A robust measure of skewness

September 27, 2002

Abstract

The asymmetry of a univariate continuous distribution is commonly measured by the classical skewness coefficient. Since this estimator is based on the first three moments of the data set, it is strongly affected by the presence of one or more outliers. In this paper we investigate 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 data sets.

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 data set \( X_n = \{x_1, x_2, \ldots, x_n\} \) 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) data set (Kleinbaum et al., 1998) which contains the systolic blood pressure of 30 patients. From the 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 data set, $b_1 = 1.36$, but if we remove the outlier, $b_1$ drops to 0.20.

![Boxplot of the Systolic Blood Pressure data set.](image)

In [authors names removed] we introduced and empirically compared several new measures of skewness which are less sensitive to outlying values. Of these measures the medcouple (MC) arose as the overall winner taking into account its performance at uncontaminated data sets and its robustness at contaminated samples. Let us e.g. reconsider the SBP data set. Here, the medcouple is exactly zero, both at the full and the cleaned data set, which illustrates nicely its robustness towards outliers.

The present paper studies the medcouple in more detail. In Section 2 we recall the definition of the medcouple and verify that it satisfies natural requirements of a skewness measure. In Section 3 we show the robustness of the medcouple by computing its breakdown value and influence function. A computationally fast algorithm is provided in Section 4 and used in Section 5 to compare the performance and the robustness of the medcouple with other robust skewness measures. Section 6 contains an example whereas 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 \( x_1 \leq x_2 \leq \ldots \leq x_n \). Let \( m_n \) denote the median of \( X_n \), defined as usual as

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

In [authors names removed] we introduced the medcouple \( (MC_n) \) as

\[
MC_n = \med_{x_i \leq m_n \leq x_j} h(x_i, x_j)
\] (2.1)

where for all \( x_i \neq x_j \) the kernel function \( h \) is given by:

\[
h(x_i, x_j) = \frac{(x_j - m_n) - (m_n - x_i)}{x_j - x_i}.
\] (2.2)

For the special case \( x_i = x_j = m_n \), we define the kernel as follows. Let \( m_1 < \ldots < m_k \) denote the indices of the observations which are tied to the median \( m_n \), i.e. \( x_{m_l} = m_n \) for all \( l = 1, \ldots, k \). Then

\[
h(x_{m_i}, x_{m_j}) = \begin{cases} 
-1 & \text{if } i + j - 1 < k \\
0 & \text{if } i + j - 1 = k \\
+1 & \text{if } i + j - 1 > k.
\end{cases}
\] (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 e.g. more data points which are strictly larger than the median than there are who 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 towards 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 $m_F = F^{-1}(0.5)$ be the median of $F$, then the definition of $MC_F$ follows in a straightforward way from (2.1)

$$MC_F = \text{med}_{x_1 \leq m_F \leq x_2} h(x_1, x_2)$$

(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

$$H_F(u) = 4 \int_{m_F}^{+\infty} \int_{-\infty}^{m_F} I(h(x_1, x_2) \leq u) dF(x_1) dF(x_2)$$

(2.5)

we obtain the shorter formulation

$$MC_F = H_F^{-1}(0.5).$$

(2.6)

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

$$h(x_1, x_2) \leq u, \quad x_1 \leq m_F, \quad x_2 \geq m_F$$

are equivalent to $x_1 \leq \frac{x_2(u-1)+2m_F}{u+1}$ and $x_2 \geq m_F$. 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 only based 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).
**Property 1.** \( MC \) is location and scale invariant, i.e.

\[
MC(aF + b) = MC(F)
\]

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

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

\[
MC(-F) = -MC(F).
\]

**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 (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 \( G \)’

\[
F \triangleleft_{c} G \iff G^{-1}(F(x)) \text{ is convex}
\]
on the support of \( F \).

