Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gnst20

A robust scale estimator based on pairwise means

Garth Tarr [a] , Samuel Müller [a] & Neville Weber [a]
[a] School of Mathematics and Statistics, University of Sydney, Carslaw F07, Sydney, NSW, 2006, Australia
Version of record first published: 04 Oct 2011.

To cite this article: Garth Tarr , Samuel Müller & Neville Weber (2012): A robust scale estimator based on pairwise means, Journal of Nonparametric Statistics, 24:1, 187-199

To link to this article: http://dx.doi.org/10.1080/10485252.2011.621424

Journal of Nonparametric Statistics
Vol. 24, No. 1, March 2012, 187–199

Taylor & Francis
Taylor & Francis Group

A robust scale estimator based on pairwise means

Garth Tarr*, Samuel Müller and Neville Weber

School of Mathematics and Statistics, University of Sydney, Carslaw F07, Sydney, NSW 2006, Australia

(*Received 8 August 2011; final version received 4 September 2011*)

We propose a new robust scale estimator, the pairwise mean scale estimator P_n, which in its most basic form is the interquartile range of the pairwise means. The use of pairwise means leads to a surprisingly high efficiency across many distributions of practical interest. The properties of P_n are presented under a unified generalised L-statistics framework, which encompasses numerous other scale estimators. Extensions to P_n are proposed, including taking the range of the middle $\tau \times 100\%$ instead of just the middle 50% of the pairwise means as well as trimming and Winsorising both the original data and the pairwise means. Furthermore, we have implemented a method using adaptive trimming, which achieves a maximal breakdown value. We investigate the efficiency properties of the pairwise mean scale estimator relative to a number of other established robust scale estimators over a broad range of distributions using the corresponding maximum likelihood estimates as a common base for comparison.

Keywords: robust statistics; generalised L-statistics; scale estimation; Hodges–Lehmann estimator; random trimming

AMS Subject Classification: 62G05; 62G35

1. Introduction

Numerous robust estimates of scale have been proposed; however, no major advancements have been made in the past 15 years. The standard robust scale estimators are still Q_n (Rousseeuw and Croux 1993) and the median absolute deviation from the median (MAD) when efficiency is not of concern. Robust estimates of scale are important for a range of applications, from true scale problems to outlier identification, and as auxiliary parameters for more involved analyses. Recent work concerning robust scale estimation includes Boente, Ruiz, and Zamar (2010) and Wu and Zuo (2008).

In this article, we propose a new robust scale estimator, the pairwise mean scale estimator P_n, which combines familiar features from a number of commonly used robust estimators and possesses surprising efficiency properties. In contrast to Q_n, which utilises pairwise differences, P_n is based on pairwise means. Analogously to the interquartile range, the most basic form of P_n is calculated as the interquartile range of the pairwise means, yielding a scale estimator that can be viewed as a natural complement to the Hodges–Lehmann location estimator (Hodges and

*Corresponding author. Email: garth.tarr@sydney.edu.au

ISSN 1048-5252 print/ISSN 1029-0311 online

http://dx.doi.org/10.1080/10485252.2011.621424
http://www.tandfonline.com

Lehmann 1963). A generalisation, $P_n(\tau)$, considers the distance between the $(1 \pm \tau)/2$ quantiles of the empirical distribution of the pairwise means. Unless otherwise specified, the notation P_n is equivalent to $P_n(0.5)$. Extensions that are based on Winsorising and trimming are investigated. We also implement a form of adaptive trimming which is shown to achieve the maximal breakdown value of 50%.

Our investigation into the efficiency properties of the pairwise mean scale estimator will be based on the work of Randal (2008), who calculated efficiencies of estimators relative to the corresponding maximum likelihood (ML) estimators at a particular distribution. This method facilitates easier comparison than the method used in the seminal study by Lax (1985).

The pairwise mean scale estimator fits into the family of generalised L-statistics (GL-statistics) which encompasses broad classes of statistics of interest in nonparametric estimation, in particular, L-statistics, U-statistics and U-quantile statistics (Serfling 1984). Thus, a wide range of statistics related to scale estimation can be embedded into a single unified class. The interquartile range, the difference of two quantiles, fits into the family of GL-statistics, as does the variance which can be written as a U-statistic with kernel $h(x, y) = (x - y)^2/2$. The trimmed and Winsorised variances also fall into the class of GL-statistics. Furthermore, Q_n can be represented in terms of a U-quantile statistic, specifically the kth-order statistic of the $\binom{n}{2}$ kernels $\{h(x_i, x_j) = |x_i - x_j|; i < j\}$, where $k = \binom{h}{2}$ and $h = \lfloor n/2 \rfloor + 1$.

M-estimators are an important exception to the class of generalised linear statistics. The most prominent robust scale estimator that sits under the M-estimator umbrella is the MAD. An advantage of P_n over M-estimators of scale is that it does not require a location estimate.

In Section 2, P_n is formally defined along with possible generalisations and its breakdown value is found. In Section 3, correction factors are found to ensure consistency for the standard deviation at the normal. Also presented in Section 3 are the influence function and asymptotic normality of P_n. In addition to being an intuitive estimate of scale, one of the primary advantages of P_n is its high efficiency over a broad range of distributions. The results of our simulation study will be given in Section 4, which show how P_n compares favourably with other robust estimates of scale.

