion of the robustness properties (sensitivity curves), the choice the robustness tuning constants and bias. Thus, this article does not yet investigate the efficiency and the variance estimation of the various proposals. The variance of the most promising estimators will be evaluated in the AMELI project by extensive simulations with SILC data. Our analysis of the proposed estimators in the paper at hand is based on the public use sample (PUS) of AT-SILC 2004. The PUS data set is a simple random sample (sampling fraction 50%) from the full AT-SILC 2004 (Statistics Austria, 2007a).

It is evident from Figure 2 that the usual estimator for the QSR is not robust, in the sense that its sensitivity curve and influence function are not bounded from above (see also Figure 1). This formally means that a single observation, provided it is sufficiently large, can drive the estimated indicator arbitrarily large. These results are in line with the simulation results by van Kerm (2007). The simplest approach with an intuitive appeal to prevent from an unbounded influence, is trimming, that is the TQSR estimator. Cowell and Victoria-Feser (2003) and (2006) derived the influence function for several statistics (including the cumulative mean functionals) for trimmed samples. Hence, one can easily show that the influence function of the TQSR is bounded.⁷ But the effect of trimming (either one-sided or two-sided trimming) can seriously bias the estimate. This is illustrated in Table 1, where α_l is set to

⁷ Assuming that the incomes are positive or at least bounded from below (what is usually the case

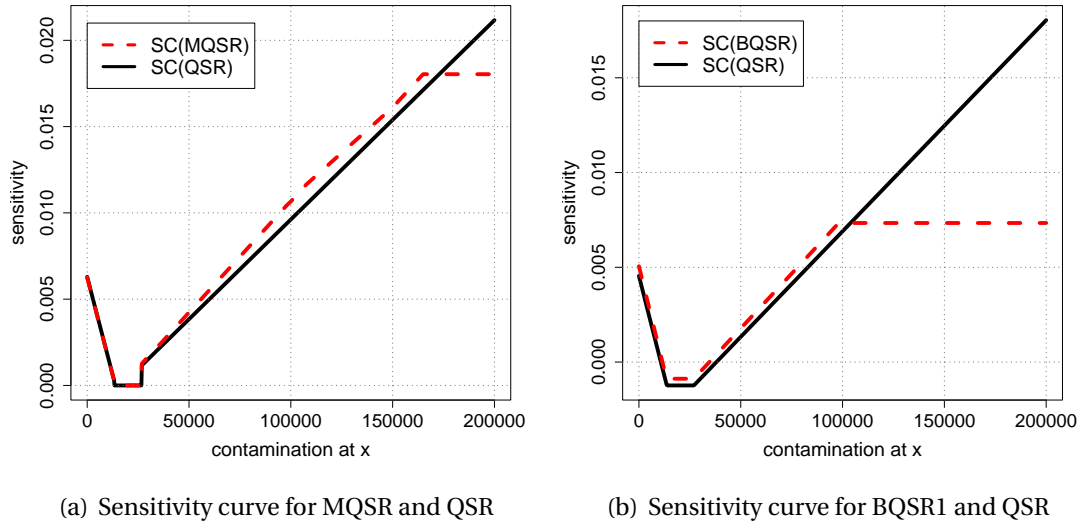


Figure 2: Sensitivity curves for the MQSR ($k=13.8$) and the BQSR1 ($\alpha_u = 0.002$, $\alpha_l = 0.0045$) compared to the standard QSR (data: public use sample AT-SILC 2004).

Table 1: Bias (in %) of TQSR and BQSR1 with particular choices of α_l and α_u

$(\alpha_l) \nabla$	TQSR (α_u)					BQSR1 (α_u)				
	0.001	0.002	0.005	0.01	0.02	0.001	0.002	0.005	0.01	0.02
0	-1.72	-3.18	-5.31	-7.56	-10.25	-1.72	-3.18	-5.31	-7.56	-10.25
0.005	-4.02	-5.44	-7.52	-9.72	-12.37	-0.65	-2.12	-4.28	-6.56	-9.30
0.010	-5.95	-7.35	-9.39	-11.55	-14.15	0.48	-1.01	-3.19	-5.49	-8.27
0.020	-9.15	-10.49	-12.46	-14.55	-14.06	2.73	1.21	-1.02	-3.37	-6.21
0.050	-15.35	-16.60	-18.44	-20.38	-22.72	11.05	9.40	6.99	4.45	1.38
0.100	-21.96	-23.12	-24.81	-26.60	-28.75	32.92	30.96	28.07	25.02	21.35

Data: public use sample AT-SILC

zero (i.e. one-sided trimmed TQSR), the relative bias (in %) increases with the choice of α_u . In case of the TQSR with trimming at both ends of the sample the bias is even larger.

