Robustifying Robust Estimators

David J. Olive and Douglas M. Hawkins [*]

Southern Illinois University and University of Minnesota

September 27, 2005

Abstract

The algorithm implementations of high breakdown estimators for regression and multivariate location and dispersion tend to be impractical to compute or to be zero breakdown inconsistent estimators. Hence the "robust estimators" used in practice are often not robust. A simple modification of existing concentration algorithms for multiple linear regression and multivariate location and dispersion results in high breakdown robust $\sqrt{n}$ 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.

[*] 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 0306304, DMS 9803622 and ACI 9619020.

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 observations $\boldsymbol{x}_i$ for $i = 1, ..., n$ are collected in an $n \times p$ matrix $\boldsymbol{W}$ with n rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$.

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

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \qquad (1.1)$$

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

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

Assume that $(T, \boldsymbol{C})$ is the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to some subset J of $c_n \approx n/2$ cases of the data. The volume of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}}_J)^T \boldsymbol{S}_J^{-1}(\boldsymbol{z} - \overline{\boldsymbol{x}}_J) \le d^2\} \qquad (1.2)$$

is equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} d^p \sqrt{det(\boldsymbol{S}_J)}, \tag{1.3}$$

and this volume will be positive unless extreme degeneracy is present among the c_n cases. See Johnson and Wichern (1988, pp. 103-104).

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 (HB) 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. Concentration algorithms for multivariate data have been suggested for the MCD criterion. For multiple linear regression, concentration algorithms have been suggested for the least trimmed sum of squares (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 concentration algorithms are described in Ruppert (1992), Víšek (1996), Hawkins and Olive (1999, 2002) and Rousseeuw and Van Driessen (1999, 2000, 2002). The DGK multivariate location and dispersion estimator (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 (2004a) *median ball algorithm* (MBA) estimator of MLD uses a typical start $(T_{0,M}, \boldsymbol{C}_{0,M}) = (\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M})$ that is the classical estimator applied after trimming the M% of cases furthest in Euclidean distance from the coordinatewise median MED($\boldsymbol{W}$) where $M \in \{0, 50\}$ (or use, e.g., $M \in \{0, 50, 60, 70, 80, 90, 95, 98\}$). Then concentration steps are performed resulting in the Mth attractor $(T_{k,M}, \boldsymbol{C}_{k,M}) = (\overline{\boldsymbol{x}}_{k,M}, \boldsymbol{S}_{k,M})$, and the $M = 0$ attractor is the DGK estimator. 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

$$\boldsymbol{C}_{MBA} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \qquad (1.4)$$

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).

In the literature there are many HB estimators for MLR and MLD that are impractical to compute such as the CM, maximum depth, GS, LQD, LMS, LTS, LTA, MCD, MVE, projection, repeated median and S estimators. Two stage estimators that use an initial high breakdown estimator from the above list are even less practical to compute. These estimators include the cross–checking, MM, one step GM, one step GR, REWLS, tau and t type estimators.

The "robust" estimators available from the software are often practical to compute, but they tend to be zero breakdown and inconsistent. Hence these "robust" estimators are not actually robust. *A very common error in the literature* is to plug in an inconsistent zero breakdown estimator in place of the HB $\sqrt{n}$ consistent estimator that is impractical to compute (e.g., use PROGRESS, SURREAL, FLTS, FMCD, the elemental algorithms described above, or the feasible solution algorithms with a fixed number of elemental starts).

As an 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. Often an inconsistent zero breakdown MCD 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 $n \to \infty$.

What is needed to make robust statistics rigorous are easily computed HB $\sqrt{n}$ consistent estimators that work well on many of the most important outlier configurations. Algorithm estimators such as the Rousseeuw and Van Driessen (1999, 2002) FMCD and FLTS estimators seem very attractive since they are easy to compute, work well on several of the most important outlier configurations, and perform well in simulations. Nevertheless, results from Hawkins and Olive (2002) suggest that these estimators as well as the widely used elemental basic resampling and concentration algorithms produce zero breakdown inconsistent estimators.

