

Reducing the mean squared error of quantile-based estimators by smoothing

Mia Hubert, Irène Gijbels and Dina Vanpaemel

May 2, 2012

Abstract

Many univariate robust estimators are based on quantiles. As already theoretically pointed out by Fernholz (1997), smoothing the empirical distribution function with an appropriate kernel and bandwidth can reduce the variance and mean squared error (MSE) of some quantile-based estimators in small data sets. In this paper we apply this idea on several robust estimators of location, scale and skewness. We propose a robust bandwidth selection and bias reduction procedure. We show that the use of this smoothing method indeed leads to smaller MSEs, also at contaminated data sets. In particular we obtain better performances for the medcouple which is a robust measure of skewness that can be used for outlier detection in skewed distributions.

1 Introduction

The goal of this paper is to construct methods for reducing the variance and the mean squared error (MSE) of different univariate robust estimators that are based on quantiles. In order to achieve this goal, the estimators are based on a kernel smoothed distribution function instead of the empirical distribution function. Smoothing the empirical distribution function is in particular advantageous in case of an underlying continuous distribution function. The first proposals to use kernel smoothing for distribution estimates date back to Nadaraya (1964) and Azzalini (1981). As usual, an appropriate choice of the bandwidth is of major importance, as over- or undersmoothing highly affects the bias and variance of

the estimators. It is shown by Fernholz (1997) that smoothing the empirical distribution function with an appropriate kernel and bandwidth can reduce the variance and MSE of estimators. This is most beneficial for estimators with a discontinuous influence function. They include the median for location, the interquartile range (IQR) for scale and the medcouple (MC) for estimating skewness (Brys et al., 2004). As the medcouple is very useful for outlier detection in skewed data (Hubert and Vandervieren, 2008; Hubert and Van der Veecken, 2008, 2010) it is our particular interest to reduce its MSE at small data sets. Our work is also motivated by the nonparametric regression method proposed in Čížek et al. (2008), which is based on smoothing the conditional distribution function.

In Section 2, the different estimators under study are defined. Our robust bandwidth selection procedure is explained in Section 3, and a method to reduce the bias of the smoothed estimators is introduced in Section 4. The performance of this robust bandwidth selection and of the bias reduction is studied in a simulation study in Section 5. Section 6 focusses on the medcouple, more specifically we study how often the medcouple, estimated on data from a positively skewed distribution, yields a positive number, and how smoothing improves the percentage of positive estimates. We also show that the smoothing procedure improves the ability to detect outliers with the adjusted boxplot (Hubert and Vandervieren, 2008), which uses the medcouple. In Section 7 this is illustrated on European international trade data, and finally Section 8 concludes.

2 Smoothing procedure

Let $X_n = \{x_1, x_2, \dots, x_n\}$ be an independent and identically distributed random sample drawn from an absolutely continuous distribution function $F(x)$ with density $f(x)$. The population quantile function is defined as

$$Q_p = \inf \{x : F(x) \geq p\} \quad (0 < p < 1).$$

Accordingly, the empirical quantile is given by

$$\hat{Q}_p = \inf \{x : F_n(x) \geq p\} \tag{1}$$

with $F_n(x)$ the empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(x \leq x_i)$$

with $I(\cdot)$ the indicator function.

As a location estimator we will study the sample median med_n of X_n :

$$\text{med}_n = \begin{cases} (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})/2 & \text{if } n \text{ is even} \\ x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \end{cases}$$

where $x_{(i)}$ denotes the i -th order statistic of X_n . Note that for n odd, med_n coincides with $\hat{Q}_{0.5}$, whereas for n even, $\text{med}_n = \frac{1}{2}(\hat{Q}_{lm} + \hat{Q}_{um})$ with $lm = \frac{1}{n}(\frac{n}{2}) = 0.5$ and $um = \frac{1}{n}(\frac{n}{2} + 1)$. For a scale estimator we look at the interquartile range

$$\text{IQR}_n = \hat{Q}_{0.75} - \hat{Q}_{0.25}.$$

To robustly estimate skewness, we consider the quartile skewness QS_n and octile skewness OS_n (Brys et al., 2003):

$$\text{QS}_n = \frac{(\hat{Q}_{0.75} - \text{med}_n) - (\text{med}_n - \hat{Q}_{0.25})}{\hat{Q}_{0.75} - \hat{Q}_{0.25}}$$

and

$$\text{OS}_n = \frac{(\hat{Q}_{0.875} - \text{med}_n) - (\text{med}_n - \hat{Q}_{0.125})}{\hat{Q}_{0.875} - \hat{Q}_{0.125}}.$$

We also study the medcouple (Brys et al., 2004) defined as:

$$\text{MC}_n = \text{med}_{x_i < \text{med}_n < x_j} g(x_i, x_j) \quad (2)$$

where for all $x_i \neq x_j$, the function g is given by:

$$g(x_i, x_j) = \frac{(x_j - \text{med}_n) - (\text{med}_n - x_i)}{x_j - x_i}. \quad (3)$$

As all the above mentioned estimates are based on quantiles, it follows from (1) that they can be computed from the empirical c.d.f. $F_n(x)$. Since $F_n(x)$ is discontinuous with (at most) n discontinuity points, it is not a very good estimator of the underlying continuous c.d.f. $F(x)$ when the sample size is small. To estimate the distribution function F in a smoother way, we can use the (continuous) kernel-based estimator (Nadaraya, 1964):

$$\tilde{F}_{n,h}(x) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

with $K(t)$ a distribution function having a density $k(t)$ that is symmetric around zero and h a bandwidth that controls the degree of smoothness. Since the choice of K is much less

important than the choice of a suitable bandwidth, we will only consider the integral of the Epanechnikov kernel, which is given by

$$K(t) = \begin{cases} 0 & t \leq -\sqrt{5} \\ \frac{3}{4\sqrt{5}}t - \frac{1}{20\sqrt{5}}t^3 + \frac{1}{2} & |t| < \sqrt{5} \\ 1 & t \geq \sqrt{5}. \end{cases}$$

Under the condition that $F(x)$ has continuous derivatives $f(x)$ and $f'(x)$, it can be shown (Azzalini, 1981) that as $n \rightarrow \infty$, $h \rightarrow 0$ and $nh \rightarrow +\infty$

$$\mathbb{E}(\tilde{F}_{n,h}(x)) = F(x) + \frac{1}{2}h^2 f'(x)\mu_2(k) + o(h^2) \quad (4)$$

and

$$\text{Var}(\tilde{F}_{n,h}(x)) = \frac{F(x)(1-F(x))}{n} - \frac{2hf(x)c}{n} + o\left(\frac{h}{n}\right) \quad (5)$$

where $\mu_2(k) = \int_{-\infty}^{+\infty} t^2 k(t) dt$ and $c = \int_{-\infty}^{+\infty} tk(t)K(t)dt$. For the Epanechnikov kernel, it holds that $\mu_2(k) = 1$ and $c = 0.2875$.

