Journal of Modern Applied Statistical Methods

Volume 9 | Issue 1 Article 5

5-1-2010

The Small-Sample Efficiency of Some Recently Proposed Multivariate Measures of Location

Marie Ng
University of Hong Kong, marieng@uw.edu

Rand R. Wilcox
University of Southern California, rwilcox@usc.edu

Journal of Modern Applied Statistical Methods
May 2010, Vol. 9, No. 1, 28-42

1538 – 9472/10/$95.00

The Small-Sample Efficiency of Some Recently Proposed Multivariate Measures of Location

Marie Ng
University of Hong Kong

Rand R. Wilcox
University of Southern California

Numerous multivariate robust measures of location have been proposed and many have been found to be unsatisfactory in terms of their small-sample efficiency. Several new measures of location have recently been derived, however, nothing is known about their small-sample efficiency or how they compare to the sample mean under normality. This research compared the efficiency for p = 2, 5, and 8 with sample sizes n = 20 and 50 for p-variate data. Although previous studies indicate that so-called skipped estimators are efficient, this study found that variations of this approach can perform poorly when n is small and p exceeds 5. One of the best estimators was found to be a skipped estimator where outliers detected by a projection method are eliminated. The TBS, OGK and RMBA estimators were included and; in some cases, they performed well, however, serious exceptions were identified suggesting that a skipped estimator based on a projection-type outlier detection method is preferable based on efficiency.

Key words: Robust methods, OGK estimator, TBS estimator, median ball algorithm, minimum generalized variance technique, projection methods, skipped estimators of location.

Introduction

A fundamental goal of this research is estimating some appropriate measure of location based on a random sample from some p-variate distribution. From basic principles, the sample mean has various optimal properties under normality; however, slight departures from normality can render it highly atypical and relatively inefficient. This has led to a variety of robust estimators, many of which are known to have relatively poor small-sample efficiency (Masse & Plante, 2003); thus study expands on Masse and Plante in several ways. First, recently proposed estimators are considered, next so-called skipped estimators are included, and lastly the present study is not limited to the bivariate case. In particular, the small-sample efficiency of the OGK estimator proposed by Maronna and Zamar (2002), the TBS (translated biweight) estimator derived by Rocke (1996) and the RMBA (median ball algorithm) suggested by Olive (2004, 2007) are examined. Skipped estimators simply mean that some appropriate multivariate outlier detection method is applied, any outliers found are removed and the mean of the remaining values is used as a measure of location.

This study considered two types of outlier detection methods. The first is based on a robust analog of Mahalanobis distance where the usual mean and covariance matrix are replaced by some robust measure of location and scale, respectively; in this case, the OGK, TBS and RMBA are considered. The second type does not use the Mahalanobis distance. One of the alternative strategies is based on a particular set of data projections in which a point is declared an outlier if it is flagged as an outlier by any projection. The other method, called the MGV method, belongs to this second class of techniques and assigns a measure of depth to points based in part on generalized variances of subsets of the data.

Marie Ng is an Assistant Professor in the Faculty of Education. Email: marieng@uw.edu. Rand R. Wilcox is a Professor of Psychology. Email: rwilcox@usc.edu.

Multivariate Outlier Detection Methods

Multivariate outlier detection methods play an integral role when using some of the

location estimators. Some basic concerns and results about multivariate outlier detection techniques are reviewed, and a description of the methods used in this research is provided. (See Wilcox (2008) for a more detailed comparison of the outlier detection methods.)

When choosing a multivariate outlier detection technique method at least two fundamental properties are of interest. The first is the outside rate per observation, which is the expected proportion of outliers among a sample of size n, for example, p_n. When sampling from a multivariate normal distribution, it is generally desirable to have a reasonably small p_n, for example 0.05; often methods are tuned to achieve this goal, at least when n is large (Rousseeuw & van Zomeren, 1990).

A second fundamental goal is to avoid masking. Roughly, a method is said to suffer from masking if the very presence of outliers causes them to be missed. Let M be some multivariate measure of location based on data randomly sampled from some p-variate distribution and let C be some measure of scatter. If M is the usual sample mean and C the usual covariance matrix based on $X_1,\ldots,X_n$, then a classic approach is to use the Mahalanobis distance

$$D_i = \sqrt{(X_i - M)C^{-1}(X_i - M)'} \qquad (1)$$

and declare X_i an outlier if D_i is sufficiently large. In particular, if the goal is to have $p_n = \alpha$, then X_i is declared an outlier if

$$D_i \geq \sqrt{\chi^2_{1-\alpha/2,p}}, \qquad (2)$$

the square root of the $1-\alpha/2$ quantile of a Chi-squared distribution with p degrees of freedom. It is known, however, that this method suffers from masking (Rousseeuw & Leroy, 1987), roughly because the usual sample mean and covariance matrix are not robust, that is, outliers can greatly influence their values thus causing D_i to be small even when X_i is highly atypical.