**Property 4.** If \( F \triangleleft_{c} G \), then \( MC(F) \leq MC(G) \).

As an example of a class of distributions that satisfy this \( c \)-ordering, we will consider in this paper Tukey’s class of \( g \)-distributions (Hoaglin et al., 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 \( g \in \mathbb{R} \). 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 only consider the right-skewed distributions for which \( g > 0 \). In Figure 2 we have plotted the density functions of \( G_0, G_{0.5} \) and \( G_{0.9} \). It is easy to show that \( G_{g_1} \) \( c \)-precedes \( G_{g_2} \) for any \( g_1 < g_2 \). From Figure 3 it is clear that Property 4 is indeed satisfied by the medcouple because it shows a monotone and even an almost linear relation between \( g \) and \( MC(G_g) \).

Throughout we will also compare the medcouple with the quartile skewness

\[
QS(F) = \frac{(Q_{0.75} - Q_{0.5}) - (Q_{0.5} - Q_{0.25})}{Q_{0.75} - Q_{0.25}}
\]  \hspace{1cm} (2.7)
and the octile skewness
\[ OS(F) = \frac{(Q_{0.875} - Q_{0.5}) - (Q_{0.5} - Q_{0.125})}{Q_{0.875} - Q_{0.125}}, \]  
(2.8)
which are entirely based 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 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 [authors names removed] they appeared to be good and fast alternatives to the medcouple.

On Figure 3 we have also drawn \( QS(G_g) \) and \( OS(G_g) \). It can be clearly seen that the functional MC lies between OS and QS. Hence, the three finite-sample measures \( MC_n, OS_n \) and \( QS_n \) 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.

![Figure 2: Density of the \( g \)-distribution for \( g = 0.1 \) (full line), \( g = 0.5 \) (dotted line), and \( g = 0.9 \) (dashed line).](image)

### 3 Robustness properties

In this section we compute 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 data sets 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, e.g. 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

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

where the data set $X'_n$ is obtained by replacing $m$ observations from $X_n$ by arbitrary values.

**Theorem 1.** If the data set $X_n$ is in general position, i.e. no two data points coincide, then

$$
\frac{1}{n} \left( \left\lfloor \frac{n}{4} \right\rfloor - 1 \right) \leq \varepsilon_n^*(MC_n; X_n) \leq \frac{1}{n} \left\lfloor \frac{n}{4} \right\rfloor.
$$

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 et al., 1986). If $\Delta_x$ is the point mass in $x$, then the influence function is defined as:

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

(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

$$MC_F = H_F^{-1}(0.5)$$

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

$$g_1(v) = \frac{v(MC_F - 1) + 2m_F}{MC_F + 1}$$  

(3.2)

and

$$g_2(v) = \frac{v(MC_F + 1) - 2m_F}{MC_F - 1}.$$  

(3.3)

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

$$IF(x, MC, F) = \frac{1}{H_F'(MC_F)} \left[ 1 - 4F(g_1(x))I(x > m_F) - 4(F(g_2(x)) - 0.5)I(x < m_F) \right. \right.$$

$$+ \left. \text{sgn}(x - m_F) \left( 1 - \frac{4}{f(m_F)(MC_F + 1)} \int_{m_F}^{+\infty} f(g_1(w))dF(w) \right) \right]$$  

(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

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

(3.5)

from which the gross-error sensitivity $\gamma^*(MC, \Phi) = \sup_x |IF(x, MC, \Phi)| = \frac{\pi}{\sqrt{2}} = 2.22$ follows. Moreover, we have $\gamma^*(QS, \Phi) = 1.86$ and $\gamma^*(OS, \Phi) = 1.09$. 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 $\gamma^*(QS)$ 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.

![Influence function of $b_1$, QS, OS and MC at the standard gaussian distribution.](image)

Figure 4: Influence function of $b_1$, QS, OS and MC at the standard gaussian distribution.