Based on the smoothed c.d.f. $\tilde{F}_{n,h}(x)$ we can consider the quantiles, which we denote by $\tilde{Q}_{p,n,h}$, for each $0 < p < 1$. To simplify the notation, we will mostly omit the dependence of the smoothed quantiles on the sample size and the bandwidth and just denote them by \tilde{Q}_p . To compute these quantiles in practice, the smoothed distribution function is computed in 200 equidistant points in the range of the data $[x_{min}, x_{max}]$. Then an extra grid point $x_{min} - (x_{max} - x_{min})/199$ is added for which $\tilde{F}_{n,h}(x)$ is set to zero, as well as a grid point $x_{max} + (x_{max} - x_{min})/199$ for which we set $\tilde{F}_{n,h}(x) = 1$. The desired quantile \tilde{Q}_p is then obtained by linear interpolation.

The smoothed version of med_n is defined by $\tilde{Q}_{0.5}$. For the skewness measures QS_n and OS_n we replace all the quantiles in their definition by the corresponding smoothed quantiles. The resulting estimators are denoted as $\tilde{Q}_{0.5}$, $\tilde{\text{QS}}_n$ and $\tilde{\text{OS}}_n$. For the computation of the IQR and the medcouple, the smoothed quantiles are computed on a grid of $m = 2n - 1$ equidistant percentages between 0 and 1. These quantiles can be considered as a new artificial sample on which the original IQR and medcouple are computed. More precisely, these smoothed estimators $\widetilde{\text{IQR}}_n = \widetilde{\text{IQR}}(X_n)$ and $\widetilde{\text{MC}}_n = \widetilde{\text{MC}}(X_n)$ are defined as $\widetilde{\text{IQR}}(X_n) = \text{IQR}_m(Y_m)$ and $\widetilde{\text{MC}}(X_n) = \text{MC}_m(Y_m)$ with

$$Y_m = \left\{ \tilde{Q}_{j/2n}; j = 1, 2, \dots, 2n - 1 \right\}.$$

Note that we only consider $m = 2n - 1$ quantiles since simulations have shown that taking more quantiles did not change the MSE significantly and only resulted in an increase in computation time.

Remark 1 An alternative for a smoothed IQR would be the straightforward formula $\tilde{Q}_{0.75} - \tilde{Q}_{0.25}$, but simulation results showed that this estimator has a larger bias than the proposed $\widetilde{\text{IQR}}_n$.

Remark 2 For the medcouple, we could alternatively replace the sample median med_n in (2) and (3) by the smoothed median $\tilde{Q}_{0.5}$ (Van der Veeken, 2010). This estimator, however providing satisfying results in our simulation study, was always outperformed by the smoothed medcouple $\widetilde{\text{MC}}_n$. It is also possible to replace the median of the $g(x_i, x_j)$ values in (2) by their smoothed median. However, this has a very small influence on the MSE of MC_s since the number of $g(x_i, x_j)$ -values is large ($O(n^2)$) and their smoothed median is almost similar to their finite-sample median. Moreover, this would involve another smoothing procedure and an additional choice of the bandwidth.

3 Data-driven bandwidth selection

In this section we propose a data-driven procedure to estimate the bandwidth. As all the quantile-based estimators considered in this paper are robust against outliers we also aim to construct a bandwidth selection method that can cope with possible outliers in the data.

From (4) and (5) we deduce that increasing the bandwidth results in a smaller variance but also in an increase in absolute bias. Hence it is common practice to use the mean integrated squared error (MISE) as a global measure of performance. The MISE is defined as

$$\text{MISE}(h) = E \int_{-\infty}^{+\infty} (\tilde{F}_{n,h}(x) - F(x))^2 dx.$$

An optimal smoothing parameter can then be defined as the value that minimizes this MISE. From (4) and (5) it follows for the Epanechnikov kernel that asymptotically, if $n \rightarrow \infty, h \rightarrow 0$ and $nh \rightarrow \infty$

$$\text{AMISE}(h) = \frac{\int_{-\infty}^{+\infty} F(x)(1 - F(x))dx}{n} - \frac{2hc}{n} + \frac{h^4 R}{4} + o(h^4) \quad (6)$$

where R is the roughness of $f(x)$:

$$R = \int_{-\infty}^{+\infty} (f'(x))^2 dx.$$

Ignoring the last term in (6), the AMISE is then minimized by setting h equal to

$$h_0 = \left(\frac{2c}{R} \right)^{1/3} n^{-1/3}. \quad (7)$$

Since the optimal bandwidth is inverse proportional to the roughness R , it holds that the less rough the distribution is, the larger the optimal bandwidth will be.

The optimal asymptotic MISE is then given by

$$\text{AMISE}(h_0) = \frac{\int_{-\infty}^{+\infty} F(x)(1 - F(x))dx}{n} - \frac{3c^{4/3}}{4^{1/3}n^{4/3}R^{1/3}} \quad (8)$$

which is lower than that of the empirical distribution function, which is equal to the first term of expression (8). The improvement over the empirical distribution function disappears as $n \rightarrow \infty$ at a rate of $n^{-4/3}$. This suggests that smoothing with the optimal bandwidth results in a considerable improvement in AMISE in case of small samples. Moreover the improvement is inverse proportional to the roughness R . This means that smaller gains are expected for rough density functions.

From equation (7) it follows that the optimal bandwidth depends on the unknown roughness R . To estimate R we use that

$$R = - \int_{-\infty}^{+\infty} f^{(2)}(x)f(x)dx = -E(f^{(2)}(X))$$

with $f^{(2)}(x)$ the second derivative of the density function $f(x)$. This implies that we can estimate R as

$$\hat{R} = -\frac{1}{n} \sum_{i=1}^n \tilde{f}^{(2)}(x_i) \quad (9)$$

where $\tilde{f}^{(2)}(x)$ is an appropriate estimate of $f^{(2)}(x)$. In our framework it is quite natural to estimate $f^{(2)}(x)$ based on a kernel density estimate (see also Delaigle and Gijbels (2002)). Since $f(x)$ can be estimated using the Epanechnikov kernel:

$$\tilde{f}(x) = \frac{1}{h_d n} \sum_{i=1}^n k\left(\frac{x - x_i}{h_d}\right)$$

with h_d an optimal bandwidth for density estimation, estimators for the first derivative $f'(x)$ and second derivative $f^{(2)}(x)$ are given by

$$\tilde{f}'(x) = \frac{1}{h_d^2 n} \sum_{i=1}^n k' \left(\frac{x - x_i}{h_d} \right) \quad (10)$$

and

$$\tilde{f}^{(2)}(x) = \frac{1}{h_d^3 n} \sum_{i=1}^n k^{(2)} \left(\frac{x - x_i}{h_d} \right) \quad (11)$$

which can be computed analytically. It is common practice to use a plug-in bandwidth (Silverman, 1986)

$$h_d = 2.34 \min \left(\hat{\sigma}_n, \frac{\text{IQR}_n}{1.349} \right) n^{-1/5} \quad (12)$$

with $\hat{\sigma}_n$ the sample standard deviation.

