Robustifying Robust Estimators

David J. Olive and Douglas M. Hawkins *

Southern Illinois University and University of Minnesota

January 12, 2007

Abstract

In the literature, estimators for regression or multivariate location and dispersion that have been shown to be both consistent and high breakdown are impractical to compute. This paper shows that a simple modification to existing concentration algorithms for multiple linear regression and multivariate location and dispersion results in high breakdown robust

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. Douglas M. Hawkins is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455-0493, USA. Their work was supported by the National Science Foundation under grants DMS 0202922, DMS 0600933, DMS 0306304, DMS 9803622 and ACI 9619020.

$\sqrt{n}$ consistent estimators that are easy to compute, and the applications for these estimators are numerous.

KEY WORDS: minimum covariance determinant estimator, multivariate location and dispersion, outliers, robust regression.

1 Introduction

The *multiple linear regression (MLR) model* is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients and $\boldsymbol{e}$ is an $n \times 1$ vector of errors. The ith case $(\boldsymbol{x}_i^T, y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$.

A *multivariate location and dispersion (MLD) model* is a joint distribution for a $p \times 1$ random vector $\boldsymbol{x}$ that is completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. The multivariate normal distribution is an important MLD model. The observations $\boldsymbol{x}_i$ for $i = 1, ..., n$ are collected in an $n \times p$ matrix $\boldsymbol{W}$ with n rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$.

In the literature there are many estimators for MLR and MLD that have been shown to be consistent and high breakdown (HB), but to our knowledge, none of these estimators is practical, computationally. Conversely, if the "robust estimator" for MLR or MLD is practical to compute, then it has not been shown to be both consistent and HB.

For example, the "state of the art" for robust MLD estimators are the FMCD estimator of Hawkins and Olive (1999), the Fast–MCD estimator of Rousseeuw and Van Driessen (1999), the OGK estimator of Maronna and Zamar (2002) and

the MBA estimator of Olive (2004). Hawkins and Olive (2002) proved that simplified versions of the FMCD and Fast–MCD estimators are inconsistent with zero breakdown. Maronna and Zamar (2002, p. 309) claim that it is straightforward to prove that the OGK estimator is consistent and HB, but fail to provide the proofs.

As another illustration, consider the cross checking estimator that uses a classical asymptotically efficient estimator if it is "close" to a consistent high breakdown robust estimator and uses the robust estimator otherwise. The resulting estimator is a high breakdown asymptotically efficient estimator. He and Wang (1997) show that the all elemental subset approximation to S estimators for MLD is consistent for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ for some constant $a > 0$. This estimator could be used as the robust estimator, but then the cross checking estimator is impractical. If the inconsistent zero breakdown FMCD algorithm is used as the robust estimator, then the resulting estimator is zero breakdown since both the "robust estimator" and the classical estimator are zero breakdown. This cross checking estimator is inconsistent since the probability that the "robust" and classical estimators are "close" does not go to one as the sample size $n \to \infty$.

Section 2 reviews the elemental basic resampling and concentration algorithms. Section 3 proves that the MBA estimator is a HB $\sqrt{n}$ consistent estima-

tor. The FMCD, Fast–MCD and least trimmed sum of squares (LTS) concentration algorithms are also modified so that they are HB $\sqrt{n}$ consistent estimators. Section 4 provides some examples.

2 Concentration Algorithms

Some notation is needed before describing concentration and elemental algorithms. Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \qquad (2.1)$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. The classical Mahalanobis distance corresponds to the sample mean and sample covariance matrix

$$T(\boldsymbol{W}) = \overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \text{ and } \boldsymbol{C}(\boldsymbol{W}) = \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})(\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})^{\mathrm{T}}.$$

Robust estimators are often computed by applying the classical estimator to a subset of the data. Consider the subset J_o of $c_n \approx n/2$ observations whose sample covariance matrix has the minimum determinant among all $C(n, c_n)$ subsets of

size c_n. Let T_{MCD} and $\boldsymbol{C}_{MCD}$ denote the sample mean and sample covariance matrix of the c_n cases in J_o. Then the minimum covariance determinant MCD(c_n) estimator is $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$. See Rousseeuw (1984).

Many high breakdown robust estimators are impractical to compute, so algorithm estimators are used instead. The "elemental basic resampling" algorithm for robust estimators uses K_n "elemental starts." For MLR an elemental set consists of p cases while an elemental set for MLD is a subset of $p+1$ cases where p is the number of variables. The jth elemental fit is a classical estimator ($\boldsymbol{b}_j$ or $(T_j, \boldsymbol{C}_j)$) computed from the jth elemental set. This fit is the jth start, and for each fit a criterion function that depends on all n cases is computed. Then the algorithm returns the elemental fit that optimizes the criterion.

Another important algorithm technique is *concentration*. Starts are again used, but they are not necessarily elemental. For multivariate data, let $(T_{0,j}, \boldsymbol{C}_{0,j})$ be the jth start and compute all n Mahalanobis distances $D_i(T_{0,j}, \boldsymbol{C}_{0,j})$. At the next iteration, the classical estimator $(T_{1,j}, \boldsymbol{C}_{1,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k steps resulting in the sequence of estimators $(T_{0,j}, \boldsymbol{C}_{0,j}), (T_{1,j}, \boldsymbol{C}_{1,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j}) = (\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j})$ is called the jth attractor. For MLR, let $\boldsymbol{b}_{0,j}$ be the jth start and compute all n residuals $r_i(\boldsymbol{b}_{0,j}) = y_i - \boldsymbol{b}_{0,j}^T \boldsymbol{x}_i$.