A seemingly natural approach to avoid masking is to take M and C to be some robust measure of location and scatter in equation (1) and then use equation (2). Campbell (1980) proposed using a particular M-estimator. The M-estimator Campbell used has a rather unsatisfactory breakdown point, however; the breakdown point of an estimator is the smallest proportion of points that must be altered to make it arbitrarily large or small. The M-estimator has a breakdown point of only 1/(p+1): this means that masking can be a problem - particularly as p gets large. Consequently, Rousseeuw and van Zomeren (1990) suggested using the minimum volume ellipsoid (MVE) estimator introduced by Rousseeuw (1985) and discussed in detail by Rousseeuw and Leroy (1987).

It appears that this method performs well in terms of achieving $p_n \approx .05$ (Wilcox, 2005); however, serious concerns have been expressed by Olive (2004) and Hawkins and Olive (2002). In addition, Fung (1993) described conditions where MVE can declare too many points outliers. Rousseeuw and van Driessen (1999) suggested replacing the MVE estimator with the fast minimum covariance determinant (FMCD) estimator, but with small to moderate sample sizes p_n becomes unstable and might exceed 0.05 by an unacceptable amount (Wilcox, 2005). At least three alternatives to the MVE and FMCD estimators exist and might be used instead.

The OGK Estimator

In its general form, the orthogonal Gnanadesikan-Kettenring (OGK) estimator, derived by Maronna and Zamar (2002), is applied as follows. Let $\sigma(X)$ and $\mu(X)$ be any measures of dispersion and location, respectively. The method proposed by Gnanadesikan and Kettenring (1972) begins with the robust covariance between any two variables, for example X and Y, is:

$$\mathrm{cov}(X,Y) = \frac{1}{4}(\sigma(X+Y)^2 - \sigma(X-Y)^2) \qquad (3)$$

When $\sigma(X)$ and $\mu(X)$ are the usual standard deviation and mean, the usual covariance between X and Y results. Following Maronna and Zamar (2002), $\sigma(X)$ is taken to be the tau scale of Yohai and Zamar (1988). Let

$$W_c(x) = \left(1 - \left(\frac{x}{c}\right)^2\right)^2 I(|x| \le c)$$

and

$$\rho_c(x) = \min(x^2, c^2),$$

where the indicator function $I(|x| \le c) = 1$ if $|x| \le c$ and 0 otherwise. For the univariate sample $X_1, \ldots, X_n$, let MAD(X) be the median of $|X_1 - M_x|, \ldots, |X_n - M_x|$, where M_x is the usual median of $X_1, \ldots, X_n$, and let

$$w_i = W_{4.5}\left(\frac{X_i - M_x}{MAD(X)}\right).$$

Again, following Maronna and Zamar (2002) the location and scale statistics are defined as

$$\mu(X) = \frac{\sum w_i X_i}{\sum w_i}$$

and

$$\sigma(X)^2 = \frac{MAD(X)}{n} \sum \rho_3\left(\frac{X_i - \mu(X)}{MAD(X)}\right)$$

Using the measure of scale in (3), the resulting measure of covariance will be denoted by $v(X, Y)$.

Following the notation in Maronna and Zamar (2002), let $\mathbf{x}_i$ be the i[th] row of the $n \times p$ matrix $\mathbf{X}$, a scatter matrix $\mathbf{V}(\mathbf{X})$ and a location vector $\mathbf{t}(\mathbf{X})$ are defined as follows:

1. Let $\mathbf{D} = diag(\sigma(X_1), \ldots, \sigma(X_p))$ and $\mathbf{y}_i = \mathbf{D}^{-1} x_i$, $i = 1, \ldots, n$.
2. Compute $\mathbf{U} = (U_{jk})$ by applying v to the $\mathbf{Y}$ columns.
3. Compute the eigenvectors $\mathbf{e}_j$ of $\mathbf{U}$ and let $\mathbf{E}$ be the matrix whose columns are the $\mathbf{e}_j$'s.
4. Let $\mathbf{A} = \mathbf{DE}$, $\mathbf{z}_i = \mathbf{A}^{-1}\mathbf{x}_i$, in which case $\mathbf{V} = \mathbf{A}\Gamma\mathbf{A}'$ and $\mathbf{t}(\mathbf{X}) = \mathbf{A}\upsilon$, where $\Gamma = diag(\sigma(Z_1)^2, \ldots, \sigma(Z_p))$ and $\upsilon = (\mu(Z_1), \ldots, \mu(Z_p))$.

Maronna and Zamar (2002) noted that the above procedure can be iterated and they report results suggesting that a single iteration be used. More precisely, compute $\mathbf{V}$ and $\mathbf{t}$ for $\mathbf{Z}$ (the matrix corresponding to $\mathbf{z}_i$ computed in step 4) and express them in the original coordinate system, namely, $\mathbf{V_2} = \mathbf{AV(Z)A}'$ and $\mathbf{t_2(X)} = \mathbf{At(Z)}$.