Cowell and Victoria-Feser (2006) propose to choose a trimming proportion according to efficiency (say, 85%) of the estimator at a known parametric distribution. Here we propose to use non-parametric estimators, BQSR1, BQSR2 or MQSR, respectively. All three candidates have a bounded sensitivity curve (and influence function) – see Figure 2 (the graph for BQSR2 is not shown) – and compensate (to some extent) for the arising bias. For BQSR1 we give preliminary rules for the choice of the constants (see Table 1). That is, if the statistician trims his data (according to an educated guess concerning the amount of outlying observations), say by $\alpha_u = 0.001$ (0.1%), he

in practical applications), the influence function of the TQSR is composed of two elements: firstly the influence function of the quintiles and secondly the influence functions of the (one-sided) trimmed mean. Both components are bounded, ergo $IF(TSQR, \bullet)$ is bounded, too.

Table 2: Comparison of bias for BQSR2, (one-sided) TQSR and MQSR

	trimming proportion (α_u)							
	0	0.001	0.002	0.003	0.005	0.01	0.02	0.05
BQSR2 relative bias (in %)	0	-0.24	-1.34	-2.63	-3.03	-4.76	-6.39	-10.01
TQSR relative bias (in %)	0	-1.72	-3.18	-4.28	-5.31	-7.56	-10.28	-15.45
MQSR* relative bias (in %)	-	-1.22	-2.09	-2.57	-3.60	-	-	-

Note: * for MQSR k was chosen such, that the number of declared outliers is equivalent to the trimming proportion of the other estimators; Data: public use sample AT-SILC