At the next iteration, a classical estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals. This iteration can be continued for k steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$. The result of the iteration $\boldsymbol{b}_{k,j}$ is called the jth attractor. The final concentration algorithm estimator is the attractor that optimizes the criterion. Using $k = 10$ concentration steps often works well, and the basic resampling algorithm is a special case with $k = 0$.

These algorithms are widely used in the literature, and the basic resampling algorithm can be used as long as the criterion can be computed. For most implementations, the number of elemental sets $K_n \equiv K$ does not depend on n. For a fixed data set with small p and an outlier proportion $\gamma < 0.5$, the probability that a clean elemental set is selected will be high if $K_n \equiv K \geq 3(2^d)$ where $d = p$ for MLR and $d = p + 1$ for MLD. Such estimators are sometimes called "high breakdown with high probability," although Hawkins and Olive (2002) showed that if $K_n \equiv K$, then the resulting elemental basic resampling estimator is inconsistent with zero breakdown, *regardless of the criterion.* Many authors, including Maronna and Yohai (2002) and Singh (1998), have mistaken "high breakdown with high probability" for "high breakdown."

Concentration algorithms for multivariate data have been suggested for the

MCD criterion. For multiple linear regression, concentration algorithms have been suggested for the LTS, least trimmed sum of absolute deviations (LTA) and least median of squares (LMS) criteria. The classical estimators used for these concentration algorithms are the ordinary least squares (OLS), least absolute deviations (L_1) and Chebyshev (L_∞) estimators, respectively. The notation CLTS, CLMS, CLTA and CMCD will be used to denote concentration algorithms for LTS, LMS, LTA and MCD, respectively. If $k > 1$, the jth attractor $\boldsymbol{b}_{k,j}$ has a criterion value at least as small as the criterion value for $\boldsymbol{b}_{1,j}$ for the CLTS, CLTA and CLMS algorithms. Rousseeuw and Van Driessen (1999) proved the corresponding result for the CMCD algorithm.

Some LTS concentration algorithms are described in Rousseeuw and Van Driessen (2000, 2002, 2006) and Víšek (1996). Salibian-Barrera and Yohai (2006) give a concentration type algorithm for S-estimators. Ruppert (1992) gives concentration estimators for LTS and LMS and also considers algorithms for regression S-estimators, the minimum volume ellipsoid (MVE) and MLD S-estimators. Hawkins and Olive (1999) suggest concentration algorithms for LMS, LTA, LTS, and MCD. A faster MVE algorithm is also given. Rousseeuw and Van Driessen (1999) give an MCD concentration algorithm. None of these algorithm estimators have been shown to be consistent or high breakdown.

The DGK estimator of MLD (Devlin, Gnanadesikan, and Kettenring 1975, 1981) uses the classical estimator computed from all n cases as the only start and Gnanadesikan and Kettenring (1972, pp. 94–95) provide a similar algorithm. The Olive (2004) *median ball algorithm* (MBA) estimator of MLD uses two starts $(T_{0,M}, \boldsymbol{C}_{0,M}) = (\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M})$ where $(\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M})$ is the classical estimator applied after trimming the $M\%$ of cases furthest in Euclidean distance from the coordinatewise median MED($\boldsymbol{W}$) where $M \in \{0, 50\}$. Then concentration steps are performed resulting in the Mth attractor $(T_{k,M}, \boldsymbol{C}_{k,M}) = (\overline{\boldsymbol{x}}_{k,M}, \boldsymbol{S}_{k,M})$. The $M = 0$ start is the classical estimator and the attractor is the DGK estimator. The $M = 50$ attractor is HB but generally inconsistent. Let $(T_A, \boldsymbol{C}_A)$ correspond to the attractor that has the smallest determinant. Then the MBA estimator $(T_{MBA}, \boldsymbol{C}_{MBA})$ takes $T_{MBA} = T_A$ and

$$\boldsymbol{C}_{MBA} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{2.2}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom. Olive (2002) shows that scaling the best attractor $\boldsymbol{C}_A$ results in a better estimate of $\boldsymbol{\Sigma}$ if the data is multivariate normal (MVN).

3 Practical Robust Estimators

Theorems 3, 4 and 5 below present practical HB $\sqrt{n}$ consistent estimators, but notation and preliminary results are needed. Following Lehmann (1999, pp. 53-54), recall that the sequence of random variables W_n is *tight* or *bounded in probability,* $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that $P(|W_n| \leq D_\epsilon) \geq 1 - \epsilon$ for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$. W_n has the same order as X_n in probability, written $W_n \asymp_P X_n$, if $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$.