2. Pairwise mean scale statistic

Let the data, $\mathbf{X} = (x_1, \ldots, x_n)$, be independent realisations of identically distributed random variables with distribution function F. The set of $\binom{n}{2}$ pairwise means is $\{h(x_i, x_j), 1 \le i < j \le n\}$, where $h(x_1, x_2) = (x_1 + x_2)/2$. Let H_n be the empirical distribution function of the pairwise means,

$$H_n(t) := \frac{2}{n(n-1)} \sum_{i<j} \mathbb{I}\{h(x_i, x_j) \le t\}, \quad \text{for } t \in \mathbb{R}.$$

The estimator $P_n(\tau)$ is defined as

$$P_n(\mathbf{X}, \tau) = P_n(\tau) = c_\tau \left[H_n^{-1}\left(\frac{1+\tau}{2}\right) - H_n^{-1}\left(\frac{1-\tau}{2}\right)\right], \tag{1}$$

where c_τ is a correction factor to make $P_n(\tau)$ consistent for the standard deviation when the underlying observations are Gaussian and $0 < \tau \le 1$. By this definition, $P_n(\tau)$ is the range of the middle $\tau \times 100\%$ of H_n.

The notion of working with quantiles of pairwise means is not new. The Hodges–Lehmann estimate, the median of the pairwise means, is a well-known robust estimator of location (Hodges and Lehmann 1963). The pairwise mean scale estimator, P_n, can be thought of as the scale analogue of this location estimate. Following the algorithm set out in Johnson and Mizoguchi (1978), in a similar manner to the Hodges–Lehmann estimate, P_n can be computed in $O(n \log n)$ time.

One attraction of P_n is its simplicity, and we will show that it has desirable efficiency and robustness properties. Robustness, in the form of a non-zero breakdown value, is guaranteed by taking the difference of quantiles of the resulting pairwise mean distribution.

The definition of the contamination breakdown value of an estimator T_n at $\mathbf{X}$ is

$$\varepsilon^*(\mathbf{X}, T_n) = \inf\{\varepsilon | b(\varepsilon; \mathbf{X}, T_n) = \infty\},$$

where b is the maximum bias that can be caused by ε-corruption:

$$b(\varepsilon; \mathbf{X}, T_n) = \sup |T_n(\mathbf{X}^*) - T_n(\mathbf{X})|,$$

and the supremum is taken over the set of all ε-corrupted samples $\mathbf{X}^*$. In other words, ε^* is the smallest value of ε for which the estimator, when applied to the ε-corrupted sample $\mathbf{X}^*$, can take values arbitrarily far from $T_n(\mathbf{X})$.

By this definition, $P_n(\tau)$ will break down if at least $(1-\tau)/2 \times 100\%$ of the pairwise means are contaminated. Arbitrarily changing m of the original observations leaves $n-m$ fixed and $\binom{n-m}{2} = (n-m)(n-m-1)/2$ pairwise means remain uncontaminated. Hence, $P_n(\tau)$ will remain bounded so long as more than $(1+\tau)/2 \times 100\%$ of the pairwise means are unaffected, that is,

$$\frac{1}{2}(n-m)(n-m-1) > \frac{1+\tau}{2}\binom{n}{2} = \frac{1}{4}(1+\tau)n(n-1).$$

Setting $m \approx n\varepsilon^*$, for large n, we have

$$(n - n\varepsilon^*)(n - n\varepsilon^* - 1) > \frac{1}{2}(1+\tau)(n^2 - n).$$

Thus,

$$\varepsilon^* < 1 - \sqrt{\frac{1+\tau}{2}} + O(n^{-1}).$$

The asymptotic breakdown value of $P_n(\tau)$ is $\varepsilon^* \approx 1 - \sqrt{(1+\tau)/2}$, and it is clear that when τ decreases, the breakdown value increases. At one extreme, as $\tau \to 1$, $P_n(\tau)$ converges to the range of the pairwise means and the breakdown value goes to 0. At the other extreme, as $\tau \to 0$, the breakdown point of P_n is the same as that of the Hodges–Lehmann estimate of location, which has a well-known breakdown value, $\varepsilon^* \approx 0.29$.

We will show in the simulation study that $\tau = 0.5$ performs well over a large range of distributions. Hence, we define the pairwise mean scale estimator P_n as

$$P_n = P_n(0.5) = c\left[H_n^{-1}\left(\frac{3}{4}\right) - H_n^{-1}\left(\frac{1}{4}\right)\right],$$

where $c = c_{0.5}$ is a large sample correction factor. Under this definition, P_n has an asymptotic breakdown value of $\varepsilon^* \approx 0.134$.

There are many ways to calculate H_n^{-1}, see Hyndman and Fan (1996) for a summary. When developing the theory for the pairwise mean scale estimator and in the simulation study, we use the inverse of the empirical distribution function, that is, $H_n^{-1}(p) := \inf\{t : H_n(t) \geq p\}$.

Trimming is a common technique used to increase the robustness of non-robust estimators, see, for example, Stigler (1977). Recently, Wu and Zuo (2008, 2009) have investigated adaptive trimming of location and scale estimates. They found that adaptive trimming increases efficiency over fixed trimming and can improve robustness by achieving the best possible breakdown point for a sensible choice of the tuning parameter.