The use of the IQR_n in (12) makes this bandwidth more robust towards outliers than if we would only use the standard deviation. We propose to use the Q_n estimator (Rousseeuw and Croux, 1993) instead, as it is an even more robust estimator of scale, with a breakdown value of 50% and a better efficiency at the normal model. This Q_n estimator roughly consists of the 25% quantile of all pairwise differences between two data points. Also asymptotic and finite-sample correction factors have been derived in order to make the estimator unbiased at normal samples. We thus use as bandwidth for estimating the density

$$h_d = 2.34 \min (\hat{\sigma}_n, Q_n) n^{-1/5}. \quad (13)$$

Note that in Zhang and Wang (2009), another robust scale estimator is proposed which also considers a quantile of differences between two data points, but only a restricted set of differences are considered. We prefer to use the Q_n estimator, because of its known robustness properties, its high efficiency at the normal model (82% versus 37% for the IQR), its common use in many robust procedures, and its free availability in statistical software such as R and Matlab.

Based on h_d , we then estimate the roughness using (9) and (11) and plug in \hat{R} in (7), which yields

$$h_F = \left(\frac{2c}{\hat{R}} \right)^{1/3} n^{-1/3}. \quad (14)$$

Remark 3 We also investigated whether a cross-validation approach would be appropriate to select the optimal bandwidth (Van der Vaeken, 2010). In particular we studied whether minimizing the cross-validation criterion (Bowman et al. (1998))

$$\frac{1}{n} \sum_{i=1}^n D_{x_i}(h) = \frac{1}{n} \sum_{i=1}^n \int [I(x - x_i) - \tilde{F}_{n,h;-i}(x_i)]^2 dx$$

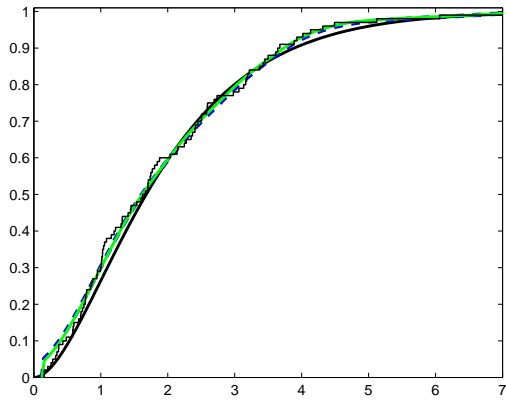
where $\tilde{F}_{n,h;-i}(x)$ is the kernel estimator computed with bandwidth h by leaving out x_i , could be used in this setting. However we found that this approach is computationally much more demanding, and it did not yield better results.

4 Reducing the bias

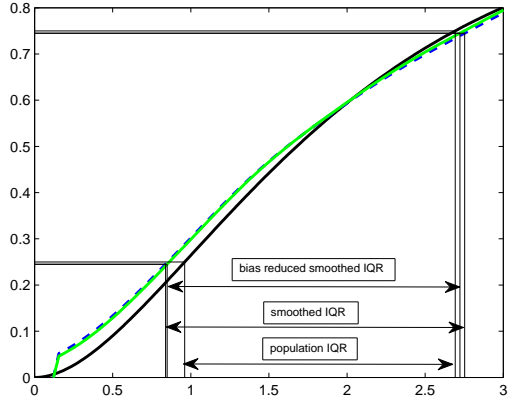
Expression (4) indicates that the bias of $\tilde{F}_{n,h}$ depends on $f'(x)$. For a unimodal distribution, $f'(x)$ is positive for x -values smaller than the mode, and negative for x -values larger than the mode. This suggests that the bias will be positive for the smaller x -values and negative for the larger x -values. This can be seen on Figure 1(a) and its detailed plot Figure 1(b) where the black solid line represents the population distribution function of a Gamma distribution $\Gamma(\alpha, \beta)$ with shape parameter $\alpha = 2$ and scale parameter $\beta = 1$. Note that the density function of a $\Gamma(\alpha, \beta)$ -distribution is given by

$$f(x; \alpha, \beta) = x^{\alpha-1} \frac{e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} \text{ for } x \geq 0 \text{ and } \alpha, \beta > 0$$

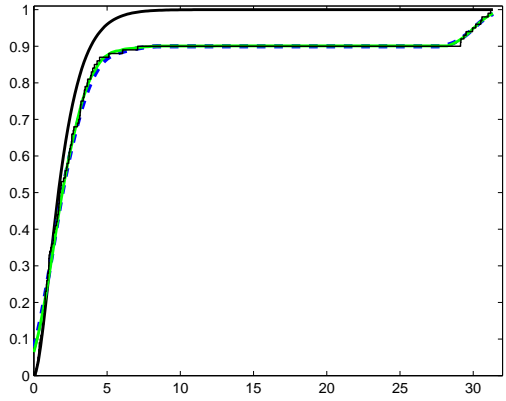
with $\Gamma(x)$ the Gamma-function. The step function in Figure 1(a) is the empirical distribution function based on a random sample of 100 observations, whereas the dashed blue line is the smoothed distribution function $\tilde{F}_{n,h}(x)$ with the bandwidth computed following (14). From Figure 1 we see that the 75th percentile is typically overestimated and the 25th percentile underestimated, so that for the smoothed IQR a double bias effect occurs. The population IQR is indicated by the bottom double arrow. The smoothed IQR is shown by the middle double arrow and is clearly larger. When 10% contamination is added by replacing 10% of the data by outliers coming from a $N(30, 1)$ -distribution, the bias is much larger as can be seen on Figures 1(c) and (d). Also notice that the smoothing procedure still yields an estimated c.d.f. which is close to the empirical c.d.f., but this empirical c.d.f. is quite different from the population c.d.f.



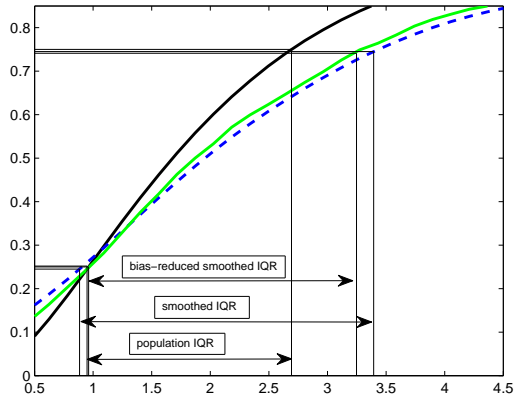
(a)



(b)



(c)



(d)

Figure 1: Population (solid black line), robustly smoothed (dashed blue line) and robustly bias-reduced smoothed (solid green line) $\Gamma(2, 1)$ -c.d.f., based on a sample of 100 observations (a) without outliers; (c) with 10% outliers coming from a $N(30, 1)$ -distribution; (b) Detail of (a); (d) Detail of (c).

