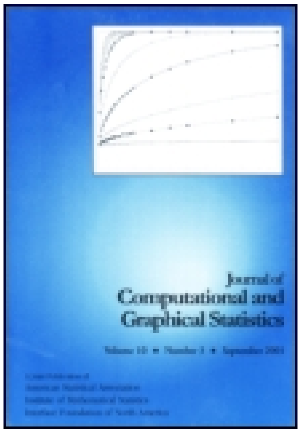

Journal of Computational and Graphical Statistics
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/ucgs20

A Robust Measure of Skewness

G. Brys is Ph.D. Student, Faculty of Applied Economics, University of Antwerp (UA), Prinsstraat 13, B-2000 Antwerp, Belgium, . M. Hubert is Associate Professor, Department of Mathematics, KU Leuven, W. De Croylaan 54, B-3001 Leuven, Belgium . A. Struyf is Postdoctoral Fellow, Department of Mathematics and Computer Science, University of Antwerp (UA), Middelheimlaan 1, B-2020 Antwerp, Belgium .

Version of record first published: 01 Jan 2012

To cite this article: (2004): A Robust Measure of Skewness, Journal of Computational and Graphical Statistics, 13:4, 996-1017

To link to this article: http://dx.doi.org/10.1198/106186004X12632

A Robust Measure of Skewness

G. Brys, M. Hubert, and A. Struyf

The asymmetry of a univariate continuous distribution is commonly measured by the classical skewness coefficient. Because this estimator is based on the first three moments of the dataset, it is strongly affected by the presence of one or more outliers. This article investigates the medcouple, a robust alternative to the classical skewness coefficient. We show that it has a 25% breakdown value and a bounded influence function. We present a fast algorithm for its computation, and investigate its finite-sample behavior through simulated and real datasets.

Key Words: Asymmetry; Influence function; Medcouple.

1. INTRODUCTION

The shape and asymmetry of a distribution can be measured by its *skewness*. A symmetric distribution has zero skewness, an asymmetric distribution with the largest tail to the right has positive skewness, and a distribution with a longer left tail has negative skewness. The classical skewness coefficient b_1 of a univariate dataset $X_n = \{x_1, x_2, \ldots, x_n\}$ sampled from a continuous distribution is defined as

$$b_1(X_n) = \frac{m_3(X_n)}{m_2(X_n)^{3/2}},$$

where m_3 and m_2 denote the third and second empirical moments of the data. However, b_1 is very sensitive to outliers in the data. One single outlier in the left tail of a symmetric or right-tailed sample can cause b_1 to become negative, whereas an outlier in the right tail of such a sample can unduly increase the classical skewness coefficient, making it hard to interpret.

As an example we consider the systolic blood pressure (SBP) dataset (Kleinbaum, Kupper, and Muller 1998) which contains the systolic blood pressure of 30 patients. From the

G. Brys is Ph.D. Student, Faculty of Applied Economics, University of Antwerp (UA), Prinsstraat 13, B-2000 Antwerp, Belgium, (E-mail: Guy.Brys@ua.ac.be). M. Hubert is Associate Professor, Department of Mathematics, KU Leuven, W. De Croylaan 54, B-3001 Leuven, Belgium (E-mail: Mia.Hubert@wis.kuleuven.ac.be). A. Struyf is Postdoctoral Fellow, Department of Mathematics and Computer Science, University of Antwerp (UA), Middelheimlaan 1, B-2020 Antwerp, Belgium (E-mail: Anja.Struyf@ua.ac.be).

Journal of Computational and Graphical Statistics, Volume 13, Number 4, Pages 996–1017
DOI: 10.1198/106186004X12632

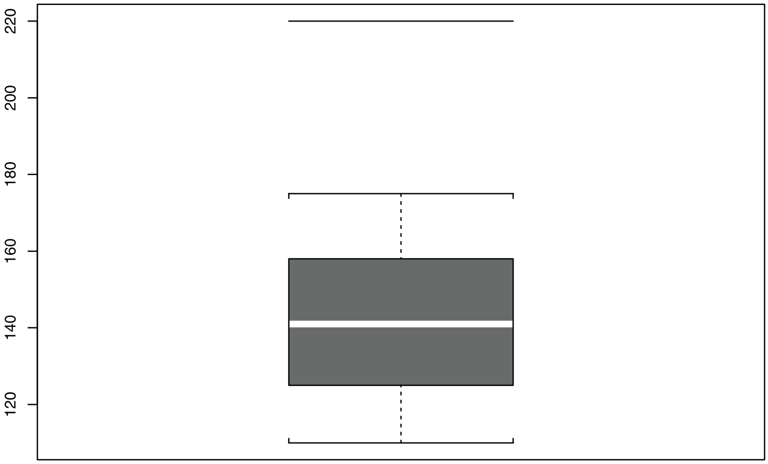

Figure 1. Boxplot of the systolic blood pressure dataset.

boxplot in Figure 1 we see that the observations are sampled from a symmetric distribution and that there is also one clear outlier. This single observation has a huge impact on the classical skewness measure b_1. At the complete dataset, $b_1 = 1.36$, but if we remove the outlier, b_1 drops to .20.

Brys, Hubert, and Struyf (2003) introduced and empirically compared several new measures of skewness that are less sensitive to outlying values. Of the proposed measures the medcouple (MC) arose as the overall winner taking into account its performance at uncontaminated datasets, its robustness at contaminated samples, and its computation time. Let us, for example, reconsider the SBP dataset. Here, the medcouple is exactly zero, at both the full and the cleaned dataset, which illustrates nicely its robustness towards outliers. Note that the repeated medtriple (RMT), also introduced by Brys et al. (2003), appeared to be a good alternative to the medcouple, and it has a breakdown value of 50%. But its computational load is very high and it showed a higher sensitivity to outliers at symmetric distributions.

This article studies the medcouple in more detail. Section 2 recalls the definition of the medcouple and verifies that it satisfies natural requirements of a skewness measure. Section 3 shows the robustness of the medcouple by computing its breakdown value and influence function. Section 4 provides a computationally fast algorithm. Section 5 uses the algorithm to compare the performance and the robustness of the medcouple with other robust skewness measures. Section 6 contains examples and Section 7 concludes. Finally, the Appendix contains all the proofs.

2. THE MEDCOUPLE

We assume that we have independently sampled n observations $X_n = \{x_1, x_2, \ldots, x_n\}$ from a continuous univariate distribution F. For notational convenience, we also assume that X_n has been sorted such that $x_1 \leq x_2 \leq \cdots \leq x_n$. Let m_n denote the median of X_n, defined as usual as

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

Brys et al. (2003) introduced the *medcouple* (MC$_n$) as

$$\mathrm{MC}_n = \underset{x_i \leq m_n \leq x_j}{\mathrm{med}} \; h(x_i, x_j), \tag{2.1}$$

where for all $x_i \neq x_j$ the kernel function h is given by

$$h(x_i, x_j) = \frac{(x_j - m_n) - (m_n - x_i)}{x_j - x_i}. \tag{2.2}$$

For the special case $x_i = x_j = m_n$, we define the kernel as follows. Let $m_1 < \cdots < m_k$ denote the indices of the observations that are tied to the median m_n; that is, $x_{m_l} = m_n$ for all $l = 1, \ldots, k$. Then

$$h(x_{m_i}, x_{m_j}) = \begin{cases} -1 & \text{if} \quad i + j - 1 < k \\ 0 & \text{if} \quad i + j - 1 = k \\ +1 & \text{if} \quad i + j - 1 > k. \end{cases} \tag{2.3}$$

Because of the denominator in (2.2) it is clear that $h(x_i, x_j)$, and hence MC$_n$, always lies between -1 and 1. The kernel (2.2) measures the (standardized) difference between the distances of x_j and x_i to the median. It is positive if x_j lies further from the median than x_i, and negative if x_i does. A zero value is attained at the symmetric case where $x_j - m_n = m_n - x_i$. When the median m_n coincides with one single data point, $h(m_n, x_j) = +1$ for all $x_j > m_n$ which expresses the fact that x_j lies infinitely farther away from the median than m_n does. Analogously, $h(x_i, m_n) = -1$ for all $x_i < m_n$. But because the number of data points which are larger than the median in this case equals the number of data points smaller than the median, we have as many $+1$ as -1, so the medcouple is not influenced by these extreme values. When several data points collapse with the median, it can happen that we have, for example, more data points that are strictly larger than the median than data points that are strictly smaller than the median, hence we will include more positive values $+1$ than negative values -1. Also notice that the number of zeros added from (2.3) equals the number of data values tied with the median. This attracts the medcouple toward zero which corresponds to the intuition that many points equal to the median decrease the skewness of a distribution. The first and third equations in (2.3) are somewhat superfluous but are added to avoid undefined kernels and to simplify the implementation of the algorithm described in Section 4.