The $P_n(\tau)$ estimator is inherently robust to a moderate number of outliers. When this may not be seen as robust enough, adaptive trimming may be implemented to achieve a 50% breakdown value. Furthermore, fixed trimming and Winsorising may be used to increase efficiency at extremely heavy-tailed distributions such as the slash or Cauchy. Symmetrically trimming a fixed proportion of the data will only increase the breakdown value if the total proportion of the original observations trimmed is greater than two times the original breakdown value.

In the context of P_n, we can trim the original data points or the pairwise means. Trimming $\gamma \times 100\%$ of the original data is equivalent to discarding $\lfloor \gamma n \rfloor$ data points. When we calculate the pairwise means from the remaining $n - \lfloor \gamma n \rfloor$ observations, we have $(n - \lfloor \gamma n \rfloor)(n - \lfloor \gamma n \rfloor - 1)/2$ pairwise means. If instead trimming occurs after the pairwise means are calculated, that is, $\alpha \times 100\%$ of the pairwise means are trimmed, $n(n-1)/2 - \lfloor \alpha n(n-1)/2 \rfloor$ pairwise means are left. Therefore, if we wish to make the proportion of pairwise means remaining comparable, we need to set

$$\alpha \approx 1 - \frac{(1-\gamma)((1-\gamma)n - 1)}{n-1},$$

which is approximately $1 - (1-\gamma)^2$ for large n.

Winsorisation may also be used on the original data points, and the resulting number of pairwise means is still equal to $\binom{n}{2}$. Winsorisation yields efficiency gains similar to trimming and, as such, it is not considered in Section 4. Of course, Winsorising the pairwise means will give identical results to $P_n(\tau)$ whenever the proportion of pairwise means Winsorised is less than $(1-\tau)/2$.

While trimming may aid in increasing efficiency at heavy-tailed distributions, adaptive trimming of P_n, denoted by $\tilde{P}_n$, is required to simultaneously achieve the maximal breakdown value while preserving high efficiency. Specifically, for preliminary high breakdown location and scale estimates, $m(\mathbf{X})$ and $s(\mathbf{X})$, respectively, an observation, x_i, is trimmed if

$$\frac{|x_i - m(\mathbf{X})|}{s(\mathbf{X})} > d, \tag{2}$$

where d is an arbitrary constant. Note that d needs to be sufficiently large such that not all the observations are trimmed. The metric on the left-hand side of Equation (2) is the absolute value of the generalised scaled deviation, as defined in Wu and Zuo (2009). Simulations suggest that a value of $d = 5$ represents a good trade-off between achieving high efficiency at heavy-tailed distributions while maintaining high efficiency at light-tailed distributions. This is in agreement with Wu and Zuo (2009), who recommended a value for the tuning parameter of between 4 and 7.

The estimator $\tilde{P}_n$ will inherit its breakdown value from the minimum breakdown value of the preliminary estimates:

$$\varepsilon^*(\mathbf{X}, \tilde{P}_n) = \min\{\varepsilon^*(\mathbf{X}, m), \varepsilon^*(\mathbf{X}, s)\}.$$

Choosing estimators with 50% breakdown values, for example, setting $m(\mathbf{X})$ to be the median or Huber's M-estimate of location and $s(\mathbf{X})$ to be the MAD or Q_n, translates into a 50% breakdown value for $\tilde{P}_n$. Adaptive trimming of the pairwise means is another alternative to the standard P_n statistic. Using auxiliary estimates of location and scale, each with a breakdown value of 50%, to adaptively trim the kernels yields a pairwise mean scale statistic with a breakdown value of 0.29, the same as that for the Hodges–Lehmann estimate of location.

3. Properties of P_n

This section considers some of the properties of P_n. We begin by finding the limiting value of $P_n(\tau)$ defined in Equation (1) as $n \to \infty$ and provide correction factors to ensure consistency

for the standard deviation in large samples when the distribution of the underlying observations is Gaussian. We also show evidence of finite sample bias and suggest finite sample correction factors for P_n.

By exploiting the *GL*-statistic structure of $P_n(\tau)$, we find the influence curve and infer related properties such as the asymptotic efficiency and gross-error sensitivity for $P_n(\tau)$. We also establish the asymptotic normality of $P_n(\tau)$.

3.1. *Correction factors*

As noted in Equation (1), a correction factor, c_τ, is required to ensure that $P_n(\tau)$ is consistent for the standard deviation. Without loss of generality, let F, the distribution of the underlying observations, be centred at zero. The cumulative distribution function (cdf) of the pairwise means, $(X_i + X_j)/2$, is given by

$$\begin{aligned} H_F(t) &= \int_{-\infty}^{\infty} \int_{-\infty}^{2t-u} f(x) f(u)\, \mathrm{d}x\, \mathrm{d}u \\ &= \int_{-\infty}^{\infty} F(2t-u) f(u)\, \mathrm{d}u. \end{aligned} \tag{3}$$

When the underlying data follow a Gaussian distribution with cdf Φ and density $\phi = \Phi'$, Equation (3) can be written as

$$H_\Phi(t) = \int_{-\infty}^{\infty} \Phi(2t-x)\phi(x)\, \mathrm{d}x.$$

The correction factor for a given $\tau \in (0,1)$ is

$$\frac{1}{c_\tau} = H_\Phi^{-1}\left(\frac{(1+\tau)}{2}\right) - H_\Phi^{-1}\left(\frac{(1-\tau)}{2}\right). \tag{4}$$

The expression in Equation (4) can easily be obtained using numerical integration. In particular, for P_n, the corresponding asymptotic correction factor is $c_{0.5} = c \approx 1/0.9539$.