This paper offers remedies. Section 2 derives some of the large sample and breakdown theory for the basic resampling and concentration algorithm estimators. Section 3 shows that the MBA estimator is robust and that it is simple to fix the FMCD and FLTS estimators: adding the classical estimator and an easily computed but biased HB start (based on a carefully chosen half set) results in easily computed HB $\sqrt{n}$ consistent CMCD and CLTS estimators. The elemental basic resampling PROGRESS `lmsreg` estimator can also be modified so that it is asymptotically equivalent to the OLS estimator. Sections 4 and 5 gives examples and applications. For example, using the robust estimators from Section 3 results in a practical robust cross checking estimator.

2 PROPERTIES OF CONCENTRATION ALGORITHMS

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 $\text{Cov}(\boldsymbol{x}) = a_X \boldsymbol{\Sigma}$ for some constant $a_X > 0$. See Johnson (1987, pp. 107-108).

Then the *population squared Mahalanobis distance*

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{2.1}$$

has density

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u) \tag{2.2}$$

and the 50% highest density region has the form of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \le U_{0.5}\}$$

where $U_{0.5}$ is the median of the distribution of U. For example, if the $\boldsymbol{x}$ are MVN, then U has the χ^2_p distribution.

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 the $M\%$ (where $0 < M < 100$) of 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 $a > 0$. Suppose that $(T_1, \boldsymbol{C}_1), ..., (T_K, \boldsymbol{C}_K)$ are K consistent estimators of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ 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 $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ 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 (2004a): $(\overline{\boldsymbol{x}}_{0,50}, \boldsymbol{S}_{0,50})$ is a high breakdown estimator. If the data distribution is EC but not spherically symmetric, then for $m \geq 0$, $\boldsymbol{S}_{m,50}$ underestimates 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. The following proposition shows that if $(T, \boldsymbol{C})$ is a consistent start, then the attractor is a consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ is the population parameter estimated by MLD concentration algorithms.

Proposition 1 (See appendix for proof). 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. Hence $\text{MED}(D_i^2(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j}))$ is a consistent estimator of $U_{0.5}/a_{MCD}$.

The following proposition proves that the elemental and "h–set" basic resampling algorithms produce inconsistent zero breakdown estimators and strongly suggests that concentration algorithms that use K starts of size h also perform poorly. The basic resampling result is remarkable since it is free of the criterion. If someone invents a new high breakdown, highly efficient estimator, we immediately know that the elemental basic resampling algorithm approximation that uses K starts will be inconsistent with zero breakdown. Assume that $h \geq p$ for MLR and that $h \geq p+1$ for multivariate location and dispersion. Hawkins and Olive (2002) proved the following result for elemental sets, and a similar result holds if the size of the jth start h_j depends on j but the sizes are bounded: $h_j \leq B$ for $j = 1, ..., K$ for some fixed positive integer B.

Proposition 2 (See appendix for proof). Suppose that each start uses h randomly selected cases and that the number of starts $K_n \equiv K$ does not depend on n (e.g., $K = 3000$). Then i) the ("h-set") basic resampling estimator is inconsistent.
ii) The k–step concentration algorithms for CLTS and CMCD are inconsistent.
iii) For equivariant MLR estimators the breakdown value is bounded above by K/n, and for CMCD the breakdown value is bounded above by $K(h-p)/n$.

Notice that for a fixed data set, $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} + \boldsymbol{u}$ where the bias vector $\boldsymbol{u} = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$. As long as $\boldsymbol{u}$ is small compared to $\hat{\boldsymbol{\beta}}$, the robust MLR estimator should be useful. Lemma 3 below suggests that $\boldsymbol{u}$ will be small for many small data sets when the basic resampling algorithm is used. The bias vector should be even smaller after concentration.