Note that the medcouple belongs to the class of incomplete generalized L-statistics (Hössjer 1996) because the kernel function h in (2.1) is not applied to all couples (x_i, x_j) from X_n, but only to those for which $x_i \leq m_n$ and $x_j \geq m_n$.

We can also consider the functional form of the medcouple, defined at any continuous distribution F. We will refer to it as MC(F) or MC$_F$. Let $m_F = F^{-1}(.5)$ be the median of F, then the definition of MC$_F$ follows in a straightforward way from (2.1)

$$\mathrm{MC}_F = \operatorname*{med}_{x_1 \le m_F \le x_2} h(x_1, x_2), \tag{2.4}$$

with x_1 and x_2 being sampled from F. The kernel h in (2.4) is the same as in (2.2) if we replace the finite-sample median m_n by m_F. Let I be the indicator function, then with

$$H_F(u) = 4 \int_{m_F}^{+\infty} \int_{-\infty}^{m_F} I\left(h(x_1, x_2) \le u\right) dF(x_1) dF(x_2) \tag{2.5}$$

we obtain the shorter formulation

$$\mathrm{MC}_F = H_F^{-1}(.5). \tag{2.6}$$

Note that the domain of H_F is $[-1, 1]$ and that the conditions

$$h(x_1, x_2) \le u, \qquad x_1 \le m_F, \qquad x_2 \ge m_F$$

are equivalent to $x_1 \le \frac{x_2(u-1)+2m_F}{u+1}$ and $x_2 \ge m_F$. Therefore (2.5) can be simplified to

$$H_F(u) = 4 \int_{m_F}^{+\infty} F\left(\frac{x_2(u-1) + 2m_F}{u+1}\right) dF(x_2).$$

Note that MC$_n$ can be seen as an estimator of MC(F).

Besides its robustness, which we will show in Section 3, the medcouple has another attractive property which the classical skewness measure b_1 lacks: because it is based only on ranks, it can also be computed at distributions without finite moments.

The following properties show that the functional medcouple possesses natural requirements of a skewness measure, as defined by van Zwet (1964) and Oja (1981). Let the random variable X have a continuous distribution F_X.

Property 1. *MC is location and scale invariant, that is,*

$$MC(F_{aX+b}) = MC(F_X)$$

for any $a > 0$ and $b \in \mathbb{R}$.

Property 2. *If we invert a distribution, the medcouple is inverted as well:*

$$MC(F_{-X}) = -MC(F_X).$$

Property 3. *If F is symmetric, then $MC(F) = 0$.*

Properties 1 and 2 follow immediately from the definitions, and imply Property 3. Property 4 tells us that the MC respects the ordering of distributions as defined by van Zwet

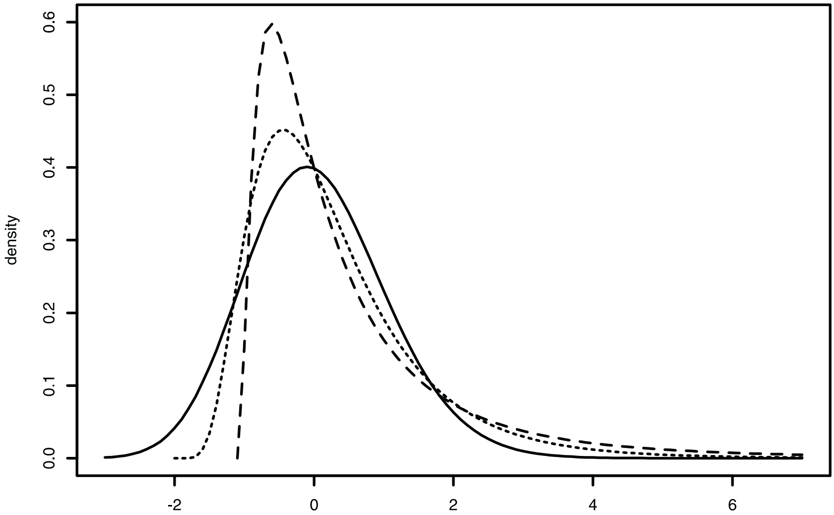

Figure 2. Density of the g-distribution for g = .1 (full line), g = .5 (dotted line), and g = .9 (dashed line).

(1964). Let F and G be continuous distributions with interval support, then it is said that G is at least as skew to the right as F, or "F c-precedes G"

$$F <_c G \Leftrightarrow G^{-1}(F(x)) \quad \text{is convex}$$

on the support of F.

Property 4. *If $F <_c G$, then $MC(F) \leq MC(G)$.*

As an example of a class of distributions that satisfy this c-ordering, we will consider in this article Tukey's class of g-distributions (Hoaglin, Mosteller, and Tukey 1985). When a random variable Z is Gaussian distributed, then

$$Y_g = \frac{(e^{gZ} - 1)}{g}$$

is said to follow a g-distribution G_g with parameter $g \in I\!R$. For $g = 0$ we set $Y_0 \equiv Z$ and thus we have zero skewness. It is clear that $G_{-g}(x) = 1 - G_g(-x)$, hence we will consider only the right-skewed distributions for which $g > 0$. In Figure 2 we have plotted the density functions of $G_{.1}, G_{.5}$, and $G_{.9}$. It is easy to show that G_{g_1} c-precedes G_{g_2} for any $g_1 < g_2$.

Throughout we will also compare the medcouple with the *quartile skewness*

$$\text{QS}(F) = \frac{(Q_{.75} - Q_{.5}) - (Q_{.5} - Q_{.25})}{Q_{.75} - Q_{.25}} \tag{2.7}$$

and the *octile skewness*

$$\text{OS}(F) = \frac{(Q_{.875} - Q_{.5}) - (Q_{.5} - Q_{.125})}{Q_{.875} - Q_{.125}}, \tag{2.8}$$

which are based entirely on certain quantiles $Q_p = F^{-1}(p)$ of the distribution F. Both QS and OS belong to the class of skewness measures introduced by Hinkley (1975), they

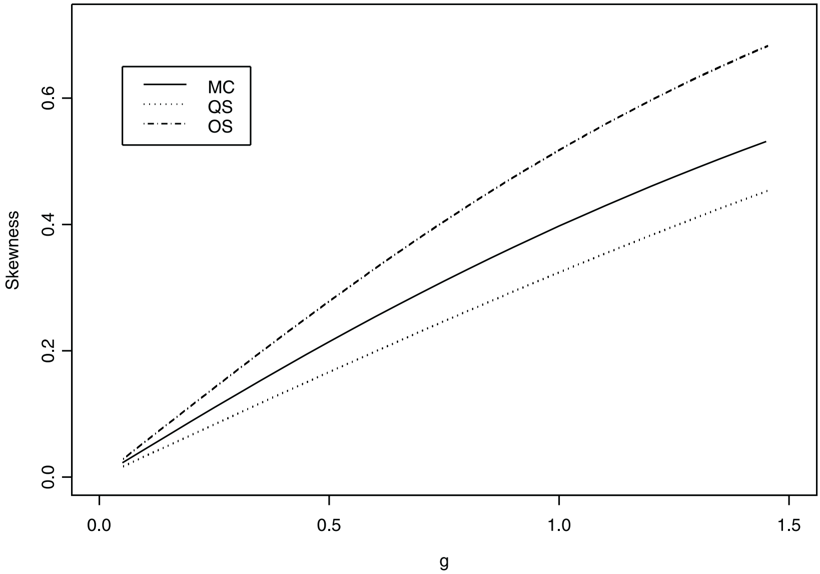

Figure 3. Monotone relation between g and the medcouple, the quartile skewness, and the octile skewness at the G_g distribution.

are bounded by $[-1, 1]$, and satisfy Properties 1 to 4. The definition of their finite-sample versions QS_n and OS_n is straightforward. From Brys et al. (2003) they appeared to be good and fast alternatives to the medcouple.