If $W_n = \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| \asymp_P n^{-\delta}$ for some $\delta > 0$, then we say that both W_n and $\hat{\boldsymbol{\beta}}_n$ **have rate** n^δ. Similar notation is used for a $k \times r$ matrix $\boldsymbol{A} = [a_{i,j}]$ if each element $a_{i,j}$ has the desired property. For example, $\boldsymbol{A} = O_P(n^{-1/2})$ if each $a_{i,j} = O_P(n^{-1/2})$. Notice that if $W_n = O_P(n^{-\delta})$, then n^δ is a lower bound on the rate of W_n. As an example, if LMS, OLS or L_1 is used for $\hat{\boldsymbol{\beta}}$, then $W_n = O_P(n^{-1/3})$, but $W_n \asymp_P n^{-1/3}$ for LMS while $W_n \asymp_P n^{-1/2}$ for OLS and L_1.

Assumption (E1): Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with probability density function

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})]$$

where $k_p > 0$ is some constant, $\boldsymbol{\mu}$ is a $p \times 1$ location vector and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite matrix and g is some known function. Also assume that $\mathrm{Cov}(\boldsymbol{x}) = a_X \boldsymbol{\Sigma}$

for some constant $a_X > 0$. See Johnson (1987, pp. 107-108).

We will say that $\boldsymbol{x}$ is "spherical about μ" if $\boldsymbol{x}$ has an $EC_p(\boldsymbol{\mu}, c\boldsymbol{I}_p, g)$ distribution where $c > 0$ is some constant.

Remark 1. The following results from the literature will be useful for examining the properties of MLD and MLR estimators.

a) Butler, Davies and Jhun (1993): The MCD(c_n) estimator is a HB $\sqrt{n}$ consistent estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ where the constant $a_{MCD} > 0$ depends on the EC distribution.

b) Lopuhaä (1999): If $(T, \boldsymbol{C})$ is a consistent estimator for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate n^δ where the constants $a > 0$ and $\delta > 0$, then the classical estimator $(\overline{\boldsymbol{x}}_M, \boldsymbol{S}_M)$ computed after trimming $M\%$ (where $0 < M < 100$) of the cases with the largest distances $D_i(T, \boldsymbol{C})$ is a consistent estimator for $(\boldsymbol{\mu}, a_M\boldsymbol{\Sigma})$ with the same rate n^δ where $a_M > 0$ is some constant. Notice that applying the classical estimator to the $c_n \approx n/2$ cases with the smallest distances corresponds to $M = 50$. In the MLR setting, He and Portnoy (1992) consider applying OLS to the cases with the smallest squared residuals. Again the resulting estimator has the same rate as the start. Also see Ruppert and Carroll (1980, p. 834), Dollinger and Staudte (1991, p. 714) and Welsh and Ronchetti (2002).

c) Rousseeuw and Van Driessen (1999): Assume that the classical estimator

$(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$ is computed from c_n cases and that the n Mahalanobis distances $D_i \equiv D_i(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$ are computed. If $(\overline{\boldsymbol{x}}_{m+1,j}, \boldsymbol{S}_{m+1,j})$ is the classical estimator computed from the c_n cases with the smallest Mahalanobis distances D_i, then the MCD criterion $\det(\boldsymbol{S}_{m+1,j}) \leq \det(\boldsymbol{S}_{m,j})$ with equality iff $(\overline{\boldsymbol{x}}_{m+1,j}, \boldsymbol{S}_{m+1,j}) = (\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$.

d) Pratt (1959): Let K be a fixed positive integer and let the constant $d > 0$. Suppose that $(T_1, \boldsymbol{C}_1), ..., (T_K, \boldsymbol{C}_K)$ are K consistent estimators of $(\boldsymbol{\mu}, d\ \boldsymbol{\Sigma})$ each with the same rate n^δ. If $(T_A, \boldsymbol{C}_A)$ is an estimator obtained by choosing one of the K estimators, then $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, d\ \boldsymbol{\Sigma})$ with rate n^δ. Similarly, suppose that $\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_K$ are K consistent estimators of $\boldsymbol{\beta}$ each with the same rate n^δ. If $\hat{\boldsymbol{\beta}}_A$ is an estimator obtained by choosing one of the K estimators, then $\hat{\boldsymbol{\beta}}_A$ is a consistent estimator of $\boldsymbol{\beta}$ with rate n^δ.

e) Olive (2002): Suppose that $(T_i, \boldsymbol{C}_i)$ are consistent estimators for $(\boldsymbol{\mu}, a_i\boldsymbol{\Sigma})$ where $a_i > 0$ for $i = 1, 2$. Let $D_{i,1}$ and $D_{i,2}$ be the corresponding distances and let R be the set of cases with distances $D_i(T_1, \boldsymbol{C}_1) \leq \text{MED}(D_i(T_1, \boldsymbol{C}_1))$. Let r_n be the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in R. Then $r_n \to 1$ in probability as $n \to \infty$.

f) Olive (2004): $(\overline{\boldsymbol{x}}_{0,50}, \boldsymbol{S}_{0,50})$ is a high breakdown estimator. If the data distribution is EC but not "spherical about μ," then for $m \geq 0$, $\boldsymbol{S}_{m,50}$ underes-

timates the major axis and overestimates the minor axis of the highest density region. Concentration reduces but fails to eliminate this bias. Hence the estimated highest density hyperellipsoid based on the attractor is "shorter" in the direction of the major axis and "fatter" in the direction of the minor axis than estimated regions based on consistent estimators. Also, see Croux and Van Aelst (2002). Arcones (1995) and Kim (2000) showed that $\overline{\boldsymbol{x}}_{0,50}$ is a HB $\sqrt{n}$ consistent estimator of $\boldsymbol{\mu}$.