Maronna and Zamar showed that the estimate can be improved by a reweighting step. Let

$$d_i = \sum\left(\frac{z_{ij} - \mu(Z_j)}{\sigma(Z_j)}\right),$$

$$w_i = I_{d_i \le d_0},$$

and

$$d_0 = \frac{\chi^2_{p,\beta} med(d_1, \ldots, d_n)}{\chi^2_{p,.5}},$$

where $\chi^2_{p,\beta}$ is the β quantile of the Chi-squared distribution with p degrees of freedom and med denotes the median. The measure of location is now estimated to be

$$\mathbf{t}_w = \frac{\sum w_i \mathbf{x}_i}{\sum w_i},$$

and the measure of scatter is

$$\mathbf{V}_w = \frac{\sum w_i (\mathbf{x}_i - \mathbf{t}_w)(\mathbf{x}_i - \mathbf{t}_w)'}{\sum w_i}.$$

When using the OGK estimator to check for outliers in this study (2) was used. Results reported by Maronna and Zamar (2002) suggest using $\beta = .9$, but Wilcox (2008) found that this can result in p_n exceeding 0.05 by a considerable amount when n is small, moreover, p_n is unstable as a function of n. Thus,

$$\beta = \max(.95, \min(.99, 1/n + .94)),$$

was found to be more satisfactory and was therefore used in this research.

The TBS Estimator

Rocke (1996) proposed an estimator known as the translated-biweight S (TBS) estimator. Generally, S-estimators of multivariate location and scatter are values for $\hat{\theta}$ and $\mathbf{S}$ that minimize $|\mathbf{S}|$, the determinant of $\mathbf{S}$, subject to

$$\frac{1}{n}\sum \xi(((\mathbf{X}_i - \hat{\theta})'\mathbf{S}^{-1}(\mathbf{X}_i - \hat{\theta}))^{1/2}) = b_0, \quad (4)$$

where b_0 is some constant, and ξ is a non-decreasing function. However, Rocke (1996) showed that S-estimators can be sensitive to outliers even if the breakdown point is close to 0.5. He suggested an alternative approach where the function $\xi(d)$ is defined as follows: let m and c be values to be determined, then when $m \le d \le m + c$,

$$\xi(d) = \frac{m^2}{2} - \frac{m^2(m^4 - 5m^2c^2 + 15c^4)}{30c^4} + d^2\left(.5 + \frac{m^4}{2c^4} - \frac{m^2}{c^2}\right) + d^3\left(\frac{4m}{3c^2} - \frac{4m^3}{3c^4}\right) + d^4\left(\frac{3m^2}{2c^4} - \frac{1}{2c^2}\right) - \frac{4md^5}{5c^4} + \frac{d^6}{6c^4}$$

for $0 \le d < m$

$$\xi(d) = \frac{d^2}{2};$$

and for $d > m + c$,

$$\xi(d) = \frac{m^2}{2} + \frac{c(5c + 16m)}{30}.$$

The values for m and c can be chosen to achieve both the desired breakdown point and the asymptotic rejection probability, roughly referring to the probability that a point will get zero weight when the sample size is large. If the asymptotic rejection probability is γ, for example, then m and c are determined by

$$E_{\chi_p^2}(\xi(d)) = b_0$$

and

$$m + c = \sqrt{\chi^2_{p,1-\gamma}}.$$

An iterative estimation method was used to compute the measures of location and scatter (Rocke & Woodruff, 1993) which requires an initial estimate of location and scatter. Here the initial estimate is the FMCD estimator which was computed with the R function cov.mcd, but some results on using an alternative initial estimate are also mentioned herein. As with the OGK estimator, when using TBS checks for outliers are based on (2).

Median Ball Algorithm

Following Olive (2004, 2007), the median ball algorithm (RMBA) begins with two initial estimates of location and scatter, both of which are based on an iterative algorithm. The strategy is as follows. For the j[th] estimator (j = 1, 2), let $(T_{0,j}, \mathbf{C}_{0,j})$ be some starting value. Compute all n Mahalanobis distances $D_i(T_{0,j}, \mathbf{C}_{0,j})$ based on this measure location and scatter. Next estimate the usual mean and covariance matrix based on the $c_n \approx n/2$ cases corresponding to the smallest distances, this yields $(T_{1,j}, \mathbf{C}_{1,j})$. Repeating this process, which is based on $D_i(T_{1,j}, \mathbf{C}_{1,j})$, yields an updated measure of location and scatter, $(T_{2,j}, \mathbf{C}_{2,j})$; following Olive (2005, 2007) $(T_{5,j}, \mathbf{C}_{5,j})$ was used.

The first of the two starting values used by Olive takes $(T_1, \mathbf{C}_{0,1})$ to be the usual mean and covariance matrix. The other starting value, $(T_{0,2}, \mathbf{C}_{0,2})$, is the usual mean and covariance based on the c_n cases that are closest to the coordinatewise median in Euclidean distance. Let $(T_A, \mathbf{C}_A) = (T_{5,i}, \mathbf{C}_{5,i})$, where i = 1 if the determinant $|\mathbf{C}_{5,1}| \leq |\mathbf{C}_{5,2}|$, otherwise i = 2. The MBA estimator of location is T_A and the measure of scatter is

$$\mathbf{C_{MBA}} = \frac{MED(D_i^2(T_A, \mathbf{C}_A))}{\chi^2_{p,.5}} \mathbf{C}_A$$

To compute the RMBA estimate, first compute $D_i^2(T_{MBA}, \mathbf{C}_{MBA})$, then

1. Compute the classical estimator $(T, \mathbf{C})$ for the cases with $D_i^2 \leq \chi^2_{p,.975}$.
2. Scale for normality: let $T_1 = T$ and

$$\mathbf{C}_1 = \frac{MED(D_i^2(T, \mathbf{C}))}{\chi^2_{p,.5}}$$

Repeat steps 1 and 2 to obtain $(T_{RMBA}, \mathbf{C}_{RMBA})$. (The R function rmba available at www.math.siu.edu/olive/rpack.txt, computes the RMBA estimate of location and scatter and was used in the simulations.)