To reduce this bias, we reconsider equation (4). Since the bias of $\tilde{F}_{n,h}(x)$ equals $\frac{1}{2}h^2 f'(x)\mu_2(k) + o(h^2)$, we consider

$$\tilde{\tilde{F}}_{n,h_F}(x) = \tilde{F}_{n,h_F}(x) - \frac{1}{2}h_F^2 \tilde{f}'(x)\mu_2(k)$$

with $\tilde{f}'(x)$ computed as in (10), and $\tilde{F}_{n,h_F}(x)$ estimated as described in Section 3. Reducing the bias however implies subtracting a possibly positive term from the estimated c.d.f., so $\tilde{\tilde{F}}_{n,h_F}(x)$ is not guaranteed to be nondecreasing for all x . Hence, in those intervals (defined by the grid points in which the c.d.f. is computed) where $\tilde{\tilde{F}}_{n,h_F}(x)$ is decreasing,

we use linear interpolation between the two closest endpoints $[a, b]$ for which $\tilde{F}(b) \geq \tilde{F}(a)$ to ensure a nondecreasing c.d.f. estimate. This yields our robust bias-reduced c.d.f. estimator (still denoted as $\tilde{F}_{n,h_F}(x)$), which in Figure 1 is indicated by a solid green line. We see that it indeed reduces the bias of the resulting IQR estimator, especially at contaminated samples. This bias reduction was also beneficial for the other quantile estimators under study, so in the following we only report the results for this smoothing procedure.

It is important to notice that all the steps in the procedure to compute $\tilde{F}_{n,h_F}(x)$ ensure affine equivariance of the estimated quantiles, i.e. for every data set $X_n = \{x_1, \dots, x_n\}$, every $c > 0$ and $d \in \mathbb{R}$ it holds that $\hat{\theta}(cX_n + d) = c\hat{\theta}(X_n) + d$ with $\hat{\theta}$ any estimated quantile. Consequently it holds that the smoothed location and scale estimators $\tilde{Q}_{0.5}$ and $\widetilde{\text{IQR}}_n$ are affine equivariant, and the smoothed skewness estimators $\widetilde{\text{QS}}_n$, $\widetilde{\text{OS}}_n$ and $\widetilde{\text{MC}}_n$ are affine invariant (just as their empirical versions).

5 Simulation study

In order to illustrate the reduction in variance and MSE of the different quantile-based estimators, we performed a simulation study on different Gamma distributions. In particular we considered random samples of size $n = 100$ from Gamma distributions with scale parameter $\beta = 1$, and shape parameters $\alpha = 2, 5, 10$. Note that increasing the shape parameter makes the distribution more symmetric. We also considered contaminated samples. Data sets with ‘left’ contamination have 5% or 10% outliers generated from a $N(-5, 1)$ -distribution, whereas ‘right’ contamination is generated from a $N(30, 1)$ -distribution. We will only report the results in case of 5% outliers, since the results in case of 10% contamination are very comparable.

All simulations are repeated 500 times and the average estimated bias, variance and mean squared error of the different estimators are tabulated. We consider both the empirical estimators med_n , IQR_n , QS_n , OS_n and MC_n (the latter being part of the MATLAB toolbox LIBRA (Verboven and Hubert, 2005) and the library `robustbase` in R) as well as their smoothed variants, using the robust data-driven bandwidth described in Section 3 and by reducing the bias of the smoothed c.d.f. as described in Section 4. The population values of most estimators are difficult to compute analytically. Therefore, they are determined as the average over 100 random samples of size 50000 in case of the medcouple

distribution	contamination	estimator	bias	variance	MSE
$\Gamma(2, 1)$	no	med_n	0.0133	0.0243	0.0244
		$\tilde{Q}_{0.5}$	0.0379	0.0199	0.0213
	5% left	med_n	-0.0689	0.0238	0.0285
		$\tilde{Q}_{0.5}$	-0.0323	0.0193	0.0203
	5% right	med_n	0.1026	0.0272	0.0376
		$\tilde{Q}_{0.5}$	0.1736	0.0210	0.0511
$\Gamma(5, 1)$	no	med_n	0.0068	0.0719	0.0718
		$\tilde{Q}_{0.5}$	0.0326	0.0550	0.0560
	5% left	med_n	-0.1329	0.0701	0.0876
		$\tilde{Q}_{0.5}$	-0.1015	0.0536	0.0638
	5% right	med_n	0.1456	0.0763	0.0974
		$\tilde{Q}_{0.5}$	0.2095	0.0577	0.1015
$\Gamma(10, 1)$	no	med_n	0.0022	0.1628	0.1625
		$\tilde{Q}_{0.5}$	0.0196	0.1389	0.1390
	5% left	med_n	-0.1969	0.1605	0.1990
		$\tilde{Q}_{0.5}$	-0.1804	0.1388	0.1711
	5% right	med_n	0.1998	0.1788	0.2184
		$\tilde{Q}_{0.5}$	0.2372	0.1487	0.2047

Table 1: Bias, variance and MSE of the median and the smoothed median at different Gamma distributions.

and numerically calculated using a Newton-Raphson approximation with the standard MATLAB function `gaminv` for all other estimators.

5.1 Location and scale estimators

We first report the results for the median in Table 1 and for the IQR in Table 2. To simplify notations, we denote the estimated median and IQR based on $\tilde{F}_{n,h_F}(x)$ again as $\tilde{Q}_{0.5}$ resp. $\widetilde{\text{IQR}}_n$. From Table 1 we can see that the smoothing procedure for the median slightly reduces the variance and the MSE compared to the empirical median in almost all situations. Also for the IQR the smoothed estimator reduces the variance and MSE as can be seen from Table 2, although rather slightly.

Overall we can conclude that for the median and the IQR the smoothing is not really

distribution	contamination	estimator	bias	variance	MSE
$\Gamma(2, 1)$	no	IQR_n	-0.0101	0.0484	0.0484
		$\widetilde{\text{IQR}}_n$	-0.0019	0.0338	0.0338
	5% left	IQR_n	0.0294	0.0461	0.0469
		$\widetilde{\text{IQR}}_n$	0.0793	0.0340	0.0402
	5% right	IQR_n	0.1822	0.0601	0.0932
		$\widetilde{\text{IQR}}_n$	0.2283	0.0367	0.0888
$\Gamma(5, 1)$	no	IQR_n	-0.0006	0.1262	0.1260
		$\widetilde{\text{IQR}}_n$	0.0253	0.0920	0.0924
	5% left	IQR_n	0.1002	0.1317	0.1415
		$\widetilde{\text{IQR}}_n$	0.1969	0.0908	0.1294
	5% right	IQR_n	0.2392	0.1467	0.2037
		$\widetilde{\text{IQR}}_n$	0.3623	0.0963	0.2274
$\Gamma(10, 1)$	no	IQR_n	-0.0049	0.2580	0.2575
		$\widetilde{\text{IQR}}_n$	0.0368	0.1965	0.1965
	5% left	IQR_n	0.1903	0.2663	0.3020
		$\widetilde{\text{IQR}}_n$	0.1711	0.2602	0.2675
	5% right	IQR_n	0.3299	0.2826	0.3909
		$\widetilde{\text{IQR}}_n$	0.4088	0.2168	0.3835

Table 2: Bias, variance and MSE of the IQR and the smoothed IQR at different Gamma distributions.

harmful, but neither extremely helpful for reducing the MSE.