The finite sample correction factors are applied after the large sample correction has been made. Finite sample correction factors for P_n, $c_{n,0.5}$, have been found analytically for $n = 3$ and 4: $c_{3,0.5} = 1.13$ and $c_{4,0.5} = 1.30$. The finite sample correction factors exhibit jumps and are not monotonically increasing in n, instead they exhibit a periodic pattern attributable to the order statistic method used to find the quantiles. For $5 \le n < 40$, finite sample correction factors have been found using simulation, and a sample of these is given in Table 1. For $n \ge 40$, the small sample correction factor for P_n is well approximated by $c_{n,0.5} = n/(n-0.7)$.

Table 1. Finite sample correction factors applied to P_n at the normal to ensure approximate unbiasedness.

n	$c_{n,0.5}$	n	$c_{n,0.5}$
5	1.108	15	1.061
6	1.064	20	1.036
7	1.165	25	1.029
8	1.103	30	1.021
9	1.087	35	1.018
10	1.105	40	1.018

3.2. *Influence curve*

Following Hampel (1974), the influence curve for a functional T at the distribution F is defined by

$$\mathrm{IC}(x; T, F) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon},$$

where δ_x has all its mass at x.

Noting that $P_n(\tau)$ is a GL-statistic, using Equation (2.15) from Serfling (1984) and assuming that F has derivative $f > 0$ on $[F^{-1}(\epsilon), F^{-1}(1-\epsilon)]$ for all $\epsilon > 0$, the influence curve for $P_n(\tau)$ is

$$\mathrm{IC}(x; P_n(\tau), F) = c_\tau \left[\frac{(1+\tau)/2 - F(2H_F^{-1}((1+\tau)/2) - x)}{\int f(2H_F^{-1}((1+\tau)/2) - x) f(x)\,\mathrm{d}x} - \frac{(1-\tau)/2 - F(2H_F^{-1}((1-\tau)/2) - x)}{\int f(2H_F^{-1}((1-\tau)/2) - x) f(x)\,\mathrm{d}x} \right].$$

Figure 1 plots the influence curves for P_n, Q_n, the MAD and the sample standard deviation when the underlying data are Gaussian. Alongside the breakdown value, another important measure of robustness is the gross-error sensitivity:

$$\gamma^* = \sup_x |\mathrm{IC}(x, T, F)|,$$

which measures the worst approximate influence a fixed amount of contamination can have on the value of the estimator, that is, it represents an approximate bound for the bias of the estimator. Figure 1 shows that P_n has a slightly higher gross-error sensitivity of $\gamma^* = 2.33$ than Q_n, which has $\gamma^* = 2.07$ at the normal. As Hampel (1974) noted, the asymptotic variance of an estimator approaches its minimum as the influence curve approaches a multiple of the log-likelihood derivative. Hence, when the underlying observations are normally distributed, the closer an estimator's influence curve is to that of the standard deviation, the more efficient it will be. In Figure 1, the influence curve of P_n is almost always closer to that of the standard deviation

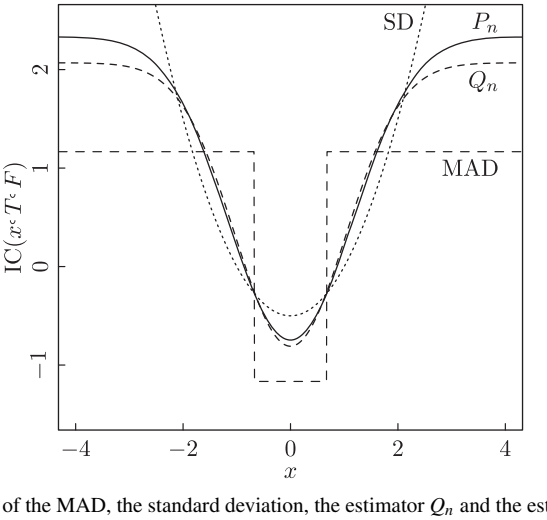

Figure 1. Influence curves of the MAD, the standard deviation, the estimator Q_n and the estimator P_n when the model distribution is Gaussian.

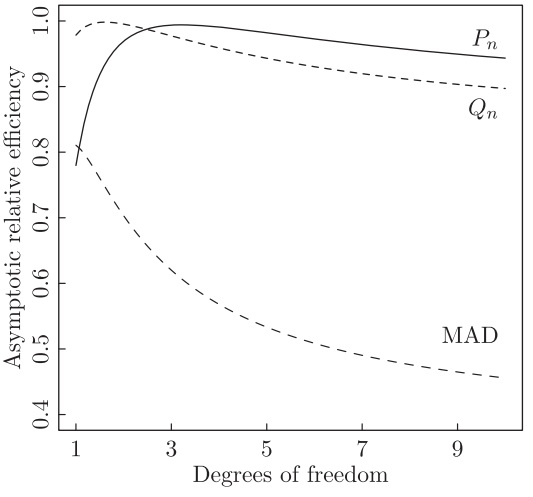

Figure 2. Asymptotic relative efficiency of the MAD, the estimator Q_n and the estimator P_n for t distributions with degrees of freedom ranging between 1 and 10.

than Q_n. This is reflected in our calculation of the asymptotic variance, which can be found as the expected square of the influence curve. Numerical integration yields

$$V = \int \mathrm{IC}(x, P_n, \Phi)^2 \, \mathrm{d}\Phi(x) = 0.579.$$

This results in an asymptotic efficiency of 0.86 when compared with 0.82 for Q_n and 0.37 for the MAD at the normal. Thus, despite having a higher gross-error sensitivity, P_n is a more efficient estimator than Q_n at the normal.