Wilcox (2008) found that if the Mahalanobis distance is computed using the RMBA estimator, and points are declared outliers using (2) with α = 0.975, the outside rate per observation is reasonably close to 0.05 under normality, provided that $n/p \geq 10$, at least for $2 \leq p \leq 12$; otherwise the outside rate per observation can be very unsatisfactory. For example, with n = 20 and p = 5 it was estimated to exceed 0.24 regardless of the correlation among the variables.

Thus, this approach is not as satisfactory compared to the OGK and TBS methods, but it was included for two reasons. First, the efficiency of the RMBA estimate of location, relative to the other methods considered, is unknown. Second, when applying the MGV method, an initial estimate of the center of a data cloud is required, and using RMBA appears to have a practical advantage in terms of controlling the outside rate per observation.

The Minimum Generalized Variance Method

From basic multivariate techniques, the generalized variance is the determinant of the usual covariance matrix; it reflects how tightly a cloud of points is clustered together. The minimum generalized variance (MGV) method is based on the fact that the generalized variance is not robust; a single unusual point can greatly inflate its value. The MGV method is applied as follows:

1. Initially, all n points are described as belonging to set A.
2. Find the p points that are most centrally located (many options exist to accomplish this). Based on results in Wilcox (2008), the approach used here takes the p most centrally located points to be the p points having the smallest Mahalanobis distance based on the RMBA estimators, T_A and $\mathbf{C}_{RMBA}$.
3. Remove the p centrally located points from set A and put them into set B. At this step, the generalized variance of the points in set B is zero.
4. If among the points remaining in set A, the i^{th} point is placed in set B, then the generalized variance of the points in set B will be changed to some value labeled s^2_{gi}, that is associated with every point remaining in A. The value s^2_{gi}, is the resulting generalized variance when it - and it only - is placed in set B. Compute s^2_{gi} for every point in A.
5. Among the s^2_{gi} values computed in the previous step, permanently remove the point associated with the smallest s^2_{gi} value from set A and put it in set B. That is, find the point in set A which is most tightly clustered

together with the points in set B; after this point is identified, permanently remove it from A and place it in B.

6. Repeat steps 4 and 5 until all points are now in set B.

The first p points removed from set A have a generalized variance of zero, this is labeled $s^2_{g(1)} = \cdots = s^2_{g(p)} = 0$. When each point is removed from A and put into B (using steps 3 and 4), the resulting generalized variance of set B is labeled $s^2_{g(p+1)}$, as this process continues each point has associated with it some generalized variance when it is put into set B. Based on this process, the i[th] point has associated with it one of the generalized variances computed. For convenience, this generalized variance associated with the i[th] point, $s^2_{g(j)}$, is labeled C_i.

The p deepest points have C values of zero. Points located at the edges of a scatterplot have the highest C values meaning that they are relatively far from the center of the cloud of points. A strategy for detecting outliers is simply applying some good univariate outlier rule to the C_i values. Note that a point would be declared only if an outlier C_i is large.

In terms of maintaining an outside rate per observation that is both stable as a function of n and p, and approximately equal to 0.05 under normality, a boxplot rule for detecting outliers seems best when p = 2, and for p > 2 a slight generalization of Carling's (2002) modification of the boxplot rule appears to perform well. In particular, if p = 2, then the i[th] point is declared an outlier if

$$C_i > q_2 + 1.5(q_2 - q_1), \qquad (5)$$

where q_1 and q_2 are the ideal fourths based on the C_i values. For p > 2 variables, the i[th] point is declared an outlier if

$$C_i > M_C + \sqrt{\chi^2_{.975,p}}(q_2 - q_1), \qquad (6)$$

where M_C is the usual median of the C_i values. (Thus, the inverse of a covariance matrix and Mahalanobis distance do not play a role when checking for outliers.)

A criticism, when detecting outliers among the C_i values, is that the interquartile range has a breakdown point of 0.25. Ideally, a univariate outlier detection method would have a breakdown point of 0.5, the highest possible value. This can be achieved with a commonly used MAD-median rule. When p = 2, for example, it means that a point $\mathbf{X}_i$ is declared an outlier if

$$\frac{|C_i - M_C|}{MAD_C} > 2.24, \qquad (7)$$

where MAD_C is the value of MAD based on the C values. The concern with this approach is that the outside rate per observation is no longer stable as a function of n and no method for correcting this problem is available at this time.