For MLR, if the start is a consistent estimator for $\boldsymbol{\beta}$, then so is the attractor if OLS is used. Hence He and Portnoy (1992) can be used with Pratt (1959) to provide simple proofs for MLR concentration algorithms. The following proposition shows that if $(T, \boldsymbol{C})$ is a consistent start, then the attractor is a consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Since $a_M \equiv a_{MCD}$ does not depend on a, the population parameter estimated by MLD concentration algorithms is $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence Lopuhaä (1999) can be used with Pratt (1959) with $d = a_{MCD}$ to provide simple proofs for MLD concentration algorithms.

Proposition 1. *Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate n^{δ} where the constants $a > 0$ and $\delta > 0$, then the classical estimator $(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$ computed after trimming the $c_n \approx n/2$ of cases with the largest distances $D_i(T, \boldsymbol{C})$ is a consistent estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with*

the same rate n^{δ}.

Proof. The result follows by Remark 1b if $a_{50} = a_{MCD}$. But by Remark 1e the overlap of cases used to compute $(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$ and $(T_{MCD}, \boldsymbol{C}_{MCD})$ goes to 100% as $n \to \infty$. Hence the two sample covariance matrices $\boldsymbol{S}_{m,j}$ and $\boldsymbol{C}_{MCD}$ both estimate the same quantity $a_{MCD}\boldsymbol{\Sigma}$. □

The following proposition shows that it is very difficult to drive the determinant of the dispersion estimator from a concentration algorithm to zero. Olive (2004) showed that the largest eigenvalue λ_1 of $\boldsymbol{S}_{0,50}$ is bounded above by $p \max |s_{i,j}|$ where $s_{i,j}$ is the (i,j) entry of $\boldsymbol{S}_{0,50}$. Hence the smallest eigenvalue λ_p is bounded below by $\det(\boldsymbol{C}_{\mathrm{MCD}})/\lambda_1^{\mathrm{p}-1}$.

Proposition 2. *Consider the CMCD and MCD estimators that both cover* c_n *cases. For multivariate data, if at least one of the starts is nonsingular, then the CMCD estimator* $\boldsymbol{C}_A$ *is less likely to be singular than the high breakdown MCD estimator* $\boldsymbol{C}_{MCD}$.

Proof. If all of the starts are singular, then the Mahalanobis distances cannot be computed and the classical estimator can not be applied to c_n cases. Suppose that at least one start was nonsingular. Then $\boldsymbol{C}_A$ and $\boldsymbol{C}_{MCD}$ are both sample covariance matrices applied to c_n cases, but by definition $\boldsymbol{C}_{MCD}$ minimizes the determinant of such matrices. Hence $0 \leq \det(\boldsymbol{C}_{\mathrm{MCD}}) \leq \det(\boldsymbol{C}_{\mathrm{A}})$. □

The following theorem shows that the MBA estimator has good statistical properties.

Theorem 3. *Suppose (E1) holds. If* $(T_A, \boldsymbol{C}_A)$ *is the attractor that minimizes* $det(\boldsymbol{S}_{k,M})$*, then* $(T_A, \boldsymbol{C}_A)$ *is a HB* $\sqrt{n}$ *consistent estimator of* $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$*. Hence the MBA estimator is a HB* $\sqrt{n}$ *consistent estimator.*

Proof. The estimator is HB since $0 < \det(\boldsymbol{C}_{\mathrm{MCD}}) \leq \det(\boldsymbol{C}_{\mathrm{A}}) \leq \det(\boldsymbol{S}_{0,50}) < \infty$ if up to nearly 50% of the cases are outliers. If the distribution is spherical about μ, then the result follows from Remark 1b since both starts are $\sqrt{n}$ consistent. Otherwise, the estimator with $M = 50$ trims too much data in the direction of the major axis and hence the resulting attractor is not estimating the highest density region. Hence $\boldsymbol{S}_{k,50}$ is not estimating $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{S}_{k,0}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ and $\|\boldsymbol{C}_{MCD} - \boldsymbol{S}_{k,0}\| = O_P(n^{-1/2})$. Thus the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0})$. □

The following theorem shows that fixing the inconsistent zero breakdown elemental CMCD algorithm is simple. Just add the two MBA starts.