Figure 2 shows the relative efficiency of P_n, Q_n and the MAD at t distributions with degrees of freedom ranging between 1 (the Cauchy distribution) and 10. The asymptotic relative efficiencies are calculated as the product of the inverse of the Fisher information and the asymptotic variance. For t distributions with degrees of freedom more than approximately 2.5, P_n is more efficient than Q_n. At extremely heavy-tailed t distributions, including the Cauchy distribution, Q_n is more efficient than P_n. The MAD also performs better than P_n at the Cauchy; however, its efficiency decays substantially as the degrees of freedom increase.

At the exponential distribution, the asymptotic relative efficiency of P_n to Q_n is 0.77. Furthermore, the gross-error sensitivity for P_n is $\gamma^* = 3.968$, which compares with $\gamma^* = 2.317$ for Q_n at the exponential. However, it is not the case that P_n is necessarily worse than Q_n for skewed distributions. When the underlying distribution of the data is χ^2_1, the asymptotic relative efficiency of P_n to Q_n is 1.43.

The performance of P_n is more attractive than that of Q_n for discrete distributions. In the limit, Q_n will be equal to zero, and therefore fails to provide a valid estimate of scale, whenever more than 25% of the pairwise differences are equal to zero. For a discrete distribution with k distinct possible outcomes, $x_1, x_2, \ldots, x_k$, and probability mass function $P(X = x_j) = p_j$, for $j = 1, 2, \ldots, k$, the expected proportion of pairwise differences equal to zero is $\sum_j p_j^2$. In particular,

$$\sum_j p_j^2 > 0.25 \iff \lim_{n\to\infty} P(Q_n = 0) = 1.$$

For example, for the Poisson distribution with expected value 1, $\sum_j p_j^2 = 0.31$ and hence $Q_n \xrightarrow{p} 0$. In contrast, for P_n to return a scale estimate of zero, the interquartile range of the pairwise means

must be equal to zero, that is, more than 50% of the pairwise means must be equal. Except for trivial two-point distributions, it can be shown that $P_n = 0$ implies $Q_n = 0$. The pairwise averaging process helps smooth the underlying discrete distribution which results in P_n being a better robust estimator than Q_n in these situations.

3.3. *Asymptotic normality*

Note that P_n is a linear combination of two U-quantile statistics. To begin, consider the U-statistic

$$\binom{n}{2}^{-1} \sum_{i<j} h(X_i, X_j)$$

with symmetric kernel function $h : \mathbb{R}^2 \to \mathbb{R}$. A U-quantile statistic is a quantile of the distribution function of the kernels of the U-statistic. As in Section 2, let $H(t) = \mathrm{P}(h(X_i, X_j) \leq t)$ be the cdf of the kernels with the corresponding empirical distribution function,

$$H_n(t) := \frac{2}{n(n-1)} \sum_{i<j} \mathbb{I}\{h(X_i, X_j) \leq t\}, \quad \text{for } t \in \mathbb{R}. \tag{5}$$

For $0 < p < 1$, the corresponding U-quantile is

$$\xi_p = H^{-1}(p) := \inf\{t : H(t) \geq p\}.$$

The pth sample quantile $H_n^{-1}(p) := \inf\{t : H_n(t) \geq p\}$ is the pth quantile of the empirical distribution function of the set of dependent random variables

$$\{h(X_i, X_j), \quad 1 \leq i < j \leq n\}.$$

Note that Equation (5) is a U-statistic with kernel $\mathbb{I}\{h(X_i, X_j) \leq t\}$. The empirical U-process is defined as

$$\left(\sqrt{n}(H_n(t) - H(t))\right)_{t \in \mathbb{R}}.$$

Silverman (1976, Theorem B) proved that in this context, $\sqrt{n}(H_n(\cdot) - H(\cdot))$ converges weakly in the Skorohod topology to an almost sure continuous zero-mean Gaussian process W with covariance function

$$\mathbb{E}W(s)W(t) = 4\mathrm{P}(h(X_1, X_2) \leq s, h(X_1, X_3) \leq t) - 4H(s)H(t) \tag{6}$$

for all $s, t \in \mathbb{R}$.

Hence, for $0 < p < q < 1$, if H', the derivative of H, is strictly positive on the interval $[H^{-1}(p) - \varepsilon, H^{-1}(q) + \varepsilon]$ for some $\varepsilon > 0$, then we can use the inverse map to show

$$\sqrt{n}(H_n^{-1}(\cdot) - H^{-1}(\cdot)) \xrightarrow{\mathcal{D}} \frac{W(H^{-1}(\cdot))}{H'(H^{-1}(\cdot))},$$

where W is a mean zero Gaussian process with covariance function defined in Equation (6). See, for example, van der Vaart and Wellner (1996, Section 3.9.4.2). Stronger results concerning empirical processes of U-statistic structure are summarised in Serfling (2002).