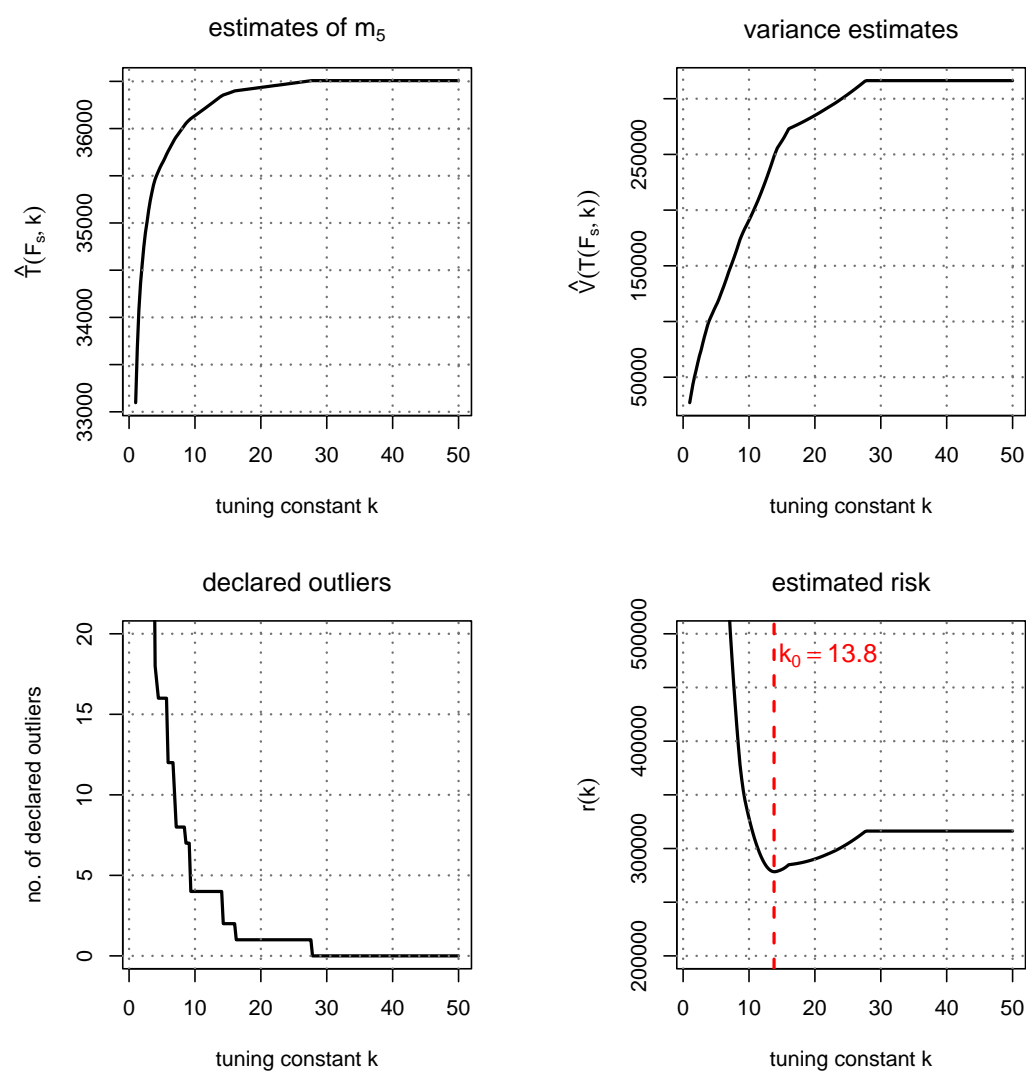


Figure 3: Visual display of the performance (i.e. estimates of m_5 , variance estimates, estimated risk and number of declared outliers) of MQSR. The Minimum Estimated Risk is attained with $k_0 = 13.8$. (data: public use sample AT-SILC 2004).

is well advised to choose (from a rule of thumb) $\alpha_l = 0.01$ (1%), to get a bias of 0.48% (see Table 1). Without the adaptation (i.e. $\alpha_l = 0$) the bias would be about 3.5 times higher (note that $TQSR(0, \alpha_u) = BQSR1(0, \alpha_u)$). With small upper trimming proportions BQSR1 can compensate for the bias. For larger trimming proportions there is no α_u which compensates the bias, since α_u should be considerably smaller than 0.2.

Alternatively BQSR2 allows another compensation, where the statistician has to choose only an adequate upper trimming proportion, α_u . In Table 2 we computed the relative bias (in %) for different choices of α_u . In addition the corresponding values of the one-sided trimmed QSR (TQSR) are added as benchmark. It is evident, that the bias of BQSR2 is substantial smaller, for small α_u . In other words, BQSR2 allows for moderately small trimming proportions a reasonable bias compensation.

An asymmetric M-estimator with a comparable tuning constant has a smaller bias than the one-sided TQSR since it downweights the outliers less (see Table 2). The performance of the MQSR is displayed in Figure 3.

4 Conclusions

This article does not yet investigate the efficiency and the variance estimation of the various proposals. The purpose is to explore the potential of these estimators. There is hope that some robustness can be gained without inducing too large bias. The most promising proposals will be evaluated in the AMELI project by extensive simulations with SILC data.

References

- Atkinson, T., B. Cantillon, E. Marlier, and B. Nolan (2002). *Social Indicators: The EU and Social Inclusion*. Oxford: Oxford University Press.
- Cowell, F. A. (2003). Measurement of inequality. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution* (2 ed.), Volume 1, pp. 87–166. Amsterdam: Elsevier.
- Cowell, F. A. and M.-P. Victoria-Feser (2003). Distribution-free inference for welfare indices under complete and incomplete information. *Journal of Economic Inequality* 1, 191–219.
- Cowell, F. A. and M.-P. Victoria-Feser (2006). Distributional dominance with trimmed data. *Journal of Business & Economic Statistics* 24(3), 291–300.
- Cowell, F. A. and M.-P. Victoria-Feser (2007). Robust stochastic dominance: A semi-parametric approach. *Journal of Economic Inequality* 5, 21–37.
- 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.
- Fuller, W. A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica* 1, 137–158.
- Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York et al.: 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), 79–87.
- Marazzi, A. and C. Ruffieux (1999). The truncated mean of an asymmetric distribution. *Computational Statistics & Data Analysis* 32, 79–100.
- 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), 1200–1204.
- Statistics Austria (2006). Einkommen, Armut und Lebensbedingungen. Ergebnisse aus EU-SILC 2004. Technical report, Statistics Austria, Vienna.
- Statistics Austria (2007a). Erläuterungen: Mikrodaten-Subsample für externe User. Technical report, Statistics Austria, Vienna.
- Statistics Austria (2007b). Standard-Dokumentation, Metainformationen (Definitionen, Erläuterungen, Methoden und Qualität) zu EU-SILC 2004. Technical report, Statistics Austria, Vienna.
- van Kerm, P. (2007). Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. *IRISS-C/I Working Paper*.
- Victoria-Feser, M.-P. (2000). Robust methods for the analysis of income distributions, inequality and poverty. *International Statistical Review* 68(3), 277–293.