Theorem 4. *Suppose (E1) holds and that the CMCD algorithm uses* $K_n \equiv K$ *randomly selected elemental starts (e.g.,* $K = 200$ *or 0), the start* $(T_{0,0}, \boldsymbol{C}_{0,0})$ *and*

the start $(T_{0,50}, \boldsymbol{C}_{0,50})$. *Then this CMCD estimator is a HB* $\sqrt{n}$ *consistent estimator. If the EC distribution is not spherical about* μ*, then the CMCD estimator is asymptotically equivalent to the DGK estimator.*

Proof. The estimator is HB since $0 < \det(\boldsymbol{C}_{\mathrm{MCD}}) \leq \det(\boldsymbol{C}_{\mathrm{CMCD}}) \leq \det(\boldsymbol{S}_{0,50}) < \infty$ if up to nearly 50% of the cases are outliers. Notice that the DGK estimator is the attractor for $(T_{0,0}, \boldsymbol{C}_{0,0})$. Under (E1), the probability that the attractor from a randomly drawn elemental set gets arbitrarily close to the MCD estimator goes to zero as $n \to \infty$. But $DGK - MCD = O_P(n^{-1/2})$. Since the number of randomly drawn elemental sets K does not depend on n, the probability that the DGK estimator has a smaller criterion value than that of the best elemental attractor also goes to one. Hence if the distribution is spherical about μ, then (with probability going to one) one of the MBA attractors will minimize the criterion value and the result follows. If (E1) holds and the distribution is not spherical about μ, then the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_{CMCD}, \boldsymbol{C}_{CMCD})$ is asymptotically equivalent to the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0})$. □

The following theorem shows that the inconsistent zero breakdown elemental CLTS estimator can be fixed by adding two carefully chosen attractors. Hawkins and Olive (2002) suggested adding a classical estimator as a start and Maronna

and Yohai (2002) claim (without proof) that the resulting estimator is consistent. If the algorithm evaluates the criterion on trial fits, then these fits will be called the attractors. Using OLS as an attractor instead of a start results in estimator with 100% Gaussian asymptotic efficiency. Let $\boldsymbol{b}_k$ be the attractor from the start consisting of OLS applied to the c_n cases with y's closest to the median of the y_i and let $\hat{\boldsymbol{\beta}}_{k,B} = 0.99\boldsymbol{b}_k$. Then $\hat{\boldsymbol{\beta}}_{k,B}$ is a HB biased estimator of $\boldsymbol{\beta}$ (biased if $\boldsymbol{\beta} \neq \mathbf{0}$, see Olive 2005).

Theorem 5. *Suppose that the CLTS algorithm uses $K_n \equiv K$ randomly selected elemental starts (e.g., $K = 500$) and the attractors $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{k,B}$. Then the resulting estimator is a HB $\sqrt{n}$ consistent estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{OLS}$.*

Proof. Olive (2005) showed that an MLR estimator is high breakdown if the median absolute residual stays bounded under high contamination. (Notice that if $\|\hat{\boldsymbol{\beta}}\| = \infty$, then the MED$(|r_i|) = \infty$, and if $\|\hat{\boldsymbol{\beta}}\| = M$ then MED$(|r_i|)$ is bounded if fewer than half of the cases are outliers.)

Concentration insures that the criterion function of the $c_n \approx n/2$ absolute residuals gets smaller. Hence LTS concentration algorithms that use a HB start are HB, and $\hat{\boldsymbol{\beta}}_{k,B}$ is a HB estimator.

The LTS estimator is consistent by Mašiček (2004), Čížek (2006) or Víšek

(2006). The LTS criterion is $Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b})$ where $r^2_{(i)}$ are the ordered squared residuals. As $n \to \infty$, consistent estimators $\hat{\boldsymbol{\beta}}$ satisfy $Q_{LTS}(\hat{\boldsymbol{\beta}})/n - Q_{LTS}(\boldsymbol{\beta})/n \to 0$ in probability. Since $\hat{\boldsymbol{\beta}}_{k,B}$ is a biased estimator of $\boldsymbol{\beta}$, with probability tending to one, OLS will have a smaller criterion value. With probability tending to one, OLS will also have a smaller criterion value than the criterion value of the attractor from a randomly drawn elemental set (by He and Portnoy 1992, also see Remark 4 in Hawkins and Olive 2002). Since K random elemental sets are used, the CLTS estimator is asymptotically equivalent to OLS. □

Remark 2. a) Basic resampling algorithms that uses a HB MLR criterion that is minimized by a consistent estimator for $\boldsymbol{\beta}$ (e.g., for LMS or LTS) can be fixed. Assume that the new algorithm uses $K_n \equiv K$ randomly selected elemental starts, the start $\hat{\boldsymbol{\beta}}_{OLS}$ and the start $\hat{\boldsymbol{\beta}}_{k,B}$. The resulting HB estimator is asymptotically equivalent to the OLS estimator if the OLS estimator is a consistent estimator of $\boldsymbol{\beta}$. The proof is nearly identical to that of the proof for Theorem 5. Note that the resulting LMS algorithm estimator is asymptotically equivalent to OLS, not to LMS.

b) From the proof of the Theorem 5, it can be seen that the OLS attractor can be replaced by any $\sqrt{n}$ consistent estimator, say $\hat{\boldsymbol{\beta}}_D$, and the resulting estimator will be a HB $\sqrt{n}$ consistent estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_D$. To

obtain an easily computed HB MLD estimator with 100% Gaussian asymptotic efficiency, use the MBA and classical estimators in the cross checking estimator.