Figure 3 shows $\text{MC}(G_g)$, $\text{QS}(G_g)$, and $\text{OS}(G_g)$ for g ranging from 0 to 1.5. From this figure it is clear that Property 4 is satisfied by the three skewness measures because the three curves are monotone increasing. It can also be seen that the functional MC lies between OS and QS. Hence, the three finite-sample measures MC_n, QS_n, and OS_n are not estimating the same quantity, although they all reflect the degree of (a)symmetry in the data. We should keep this in mind when we make a comparative study as in Section 5.

3. ROBUSTNESS PROPERTIES

This section computes the breakdown value and the influence function of the medcouple. From the latter we will derive its asymptotic variance and compare it with finite-sample variances attained at datasets of different sizes.

3.1 Breakdown Value

The breakdown value of an estimator T_n at a sample X_n measures how many observations of X_n need to be replaced to make the estimate worthless (Rousseeuw and Leroy 1987). For a univariate location estimator, for example, this means that the absolute value of the estimate becomes arbitrarily large, whereas we say that a scale estimator breaks if the estimate becomes arbitrarily large or close to zero. Because the medcouple is bounded by $[-1,1]$, we define its finite-sample breakdown value as

$$\varepsilon_n^*(\mathrm{MC}_n;X_n)=\min\left\{\frac{m}{n};\sup_{X_n'}|\mathrm{MC}_n(X_n')|=1\right\},$$

where the dataset X_n' is obtained by replacing m observations from X_n by arbitrary values.

Theorem 1. *If the dataset X_n is in general position, that is, no two data points coincide, then*

$$\frac{1}{n}\left(\left\lceil\frac{n}{4}\right\rceil-1\right)\leq\varepsilon_n^*(\mathrm{MC}_n;X_n)\leq\frac{1}{n}\left(\left\lceil\frac{n}{4}\right\rceil+1\right).$$

The MC can thus resist up to 25% outliers in the data, which is the same as for the quartile skewness QS. The breakdown value of the octile skewness is only 12.5%.

3.2 Influence Function

The influence function of an estimator T at some distribution F measures the effect on T when adding a small probability mass at the point x (Hampel, Ronchetti, Rousseeuw, and Stahel 1986). If Δ_x is the point mass in x, then the influence function is defined as

$$\mathrm{IF}(x,T,F)=\lim_{\varepsilon\downarrow 0}\frac{T((1-\varepsilon)F+\varepsilon\Delta_x)-T(F)}{\varepsilon}. \tag{3.1}$$

As we have pointed out in (2.6), at any continuous distribution F with median m_F, the functional MC is equal to

$$\mathrm{MC}_F=H_F^{-1}(.5),$$

with

$$H_F(u)=4\int_{m_F}^{+\infty}F\left(\frac{x_2(u-1)+2m_F}{u+1}\right)dF(x_2).$$

To derive the influence function of the medcouple we will also use the functions

$$g_1(v)=\frac{v(\mathrm{MC}_F-1)+2m_F}{\mathrm{MC}_F+1} \tag{3.2}$$

and

$$g_2(v)=\frac{v(\mathrm{MC}_F+1)-2m_F}{\mathrm{MC}_F-1}. \tag{3.3}$$

Theorem 2. *Assume that F is an absolute continuous distribution with density f such that $MC_F\neq 1, f(m_F)\neq 0$ and $H_F'(MC_F)\neq 0$, then*

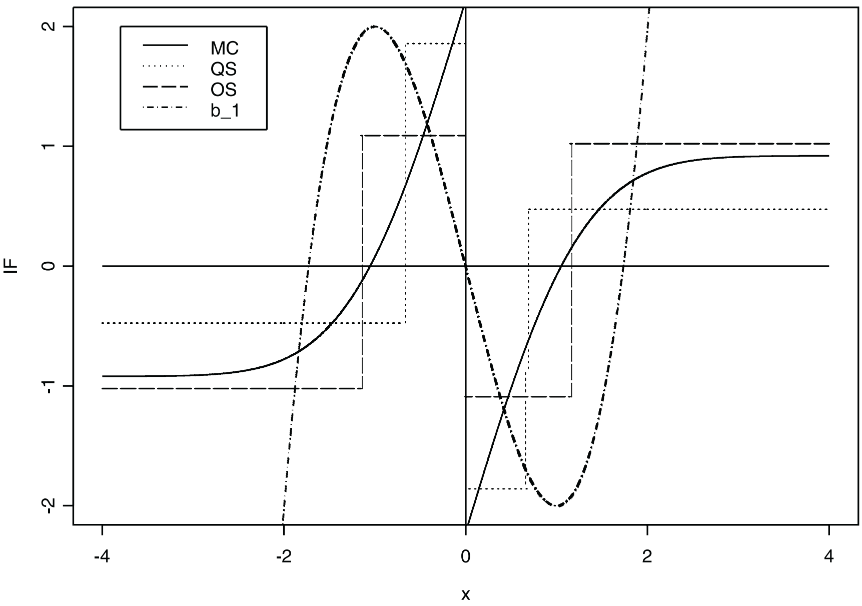

Figure 4. *Influence function of* b_1*, QS, OS, and MC at the standard Gaussian distribution.*

IF(x,MC,F)

$$= \frac{1}{H_F'(MC_F)}\left[1 - 4F(g_1(x))I(x > m_F) - 4(F(g_2(x)) - .5)I(x < m_F)\right.$$
$$\left. + sgn(x - m_F)\left(1 - \frac{4}{f(m_F)(MC_F + 1)}\int_{m_F}^{+\infty} f(g_1(w))dF(w)\right)\right]. \tag{3.4}$$

From Theorem 2 it follows that the medcouple has a bounded influence function, in contrast to the classical skewness measure b_1 (Groeneveld 1991). The influence functions of QS and OS were also derived by Groeneveld (1991) and are bounded as well. Figure 4 shows the influence function of these four estimators at the standard Gaussian distribution $F = \Phi$. For the medcouple we obtain

$$\text{IF}(x, \text{MC}, \Phi) = \pi(2\Phi(x) - 1 - \frac{1}{\sqrt{2}}\text{sgn}(x)) \tag{3.5}$$

from which the gross-error sensitivity $\gamma^*(\text{MC}, \Phi) = \sup_x |\text{IF}(x, \text{MC}, \Phi)| = \frac{\pi}{\sqrt{2}} = 2.22$ follows. Moreover, we have $\gamma^*(\text{QS}, \Phi) = 1.86$ and $\gamma^*(\text{OS}, \Phi) = 1.09$. We see that the influence functions of QS and OS are step functions, whereas the influence function of MC is continuous (except in the median). The IF of the medcouple is like a smoothed version of IF(QS) and IF(OS). Its gross-error sensitivity is close to $\gamma^*(\text{QS})$ and is obtained by inliers close to zero. The influence of outliers at infinity is smaller and comparable to the influence of outliers on OS.

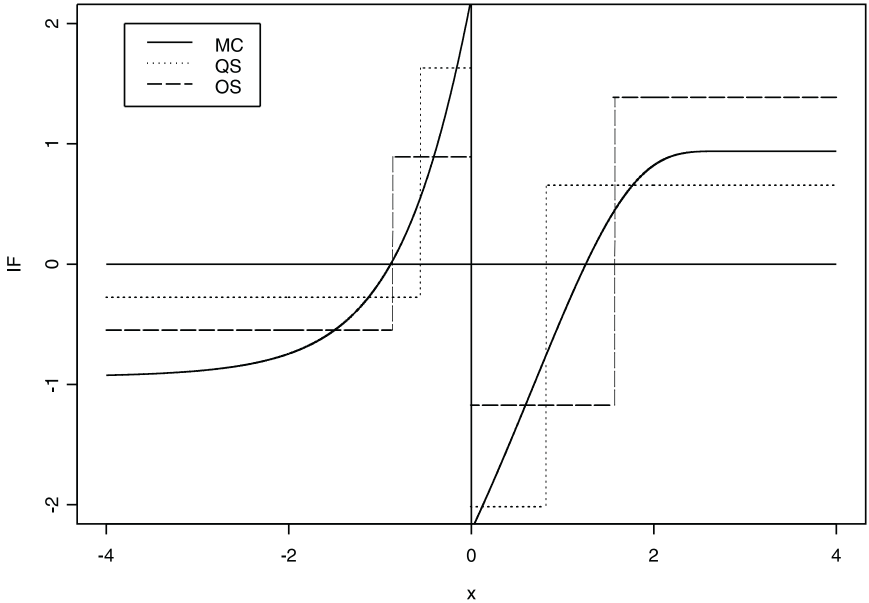

Figure 5. Influence function of b_1, QS, OS, and MC at the $G_{.5}$ distribution.