Figure 5 shows the influence functions of MC, QS and OS at the asymmetric distribution $G_{0.5}$, 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 $\gamma^*(MC,G_{0.5}) = 2.21$ is only slightly larger than $\gamma^*(QS,G_{0.5}) = 2.01$. 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):

$$V(T, F) = \int IF(x, T, F)^2 dF(x).$$  \hspace{1cm} (3.6)

The asymptotic normality of QS and OS is proven in Moors et al. (1996). For the medcouple we constructed QQ-plots in [authors names removed] which suggest 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 $V(MC, \Phi) = \frac{2}{5}(5 - 3\sqrt{2}) = 1.25$, whereas $V(QS, \Phi) = 1.84$ and $V(OS, \Phi) = 1.15$. For the $G_{0.5}$ 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 $g = 0.5$. Table 1 lists the average over the $M$ runs of $n\text{Var}(T_n)$ 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 at large data sets. Here we present an algorithm which only needs $O(n \log n)$ time. Assume $X_n = \{x_1, \ldots, x_n\}$ the observed data set 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

Figure 5: Influence function of $b_1, QS, OS$ and MC at the $G_{0.5}$ distribution.
<table>
<thead>
<tr>
<th></th>
<th>MC</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$G_0$</td>
<td>$G_{0.5}$</td>
<td>$G_0$</td>
<td>$G_{0.5}$</td>
<td>$G_0$</td>
<td>$G_{0.5}$</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.696</td>
<td>0.736</td>
<td>1.210</td>
<td>1.235</td>
<td>0.955</td>
<td>0.975</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.962</td>
<td>0.990</td>
<td>1.488</td>
<td>1.503</td>
<td>1.020</td>
<td>1.010</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1.108</td>
<td>1.132</td>
<td>1.681</td>
<td>1.638</td>
<td>1.091</td>
<td>1.058</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>1.178</td>
<td>1.180</td>
<td>1.733</td>
<td>1.697</td>
<td>1.119</td>
<td>1.054</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>1.175</td>
<td>1.203</td>
<td>1.743</td>
<td>1.719</td>
<td>1.102</td>
<td>1.077</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.205</td>
<td>1.246</td>
<td>1.763</td>
<td>1.732</td>
<td>1.138</td>
<td>1.102</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1.216</td>
<td>1.248</td>
<td>1.784</td>
<td>1.820</td>
<td>1.142</td>
<td>1.067</td>
<td></td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.246</td>
<td>1.261</td>
<td>1.839</td>
<td>1.861</td>
<td>1.151</td>
<td>1.023</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Finite sample variance ($n$ times the variance) and asymptotic variance of MC, QS and OS at the standard gaussian and the $G_{0.5}$ distribution.

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 w.l.o.g. because the MC is location invariant. Let $Z_n = X_n - m_n$ denote the shifted data set, 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

$$Z^- = \{ z^-_i := z_k \in Z_n; z_k \leq 0 \}$$

$$Z^+ = \{ z^+_i := z_l \in Z_n; z_l \geq 0 \}$$

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 $i$-th column and the $j$-th 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:

\[
\begin{pmatrix}
  h(z_1^-, z_1^+) \geq h(z_p^-, z_1^+) \\
  \geq & \geq & \geq \\
  h(z_1^-, z_q^+) \geq h(z_p^-, z_q^+)
\end{pmatrix}
\]

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 e.g. that four data points coincide with the median. Then we obtain the following matrix:

\[
\begin{pmatrix}
  +1 & +1 & +1 & +1 & h(z_1^-, z_1^+) \cdots h(z_p^-, z_1^+) \\
  +1 & +1 & +1 & +1 & h(z_1^-, z_q^+) \cdots h(z_p^-, z_q^+) \\
  +1 & +1 & +1 & 0 & -1 \cdots -1 \\
  +1 & +1 & 0 & -1 & -1 \cdots -1 \\
  +1 & 0 & -1 & -1 & -1 \cdots -1 \\
  0 & -1 & -1 & -1 & -1 \cdots -1
\end{pmatrix}
\]
<table>
<thead>
<tr>
<th>n</th>
<th>naive MC</th>
<th>fast MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.0221</td>
<td>0.0197</td>
</tr>
<tr>
<td>500</td>
<td>0.2014</td>
<td>0.0267</td>
</tr>
<tr>
<td>1000</td>
<td>1.5951</td>
<td>0.0404</td>
</tr>
<tr>
<td>2000</td>
<td>8.3790</td>
<td>0.0570</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td>0.1317</td>
</tr>
<tr>
<td>10000</td>
<td></td>
<td>0.2440</td>
</tr>
<tr>
<td>50000</td>
<td></td>
<td>1.3212</td>
</tr>
</tbody>
</table>

Table 2: Average CPU times (in seconds) of the naive and the fast algorithm of MC based on 100 random samples.