5.2 Skewness estimators

For the estimators of skewness the situation is different. The simulation results in Tables 3, 4 and 5 show that a considerable reduction in variance and MSE is achieved by the smoothed skewness estimators. Only in one specific situation ($\Gamma(2, 1)$ with 5% left contamination) the MSE of the $\widetilde{\text{OS}}_n$ is slightly larger compared to OS_n .

We also show the effect of smoothing the medcouple on smaller and larger sample sizes. The results for a $\Gamma(5, 1)$ -distribution are shown in the boxplots in Figure 2. The left (blue) boxplot in Figure 2(a) shows the MC_n estimates for 500 data sets, and the right (green) boxplot its smoothed counterpart for sample sizes 25, 50, 100, 250 and 500.

distribution	contamination	estimator	bias	variance	MSE
$\Gamma(2, 1)$	no	QS_n	-0.0162	0.0182	0.0185
		\widetilde{QS}_n	-0.0283	0.0044	0.0052
	5% left	QS_n	-0.0272	0.0185	0.0193
		\widetilde{QS}_n	-0.0545	0.0028	0.0057
	5% right	QS_n	0.0128	0.0179	0.0181
		\widetilde{QS}_n	-0.0393	0.0013	0.0028
$\Gamma(5, 1)$	no	QS_n	-0.0064	0.0189	0.0189
		\widetilde{QS}_n	-0.0164	0.0047	0.0050
	5% left	QS_n	-0.0260	0.0186	0.0193
		\widetilde{QS}_n	-0.0413	0.0031	0.0048
	5% right	QS_n	0.0223	0.0180	0.0185
		\widetilde{QS}_n	-0.0006	0.0019	0.0019
$\Gamma(10, 1)$	no	QS_n	-0.0009	0.0176	0.0176
		\widetilde{QS}_n	-0.0045	0.0073	0.0073
	5% left	QS_n	-0.0242	0.0180	0.0185
		\widetilde{QS}_n	-0.0286	0.0062	0.0070
	5% right	QS_n	0.0325	0.0174	0.0184
		\widetilde{QS}_n	0.0234	0.0058	0.0063

Table 3: Bias, variance and MSE of the Quantile Skewness and the smoothed Quantile Skewness at different Gamma distributions.

Also the population value for the medcouple (0.136) is shown by the red horizontal line. In all cases, it can be noticed that the variability of the smoothed medcouple decreases a lot compared to the empirical ones, whereas the median stays close to the population value (although with a small negative bias), which makes the smoothed estimates more reliable. In Figure 2(b) 5% of the data has been replaced with data coming from a $N(30, 1)$ -distribution. Here we see that the variability and the bias decreases, which is in line with the numerical output from Table 5.

Table 6 shows the average computation times (in seconds) over 500 simulations of the empirical and smoothed medcouple for different values of n . Although the computation time for \widetilde{MC}_n is considerably larger than for MC_n , it is still very reasonable for all n .

distribution	contamination	estimator	bias	variance	MSE
$\Gamma(2, 1)$	no	OS_n	-0.0353	0.0105	0.0117
		\widetilde{OS}_n	-0.0421	0.0053	0.0071
	5% left	OS_n	-0.0604	0.0106	0.0142
		\widetilde{OS}_n	-0.1051	0.0043	0.0153
	5% right	OS_n	0.0310	0.0103	0.0113
		\widetilde{OS}_n	-0.0303	0.0038	0.0047
$\Gamma(5, 1)$	no	OS_n	-0.0333	0.0115	0.0126
		\widetilde{OS}_n	-0.0253	0.0050	0.0056
	5% left	OS_n	-0.0789	0.0116	0.0178
		\widetilde{QS}_n	-0.0817	0.0041	0.0108
	5% right	OS_n	0.0387	0.0103	0.0118
		\widetilde{OS}_n	0.0216	0.0035	0.0040
$\Gamma(10, 1)$	no	OS_n	-0.0297	0.0107	0.0116
		\widetilde{OS}_n	-0.0076	0.0065	0.0065
	5% left	OS_n	-0.0804	0.0105	0.0169
		\widetilde{OS}_n	-0.0590	0.0060	0.0095
	5% right	OS_n	0.0373	0.0105	0.0119
		\widetilde{OS}_n	0.0576	0.0063	0.0096

Table 4: Bias, variance and MSE of the Octile Skewness and the smoothed Octile Skewness at different Gamma distributions.

6 Properties of the smoothed medcouple

The medcouple is very useful for the analysis of skewed data. In this Section we focus on a few properties of the smoothed medcouple \widetilde{MC}_n and compare it to the empirical medcouple MC_n . First we discuss how often the estimators return a positive value when estimating the medcouple for a positively skewed distribution, and next we compare the outlier detection capacity of the adjusted boxplot (Hubert and Vandervieren, 2008) when the empirical and the smoothed medcouple and quantile estimators are used.

distribution	contamination	estimator	bias	variance	MSE
$\Gamma(2, 1)$	no	MC_n	-0.0102	0.0119	0.0120
		\widetilde{MC}_n	-0.0353	0.0037	0.0049
	5% left	MC_n	-0.0599	0.0133	0.0169
		\widetilde{MC}_n	-0.1045	0.0024	0.0133
	5% right	MC_n	0.0335	0.0119	0.0130
		\widetilde{MC}_n	-0.0320	0.0017	0.0027
$\Gamma(5, 1)$	no	MC_n	-0.0013	0.0126	0.0126
		\widetilde{MC}_n	-0.0196	0.0036	0.0039
	5% left	MC_n	-0.0549	0.0134	0.0164
		\widetilde{MC}_n	-0.0774	0.0026	0.0086
	5% right	MC_n	0.0513	0.0122	0.0148
		\widetilde{MC}_n	0.0121	0.0019	0.0020
$\Gamma(10, 1)$	no	MC_n	0.0040	0.0119	0.0119
		\widetilde{MC}_n	-0.0057	0.0050	0.0050
	5% left	MC_n	-0.0511	0.0116	0.0142
		\widetilde{MC}_n	-0.0596	0.0044	0.0079
	5% right	MC_n	0.0602	0.0127	0.0163
		\widetilde{MC}_n	0.0417	0.0043	0.0060

Table 5: Bias, variance and MSE of the medcouple and the smoothed medcouple at different Gamma distributions.

estimator	$n = 25$	$n = 50$	$n = 100$	$n = 250$	$n = 500$
MC_n	0.0002	0.0003	0.0006	0.0008	0.0017
\widetilde{MC}_n	0.0089	0.0105	0.0128	0.0204	0.0321

Table 6: Average computation times (in seconds) for the empirical and smoothed medcouple for a $\Gamma(5, 1)$ distribution.