Figure 5 shows the influence functions of MC, QS, and OS at the asymmetric distribution $G_{.5}$, the unbounded classical skewness being removed from the plot for clarity. The IF of the medcouple is again continuous (except in the median), and its gross-error sensitivity $\gamma^*(\text{MC}, G_{.5}) = 2.21$ is only slightly larger than $\gamma^*(\text{QS}, G_{.5}) = 2.01$. The influence of far outliers to the left of the median is larger than for QS and OS, but the influence of far right outliers is smaller than for OS.

3.3 Asymptotic Variance

If an estimator T is asymptotically normal at a distribution F, its asymptotic variance $V(T, F)$ is given by (Hampel et al. 1986)

$$V(T, F) = \int \text{IF}(x, T, F)^2 dF(x). \tag{3.6}$$

The asymptotic normality of QS and OS was proven by Moors et al. (1996). For the medcouple, Brys et al. (2003) constructed QQ-plots which suggested its asymptotic normal behavior. Moreover, we expect MC to be asymptotically normal because it belongs to the class of incomplete generalized L-statistics of Hössjer (1996).

At the normal distribution we use (3.5) to derive $V(\text{MC}, \Phi) = \frac{\pi^2}{6}(5 - 3\sqrt{2}) = 1.25$, whereas $V(\text{QS}, \Phi) = 1.84$ and $V(\text{OS}, \Phi) = 1.15$. For the $G_{.5}$ distribution we used numerical integration to obtain the asymptotic variances given in Table 1. To illustrate the convergence of the finite-sample variance of MC_n, QS_n, and OS_n to their asymptotic variance, $M = 10{,}000$ samples of size n were drawn from a G_g distribution for $g = 0$ and

Table 1. Finite Sample Variance (*n* times the variance) and Asymptotic Variance of MC, QS, and OS at the Standard Gaussian and the $G_{.5}$ Distribution

	MC		*QS*		*OS*	
n	G_0	$G_{.5}$	G_0	$G_{.5}$	G_0	$G_{.5}$
10	.696	.736	1.210	1.235	.955	.975
20	.962	.990	1.488	1.503	1.020	1.010
40	1.108	1.132	1.681	1.638	1.091	1.058
60	1.178	1.180	1.733	1.697	1.119	1.054
80	1.175	1.203	1.743	1.719	1.102	1.077
100	1.205	1.246	1.763	1.732	1.138	1.102
200	1.216	1.248	1.784	1.820	1.142	1.067
∞	1.246	1.261	1.839	1.861	1.151	1.023

$g = .5$. Table 1 lists the average over the M runs of $n\mathrm{var}(T_n)$ for the three skewness measures, for data sizes ranging from $n = 10$ to $n = 200$. We see that they all converge to their asymptotic variance fairly well.

4. FAST ALGORITHM

The naive algorithm of the medcouple evaluates the kernel function $h(x_i, x_j)$ for each couple (x_i, x_j) with $x_i \leq m_n$ and $x_j \geq m_n$. Therefore this algorithm needs $O(n^2)$ time, which is too slow for large datasets. Here we present an algorithm that needs $O(n \log n)$ time. Assume $X_n = \{x_1, \ldots, x_n\}$ the observed dataset sampled from a continuous univariate distribution. The pseudo-code of the algorithm is then as follows:

1. Order the observations from largest to smallest. With a suitable algorithm, this can be done in $O(n \log n)$ time.
2. For ease of notation and for numerical stability, transform the data by subtracting the median m_n of X_n. This can be done without loss of generality because the MC is location invariant. Let $Z_n = X_n - m_n$ denote the shifted dataset, then the kernel $h(z_i, z_j)$ reduces to

$$h(z_i, z_j) = \frac{z_j + z_i}{z_j - z_i}.$$

Let Z^- and Z^+ be defined as

$$\begin{aligned} Z^- &= \left\{z_i^- := z_k \in Z_n; z_k \leq 0\right\} \\ Z^+ &= \left\{z_j^+ := z_l \in Z_n; z_l \geq 0\right\} \end{aligned}$$

whereby Z^- and Z^+ remain sorted in descending order. Further, let p (resp. q) be the size of Z^- (resp. Z^+).

3. Assume first that we have no observations tied up at the median. Consider then the following $q \times p$ matrix which for each $i = 1, \ldots, p$ and $j = 1, \ldots, q$ contains $h(z_i^-, z_j^+)$ at the ith column and the jth row:

$$\begin{pmatrix} h(z_1^-, z_1^+) & \cdots & h(z_p^-, z_1^+) \\ \vdots & \ddots & \vdots \\ h(z_1^-, z_q^+) & \cdots & h(z_p^-, z_q^+) \end{pmatrix}$$

Using the definition of the kernel, and the ordering of Z^- and Z^+ it is easy to verify that

$$h(z_i^-, z_j^+) \geq h(z_{i+1}^-, z_j^+)$$

for each $i = 1, \ldots, p-1$ and $j = 1, \ldots, q$ and

$$h(z_i^-, z_j^+) \geq h(z_i^-, z_{j+1}^+)$$

for $i = 1, \ldots, p$ and $j = 1, \ldots, q-1$. Hence, we obtain the following scheme:

$$\left(\begin{tikzcd} h(z_1^-, z_1^+) \arrow[r, "\geq"] \arrow[d, "\geq"] \arrow[dr, "\geq"] & h(z_p^-, z_1^+) \arrow[d, "\geq"] \\ h(z_1^-, z_q^+) \arrow[r, "\geq"] & h(z_p^-, z_q^+) \end{tikzcd}\right)$$

Note that we do not need to compute all the values in this table which would again be of $O(n^2)$, but only those who are specifically needed in Step 4 of the algorithm. When some observations are tied with the median, the monotonicity of the table still holds because of the definition of the kernel in that case, see (2.3). Assume, for example, that four data points coincide with the median. Then we obtain the following matrix:

$$\left(\begin{array}{cccccccc} +1 & +1 & +1 & +1 & h(z_1^-, z_1^+) & \cdots & h(z_p^-, z_1^+) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +1 & +1 & +1 & +1 & h(z_1^-, z_q^+) & \cdots & h(z_p^-, z_q^+) \\ +1 & +1 & +1 & 0 & -1 & \cdots & -1 \\ +1 & +1 & 0 & -1 & -1 & \cdots & -1 \\ +1 & 0 & -1 & -1 & -1 & \cdots & -1 \\ 0 & -1 & -1 & -1 & -1 & \cdots & -1 \end{array}\right).$$

Table 2. Average CPU Times (in seconds) of the Naive and the Fast Algorithm of MC Based on 100 Random Samples

n	*Naive MC*	*Fast MC*
100	.0221	.0197
500	.2014	.0267
1,000	1.5951	.0404
2,000	8.3790	.0570
5,000	—	.1317
10,000	—	.2440
50,000	—	1.3212

4. Apply the algorithm proposed by Johnson and Mizoguchi (1978). This algorithm finds in $O(n \log n)$ time the kth order statistic in a table $[X_i + Y_j]_{i,j}$ with ordered vectors X_i and Y_j. Essentially they use only the monotonicity of the table in the rows, the columns and the diagonals. This condition is fulfilled in the table constructed in Step 3, so we find the median in $O(n \log n)$ time.

Because we first sorted the observations in $O(n \log n)$ time and then applied the algorithm of Johnson and Mizoguchi (1978), the whole procedure needs $O(n \log n)$ time.

In Table 2 the average CPU times (in seconds) for the computation of the medcouple on 100 random samples of different sizes n are given. Computations were performed on a Pentium II 450 Mhz processor, using S-Plus with an interface to C. Clearly, the fast algorithm is a huge improvement on the naive algorithm, especially for large datasets. For $n \geq 5{,}000$ the naive computation is even not performed because it took too long.

The S-Plus and MATLAB source code of this fast algorithm is available at the Web sites http://www.agoras.ua.ac.be and http://www.wis.kuleuven.ac.be/stat/robust.html.

5. FINITE-SAMPLE BEHAVIOR

This section compares the finite-sample behavior of MC_n, QS_n, and OS_n at uncontaminated as well as contaminated datasets. As we have discussed before, we find this comparison appropriate because these three measures are bounded by $[-1, 1]$, and they all have a positive breakdown point and a bounded influence function.