A Projection Method

Consider any projection of data onto a straight line. A projection-type method for detecting outliers among multivariate data is based on the idea that, if a point is an outlier, then it should be an outlier for some projection of the n points. Thus, if it were possible to consider all possible projections and, if for some projection a point is an outlier, then the point is declared an outlier. Not all projections can be considered, hence, following Wilcox (2005), the strategy is to orthogonally project the data onto all n lines formed by the center of the data cloud, as represented by $\hat{\xi}$, and each $\mathbf{X}_i$. Here, $\hat{\xi}$ was taken to be the RMBA measure of location. (Checks suggest that other choices for $\hat{\xi}$ have no practical value for the problem considered herein.)

The computational details are as follows. Fix i, and for the point $\mathbf{X}_i$, orthogonally project all n points onto the line connecting $\hat{\xi}$ and $\mathbf{X}_i$, and let D_{ij} be the

distance between $\hat{\xi}$ and $\mathbf{X}_j$ based on this projection. Let

$$\mathbf{A}_i = \mathbf{X}_i - \hat{\xi},$$

and

$$\mathbf{B}_j = \mathbf{X}_j - \hat{\xi},$$

where both $\mathbf{A}_i$ and $\mathbf{B}_j$ are column vectors having length p. Next let

$$\mathbf{C}_j = \frac{\mathbf{A}'\mathbf{B}_j}{\mathbf{B}_j},$$

where $j = 1, \ldots, n$. Then when projecting the points onto the line between $\mathbf{X}_i$ and $\hat{\xi}$, the distance of the j[th] point from $\hat{\xi}$ is

$$D_{ij} = \| \mathbf{C}_j \|,$$

where

$$\| \mathbf{C} \| = \sqrt{C_{1p}^2 + \cdots C_{jp}^2}.$$

Here, an extension of Carling's modification of the boxplot rule (similar to the modification used by the MGV method) is used to check for outliers among D_{ij} values. Let $\ell = [n/4 + 5/12]$, where [.] is the greatest integer function and let

$$h = \frac{n}{4} + \frac{5}{12} - \ell.$$

For fixed i, let $D_{i(1)} \leq \cdots \leq D_{i(n)}$ be the n distances written in ascending order.

If the ideal fourths associated with the D_{ij} values are

$$q_1 = (1-h)D_{i(\ell)} + hD_{i(\ell+1)}$$

and

$$q_2 = (1-h)D_{i(k)} + hD_{i(k-1)},$$

where $k = n - j + 1$, then the j[th] point is declared an outlier if

$$D_{ij} > M_D + \sqrt{\chi^2_{.975,p}}(q_2 - q_1), \qquad (8)$$

where M_D is the usual sample median based on $D_{i1}, \ldots, D_{in}$.

The process described is for a single projection; for fixed i, points are projected onto the line connecting $\mathbf{X}_i$ to $\hat{\xi}$. Repeating this process for each i, $i = 1, \ldots, n$, a point is declared an outlier if for any of these projections, it satisfies equation (8). This will be called method OP, which has certain similarities with a projection method suggested by Pena and Prieto (2001). One important difference is that the method used Pena and Prieto is based on the usual sample mean, which is not robust and could result in masking.

As was the case with the MGV method, a simple and seemingly desirable modification of the method described is to replace the interquartile range with the median absolute deviation (MAD) measure of scale based on the values $D_{i1}, \ldots, D_{in}$. Thus, if MAD is the median of the values $|D_{i1} - M_D|, \ldots, |D_{in} - M_D|$, which is denoted by MAD_i, then the j[th] point is declared an outlier if for any i,

$$D_{ij} > M_D + \sqrt{\chi^2_{.95,p}}\,\frac{MAD_i}{.6745} \qquad (9)$$

(Similar to the MGV method, equation (2) is not used when checking for outliers.) Equation (9) represents an approximation of the method given by Donoho and Gasko (1992).

An appealing feature of MAD is that it has a higher finite sample breakdown point than the interquartile range; however, a negative feature of equation (9) is that the outside rate per observation appears to be less stable as a function of n. In the bivariate case, for example, it is approximately 0.09 with n = 10 and drops below 0.02 as n increases. For the same situations, the outside rate per observation using equation (9) ranges, approximately, between 0.043 and 0.038.

Summary of the Estimators

In summary, eight alternatives to the sample mean were considered. The first three were RMBA, OGK and TBS. The remaining five are skipped estimators where outliers are removed after which the mean of the remaining data is computed. Three of these five estimators use (2) in conjunction with MVE, MCD and TBS and are denoted by MVE(S), MCD(S) and TBS(S); the other two use the MGV and OP outlier detection methods with the initial measure of location given by RMBA. For convenience, the estimators RMBA, OGK, MCD(S), OP, MVE(S), MGV and TBS(S) are labeled $\hat{\eta}_1,\ldots,\hat{\eta}_8$, respectively. The usual sample mean is labeled $\hat{\eta}_0$.