c) To robustify the Rousseeuw and Van Driessen (1999, 2006) Fast–MCD and Fast–LTS algorithms, which use 500 starts, partitioning, iterates 5 starts to convergence, and then a reweight for efficiency step, consider the following argument. Add the consistent and high breakdown biased attractors to the algorithm. Suppose the data set has n_D cases. Then the maximum number of concentration steps until convergence is bounded by k_D, say. Assume that for $n > n_D$, no more than k_D concentration steps are used. (This assumption is reasonable. Asymptotic theory is meant to simplify matters, not to make things more complex. Also the algorithm is supposed to be fast. Letting the maximum number of concentration steps increase to ∞ would result in an impractical algorithm.) Then the elemental attractors are inconsistent so the probability that the MCD or LTS criterion picks the consistent estimator goes to one. The "reweight for efficiency step" does not change the $\sqrt{n}$ rate by Lopuhaä (1999) or He and Portnoy (1992).

d) The notation "CLTS" means that the attractors were evaluated using the LTS criterion; however, the CLTS estimator in not estimating the LTS estimator, but is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{OLS}$. Similarly, the CMCD estimator given by Theorem 4 is not estimating the MCD estimator, but is asymptotically

equivalent to the DGK estimator.

4 Examples and Simulations

Suppose that the concentration algorithm covers c_n cases. Then Hawkins and Olive (2002) suggested that concentration algorithms using K starts each consisting of h cases can handle roughly a percentage γ_o of huge outliers where

$$\gamma_o \approx \min(\frac{n-c_n}{n}, 1-[1-(0.2)^{1/K}]^{1/h})100\% \tag{3.1}$$

if n is large. Empirically, this value seems to give a rough approximation for many simulated data sets.

However, if the data set is multivariate and the bulk of the data falls in one compact ellipsoid while the outliers fall in another hugely distant compact ellipsoid, then a concentration algorithm using a single start can sometimes tolerate nearly 25% outliers. For example, suppose that all $p+1$ cases in the elemental start are outliers but the covariance matrix is nonsingular so that the Mahalanobis distances can be computed. Then the classical estimator is applied to the $c_n \approx n/2$ cases with the smallest distances. Suppose the percentage of outliers is less than 25% and that all of the outliers are in this "half set." Then the sample mean applied to the c_n cases should be closer to the bulk of the data than to the

cluster of outliers. Hence after a concentration step, the percentage of outliers in the c_n cases will be reduced if the outliers are very far away. After the next concentration step the percentage of outliers will be further reduced and after several iterations, all c_n cases will be clean. (For outliers of this type, using $c_n \approx 2n/3$ might be able to handle an outlier percentage near 33%.)

Rocke and Woodruff (1996) suggest that the hardest shape that outliers can take is when they have the same covariance matrix as the clean data but shifted mean. In simulations, estimators based on concentration estimators were much more effective on such data sets than estimators based on the basic resampling algorithm.

The Rousseeuw and Van Driessen (1999) DD plot is a plot of classical versus robust Mahalanobis distances and is very useful for detecting outliers. In a small simulation study, 20% outliers were planted for various values of p. If the outliers were distant enough, then the minimum DGK distance for the outliers was larger than the maximum DGK distance for the nonoutliers, and thus the outliers were separated from the bulk of the data in the DD plot. For example, when the clean data comes from the $N_p(\mathbf{0}, \boldsymbol{I}_p)$ distribution and the outliers come from the $N_p(2000\ \mathbf{1}, \boldsymbol{I}_p)$ distribution, the DGK estimator with 10 concentration steps was able to separate the outliers in 17 out of 20 runs when $n = 9000$ and $p = 30$.

With 10% outliers, a shift of 40, $n = 600$ and $p = 50$, 18 out of 20 runs worked. Olive (2004) showed similar results for the Rousseeuw and Van Driessen (1999) Fast–MCD algorithm and that the MBA estimator could often correctly classify up to 49% hugely distant outliers.

We examined several data sets from (www.math.siu.edu/olive/ol-bookp.htm) to illustrate the DGK, MBA and Fast–MCD estimators. For each data set the d outliers were deleted and then made the first d cases in the data set. Then the last $n - m$ cases were deleted so that the outliers could not be detected in the DD plot. The Buxton (1920) data `cyp.lsp` consists of measurements *bigonal breadth*, *cephalic index*, *head length*, *height* and *nasal height*. There were 76 cases and cases 61–65 had heights about 0.75 inches with head lengths well over 5 feet. The DGK, Fast–MCD and MBA estimators failed when there were 21, 14 and 10 cases remaining, respectively.

The Gladstone (1905-6) data consists of the variables *age*, *ageclass*, *breadth*, *brnweight*, *cause*, *cephalic*, *circum*, *head height*, *height*, *length*, *sex* and *size*. There were 267 cases and cases 230, 254, 255, 256, 257 and 258 were outliers corresponding to infants. The variables *ageclass*, *cause* and *sex* were categorical and caused the Fast–MCD estimator to be singular. Hence these three variables were deleted and there were 6 outliers and 9 variables. The DGK, Fast–MCD and MBA esti-

mators failed when there were 30, 20 and 18 cases remaining, respectively.

The Schaaffhausen (1878) data `museum.lsp` consists of the variables *head length*, *head breadth*, *head height*, *lower jaw length*, *face length*, *upper jaw length*, *height of lower jaw*, *eye width*, *traverse diagonal length* and *cranial capacity*. There were 60 cases and the first 47 were humans while the remaining 13 cases were apes (outliers). The DGK, Fast–MCD and MBA estimators failed when there were 38, 34 and 26 cases remaining, respectively.