5.1 Performance at Uncontaminated Distributions

Let us first concentrate on the behavior of the estimators at a *symmetric* distribution. For this, we have generated 1,000 samples of each $n = 1{,}000$ observations from the Gaussian distribution G_0 and from the fat-tailed Cauchy distribution. In Table 3 we have listed the average estimated skewness and the standard error of the different estimators. We see that the average estimate is close to zero for all of them and that their variability is very comparable.

At *right-tailed* distributions we expect to have a positive skewness estimate. Therefore we now focus on simulations for distributions G_g with $g > 0$. We generated 1,000 samples of different data sizes ($n = 50, 100, 500,$ and 1,000) and computed for each estimator the

Table 3. Average Estimated Skewness and Standard error of MC_n, OS_n, and QS_n at the Symmetric Gaussian Distribution G_0 and at the Symmetric Fat-Tailed Cauchy Distribution, Computed for 1,000 Samples of Size $n = 1{,}000$

	G_0		*Cauchy*	
Estimator	*Ave*	*St. error*	*Ave*	*St. error*
OS_n	.00027	.00106	.00186	.00156
QS_n	.00212	.00135	.00247	.00195
MC_n	.00113	.00112	.00000	.00138

frequency of strictly positive values. Table 4 shows the results for $g = .1, .2, .3$, and .4. We also sampled from distributions with $g > .4$, but at the larger sample sizes all the measures then behave in a perfect way (i.e., the frequencies were overall equal to 1). From the table we can conclude that OS_n is most capable of detecting small positive skewness, followed by MC_n and QS_n. It is not surprising that OS_n outperforms QS_n, because OS_n uses more information from the tails. The medcouple, which has the same breakdown value as QS_n and approximately the same gross-error sensitivity, yields much better results than QS_n and thus is much less conservative than QS_n in detecting skewness. This will also be illustrated in the example in Section 6.2.

5.2 Performance at Distributions With Contamination

Let us now compare the robustness of the estimators against contamination. For this, we have generated 1,000 samples of each $n = 100$ observations from G_g distributions with g varying between 0 and .5. We thus considered symmetric as well as right-skewed distributions. Then we replaced 5% and 15% of the data with outliers spread out far in the tail of the distribution, and computed the absolute value of the average difference between the estimated skewness of the contaminated and of the original dataset. Figures 6(a) and (b)

Table 4. Fraction of Strictly Positive Skewness Estimates for 1,000 Samples of Different Data Sizes n from Several Distributions G_g

n	*Estimator*	*$G_{.1}$*	*$G_{.2}$*	*$G_{.3}$*	*$G_{.4}$*
50	OS_n	.666	.761	.844	.938
	QS_n	.593	.643	.671	.774
	MC_n	.613	.711	.776	.872
100	OS_n	.718	.845	.946	.979
	QS_n	.626	.677	.761	.839
	MC_n	.675	.789	.890	.936
500	OS_n	.885	.994	1.000	1.000
	QS_n	.713	.863	.954	.987
	MC_n	.814	.965	.994	1.000
1,000	OS_n	.957	.999	1.000	1.000
	QS_n	.793	.951	.994	1.000
	MC_n	.889	.995	.999	1.000

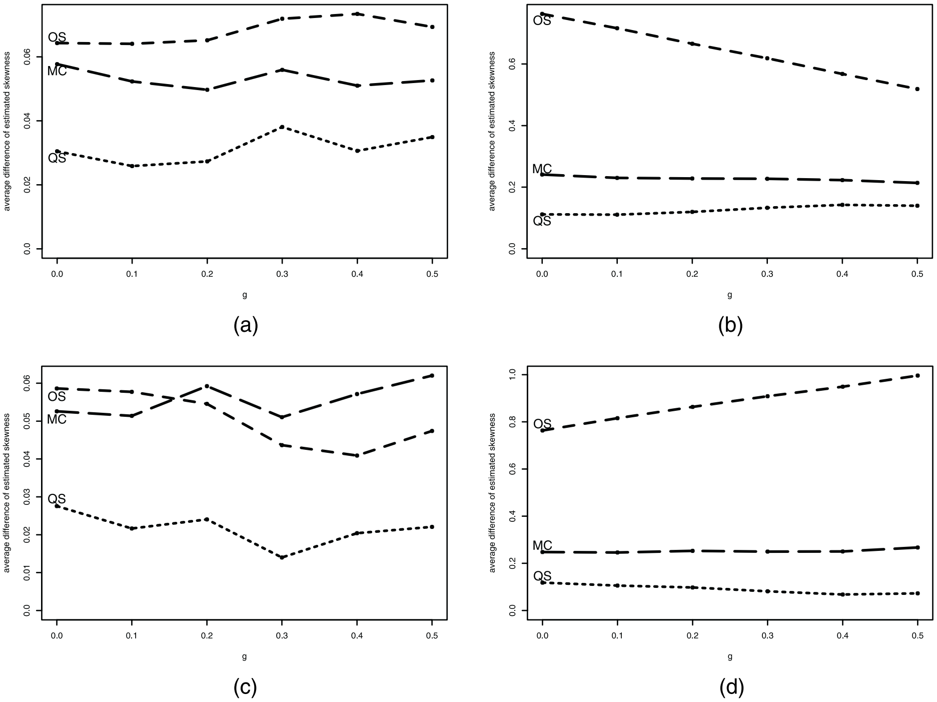

Figure 6. Absolute value of the average difference between the skewness estimate at contaminated and at uncontaminated data, for different values of g (a) with 5% of right contamination, (b) with 15% of right contamination, (c) with 5% of left contamination, (d) with 15% of left contamination.

contain the results for contamination in the right tail of the distribution, whereas Figures 6(c) and (d) are obtained by putting the outliers in the left tail.

These figures tell us that all three estimators perform well with a relatively small amount of contamination. Their bias is always very low, the smallest one being obtained by QS_n because it is based only on the middle part of the data. With 15% of contamination the three measures show more bias. The octile skewness clearly fails to give precise estimates, because its breakdown value is only 12.5%.

Note that the curves for QS and MC in Figure 6(b) and Figure 6(d) do not vary much with increasing g, whereas the OS curve is clearly decreasing with right contamination and increasing for left contamination. This is caused by the fact that we have put the outliers at the same values for any g. Let Q'_p denote the quantiles of the contaminated data, and assume for simplicity that quantiles always coincide with data points. With 15% contamination, $Q'_{.875}$ is an outlier, which explains the large bias of OS. But if $g_1 \leq g_2$, the outliers we have constructed are lying further in the right tail of G_{g_1} than in the right tail of G_{g_2}. For left contamination as in Figure 6(d) the inverse holds because the outliers are then lying further in the left tail of G_{g_2} than in the left tail of G_{g_1}.

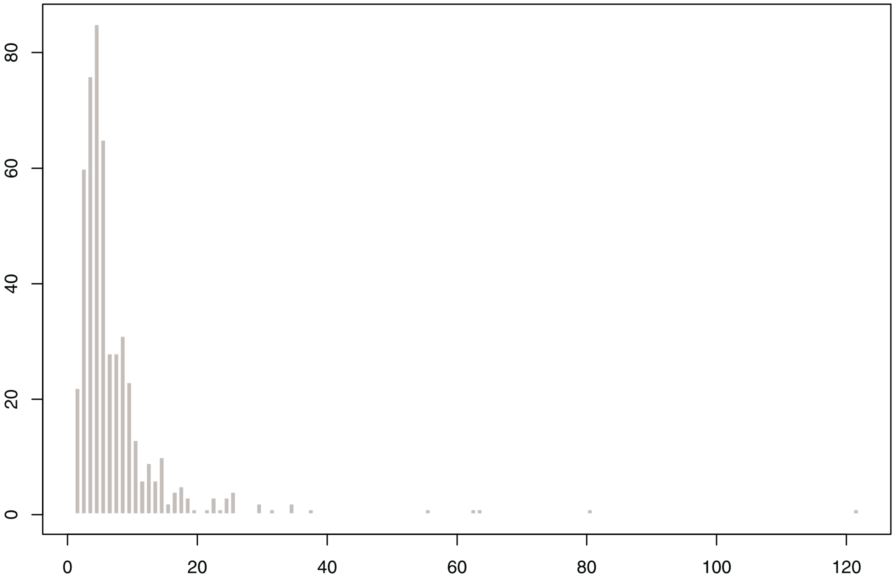

Figure 7. Histogram of the "length of stay" data.