6.1 Positive skewness

Since the medcouple is a measure of skewness, one would expect it to be positive for positively skewed distributions such as Gamma distributions. For a $\Gamma(\alpha, \beta)$ -distribution, the third standardized moment is given by $\frac{2}{\sqrt{\alpha}}$ (and hence always positive). So the smaller α is, the more skewed the distribution, and the more estimates of the medcouple we expect

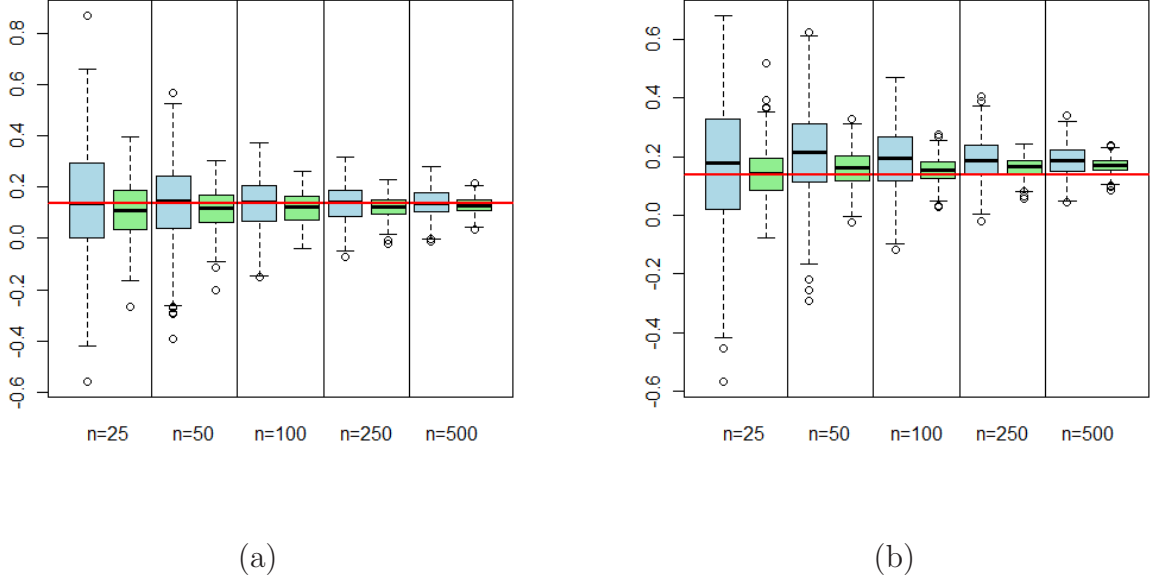


Figure 2: Boxplots of the empirical (left, blue) and smoothed (right, green) medcouple for different sample sizes, and the population medcouple (red horizontal line) for a $\Gamma(5, 1)$ distribution (a) without outliers, and (b) with 5% right contamination.

to be positive. We simulated 500 data sets of size 100 coming from a $\Gamma(2, 1)$, $\Gamma(5, 1)$ and $\Gamma(10, 1)$ -distribution with and without 5% left and right contamination as in Section 5. In Table 7 we report the proportion of positive values attained by the empirical MC_n and the smoothed \widetilde{MC}_n at these 500 data sets.

For the very skewed $\Gamma(2, 1)$ -distribution, both MC_n and \widetilde{MC}_n have a high percentage of positive estimates, which drops a bit when 5% left contamination is added for MC_n , but hardly for \widetilde{MC}_n . Note that the population medcouple is equal to 0.223. It can be noticed from Table 7 that \widetilde{MC}_n performs better since it always provides more positive estimates than MC .

The $\Gamma(5, 1)$ -distribution is a little more symmetric, but still fairly skewed with a population medcouple of 0.136. As expected we see from Table 7 that the percentage of positive estimates for the medcouple is lower than for the $\Gamma(2, 1)$ -distribution. It is mainly sensitive to left contamination, but in all situations considered, \widetilde{MC}_n takes more positive values than MC_n .

The $\Gamma(10, 1)$ -distribution is almost symmetric, with a population medcouple of 0.095. In this case, both estimators yield a lower percentage of positive estimates, again mainly

distribution	contamination	proportion of positive estimates	
		MC_n	\widetilde{MC}_n
$\Gamma(2, 1)$	no	0.976	1
	5% left	0.916	0.998
	5% right	0.982	1
$\Gamma(5, 1)$	no	0.898	0.974
	5% left	0.782	0.882
	5% right	0.954	1
$\Gamma(10, 1)$	no	0.804	0.886
	5% left	0.662	0.706
	5% right	0.914	0.980

Table 7: Proportion of positive estimates for the medcouple and the smoothed medcouple at different Gamma distributions.

when 5% left contamination is added. Also in this situation the sign of the medcouple is more often estimated correctly by \widetilde{MC}_n than by MC_n .

6.2 Outlier detection

One of our motivations to study smoothed variants of the medcouple is to increase the ability of detecting outliers at skewed distributions. In Hubert and Vandervieren (2008); Hubert and Van der Veecken (2008, 2010) it was shown how the medcouple can be used to detect outliers in univariate and multivariate data, and how this also improves the classification of skewed multivariate data. Here, we focus on the detection of univariate skewed data. Outliers can be flagged as those observations that exceed the whiskers of the adjusted boxplot (Hubert and Vandervieren, 2008). When $MC_n \geq 0$ they are defined as

$$\hat{Q}_{0.25} - 1.5 \exp(-4MC_n)IQR_n \text{ and } \hat{Q}_{0.75} + 1.5 \exp(3MC_n)IQR_n. \quad (15)$$

For left-skewed distributions, the whiskers are analogously given by

$$\hat{Q}_{0.25} - 1.5 \exp(-3MC_n)IQR_n \text{ and } \hat{Q}_{0.75} + 1.5 \exp(4MC_n)IQR_n. \quad (16)$$

To illustrate that a more accurate outlier detection procedure can be achieved by using the smoothed estimators, we consider 500 samples of 45 observations from a $\Gamma(2, 1)$, a

$\Gamma(5, 1)$ and a $\Gamma(10, 1)$ -distribution to which 5 outliers are added. This setup thus corresponds to a quite small data set with 10% contamination. The outliers are sampled from a $N(\mu, 1)$ -distribution, with μ ranging in 21 steps from μ_0 to $\mu_0 + 20$. For the shape $\alpha = 2$ we take $\mu_0 = 15$, for $\alpha = 5$ we use $\mu_0 = 25$, whereas for $\alpha = 10$ we set $\mu_0 = 35$. Doing so, the contamination is roughly placed at the same distance to the center of the data for the three distributions.