In the case of $P_n(\tau)$, the kernel takes the form $h(X_i, X_j) = (X_i + X_j)/2$. Conditioning on $X_1 = x$, the covariance function (6) can be rewritten as follows:

$$\mathrm{Cov}(W(t), W(s)) = 4 \int F(2s - x)F(2t - x)\, \mathrm{d}F(x) - 4H(s)H(t),$$

where F is the distribution function of the underlying observations.

Recall that $H(\xi_p) = p$, the limiting covariance function for the quantiles used in the construction of $P_n(\tau)$, $n\text{Cov}(H_n^{-1}(a), H_n^{-1}(b))$, is

$$v(a,b) = 4\frac{\int F(2\xi_a - x)F(2\xi_b - x)\,\mathrm{d}F(x) - ab}{H'(\xi_a)H'(\xi_b)},$$

for $a = (1-\tau)/2$, $b = (1+\tau)/2$ and $0 < \tau < 1$. Hence,

$$\sqrt{n}(P_n(\tau) - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, c_\tau^2 V),$$

where $\theta = c_\tau(H^{-1}(b) - H^{-1}(a))$ and $V = v(a,a) + v(b,b) - 2v(a,b)$, both depend on the underlying distribution of the population.

Noting that the derivative of $H(t)$ is

$$H'(t) = \int 2f(2t-u)f(u)\,\mathrm{d}u,$$

it can be shown that the asymptotic variance found here is equivalent to the expected square of the influence function discussed previously.

Furthermore, from the general results in Serfling (2002, Section 12.3.4), the almost sure behaviour of $P_n(\tau)$ can be deduced from the Bahadur representation for the quantiles, $H_n^{-1}(\tau)$.

4. Relative efficiencies in finite samples

The asymptotic variance of P_n has been found and corresponding asymptotic efficiencies have been deduced in Section 3. In particular, it has been shown that, asymptotically, P_n is more efficient than Q_n for t distributions with more than approximately 2.5 degrees of freedom. However, it is often of practical interest to determine small sample relative efficiencies. This section considers the finite sample relative efficiency of P_n, and variants thereof, using configural polysampling. It also examines the finite relative efficiencies over the same range of t distributions considered asymptotically in Figure 2.

Under configural polysampling, estimators are evaluated at particular distributions chosen to exhibit more extreme characteristics than what might be observed in practice. For example, the Gaussian distribution has tails that die off rapidly, and the Cauchy has tails that die off extremely slowly. Such 'extreme' distributions will be referred to as *corners*. If it can be shown that an estimator performs well over all the corners considered, it is a fair assumption that the estimator will perform at least as well at intermediate distributions.

A key performance measure for estimators is the polyefficiency, the minimum efficiency that an estimator achieves over a selection of corners. Yatracos (1991) has shown that high polyefficiency over a finite selection of corners implies at least as high an efficiency at any convex combination of these corners.

Scale estimates are computed for samples of size n from each of Tukey's three corners: the Gaussian corner where observations are sampled independently from a standard normal distribution; the one wild corner, where $n-1$ observations are independent Gaussian and the remaining observation is scaled by a factor of 10; and the slash corner where observations are constructed as the ratio of an independent Gaussian random variable and an independent standard uniform random variable. The slash distribution has Cauchy-like tails, but it is considered to be more generally representative of real data as it is less peaked at the median than the Cauchy.

Improving on the methodology set out in Lax (1985), Randal (2008) proposed using ML scale estimates as the base case against which all estimators are compared. Previously, efficiencies were typically measured relative to the most efficient estimator considered for each distribution.

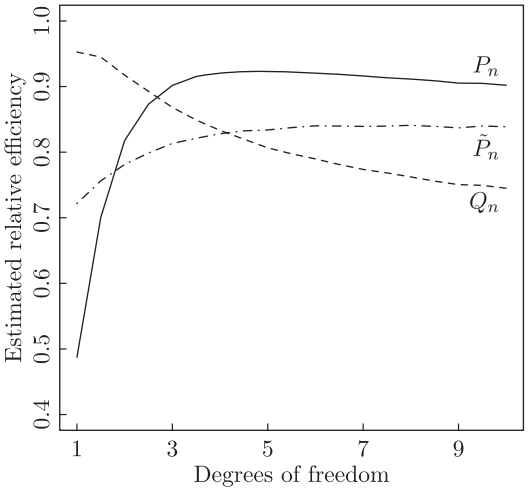

Figure 3. Finite sample relative efficiency, estimated from one million samples of size $n = 20$, for Q_n, P_n and adaptively trimmed P_n with tuning parameter $d = 5$, $\tilde{P}_n$.

ML estimates are asymptotically efficient and may be used as a common reference case in future simulation studies.

Efficiencies are estimated over m independent samples as

$$\widehat{\text{eff}}(T) = \frac{\widehat{\text{var}}(\ln \hat{\sigma}_1, \ldots, \ln \hat{\sigma}_m)}{\widehat{\text{var}}(\ln T(\mathbf{X}_1), \ldots, \ln T(\mathbf{X}_m))}, \tag{7}$$

where $\mathbf{X}_i$ are independent samples of size n, and for each $i = 1, 2, \ldots, m$, $\hat{\sigma}_i$ is the ML scale estimate and $T(\mathbf{X}_i)$ is the proposed scale estimate.