(4) Apply the algorithm proposed by Johnson and Mizoguchi (1978). This algorithm finds in $O(n \log n)$ time the $k^{th}$ order statistic in a table $[X_i + Y_j]_{i,j}$ with ordered vectors $X_i$ and $Y_j$. Essentially they only use the monotonicity of the table in the rows, the columns and the diagonals. This condition is fulfilled in the table constructed in step 4, 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, 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 data sets. For $n \geq 5000$ the naive computation is even not performed because it took too long.

## 5 Finite-sample behavior

In this section we compare the finite-sample behavior of $MC_n$, $QS_n$ and $OS_n$ at uncontaminated as well as contaminated data sets. 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.
<table>
<thead>
<tr>
<th>estimator</th>
<th>$G_0$ ave</th>
<th>st.error</th>
<th>Cauchy ave</th>
<th>st.error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$OS_n$</td>
<td>0.00027</td>
<td>0.00106</td>
<td>0.00186</td>
<td>0.00156</td>
</tr>
<tr>
<td>$QS_n$</td>
<td>0.00212</td>
<td>0.00135</td>
<td>0.00247</td>
<td>0.00195</td>
</tr>
<tr>
<td>$MC_n$</td>
<td>0.00113</td>
<td>0.00112</td>
<td>0.00000</td>
<td>0.00138</td>
</tr>
</tbody>
</table>

Table 3: Average estimated skewness and standard error of MC, OS and QS at the symmetric gaussian distribution $G_0$ and at the symmetric fat-tailed Cauchy distribution.

5.1 Performance at uncontaminated distributions

Let us first concentrate on the behaviour of the estimators at a symmetric distribution. For this, we have generated 1000 samples of each $n = 1000$ 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 1000 samples of different data sizes ($n = 50, 100, 500$ and $1000$) and computed for each estimator the frequency of strictly positive values. Table 4 shows the results for $g = 0.1, 0.2, 0.3$ and $0.4$. We also sampled from distributions with $g > 0.4$, 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 $QS_n$, since $OS_n$ 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.

5.2 Performance at distributions with contamination

Let us now compare the robustness of the estimators against contamination. For this, we have generated 1000 samples of each $n = 100$ observations from $G_g$ distributions with $g$ varying
<table>
<thead>
<tr>
<th>$n$</th>
<th>estimator</th>
<th>$G_{0.1}$</th>
<th>$G_{0.2}$</th>
<th>$G_{0.3}$</th>
<th>$G_{0.4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>$OS_n$</td>
<td>0.666</td>
<td>0.761</td>
<td>0.844</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>$QS_n$</td>
<td>0.593</td>
<td>0.643</td>
<td>0.671</td>
<td>0.774</td>
</tr>
<tr>
<td></td>
<td>$MC_n$</td>
<td>0.613</td>
<td>0.711</td>
<td>0.776</td>
<td>0.872</td>
</tr>
<tr>
<td>100</td>
<td>$OS_n$</td>
<td>0.718</td>
<td>0.845</td>
<td>0.946</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>$QS_n$</td>
<td>0.626</td>
<td>0.677</td>
<td>0.761</td>
<td>0.839</td>
</tr>
<tr>
<td></td>
<td>$MC_n$</td>
<td>0.675</td>
<td>0.789</td>
<td>0.890</td>
<td>0.936</td>
</tr>
<tr>
<td>500</td>
<td>$OS_n$</td>
<td>0.885</td>
<td>0.994</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>$QS_n$</td>
<td>0.713</td>
<td>0.863</td>
<td>0.954</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>$MC_n$</td>
<td>0.814</td>
<td>0.965</td>
<td>0.994</td>
<td>1.000</td>
</tr>
<tr>
<td>1000</td>
<td>$OS_n$</td>
<td>0.957</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>$QS_n$</td>
<td>0.793</td>
<td>0.951</td>
<td>0.994</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>$MC_n$</td>
<td>0.889</td>
<td>0.995</td>
<td>0.999</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 4: Fraction of strictly positive skewness estimates for 1000 samples of different data sizes $n$ from several distributions $G_s$. 
between 0 and 0.5. 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 average absolute difference between the estimated skewness at the contaminated and at the original data set. Figures 6(a) and (b) 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 relative small amount of contamination. Their bias is always very low, the smallest one being obtained by \( QS_n \) because it is only based 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%.