All three estimators gave similar DD plots when all of the cases were used and the DGK estimator had considerable outlier resistance. For MLD, concentration is a very effective technique even if the classical estimator is used as the only start. For two of the data sets, the MBA estimator failed when the number of outliers was equal to the number of clean cases, as might be expected from a HB estimator.

5 Conclusions

The literature on HB MLR or MLD estimators has major flaws: estimators that have been shown to be HB and consistent are impractical to compute, while estimators that are practical to compute have not been shown to be both HB and consistent. This paper has shown that it is possible to create estimators that

have attractive theory and that are easy to compute, and multivariate robust statistics can now be used for inference as well as outlier detection.

The MBA estimator has good outlier resistance, is asymptotically equivalent to a scaled DGK estimator and is about two orders of magnitude faster than the inconsistent FMCD estimator. The RMBA estimator that uses two "reweight for efficiency steps" could also be used and is HB and $\sqrt{n}$ consistent by Lopuhaä (1999).

The CLTS estimator is HB with 100% Gaussian asymptotic efficiency, but the outlier resistance is not much better than that of FLTS. The CLTS estimator is affine equivariant but not regression equivariant. The MBA estimator is not affine equivariant. The properties of computability, outlier resistance and $\sqrt{n}$ consistency are far more important than the property of equivariance, and the CLTS and CMCD estimators are asymptotically equivalent to equivariant estimators.

Estimators can easily be made affine equivariant if they are allowed to depend on the initial data collected in an $n \times p$ matrix $\boldsymbol{W}$ (and practical elemental "robust estimators" depend on $\boldsymbol{W}$ since they are not permutation invariant). To see this, let $\boldsymbol{B} = \mathbf{1}\boldsymbol{b}^T$ where $\mathbf{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{b}$ is a $p \times 1$ constant vector. Hence the ith row of $\boldsymbol{B}$ is $\boldsymbol{b}_i^T \equiv \boldsymbol{b}^T$ for $i = 1, ..., n$. Consider

the affine transformation $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix. Let $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ be the MBA estimator. Define $(T_{\boldsymbol{W}}(\boldsymbol{Z}), \boldsymbol{C}_{\boldsymbol{W}}(\boldsymbol{Z}))$ by $T_{\boldsymbol{W}}(\boldsymbol{W}) = T(\boldsymbol{W})$ and $\boldsymbol{C}_{\boldsymbol{W}}(\boldsymbol{W}) = \boldsymbol{C}(\boldsymbol{W})$, and $T_{\boldsymbol{W}}(\boldsymbol{Z}) = \boldsymbol{A}^T T(\boldsymbol{W}) + \boldsymbol{b} = \boldsymbol{A}^T T_{\boldsymbol{W}}(\boldsymbol{W}) + \boldsymbol{b}$ and $\boldsymbol{C}_{\boldsymbol{W}}(\boldsymbol{Z}) = \boldsymbol{A}^T \boldsymbol{C}(\boldsymbol{W})\boldsymbol{A} = \boldsymbol{A}^T \boldsymbol{C}_{\boldsymbol{W}}(\boldsymbol{W})\boldsymbol{A}$. Thus $(T_{\boldsymbol{W}}, \boldsymbol{C}_{\boldsymbol{W}})$ is affine equivariant.

The applications of the estimators from Theorems 3–5 and Remark 2 are numerous. For example, they can be used to robustify the "robust estimators" for multivariate techniques and generalized linear models that use Fast–MCD as a "plug in" estimator. The MBA estimator can also be used to create an easily computed HB $\sqrt{n}$ consistent Mahalanobis depth estimator.

The Maronna and Zamar (2002) OGK estimator may be a competitor to the MBA and CMCD estimators, but theory is needed. See Mehrotra (1995) for a similar estimator. Exact computation of the MCD estimator is surveyed by Bernholt and Fischer (2004).

For any given estimator, it is easy to find outlier configurations where the estimator fails. One of the most useful techniques for robust statistics is to make scatterplot matrices of residuals and of fitted values and the response y, or of Mahalanobis distances from several estimators including starts and attractors.

The *R/Splus* software has a function `cov.mcd` for computing the Fast–MCD

estimator. The group of functions `rpack.txt`, available from (www.math.siu.edu/olive/rpack.txt), contains functions `covdgk`, `covmba2` and `rmba` for computing the scaled DGK, MBA and RMBA estimators. The function `ddcomp2` makes the DD plots of the above four estimators.

6 References

Arcones, M.A., 1995. Asymptotic normality of multivariate trimmed means. Statist. Probab. Lett. 25, 43-53.

Bernholt, T., Fischer, P., 2004. The complexity of computing the MCD-estimator. Theoretical Computer Science. 326, 383-398.

Butler, R.W., Davies, P.L., Jhun, M., 1993. Asymptotics for the minimum covariance determinant estimator. Ann. Statist. 21, 1385-1400.

Buxton, L.H.D., 1920. The anthropology of Cyprus. J. Royal Anthropological Institute of Great Britain and Ireland. 50, 183-235.

Čížek, P., 2006. Least trimmed squares under dependence. J. Statist. Plann. Infer. 136, 3967-3988.