6. EXAMPLES

6.1 Length of Stay Data

We examined Belgian data of 500 patients recovering from surgical foot procedures in 1988 (Marazzi, Paccaud, Ruffieux, and Beguin 1998). The variable of interest is the length of stay in days, which is skewed, distributed with a long tail to the right, as can be seen on the histogram in Figure 7. The skewness estimates for this dataset are given in Table 5. We see that the medcouple attains an intermediate value between QS_n and OS_n, whereas the classical estimate b_1 is rather high. When we remove the five most extreme data points, whose length of stay is larger than 56 days, we obtain the estimates listed in the second row of Table 5. The classical skewness b_1 drops a lot when we remove these outliers, the octile skewness decreases slightly, whereas QS_n and MC_n remain the same. This illustrates again the strong robustness of QS_n and MC_n towards outliers.

Table 5. Skewness Estimates for the "Length of Stay" Data at the Full and a Reduced Dataset

	QS_n	MC_n	OS_n	b_1
Full dataset	.20	.33	.38	6.64
Reduced dataset	.20	.33	.33	2.48

Table 6. One-Month Relative Price Differences of Belgian CPI data (September 1978)

−.036	.328	2.216	.129	−.162	8.903	2.540	−.316	−1.819	.207
−.778	−.039	−.181	.048	−.218	1.444	.207	.485	.177	.367
.161	2.130	.245	.142	.687	1.261	.149	.169	−.049	.129
.091	.024	−.087	.792	.328	−.132	.014	.000	1.943	.311
−.096	.329	.950	−.077	−.014	.000	−.294	.071	.007	1.089
.000	2.664	.038	.109	.018	.099	−.707	.000	1.722	8.414

6.2 Belgian Consumer Price Index Data

Our next example concerns Belgian consumer price index (CPI) data of September 1978, consisting of one-month relative price differences of 60 product categories as bread, tobacco, personal transport, and so on. They are listed in Table 6. Details of the data can be found in Aucremanne et al. (2002, 2004).

The nonparametric density estimate in Figure 8(a) and the boxplot in Figure 8(b) clearly show the presence of some outliers in the right tail of the distribution. In Figure 8(a) we have superimposed the density of $G_{.3}$ which fits the empirical density quite well.

Assume that we want to test whether the data are sampled from a normal (hence, symmetric) distribution F or not. Because the skewness measures are translation and scale invariant, this implies

$$\begin{cases} H_0 : \gamma(F) = 0 \\ H_a : \gamma(F) \neq 0 \end{cases}$$

with γ being one of the skewness measures MC, OS, or QS. If γ_n is asymptotically normally distributed (which has been formally proved for OS_n and QS_n), we can use the z-statistic

$$z = \sqrt{n}\frac{\gamma_n}{\sqrt{V(\gamma, \Phi)}} \approx_{H_0} N(0, 1) \tag{6.1}$$

with $V(\gamma, \Phi)$ the asymptotic variance of γ at the standard normal distribution $\Phi = G_0$. From Table 1 we obtain $V(\mathrm{MC}, \Phi) = 1.25$, $V(\mathrm{QS}, \Phi) = 1.84$, and $V(\mathrm{OS}, \Phi) = 1.15$. The p value of this test equals $p = 2P(Z < -|z|) = 2\Phi(-|z|)$.

If we apply this test on the Belgian CPI data, we obtain the z values and p values listed in the first row of Table 7. We see that the test based on QS accepts the hypothesis of normality at the 5% significance level, while the tests based on MC and OS clearly reject the normality assumption. Otherwise said, QS is not able to detect small positive skewness.

Analogously, we can test whether the data are sampled from the $G_{.3}$ distribution:

$$\begin{cases} H_0 : \gamma(F) = \gamma(G_{.3}) \\ H_a : \gamma(F) \neq \gamma(G_{.3}) \end{cases}$$

Under H_0 it holds that

$$z = \sqrt{n}\frac{\gamma_n - \gamma(G_{.3})}{\sqrt{V(\gamma, G_{.3})}} \approx_{H_0} N(0, 1) \tag{6.2}$$

so that z can be used in the same way as above. From the second row of Table 7 we now conclude that the tests based on QS and MC accept the null hypothesis at the 5% significance

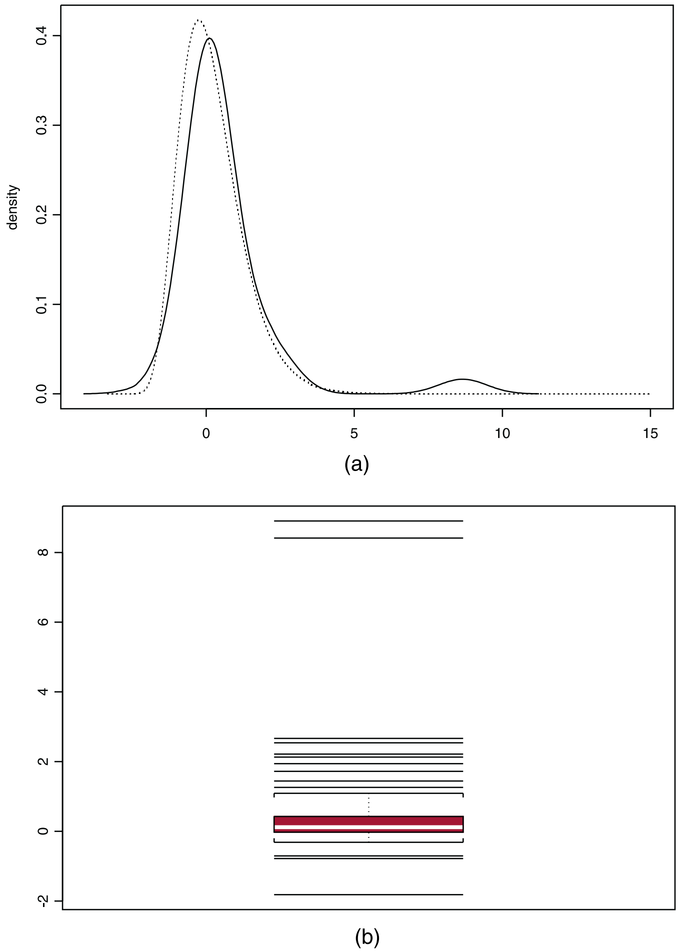

Figure 8. (a) Nonparametric density estimation of Belgian CPI data (solid line) and density of the $G_{.3}$ distribution (dotted line) superimposed; (b) boxplot of Belgian CPI data.

level, while the one using OS does not. This shows that the outliers have inflated OS too much.

7. DISCUSSION AND CONCLUSION

In this article we have studied a new robust measure of skewness, which we call the *medcouple* because it is the median over certain kernels which are defined on couples. We have proved that its breakdown value equals 25%, and that its influence function is bounded. A fast $O(n \log n)$ algorithm is provided and used to perform empirical studies at uncontaminated as well as contaminated datasets. By comparing the MC with the octile skewness and the quartile skewness, we can make the following conclusions:

1. All three measures—$\mathrm{MC}_n, \mathrm{OS}_n$, and QS_n—perform very well at symmetric (uncontaminated) distributions.

Table 7. Results of the Hypothesis Tests (6.1) and (6.2) for the Belgian CPI Data

	QS_n		MC_n		OS_n	
G	z	p	z	p	z	p
0	1.909	.056	2.616	.009	5.017	.000
0.3	1.328	.184	1.691	.091	3.864	.000

2. OS_n is most capable of detecting small positive skewness. Also MC_n does it well, but QS_n does not as it uses too little information from the tails of the distribution.
3. QS_n is the most insensitive to outliers, followed by MC_n. When the contamination is over 12.5%, OS_n shows a too large bias due to its lower breakdown value.

As an overall conclusion we can thus state that MC_n combines the strengths of OS_n and QS_n: it has the sensitivity of OS_n to detect skewness and the robustness of QS_n towards outliers. These features, together with the low computational complexity, make the medcouple an attractive, fast, and robust skewness estimator.

Moreover we can use MC_n to measure the left and right tail weight of a sample, by applying it to the left, respectively the right, half of the distribution (Brys et al. 2004). The medcouple is also useful to construct an adjusted boxplot for skewed distributions (Vandervieren and Hubert 2004).