An alternative measure, based on the standardised variances, proposed by Rousseeuw and Croux (1993) was also considered and it gave results largely in agreement with the Lax relative efficiencies and is, therefore, not reported here. The key difference between these two measures of efficiency is the presence of a log transformation in Equation (7) which acts to stabilise the variance estimates. This is not present in the alternative measure, which, therefore, heavily penalises inefficient scale estimates in heavy-tailed data. Correction factors do not play a role in determining efficiency in either measure. Therefore, so far as efficiency is concerned, it is not an issue that the estimators are defined with consistency factors for the standard normal only.

Simulations were performed over numerous sample sizes in the range $20 \leq n \leq 100$. However, in the interests of space, only selected results for samples of size $n = 20$ are presented here.

Figure 3 presents the estimated relative efficiencies over t distributions with degrees of freedom ranging from 1 to 10 in increments of 0.5. The curves for P_n and Q_n are similar to the asymptotic efficiencies in Figure 2. Of particular interest is the speed with which P_n overtakes the adaptively trimmed P_n. For t distributions with 2 or more degrees of freedom, P_n is more efficient than $\tilde{P}_n$, the adaptively trimmed form of P_n with tuning parameter $d = 5$. Furthermore, for t distributions with 3 or more degrees of freedom, P_n is more efficient than Q_n.

Similar to what has been observed asymptotically, contrasting results are found in finite samples from skewed distributions. At the exponential, in samples of size $n = 20$, P_n is 0.87 times as efficient as Q_n, which is an improvement over the asymptotic result. Also at the χ^2_1 in samples of size 20, P_n is 1.21 times more efficient than Q_n.

Figure 4 presents the relative efficiencies of a variety of scale estimators at Tukey's three corner distributions. The estimators considered are P_n, the interquartile range of the pairwise means; $\tilde{P}_n$, the adaptively trimmed P_n with tuning parameter $d = 5$; $\breve{P}_n$, the symmetric fixed trimming where

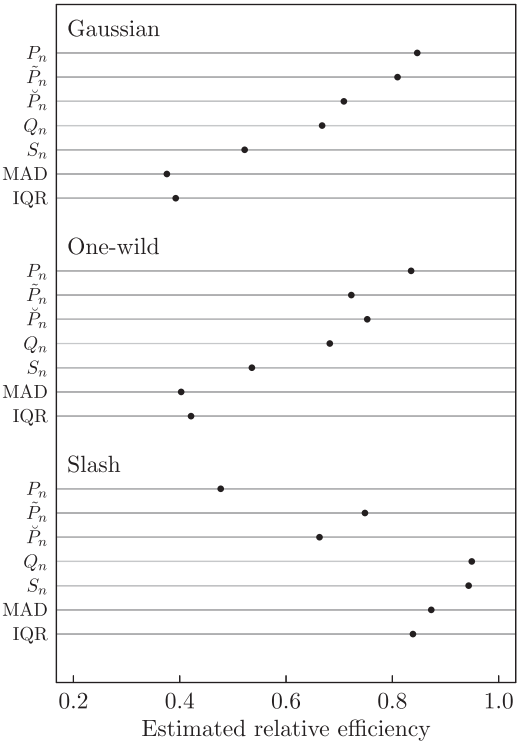

Figure 4. Estimated efficiencies relative to ML scale estimates for samples of size $n = 20$, calculated over one million independent samples. The various pairwise mean scale estimators are P_n, the interquartile range of the pairwise means; $\tilde{P}_n$, the adaptively trimmed P_n with tuning parameter $d = 5$; and $\check{P}_n$, the symmetric fixed trimming where 5% of the observations have been trimmed off from both tails of the original data.

5% of the observations have been trimmed off from both tails of the original data, for $n = 20$, this means that the maximum and minimum values have been deleted; Q_n and S_n (Rousseeuw and Croux 1993); the MAD; and the interquartile range. The estimator P_n is observed to be the most efficient at the normal and one wild corners; however, as at the Cauchy, it performs poorly at the slash corner. Among the estimators considered, $\tilde{P}_n$ with tuning parameter $d = 5$ is triefficient as it has the highest minimum efficiency across all three corners. Its lowest efficiency occurs at the one wild corner with 72%. However, as noted in Figure 3, the efficiency gain from using $\tilde{P}_n$ over P_n disappears as the tails become slightly less heavy.

A moderate amount of fixed trimming, for example, 5% at each tail as shown in Figure 4 does markedly improve the efficiency of P_n at the slash corner. However, it also compromises the efficiency of P_n at the normal. In other simulations that we have conducted, this result is not noticeably improved as the sample size increases. Therefore, adaptive trimming with a sensible choice of the tuning parameter is recommended over fixed trimming.

We have considered the impact that the tuning parameter d has on the efficiency of $\tilde{P}_n$. Intuitively, as d increases, more observations remain in the sample, which, for distributions that are not extremely heavy tailed, tends to increase efficiency. However, if the data are sampled from an extremely heavy-tailed distribution, the opposite tends be true – removing more of the observations in the tails actually increases the efficiency of the scale estimate.

Though not presented in Figure 4, random trimming of the kernels, that is, the pairwise means, achieves similar results as randomly trimming the original data at the normal and one wild corners; however, randomly trimming the original data leads to higher efficiency at the slash.