### 6 Example

We examined Belgian data of 500 patients recovering from surgical foot procedures in 1988 (Marazzi et al., 1998). The variable of interest is the length of stay in days, which is skewly distributed with a long tail to the right, as can be seen on the histogram in Figure 7. The skewness estimates for this data set 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 5 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>
<thead>
<tr>
<th>( QS_n )</th>
<th>( MC_n )</th>
<th>( OS_n )</th>
<th>( b_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>full data set</td>
<td>0.20</td>
<td>0.33</td>
<td>0.38</td>
</tr>
<tr>
<td>reduced data set</td>
<td>0.20</td>
<td>0.33</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Table 5: Skewness estimates for the ‘length of stay’ data at the full and a reduced data set.
Figure 6: Average absolute 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.

7 Discussion and conclusion

In this paper we have studied a new robust measure of skewness, which we call the \textit{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 data sets. By comparing the MC with the octile skewness and the quartile skewness, we can make the following conclusions:
Figure 7: Histogram of the ‘length of stay’ data.

(1) all three measures $MC_n$, $OS_n$ and $QS_n$ perform very well at symmetric (uncontaminated) distributions;

(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 few 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 kurtosis of a sample, by applying it to $|x_i|$. This will be subject to further research.

Appendix

Proof of Proposition 4

Proof. Without loss of generality we assume $F^{-1}(0.5) = 0$ and $G^{-1}(0.5) = 0$. Following (2.5)
we must show that $\text{med } H_F \leq \text{med } H_G$ with

$$H_F(u) = 4 \int_0^{+\infty} \int_{-\infty}^{0} I(h(x_1, x_2) \leq u) dF(x_1) dF(x_2)$$

$$H_G(u) = 4 \int_0^{+\infty} \int_{-\infty}^{0} I(h(y_1, y_2) \leq u) dG(y_1) dG(y_2)$$

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

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

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

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

In Groeneveld and Meeden (1984) it is 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 $p + q < 1$.

\[\square\]

**Proof of Theorem 1**

*Proof.* First we prove that $\varepsilon_n^* \leq ([n/4]) / n$. Because $MC_n$ is location invariant, we may assume w.l.o.g. that $m_n(X_n) = 0$. By symmetry we also assume that $MC(X_n) \geq 0$. Take any $MC(X_n) < B < 1$. We will now show that we can construct a contaminated sample $X'_n$ by replacing $[n/4]$ data points from $X_n$ such that $MC(X'_n) > B$. For this we shift the $[n/4]$ (i.e. $n - [3n/4]$) largest values of $X_n$ by a constant $k > 2 \max |x_i|/(1 - B)$, i.e. we let

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

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

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

Because

$$\frac{x_j + x_i + k}{x_j - x_i + k} > B \iff k > \frac{x_j(B - 1) - x_i(B + 1)}{1 - B}$$

(7.1)

19
if \( x_i < x_j \), we obtain that \( h(x_i, x_j') > B \) for each \( j > [3n/4] \). Since \( i \leq [n/2] \), at least \( [n/2][n/4] \) of the \( h(x_i, x_j') \) are larger than \( B \). Now, since \( X_n \) is in general position, also \( X'_n \) is, hence for \( n \) even, the medcouple is defined as the median over \( \frac{n}{2} \) numbers, whereas for \( n \) odd, the median is taken over \( \frac{n+1}{2} \) numbers. The medcouple of \( X'_n \) will thus be larger than \( B \) because it is easy to verify that at least \( [(n^2/4 + 1)/2] \) for \( n \) even, resp. \( [((n+1)^2/4 + 1)/2] \) for \( n \) odd, of the \( h(x_i, x_j') \) are larger than \( B \).