For each sample we compute the whiskers of the adjusted boxplot as in (15) and (16). Observations that exceed the whiskers are flagged as outliers. Next, we do the same by replacing all estimates $\hat{Q}_{0.25}$, $\hat{Q}_{0.75}$, IQR_n and MC_n by their smoothed variants. In Figure 3 we show the sensitivity and specificity of both outlier detection rules. The sensitivity is defined as the average percentage of observations that are correctly flagged as outliers. The specificity indicates the average percentage of regular observations that are correctly classified as such.

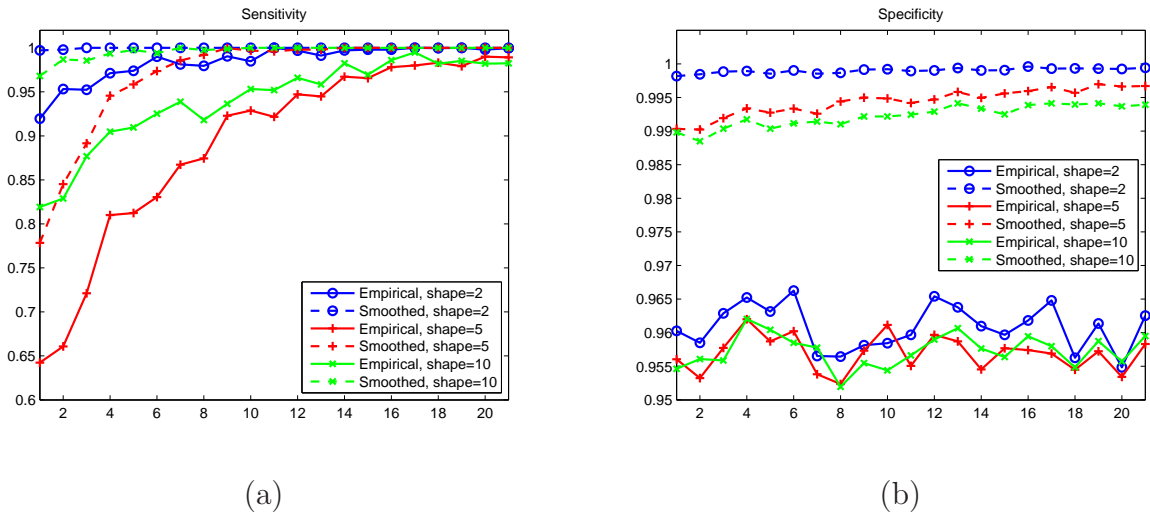


Figure 3: (a) Sensitivity and (b) specificity to outlier detection based on the adjusted boxplot computed with the empirical quantiles and the empirical medcouple, and with the smoothed quantiles and smoothed medcouple.

We see that smoothing the distribution function considerably increases both the sensitivity and the specificity at all distributions. Comparable results were obtained at different sample sizes and amounts of contamination.

7 Real data example: EU international trade data

In the Joint Research Centre of the European Commission, EU international trade data are analyzed for different purposes, in particular for detecting Customs frauds that are relevant for the budget of the EU. We do not have the raw data at our disposal, but some intermediate data that are used to robustly estimate a sort of import price for each Member State. Such fair prices are used for several purposes, and for one purpose fair prices which are abnormal need to be highlighted.

In a first example we look at import prices in 23 different EU countries. The histogram in Figure 4 shows that the data are right skewed with possibly two outliers on the right.

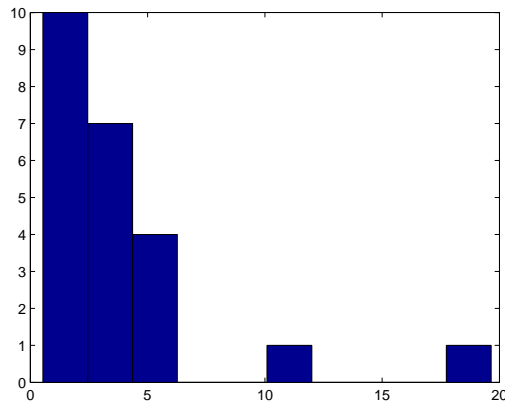


Figure 4: Histogram of import prices into 23 EU countries for the first example.

We first consider the adjusted boxplot for which empirical quantiles and the empirical medcouple (with a value of 0.2909) is used. The boxplot in Figure 5(a) shows that both the smallest and largest observation surpass the whiskers, indicating them as possible outliers. The boxplot based on the smoothed medcouple (with smaller value 0.1226) and smoothed quantiles (Figure 5(b)) however shows no data beyond the left whisker, and indicates the two largest observations as possible outliers, which is more consistent with the histogram of the data.

In a second example, we analyze a different set of fair import prices in 23 EU countries. The histogram in Figure 6 may suggest that the distribution of the import prices is slightly right skewed with two outliers on the left, and also the adjusted boxplot in Figure 7(a) shows two observations below the lower whisker (based on an estimated medcouple of 0.1083). The smoothed adjusted boxplot shows a more symmetric distribution, without

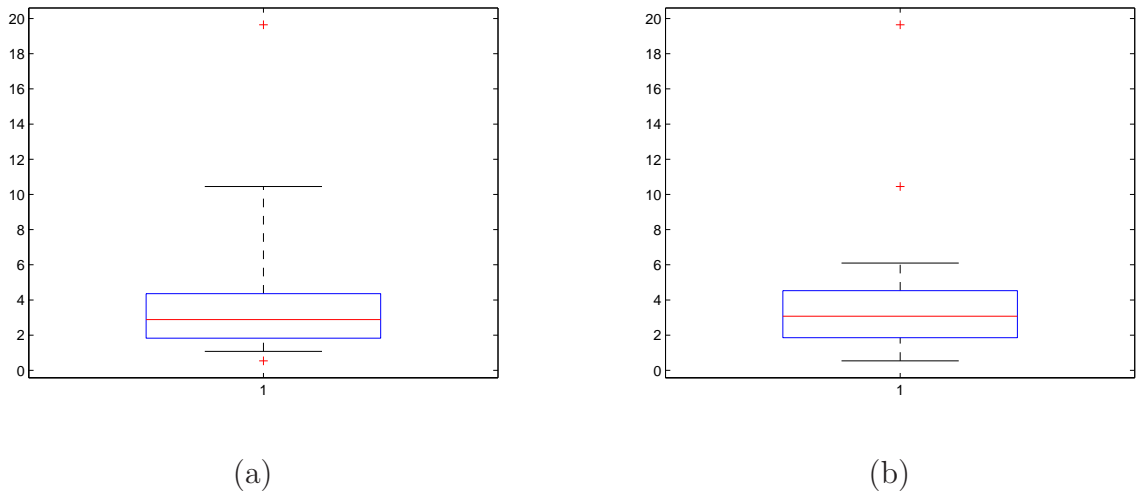


Figure 5: Adjusted boxplot for the import prices of the first example based on (a) empirical quantiles and medcouple; (b) smoothed quantiles and smoothed medcouple.

any outliers and a smoothed medcouple of 0.0846. Although we do not know which representation of the data is the most accurate, we see that the smoothed version yields a more conservative (i.e. a more symmetric) result. This seems plausible as the 'outliers' are not very much separated from the other data points.