It is of interest also to consider the efficiency properties of the more general estimator $P_n(\tau)$ for various τ. At the normal corner, the efficiency of $P_n(\tau)$ tends to increase with τ. This pattern continues until $\tau \approx 0.9$ when the efficiency starts to drop off. At the one wild corner, the pattern is largely the same, although for $n = 20$ when $\tau > 0.75$, the estimator begins to break down. In contrast to the normal and one wild corners, increasing τ at the slash tends to decrease relative efficiency. These results confirm the choice of $P_n = P_n(0.5)$ as a reasonable default robust scale estimator.

The small sample performance of P_n is also better than that of Q_n for discrete distributions. Consider, for example, the case of a binomial distribution, $X \sim \mathcal{B}(6, 0.4)$. In the limit, neither Q_n nor P_n will converge to zero. However, if samples of size $n = 20$ are drawn from this distribution, Q_n will return a scale estimate of zero, on average, 12% of the time due to the discrete nature of the data. In contrast, P_n returns zero less than 0.1% of the time. Apart from some trivial cases, if $P_n = 0$ in finite samples, this implies that $Q_n = 0$.

Importantly, P_n possesses small sample efficiency that is comparable to that obtained asymptotically. It performs particularly well at the normal and one wild corners as well as heavy-tailed distributions such as the t distribution with degrees of freedom greater than approximately 2.5. Even though $\tilde{P}_n$ with a trimming parameter of $d = 5$ is triefficient among the scale estimators considered here, P_n would still be preferred as a scale estimator.

5. Concluding remarks

In this article, we have proposed a scale estimator based on the difference of two order statistics of the empirical distribution function of the pairwise means. Hence, $P_n(\tau)$ fits into the *GL*-statistic framework, alongside many other well-known scale estimators. Choosing $\tau = 0.5$ results in the estimator P_n, which possesses reasonable robustness properties, while maintaining high efficiencies and has an intuitive interpretation: the interquartile range of the pairwise means.

We have found that P_n, in its standard form, has a breakdown value of 13%. Its influence function more closely approximates that of the standard deviation at the normal, leading to a relatively high asymptotic efficiency of 86%. We have also found the gross-error sensitivity and demonstrated the asymptotic normality of P_n. Furthermore, when the underlying distribution is discrete, P_n is more robust to repeated observations than Q_n.

In finite samples, P_n also performs admirably. In samples of size $n = 20$, at the normal, P_n is 27% more efficient than Q_n and 22% more efficient at the one wild corner. P_n maintains its efficiency advantage over Q_n even when the underlying distribution of the data has quite heavy tails.

In summary, P_n is an alternative robust scale estimator that is not tailored to be most efficient at any particular distribution, rather it maintains high efficiency over a wide range of distributions.

References

Boente, G., Ruiz, M., and Zamar, R.H. (2010), 'On a Robust Local Estimator for the Scale Function in Heteroscedastic Nonparametric Regression', *Statistics and Probability Letters*, 80, 1185–1195.

Hampel, F. (1974), 'The Influence Curve and its Role in Robust Estimation', *Journal of the American Statistical Association*, 69, 383–393.

Hodges, J., and Lehmann, E.L. (1963), 'Estimates of Location Based on Rank Tests', *The Annals of Mathematical Statistics*, 34, 598–611.

Hyndman, R., and Fan, Y. (1996), 'Sample Quantiles in Statistical Packages', *American Statistician*, 50, 361–365.

Johnson, D.B., and Mizoguchi, T. (1978), 'Selecting the Kth Element in $X + Y$ and $X_1 + X_2 + \cdots + X_m$', *SIAM Journal on Computing*, 7, 147–153.

Lax, D. (1985), 'Robust Estimators of Scale: Finite-Sample Performance in Long-Tailed Symmetric Distributions', *Journal of the American Statistical Association*, 80, 736–741.

Randal, J. (2008), 'A Reinvestigation of Robust Scale Estimation in Finite Samples', *Computational Statistics & Data Analysis*, 52, 5014–5021.
Rousseeuw, P., and Croux, C. (1993), 'Alternatives to the Median Absolute Deviation', *Journal of the American Statistical Association*, 88, 1273–1283.
Serfling, R.J. (1984), 'Generalized *L*-, *M*-, and *R*-statistics', *The Annals of Statistics*, 12, 76–86.
Serfling, R.J. (2002), 'Robust Estimation via Generalized *L*-statistics: Theory, Applications, and Perspectives', in *Advances on Methodological and Applied Aspects of Probability and Statistics*, ed. N. Balakrishnan, New York: Taylor & Francis, pp. 197–217.
Silverman, B. (1976), 'Limit Theorems for Dissociated Random Variables', *Advances in Applied Probability*, 8, 806–819.
Stigler, S. (1977), 'Do Robust Estimators Work with Real Data?', *The Annals of Statistics*, 5, 1055–1098.
van der Vaart, A., and Wellner, J. (1996), *Weak Convergence and Empirical Processes*, Springer Series in Statistics, New York: Springer.
Wu, M., and Zuo, Y. (2008), 'Trimmed and Winsorized Standard Deviations Based on a Scaled Deviation', *Journal of Nonparametric Statistics*, 20, 319–335.
Wu, M., and Zuo, Y. (2009), 'Trimmed and Winsorized Means Based on a Scaled Deviation', *Journal of Statistical Planning and Inference*, 139, 350–365.
Yatracos, Y. (1991), 'A Note on Tukey's Polyefficiency', *Biometrika*, 78, 702–3.