Secondly, we show that \( \varepsilon^*_n \geq ([n/4] - 1)/n \). Replace \( k < [\frac{n}{4}] - 1 \) data points by arbitrarily values \( x'_i \). We will show that the medcouple of this contaminated data set 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

\[
\left[ \frac{n}{4} \right] + 1 \leq \{a, b\} \leq \frac{n}{2},
\]

whereas for \( n \) odd this becomes

\[
\left[ \frac{n+1}{4} \right] + 1 \leq \{a, b\} \leq \frac{n+1}{2}.
\]

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, resp. \( [((n+1)^2/4 + 1)/2] \) for \( n \) odd, hence the medcouple is obtained as the average of one or two of these uncontaminated kernels.

\[\square\]

**Proof of Theorem 2**

**Proof.** First, we rewrite (2.5) for a contaminated distribution \( F_\varepsilon = (1 - \varepsilon)F + \varepsilon \Delta_\varepsilon \). Let \( MC_\varepsilon = MC(F_\varepsilon) \) and \( m_\varepsilon = F_\varepsilon^{-1}(0.5) \), then the following equation holds:

\[
\frac{1}{8} = \int_{m_\varepsilon}^{+\infty} \int_{-\infty}^{m_\varepsilon} I \left( \frac{x_2 + x_1 - 2m_\varepsilon}{x_2 - x_1} \leq MC_\varepsilon \right) dF_\varepsilon(x_1) dF_\varepsilon(x_2).
\]

Note that the conditions

\[
\frac{x_1 + x_2 - 2m_\varepsilon}{x_2 - x_1} \leq MC_\varepsilon, \quad x_1 \leq m_\varepsilon, \quad x_2 \geq m_\varepsilon, \quad -1 \leq MC_\varepsilon \leq 1
\]

are equivalent to

\[
x_1 \leq \frac{x_2(MC_\varepsilon - 1) + 2m_\varepsilon}{1 + MC_\varepsilon}, \quad x_2 \geq m_\varepsilon.
\]
We now introduce the functions

\[
g_1(v, \varepsilon) = \frac{v(MC_\varepsilon - 1) + 2m_\varepsilon}{MC_\varepsilon + 1}
\]

\[
g_2(v, \varepsilon) = \frac{v(MC_\varepsilon + 1) - 2m_\varepsilon}{MC_\varepsilon - 1}
\]

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

\[
\frac{1}{8} = \int_{m_\varepsilon}^{+\infty} F_\varepsilon(g_1(x_2, \varepsilon))dF_\varepsilon(x_2)
\]

\[
= \int_{m_\varepsilon}^{+\infty} [(1 - \varepsilon)F + \varepsilon \Delta_\varepsilon(g_1(x_2, \varepsilon)) d[(1 - \varepsilon)F + \varepsilon \Delta_\varepsilon(x_2)]
\]

\[
= (1 - 2\varepsilon) \int_{m_\varepsilon}^{+\infty} F(g_1(x_2, \varepsilon))dF(x_2) + \varepsilon \int_{m_\varepsilon}^{+\infty} F(g_1(x_2, \varepsilon))d\Delta_\varepsilon(x_2)
\]

\[
+ \varepsilon \int_{m_\varepsilon}^{+\infty} \Delta_\varepsilon(g_1(x_2, \varepsilon))dF(x_2) + O(\varepsilon^2).
\]

(7.2)

To compute \( IF(x, MC, F) = \frac{\partial}{\partial \varepsilon} MC(F_\varepsilon)\big|_{\varepsilon=0} \) we derive equality (7.2) with respect to \( \varepsilon \), and let \( \varepsilon \to 0 \). Since the terms in \( \varepsilon^2 \) vanish, we have to derive the first three terms only, denoted by \( T_{1,\varepsilon}, T_{2,\varepsilon} \) and \( T_{3,\varepsilon} \).