Results

Simulations were used to compare the efficiency of the sample mean to the eight alternative estimators. The efficiency of the j[th] estimator ($j = 1,\ldots,8$) was measured with

$$E = \frac{V(\hat{\eta}_j)}{V(\hat{\eta}_0)},$$

where $V(\hat{\eta}_j)$ is the generalized variance associated with the sampling distribution of $\hat{\eta}_j$. All simulations were conducted using the software R. Methods OP and MGV were applied with software from Wilcox (2005) that was downloaded from http://psychology.usc.edu/faculty_homepage.php?id=43. (The R function smean in Wilcox (2005) defaults to method OP. The R code for all estimators is available from the author upon request.)

To describe how data were generated, first consider the univariate case. An observation X from a g-and-h distribution (Hoaglin, 1985) is generated by first generating a value from a standard normal distribution yielding Z, for example, and computing

$$X = \frac{\exp(gZ)-1}{g}\exp(hZ^2/2)$$

where g and h are parameters that determine the third and fourth moments. When g = 0, this last equation is taken to be

$$X = Z\exp(hZ^2/2)$$

For the multivariate case, data were generated from a multivariate normal distribution having a common correlation, ρ, and the values of the marginal distributions were transformed to a g-and-h distribution. The four (marginal) g-and-h distributions used were the standard normal (g = h = 0), a symmetric heavy-tailed distribution (g = 0, h = 0.2), an asymmetric distribution with relatively light tails (g = 0.5, h = 0), and an asymmetric distribution with heavy tails (g = 0.5, h = 0.2). (For details about these distributions, see Hoaglin, 1985.) The values for ρ were taken to be 0, 0.5 and 0.8.

Tables 1-6 show the estimated efficiency of the eight estimators based on 1,000 replications. One method to condense the results in a useful way is to determine which robust estimator has the best efficiency among each of the 72 conditions studied. The OP estimator was best for 56 conditions and it was among the top two for 62 conditions. Another perspective considers which estimator competes best with the mean under normality; with two exceptions, this is method OP. The two exceptions occur when ρ = 0 and p = 5 or p = 8, in which case MGV is best.

With p = 5 the advantage of OP over MGV is not striking but with p=8 (and if ρ =0), MGV may have a worthwhile advantage. MGV is often among the two best estimators however, when sampling from a heavy-tailed distribution the mean can have better efficiency - sometimes strikingly so - even when other estimators beat the mean by a considerable amount. Although, RMBA, OGK and TBS do not compete well with OP in general, they can offer an advantage when p = 8, ρ = 0.5 or ρ = 0.8 and sampling is from a skewed, heavy-tailed distribution.

Conclusion

The success of the OP method is not surprising considering the results in detecting outliers

recently summarized in Wilcox (2008). Also based on results from Wilcox (2008), there was some anticipation that MGV would compete effectively with OP. Under some conditions it is a reasonable alternative, but it seems that, in terms of efficiency, the skipped estimator based on the OP outlier detection method is generally preferable, sometimes by a substantial amount. The poor performance of MGV when p = 8 and sampling is from a skewed, heavy-tailed distribution, was not expected. The OGK, TBS and RMBA estimators compete well with OP, particularly when sampling from a skewed, heavy-tailed distribution and $p \geq 5$, but for routine use, OP seems preferable and - for a variety of situations - it offers a distinct advantage.

Table 1: Estimated Efficiency for First Four Estimators, ρ =0

n	g	h	p	RMBA	OGK	TBS	MCD(S)
20	0	0	2	1.84	1.84	2.84	2.93
50	0	0	2	1.47	1.98	2.87	2.7
20	0	0	5	10.06	4.2	8.42	11.08
50	0	0	5	2.73	3.73	3.25	10.13
20	0	0	8	112.25	9.14	33.97	33.97
50	0	0	8	13.2	7.19	5.99	57.52
20	0	0.2	2	0.61	0.71	0.72	0.76
50	0	0.2	2	0.64	0.74	0.67	0.67
20	0	0.2	5	1.03	0.87	1.19	1.36
50	0	0.2	5	0.52	0.6	0.53	0.77
20	0	0.2	8	2.65	0.83	1.91	1.91
50	0	0.2	8	1.18	0.87	0.71	2.02
20	0.5	0	2	1.49	1.58	1.57	1.94
50	0.5	0	2	1.39	1.33	1.31	2.04
20	0.5	0	5	5.15	3.68	5.7	6.65
50	0.5	0	5	3.38	2.86	3.44	6.92
20	0.5	0	8	16.34	12.19	19.63	19.58
50	0.5	0	8	17.32	11.71	13.98	54.34
20	0.5	0.2	2	0.27	0.36	0.27	0.33
50	0.5	0.2	2	0.13	0.15	0.12	0.16
20	0.5	0.2	5	0.08	0.16	0.13	0.13
50	0.5	0.2	5	0.06	0.07	0.68	0.1
20	0.5	0.2	8	0.09	0.15	0.27	0.27
50	0.5	0.2	8	0.03	0.04	0.05	0.06