Croux, C., Van Aelst, S., 2002. Comment on 'Nearest-neighbor variance estimation (NNVE): robust covariance estimation via nearest-neighbor cleaning' by N. Wang and A.E. Raftery. J. Amer. Statist. Assoc. 97, 1006-1009.

Devlin, S.J., Gnanadesikan, R., Kettenring, J.R., 1975. Robust estimation and outlier detection with correlation coefficients. Biometrika. 62, 531-545.

Devlin, S.J., Gnanadesikan, R., Kettenring, J.R., 1981. Robust estimation of dispersion matrices and principal components. J. Amer. Statist. Assoc. 76, 354-362.

Dollinger, M.B., Staudte, R.G., 1991. Influence functions of iteratively reweighted least squares estimators. J. Amer. Statist. Assoc. 86, 709-716.

Gladstone, R.J., 1905-1906. A study of the relations of the brain to the size of the head. Biometrika. 4, 105-123.

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

Hawkins, D.M., Olive, D.J., 1999. Improved feasible solution algorithms for high breakdown estimation. Computat. Statist. Data Analys. 30, 1-11.

Hawkins, D.M., Olive, D.J., 2002. Inconsistency of resampling algorithms for high breakdown regression estimators and a new algorithm (with discussion). J. Amer. Statist. Assoc. 97, 136-159.

He, X., Portnoy, S., 1992. Reweighted LS estimators converge at the same rate as the initial estimator. Ann. Statist. 20, 2161-2167.

He, X., Wang, G., 1997. A qualitative robustness of S*- estimators of multivari-

ate location and dispersion. Statist. Neerlandica. 51, 257-268.

Johnson, M.E., 1987. Multivariate Statistical Simulation. Wiley, New York.

Kim, J., 2000. Rate of convergence of depth contours: with application to a multivariate metrically trimmed mean. Statist. Probab. Lett. 49, 393-400.

Lehmann, E.L., 1999. Elements of Large–Sample Theory. Springer-Verlag, New York.

Lopuhaä, H.P., 1999. Asymptotics of reweighted estimators of multivariate location and scatter. Ann. Statist. 27, 1638-1665.

Maronna, R.A., Yohai, V.J., 2002. Comment on 'Inconsistency of resampling algorithms for high breakdown regression and a new algorithm' by D.M. Hawkins and D.J. Olive. J. Amer. Statist. Assoc. 97, 154-155.

Maronna, R.A., Zamar, R.H., 2002. Robust estimates of location and dispersion for high-dimensional datasets. Technom. 44, 307-317.

Mašiček, L., 2004. Optimality of the least weighted squares estimator. Kybernetika. 40, 715-734.

Mehrotra, D.V., 1995. Robust elementwise estimation of a dispersion matrix. Biometrics. 51, 1344-1351.

Olive, D.J, 2002. Applications of robust distances for regression. Technom. 44, 64-71.

Olive, D.J., 2004. A resistant estimator of multivariate location and dispersion. Computat. Statist. Data Analys. 46, 99-102.

Olive, D.J., 2005. Two simple resistant regression estimators. Computat. Statist. Data Analys. 49, 809-819.

Pratt, J.W., 1959. On a general concept of 'in probability'. Ann. Math. Statist. 30, 549-558.

Rocke, D.M., Woodruff, D.L., 1996. Identification of outliers in multivariate data. J. Amer. Statist. Assoc. 91, 1047-1061.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Amer. Statist. Assoc. 79, 871-880.

Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technom. 41, 212-223.

Rousseeuw, P.J., Van Driessen, K., 2000. An algorithm for positive-breakdown regression based on concentration steps. In: Gaul, W., Opitz, O., Schader, M. (Eds.), Data Analysis: Modeling and Practical Application, Springer-Verlag, New York.

Rousseeuw, P.J., Van Driessen, K., 2002. Computing LTS regression for large data sets. Estadistica. 54, 163-190.

Rousseeuw, P.J., Van Driessen, K., 2006. Computing LTS regression for large

data sets. Data Mining Knowledge Discovery. 12, 29-45.

Ruppert, D., 1992. Computing S-estimators for regression and multivariate location/dispersion. J. Computat. Graphic. Statist. 1, 253-270.

Ruppert, D., Carroll, R.J., 1980. Trimmed least squares estimation in the linear model. J. Amer. Statist. Assoc. 75, 828-838.

Salibian-Barrera, M., Yohai, V.J., 2006. A fast algorithm for S-regression estimates. J. Computat. Graphic. Statist. 15, 414-427.

Schaaffhausen, H., 1878. Die anthropologische sammlung des anatomischen der Universitat Bonn. Archiv Anthropologie. 10, 1-65, Appendix.

Singh, K., 1998. Breakdown Theory for Bootstrap Quantiles. Ann. Statist. 26, 1719-1732.

Víšek, J.Á., 1996. On high breakdown point estimation. Computat. Statist. 11, 137-146.

Víšek, J.Á., 2006. The least trimmed squares - part III: asymptotic normality. Kybernetika. 42, 203-224.

Welsh, A.H., Ronchetti, E., 2002. A journey in single steps: robust one-step M-estimation in linear regression. J. Statist. Plann. Infer. 103, 287-310.