APPENDIX

Proof of Property 4

Proof: Without loss of generality we assume $F^{-1}(.5) = 0$ and $G^{-1}(.5) = 0$. Following (2.5) we must show that med $H_F \leq$ med H_G with

$$\begin{aligned} H_F(u) &= 4\int_0^{+\infty}\int_{-\infty}^{0} I\left(h(x_1,x_2) \leq u\right) dF(x_1)dF(x_2) \\ H_G(u) &= 4\int_0^{+\infty}\int_{-\infty}^{0} I\left(h(y_1,y_2) \leq u\right) dG(y_1)dG(y_2). \end{aligned}$$

As F and G have interval support, they have a strictly monotone quantile function, hence we can find for any couple (x_1, x_2) with $x_1 \leq 0 \leq x_2$ a unique couple (y_1, y_2) with $y_1 \leq 0 \leq y_2$ such that

$$x_1 = F^{-1}(p) \qquad x_2 = F^{-1}(q) \qquad y_1 = G^{-1}(p) \qquad y_2 = G^{-1}(q)$$

with $p \in [0, \frac{1}{2}]$ and $q \in [\frac{1}{2}, 1]$. It is thus sufficient to show that

$$\frac{F^{-1}(q) + F^{-1}(p)}{F^{-1}(q) - F^{-1}(p)} \leq \frac{G^{-1}(q) + G^{-1}(p)}{G^{-1}(q) - G^{-1}(p)}.$$

Groeneveld and Meeden (1984) proved that this inequality is satisfied if $F <_c G$ and $p + q = 1$. It is straightforward to see that their proof also holds for $p + q < 1$. □

Proof of Theorem 1

Proof: First we prove that $\varepsilon_n^* \leq (\lceil n/4 \rceil + 1)/n$. Because MC_n is location invariant, we may assume without loss of generality that $m_n(X_n) = 0$. By symmetry we also assume that MC(X_n) ≥ 0. Take any MC(X_n) $< B < 1$. We will now show that we can construct a contaminated sample X'_n by replacing $\lceil n/4 \rceil + 1$ data points from X_n such that MC(X'_n) $> B$. For this we shift the $\lceil n/4 \rceil + 1$ $(= n - [3n/4] + 1)$ largest values of X_n by a constant $k > 2 \max |x_i|/(1 - B)$, that is, we let

$$x'_i = \begin{cases} x_i & \text{for} \quad i = 1, \ldots, [\frac{3n}{4}] - 1 \\ x_i + k & \text{for} \quad i = [\frac{3n}{4}], \ldots, n. \end{cases}$$

Now, $m_n(X'_n) = m_n(X_n)$ and for all $x_i \leq m_n$ we have that

$$h(x_i, x'_j) = \begin{cases} h(x_i, x_j) & \text{for} \quad j = 1, \ldots, [\frac{3n}{4}] - 1 \\ \frac{x_j + x_i + k}{x_j - x_i + k} & \text{for} \quad j = [\frac{3n}{4}], \ldots, n. \end{cases}$$

Because

$$\frac{x_j + x_i + k}{x_j - x_i + k} > B \Leftrightarrow k > \frac{x_j(B - 1) - x_i(B + 1)}{1 - B} \tag{A.1}$$

if $x_i < x_j$, we obtain that $h(x_i, x'_j) > B$ for each $j \geq [3n/4]$. Because $i \leq \lceil n/2 \rceil$, at least $\lceil n/2 \rceil (\lceil n/4 \rceil + 1)$ of the $h(x_i, x'_j)$ are larger than B. Now, because X_n is in general position, also X'_n is, hence for n even, the medcouple is defined as the median over $\frac{n}{2}\frac{n}{2}$ numbers, whereas for n odd, the median is taken over $\frac{n+1}{2}\frac{n+1}{2}$ numbers. The medcouple of X'_n will thus be larger than B because it is easy to verify that at least $[n^2/8] + 1$ for n even, resp. $[(n + 1)^2/8] + 1$ for n odd, of the $h(x_i, x'_j)$ are larger than B.

Second, we show that $\varepsilon^*_n \geq (\lceil n/4 \rceil - 1)/n$. Replace $k < \lceil \frac{n}{4} \rceil - 1$ data points by arbitrary values x'_i. We will show that the medcouple of this contaminated dataset still depends completely on the original data points and consequently that its absolute value is smaller than 1. Denote the median of this new data set by m_n. We call a the number of original data points to the left of m_n, and b the number of original points to the right of m_n. It is clear that $a + b \geq [3n/4] + 2$. Moreover, if n is even, then

$$\left[\frac{n}{4}\right] + 1 \leq \min\{a, b\} \quad \text{and} \quad \max\{a, b\} \leq \frac{n}{2},$$

whereas for n odd this becomes

$$\left[\frac{n + 1}{4}\right] + 1 \leq \min\{a, b\} \quad \text{and} \quad \max\{a, b\} \leq \frac{n + 1}{2}.$$

The number of uncontaminated expressions $h(x_i, x_j)$ is $ab \geq a([3n/4] + 2 - a)$. It is easy to verify that this lower bound is strictly larger than $[(n^2/4 + 1)/2]$ for n even, and $[((n + 1)^2/4 + 1)/2]$ for n odd, hence the medcouple is obtained as the average of one or two of these uncontaminated kernels. □

Proof of Theorem 2

Proof: First, we rewrite (2.5) for a contaminated distribution $F_\varepsilon = (1 - \varepsilon)F + \varepsilon\Delta_x$. Let MC$_\varepsilon$ = MC(F_ε) and $m_\varepsilon = F_\varepsilon^{-1}(.5)$, then the following equation holds:

$$\frac{1}{8} = \int_{m_\varepsilon}^{+\infty} \int_{-\infty}^{m_\varepsilon} I\left(\frac{x_2 + x_1 - 2m_\varepsilon}{x_2 - x_1} \leq \text{MC}_\varepsilon\right) dF_\varepsilon(x_1) dF_\varepsilon(x_2).$$

Note that the conditions

$$\frac{x_1 + x_2 - 2m_\varepsilon}{x_2 - x_1} \le \mathrm{MC}_\varepsilon, \qquad x_1 \le m_\varepsilon, \qquad x_2 \ge m_\varepsilon, \qquad -1 \le \mathrm{MC}_\varepsilon \le 1$$

are equivalent to

$$x_1 \le \frac{x_2(\mathrm{MC}_\varepsilon - 1) + 2m_\varepsilon}{1 + \mathrm{MC}_\varepsilon}, \qquad x_2 \ge m_\varepsilon, \qquad -1 \le \mathrm{MC}_\varepsilon \le 1.$$

We now introduce the functions

$$\begin{aligned} g_1(v,\varepsilon) &= \frac{v(\mathrm{MC}_\varepsilon - 1) + 2m_\varepsilon}{\mathrm{MC}_\varepsilon + 1} \\ g_2(v,\varepsilon) &= \frac{v(\mathrm{MC}_\varepsilon + 1) - 2m_\varepsilon}{\mathrm{MC}_\varepsilon - 1} \end{aligned}$$

which for $\varepsilon = 0$ collapse with g_1 and g_2 as defined in (3.2) and (3.3). With these notations, we obtain

$$\begin{aligned} \frac{1}{8} &= \int_{m_\varepsilon}^{+\infty} F_\varepsilon(g_1(x_2,\varepsilon))dF_\varepsilon(x_2) \\ &= \int_{m_\varepsilon}^{+\infty} \big[(1-\varepsilon)F + \varepsilon\Delta_x\big]\big(g_1(x_2,\varepsilon)\big)\, d\big[(1-\varepsilon)F + \varepsilon\Delta_x\big](x_2) \\ &= (1-2\varepsilon)\int_{m_\varepsilon}^{+\infty} F(g_1(x_2,\varepsilon))dF(x_2) + \varepsilon\int_{m_\varepsilon}^{+\infty} F(g_1(x_2,\varepsilon))d\Delta_x(x_2) \\ &\quad + \varepsilon\int_{m_\varepsilon}^{+\infty} \Delta_x(g_1(x_2,\varepsilon))dF(x_2) + O(\varepsilon^2). \end{aligned} \tag{A.2}$$