Table 2: Estimated Efficiency for First Four Estimators, ρ =.5

n	g	h	p	RMBA	OGK	TBS	MCD(S)
20	0	0	2	1.76	1.99	2.85	2.82
50	0	0	2	1.39	1.88	2.51	2.38
20	0	0	5	8.46	3.7	8.32	10.53
50	0	0	5	3.15	4.05	3.42	12.06
20	0	0	8	126.95	11.09	41.53	1.49
50	0	0	8	13.73	7.89	6.5	65.62
20	0	0.2	2	0.62	0.73	0.7	0.76
50	0	0.2	2	0.56	0.69	0.62	0.61
20	0	0.2	5	0.67	0.51	0.64	0.69
50	0	0.2	5	0.31	0.38	0.3	0.41
20	0	0.2	8	1.88	0.57	1.53	1.52
50	0	0.2	8	0.49	0.34	0.28	0.75
20	0.5	0	2	1.13	1.14	1.11	1.42
50	0.5	0	2	1.17	1.1	0.94	1.54
20	0.5	0	5	1.51	1.43	1.95	2.05
50	0.5	0	5	1.91	1.39	1.63	2.8
20	0.5	0	8	2.5	1.57	3.78	3.8
50	0.5	0	8	1.72	0.97	1.34	2.42
20	0.5	0.2	2	0.18	0.24	0.18	0.21
50	0.5	0.2	2	0.18	0.2	0.15	0.18
20	0.5	0.2	5	0.01	0.02	0.02	0.02
50	0.5	0.2	5	0.01	0.02	0.02	0.02
20	0.5	0.2	8	<.01	<.01	<.01	<.01
50	0.5	0.2	8	<.01	<.01	<.01	<.01

Table 3: Estimated Efficiency for First Four Estimators, $\rho = .8$

n	g	h	p	RMBA	OGK	TBS	MCD(S)
20	0	0	2	1.94	2.02	2.88	2.97
50	0	0	2	1.42	2.14	2.73	2.46
20	0	0	5	11.47	4.18	9.33	12.15
50	0	0	5	2.64	3.74	3.03	10.33
20	0	0	8	119.43	9.11	36.34	36.31
50	0	0	8	11.69	6.39	6.07	63.91
20	0	0.2	2	0.52	0.69	0.59	0.59
50	0	0.2	2	0.54	0.69	0.56	0.61
20	0	0.2	5	0.5	0.36	0.45	0.51
50	0	0.2	5	0.22	0.3	0.2	0.29
20	0	0.2	8	0.56	0.2	0.42	0.42
50	0	0.2	8	0.17	0.13	0.1	0.27
20	0.5	0	2	1.18	1.29	1.16	1.33
50	0.5	0	2	1.18	1.28	0.96	1.53
20	0.5	0	5	0.87	0.85	1.14	1.17
50	0.5	0	5	0.74	0.55	0.77	1.08
20	0.5	0	8	0.61	0.39	0.79	0.7
50	0.5	0	8	0.54	0.23	0.52	0.7
20	0.5	0.2	2	0.09	0.14	0.09	0.11
50	0.5	0.2	2	0.11	0.15	0.09	0.11
20	0.5	0.2	5	0.01	0.01	0.01	0.01
50	0.5	0.2	5	<.01	<.01	<.01	<.01
20	0.5	0.2	8	<.01	<.01	<.01	<.01
50	0.5	0.2	8	<.01	<.01	<.01	<.01

Table 4: Estimated Efficiency for Four Skipped Estimators, ρ =0

n	g	h	p	OP	MVE(S)	MGV	TBS(S)
20	0	0	2	1.36	2.48	1.47	1.92
50	0	0	2	1.32	1.91	1.5	1.53
20	0	0	5	2.22	6.98	1.49	8.03
50	0	0	5	1.86	3.48	1.36	2.39
20	0	0	8	3.54	12.64	2.21	34.25
50	0	0	8	2.74	9.55	1.93	4.63
20	0	0.2	2	0.56	0.64	0.61	0.59
50	0	0.2	2	0.56	0.67	0.56	0.56
20	0	0.2	5	0.45	0.84	0.91	0.84
50	0	0.2	5	0.49	0.71	0.93	0.58
20	0	0.2	8	0.52	2.55	2.09	3.28
50	0	0.2	8	0.39	0.87	1.18	0.58
20	0.5	0	2	1.02	1.86	1.12	1.46
50	0.5	0	2	1.21	2.22	1.35	1.55
20	0.5	0	5	1.56	4.36	1.66	5.49
50	0.5	0	5	0.45	0.66	0.8	0.5
20	0.5	0	8	0.51	2.27	1.87	2.95
50	0.5	0	8	0.39	0.7	1.05	0.57
20	0.5	0.2	2	0.2	0.25	0.26	0.22
50	0.5	0.2	2	0.21	0.25	0.24	0.23
20	0.5	0.2	5	0.06	0.19	1.21	0.18
50	0.5	0.2	5	0.05	0.11	1.18	0.09