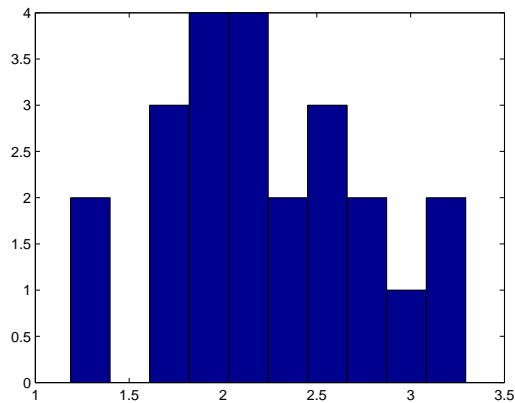


Figure 6: Histogram of import prices into 23 EU countries for the second example.

8 Conclusion

In this work, we presented a robust method to reduce the MSE of quantile-based estimators like the median, the IQR, the quantile and octile skewness and the medcouple,

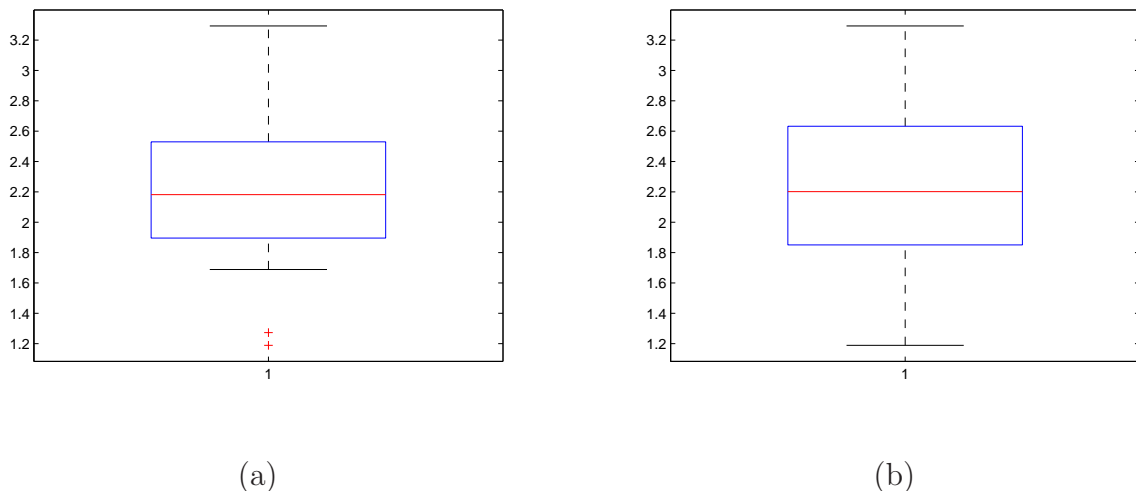


Figure 7: Adjusted boxplot for the import prices for the second example based on (a) empirical quantiles and medcouple; (b) smoothed quantiles and smoothed medcouple.

by robustly smoothing the empirical c.d.f. and reducing the bias of this smoothed c.d.f. The proposed procedure yields affine equivariant location and scale estimators, and affine invariant skewness estimators.

Simulation results show that the estimators based on the smoothed c.d.f. indeed show a reduced MSE compared to the empirical estimates. Also the variance of the smoothed estimators is smaller than their empirical counterparts, and in some cases they show a smaller bias as well.

In particular we focussed on the medcouple. Smoothing the medcouple decreases its variance, especially for small sample sizes, without seriously increasing the bias. In addition we can conclude that the smoothed medcouple returns more positive estimates in case of a positively skewed distribution than the empirical medcouple. Used in combination with the adjusted boxplot, the smoothed estimators yield a higher sensitivity to outliers and a better specificity. We also compared the smoothed adjusted boxplot to the original one using two real data examples, and noticed that the smoothed adjusted boxplot seems to represent the data better.

Even though smoothing somewhat increases the computation time, we feel that the improvement in MSE is worth the effort, especially at small sample sizes where the computational complexity is less of an issue. The programs for computing the smoothed estimators, as well as the smoothed adjusted boxplot, will be made available in LIBRA (Verboven

and Hubert, 2005).

Acknowledgements

We acknowledge the financial support by the GOA/07/04-project of the Research Fund K.U.Leuven. We are grateful to Pavel Čížek for his suggestion to decrease the variability of the medcouple by a smoothing procedure. This advice has been the start of our research. We also would like to thank Peter Rousseeuw for useful comments on an earlier draft, and Domenico Perrotta who kindly shared the import price data with us.

References

- A. Azzalini. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68:326–328, 1981.
- A.W. Bowman, P. Hall, and T. Prvan. Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85:799–808, 1998.
- G. Brys, M. Hubert, and A. Struyf. A comparison of some new measures of skewness. In R. Dutter, P. Filzmoser, U. Gather, and P.J. Rousseeuw, editors, *Developments in Robust Statistics: International Conference on Robust Statistics 2001*, volume 114, pages 98–113. Physika Verlag, Heidelberg, 2003.
- G. Brys, M. Hubert, and A. Struyf. A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13:996–1017, 2004.
- P. Čížek, J. Tamine, and W. Härdle. Smoothed L-estimation of regression function. *Computational Statistics and Data Analysis*, 52:5154–5162, 2008.
- A. Delaigle and I. Gijbels. Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 64(4):869–886, 2002.
- L.T. Fernholz. Reducing the variance by smoothing. *Journal of Statistical Planning and Inference*, 57(1):29–38, 1997. Robust statistics and data analysis, I.

- M. Hubert and S. Van der Veeken. Robust classification for skewed data. *Advances in Data Analysis and Classification*, 4:239–254, 2010.
- M. Hubert and S. Van der Veeken. Outlier detection for skewed data. *Journal of Chemometrics*, 22:235–246, 2008.
- M. Hubert and E. Vandervieren. An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52(12):5186–5201, 2008.
- E.A. Nadaraya. Some new estimates for distribution functions. *Theory of Probability and its Applications*, 9:497–500, 1964.
- P.J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283, 1993.
- B.W. Silverman. *Density Estimation For Statistics and Data Analysis*. Chapman and Hall, London, 1986.
- S. Van der Veeken. *Robust and nonparametric methods for skewed data*. PhD thesis, Katholieke Universiteit Leuven, 2010.
- S. Verboven and M. Hubert. LIBRA: a Matlab library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75:127–136, 2005.
- J. Zhang and X. Wang. Robust normal reference bandwidth for kernel density estimation. *Statistica Neerlandica*, 63(1):13–23, 2009.