\[
\frac{\partial}{\partial \varepsilon} T_{1,\varepsilon}\bigg|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon} \left[ (1 - 2\varepsilon) \int_{m_\varepsilon}^{+\infty} F(g_1(x_2, \varepsilon))dF(x_2) \right]\bigg|_{\varepsilon=0}
\]

\[
= -2 \int_{m_F}^{+\infty} F(g_1(x_2))dF(x_2) + \frac{\partial}{\partial \varepsilon} \int_{m_\varepsilon}^{+\infty} F(g_1(x_2, \varepsilon))dF(x_2)\bigg|_{\varepsilon=0}
\]

(7.3)

By definition of \( MC_F \), the first term in (7.3) equals \(-\frac{1}{4}\), whereas

\[
\frac{\partial}{\partial \varepsilon} \int_{m_\varepsilon}^{+\infty} F(g_1(x_2, \varepsilon))dF(x_2)\bigg|_{\varepsilon=0} =
\]

\[
\int_{m_\varepsilon}^{+\infty} F'(g_1(x_2, \varepsilon)) \frac{\partial}{\partial \varepsilon} g_1(x_2, \varepsilon)dF(x_2)\bigg|_{\varepsilon=0} - F(g_1(m_\varepsilon, \varepsilon))F'(m_\varepsilon) \frac{\partial}{\partial \varepsilon} m_\varepsilon\bigg|_{\varepsilon=0}
\]

Calculus yields

\[
\frac{\partial}{\partial \varepsilon} g_1(x_2, \varepsilon)\bigg|_{\varepsilon=0} = \frac{2(x_2 - m_F)IF(x, MC_F, F) + 2IF(x, m_F, F)(MC_F + 1)}{(MC_F + 1)^2}
\]

hence

\[
\frac{\partial}{\partial \varepsilon} T_{1,\varepsilon}\bigg|_{\varepsilon=0} = -\frac{1}{4} + IF(x, MC_F, F) \int_{m_F}^{+\infty} 2(x_2 - m_F) (MC_F + 1)^2 f(g_1(x_2))dF(x_2)
\]

\[
+ 2 IF(x, m_F, F) \int_{m_\varepsilon}^{+\infty} \frac{f(g_1(x_2))}{MC_F + 1} dF(x_2) - \frac{1}{2} f(m_F) IF(x, m_F, F).
\]

(7.4)
The second term $T_{2,\varepsilon}$ in equation (7.2) has partial derivative
\[
\frac{\partial}{\partial \varepsilon} T_{2,\varepsilon} \bigg|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon} \left[ \varepsilon \int_{m_F}^{+\infty} F(g_1(x_2, \varepsilon)) d\Delta_x(x_2) \right]_{\varepsilon=0}
= \int_{m_F}^{+\infty} F(g_1(x_2, \varepsilon)) d\Delta_x(x_2)_{\varepsilon=0}
= F(g_1(x)) I(x > m_F) \tag{7.5}
\]
whereas for the third term $T_{3,\varepsilon}$ we obtain
\[
\frac{\partial}{\partial \varepsilon} T_{3,\varepsilon} \bigg|_{\varepsilon=0} = \int_{m_F}^{+\infty} \Delta_x(g_1(x_2, \varepsilon)) dF(x_2)_{\varepsilon=0}
= \int_{m_F}^{+\infty} I(x < g_1(x_2, \varepsilon)) dF(x_2)_{\varepsilon=0}
= \int_{m_F}^{+\infty} I(x_2 < g_2(x, \varepsilon)) dF(x_2)_{\varepsilon=0}
= \int_{m_F}^{g_2(x, \varepsilon)} I(m_F < g_2(x, \varepsilon)) dF(x_2)_{\varepsilon=0}
= I(g_2(x) > m_F) \left[ F(g_2(x)) - \frac{1}{2} \right]
= I(x < m_F) \left[ F(g_2(x)) - \frac{1}{2} \right]. \tag{7.6}
\]
Combining equations (7.2), (7.4), (7.5) and (7.6) and using the fact that
\[
H'_F(MC_F) = 4 \int_{m_F}^{+\infty} 2f(g_1(x_2)) \left( \frac{x_2 - m_F}{(MC_F + 1)^2} \right) dF(x_2)
\]
and
\[
IF(x, m_F, F) = \frac{1}{2f(m_F)} \text{sgn}(x - m_F)
\]
finally leads to equation (3.4). \qed

References