Table 5: Estimated Efficiency for Four Skipped Estimators, ρ =.5

n	g	h	p	OP M	VE(S)	MGV	TBS(S)
20	0	0	2	1.31	2.55	1.36	0.174
50	0	0	2	1.32	2.1	1.63	1.56
20	0	0	5	1.43	6.12	1.52	8.28
50	0	0	5	1.33	3.27	1.42	2.27
20	0	0	8	1.3	11.32	2.5	38.73
50	0	0	8	1.18	7.73	1.97	3.77
20	0	0.2	2	0.52	0.65	0.54	0.54
50	0	0.2	2	0.47	0.52	0.5	0.47
20	0	0.2	5	0.33	0.69	1.07	0.71
50	0	0.2	5	0.26	0.43	0.79	0.3
20	0	0.2	8	0.26	1.16	1.83	1.15
50	0	0.2	8	0.22	0.44	1.12	0.29
20	0.5	0	2	0.98	1.46	1.09	1.3
50	0.5	0	2	0.98	1.5	1	1.22
20	0.5	0	5	0.69	1.55	0.142	1.69
50	0.5	0	5	0.74	1.76	1.56	1.45
20	0.5	0	8	0.84	3.31	1.86	3.65
50	0.5	0	8	0.7	1.84	1.93	2.18
20	0.5	0.2	2	0.16	0.2	0.18	0.17
50	0.5	0.2	2	0.12	0.13	0.14	0.13
20	0.5	0.2	5	0.02	0.04	1.19	0.03
50	0.5	0.2	5	0.01	0.01	0.92	0.01
20	0.5	0.2	8	<0.01	0.03	2.2	0.01
50	0.5	0.2	8	<0.01	0.02	1.36	<0.01

Table 6: Estimated Efficiency for Four Skipped Estimators, ρ =.8

n	g	h	p	OP	MVE(S)	MGV	TBS(S)
20	0	0	2	1.21	2.48	0.14	1.8
50	0	0	2	1.3	2.23	1.62	1.59
20	0	0	5	1.16	7.26	1.72	8.7
50	0	0	5	1.21	3.3	1.48	2.4
20	0	0	8	11.11	14.55	2.68	39.5
50	0	0	8	1.04	8.88	2.31	4.77
20	0	0.2	2	0.49	0.64	0.53	0.54
50	0	0.2	2	0.52	0.63	0.58	0.54
20	0	0.2	5	0.27	0.45	0.94	0.46
50	0	0.2	5	0.2	0.24	0.75	0.19
20	0	0.2	8	0.21	0.9	1.93	0.62
50	0	0.2	8	0.14	0.17	1.03	0.11
20	0.5	0	2	0.94	1.34	1.01	1.18
50	0.5	0	2	0.98	1.41	1.06	1.15
20	0.5	0	5	0.61	1.1	1.35	1.02
50	0.5	0	5	0.51	0.87	1.03	0.69
20	0.5	0	8	0.43	1.64	1.98	1.32
50	0.5	0	8	0.35	0.45	1.46	0.54
20	0.5	0.2	2	0.13	0.14	0.14	0.13
50	0.5	0.2	2	0.12	0.13	0.01	0.12
20	0.5	0.2	5	0.02	0.01	0.79	0.01
50	0.5	0.2	5	0.01	0.01	0.82	0.01
20	0.5	0.2	8	<.01	<.01	2.11	<.01
50	0.5	0.2	8	<.01	<.01	1.37	<.01

References

Carling, K. (2000). Resistant outlier rules and the non-Gaussian case. *Computational Statistics & Data Analysis*, *33*, 249-258.

Donoho, D. L., & Gasko, M. (1992). Breakdown properties of the location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, *20*, 1803-1827.

Fung, W. K. (1993). Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association*, *88*, 515-519.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics*, *28*, 81-124.

Hawkins, D. M., & Olive, D. (2002). Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm. *Journal of the American Statistical Association*, *97*, 136-147.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller & J. Tukey (Eds.) *Exploring Data Tables, Trends and Shapes*, 461-513. New York: Wiley.

Maronna, R. A., & Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, *44*, 307-317.

Masse, J. C., & Plante, J. F. (2003). A Monte Carlo study of the accuracy and robustness of ten bivariate location estimators. *Computational Statistics & Data Analysis*, *42*, 1-26.

Olive, D. J. (2004). A resistant estimator of multivariate location and dispersion. *Computational Statistics & Data Analysis*, *46*, 93-102.

Olive, D. J. (2007). Applied robust statistics. Unpublished manuscript. (www.math.siu.edu/olive/ol-bookp.htm).

Pena, D., & Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, *43*, 286-299.

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *Annals of Statistics*, *24*, 1327-1345.

Rocke, D. M., & Woodruff, D. L. (1993). Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica*, *47*, 27-42.

Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, *91*, 1047-1061.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug and W. Wertz (Eds), *Mathematical statistics and applications*, *B*., 283-297. Dordrecht: Reidel Publishing

Rousseeuw, P. J., & Leroy, A. M. (1987). Robust regression and outlier detection. New York: Wiley.

Rousseeuw, P. J., & Van Driesen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*, 212-223.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*, 633-639.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*, *2nd Ed*. San Diego CA: Academic Press.

Wilcox, R. R. (2008). Some small-sample properties of some recently proposed outlier detection techniques. *Journal of Statistical Computation and Simulation, 78,* 701-712

Yohai, V. J., & Zamar, R. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, *86*, 403-413.