To compute $\mathrm{IF}(x, \mathrm{MC}, F) = \frac{\partial}{\partial\varepsilon}\mathrm{MC}(F_\varepsilon)|_{(\varepsilon=0)}$ we derive equality (A.2) with respect to ε, and let $\varepsilon \to 0$. Since the terms in ε^2 vanish, we have to derive the first three terms only, denoted by $T_{1,\varepsilon}$, $T_{2,\varepsilon}$ and $T_{3,\varepsilon}$.

$$\begin{aligned} \frac{\partial}{\partial\varepsilon}T_{1,\varepsilon}\bigg|_{(\varepsilon=0)} &= \frac{\partial}{\partial\varepsilon}\left[(1-2\varepsilon)\int_{m_\varepsilon}^{+\infty} F(g_1(x_2,\varepsilon))dF(x_2)\right]\bigg|_{(\varepsilon=0)} \\ &= -2\int_{m_F}^{+\infty} F(g_1(x_2))dF(x_2) + \frac{\partial}{\partial\varepsilon}\int_{m_\varepsilon}^{+\infty} F(g_1(x_2,\varepsilon))dF(x_2)\bigg|_{(\varepsilon=0)}. \end{aligned} \tag{A.3}$$

By definition of MC_F, the first term in (A.3) equals $-\frac{1}{4}$, whereas Leibnitz' rule yields

$$\begin{aligned} &\frac{\partial}{\partial\varepsilon}\int_{m_\varepsilon}^{+\infty} F(g_1(x_2,\varepsilon))dF(x_2)\bigg|_{(\varepsilon=0)} \\ &= \int_{m_F}^{+\infty} F'(g_1(x_2,0))\frac{\partial}{\partial\varepsilon}g_1(x_2,\varepsilon)\bigg|_{(\varepsilon=0)} dF(x_2) - F(g_1(m_F,0))F'(m_F)\frac{\partial}{\partial\varepsilon}m_\varepsilon\bigg|_{(\varepsilon=0)}. \end{aligned}$$

Calculus yields

$$\frac{\partial}{\partial\varepsilon}g_1(x_2,\varepsilon)\bigg|_{(\varepsilon=0)} = \frac{2(x_2 - m_F)\mathrm{IF}(x, \mathrm{MC}_F, F) + 2\mathrm{IF}(x, m_F, F)(\mathrm{MC}_F + 1)}{(\mathrm{MC}_F + 1)^2},$$

hence

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}T_{1,\varepsilon}\bigg|_{(\varepsilon=0)} &= -\frac{1}{4}+\mathrm{IF}(x,\mathrm{MC}_F,F)\int_{m_F}^{+\infty}\frac{2(x_2-m_F)}{(\mathrm{MC}_F+1)^2}f(g_1(x_2))dF(x_2)\\ &\quad +2\,\mathrm{IF}(x,m_F,F)\int_{m_F}^{+\infty}\frac{f(g_1(x_2))}{\mathrm{MC}_F+1}dF(x_2)-\frac{1}{2}f(m_F)\mathrm{IF}(x,m_F,F).\end{aligned} \tag{A.4}$$

The second term $T_{2,\varepsilon}$ in Equation (A.2) has partial derivative

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}T_{2,\varepsilon}\bigg|_{(\varepsilon=0)} &= \frac{\partial}{\partial\varepsilon}\left[\varepsilon\int_{m_\varepsilon}^{+\infty}F(g_1(x_2,\varepsilon))d\Delta_x(x_2)\right]\bigg|_{(\varepsilon=0)}\\ &= \int_{m_\varepsilon}^{+\infty}F(g_1(x_2,\varepsilon))d\Delta_x(x_2)\bigg|_{(\varepsilon=0)}\\ &= F(g_1(x))I(x>m_F),\end{aligned} \tag{A.5}$$

whereas for the third term $T_{3,\varepsilon}$ we obtain

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}T_{3,\varepsilon}\bigg|_{(\varepsilon=0)} &= \int_{m_\varepsilon}^{+\infty}\Delta_x(g_1(x_2,\varepsilon))dF(x_2)\bigg|_{(\varepsilon=0)}\\ &= \int_{m_\varepsilon}^{+\infty}I(x<g_1(x_2,\varepsilon))dF(x_2)\bigg|_{(\varepsilon=0)}\\ &= \int_{m_\varepsilon}^{+\infty}I(x_2<g_2(x,\varepsilon))dF(x_2)\bigg|_{(\varepsilon=0)}\\ &= \int_{m_\varepsilon}^{g_2(x,\varepsilon)}I(m_\varepsilon<g_2(x,\varepsilon))dF(x_2)\bigg|_{(\varepsilon=0)}\\ &= I(g_2(x)>m_F)\left[F(g_2(x))-\frac{1}{2}\right]\\ &= I(x<m_F)\left[F(g_2(x))-\frac{1}{2}\right].\end{aligned} \tag{A.6}$$

Combining Equations (A.2), (A.4), (A.5), and (A.6) and using the fact that

$$H'_F(\mathrm{MC}_F)=4\int_{m_F}^{+\infty}2f(g_1(x_2))\left(\frac{x_2-m_F}{(\mathrm{MC}_F+1)^2}\right)dF(x_2),$$

and

$$\mathrm{IF}(x,m_F,F)=\frac{1}{2f(m_F)}\mathrm{sgn}(x-m_F)$$

finally leads to Equation (3.4). □

[Received September 2002. Revised August 2003.]

REFERENCES

Aucremanne, L., Brys, G., Hubert, M., Rousseeuw, P. J., and Struyf, A. (2002), "Inflation, Relative Prices and Nominal Rigidities," National Bank of Belgium, Working Paper No. 20, May 2002.

——— (2004), "A Study of Belgian Inflation, Relative Prices and Nominal Rigidities using New Robust Measures of Skewness and Tail Weight," in *Theory and Applications of Recent Robust Methods*, eds. M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, Statistics for Industry and Technology, Basel: Birkhauser, pp. 13–25.

Brys, G., Hubert, M., and Struyf, A. (2003), "A Comparison of Some New Measures of Skewness," in *Developments in Robust Statistics, ICORS 2001*, eds. R. Dutter, P. Filzmoser, U. Gather, and P.J. Rousseeuw, Heidelberg: Springer-Verlag, pp. 98–113.

——— (2004), "Robust Measures of Tail Weight," *Computational Statistics and Data Analysis*, to appear (doi: 10.1016/j.csda.2004.09.012).

Groeneveld, R. A. (1991), "An Influence Function Approach to Describing the Skewness of a Distribution," *The American Statistician*, 45, 97–102.

Groeneveld, R. A., and Meeden, G. (1984), "Measuring Skewness and Kurtosis," *The Statistician*, 33, 391–399.

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

Hinkley, D. V. (1975), "On Power Transformations to Symmetry," *Biometrika*, 62, 101–111.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1985), *Exploring Data Tables, Trends, and Shapes,* New York: Wiley.

Hössjer, O. (1996), "Incomplete Generalized L-Statistics," *The Annals of Statistics*, 24, 2631–2654.

Johnson, D. B., and Mizoguchi, T. (1978), "Selecting the Kth Element in $X + Y$ and $X_1 + X_2 + \cdots + X_m$," *SIAM Journal of Computing*, 7, 147–153.

Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. (1998), *Applied Regression Analysis and Other Multivariate Methods*, Belmont, CA: Duxbury Press.

Marazzi, A., Paccaud, F., Ruffieux, C., and Beguin, C. (1998), "Fitting the Distribution of Length of Stay by Parametric Models," *Medical Care*, 36, 916–927.

Moors, J. J. A., Wagemakers, R. T. A., Coenen, V. M. J., Heuts, R. M. J., and Janssens, M. J. B. T. (1996), "Characterizing Systems of Distributions by Quantile Measures," *Statistica Neerlandica*, 50, 417–430.

Oja, H. (1981), "On Location, Scale, Skewness and Kurtosis of Univariate Distributions," *Scandinavian Journal of Statistics*, 8, 154–168.

Rousseeuw, P. J., and Leroy, A. (1987), *Robust Regression and Outlier Detection,* New York: Wiley.

Vandervieren, E., and Hubert, M. (2004), "An Adjusted Boxplot for Skewed Distributions," *COMPSTAT 2004, Proceedings in Computational Statistics*, ed. J. Antoch, Heidelberg: Springer-Verlag, pp. 1933–1940.

Zwet, W. R. van (1964), *Convex Transformations of Random Variables*, Amsterdam: Mathematisch Centrum.