Lemma 3 (See appendix for proof). Suppose that $K_n \equiv K$ random starts of size h are selected and let $Q_{(1)} \leq Q_{(2)} \leq \cdots \leq Q_{(B)}$ correspond to the order statistics of the criterion values of the $B = C(n, h)$ possible starts of size h. Let R be the rank of the smallest criterion value from the K starts. If $P(R \leq R_\alpha) = \alpha$, then

$$R_\alpha \approx B[1 - (1-\alpha)^{1/K}].$$

Remark 2. If $K = 500$, then with 95% probability about 1 in 10000 elementals sets will be better than the best elemental start found from the elemental concentration algorithm. From Feller (1957, pp. 211-212),

$$E(R) \approx 1 + \frac{B}{K+1}, \text{ and Var(R)} \approx \frac{\mathrm{KB}^2}{(\mathrm{K}+1)^2(\mathrm{K}+2)} \approx \frac{\mathrm{B}^2}{\mathrm{K}^2}.$$

Remark 3. Hawkins and Olive (2002) showed that MLR algorithms that use K_n randomly selected elemental starts have a rate $\leq K_n^{1/p}$ and may have a rate $\leq K_n^{1/2p}$. *Increasing the number of elemental sets to* $K_n = n^\delta$ *for* $1 \leq \delta \leq 3$ *produces an estimator with a poor computational time and a poor convergence rate.*

We certainly prefer to use consistent estimators whenever possible. When the start subset size $h_n \equiv h$ and the number of starts $K_n \equiv K$ are both fixed, the estimator is inconsistent. The situation changes dramatically if the start subset size $h_n = g(n) \to \infty$ as $n \to \infty$. In particular, if several starts with rate $n^{1/2}$ are used, the final estimator also has rate $n^{1/2}$. The drawback to these algorithms is that they often do not have enough outlier resistance. Again the basic resampling result below is free of the criterion. The conditions in Proposition 4ii hold, for example, if the classical estimator is applied to h_n cases randomly drawn from a distribution with a covariance matrix $\mathrm{Cov}(\boldsymbol{x}) = a_X \boldsymbol{\Sigma}$. Then each of the K starts estimates $(\boldsymbol{\mu}, a_X \boldsymbol{\Sigma})$ with rate $[h_n]^{1/2}$.

Proposition 4 (See appendix for proof). Suppose $K_n \equiv K$ starts are used and that all starts have subset size $h_n = g(n) \uparrow \infty$ as $n \to \infty$. Assume that the estimator applied to the subset has rate n^δ. i) For the h_n-set basic resampling algorithm, the MLR algorithm estimator has rate $[g(n)]^\delta$.

ii) If each of the K estimators $(T_i, \boldsymbol{C}_i)$ is a $[g(n)]^\delta$ consistent estimator for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ (i.e., $a_i \equiv a$ for $i = 1, ..., K$), then the MLD h_n-set basic resampling algorithm estimator has rate $[g(n)]^\delta$.

iii) Under mild regularity conditions (e.g., given by He and Portnoy 1992), the CLTS estimator has rate $[g(n)]^\delta$.

iv) The CMCD estimator has rate $[g(n)]^\delta$ if assumption (E1) holds.

v) The DGK estimator has rate $n^{1/2}$ if assumption (E1) holds.

vi) The MBA estimator has rate $n^{1/2}$ if (E1) holds and the distribution is spherically symmetric.

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

$$\gamma_o \approx \min(\frac{n - c_n}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h})100\% \tag{2.3}$$

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 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%.)

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 (2004a) showed similar results for the Rousseeuw and Van Driessen (1999) FMCD algorithm and that the MBA estimator could often correctly classify up to 49% hugely distant outliers.

The following proposition shows that it is very difficult to drive the determinant of the dispersion estimator from a concentration algorithm to zero.

Proposition 5 (See appendix for proof). 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 $\boldsymbol{C}_{MCD}$.

Notice that concentration algorithm estimators with very good rates are easy to construct. The DGK estimator works for multivariate data. For MLR, let the start have rate $n^{1/2}$ and apply $k = 10$ OLS concentration steps. Let $\hat{\boldsymbol{\beta}}_{Q,n}$ be the robust estimator that the concentration estimator is approximating, e.g., LMS.

Proposition 6 (See appendix for proof). Suppose that the concentration estimator $\hat{\boldsymbol{\beta}}_{A,n}$ is approximating the estimator $\hat{\boldsymbol{\beta}}_{Q,n}$. If $\hat{\boldsymbol{\beta}}_{Q,n}$ has rate n^{δ_1} and $\hat{\boldsymbol{\beta}}_{A,n}$ has rate n^{δ_2}, then $\|\hat{\boldsymbol{\beta}}_{Q,n} - \hat{\boldsymbol{\beta}}_{A,n}\| = O_P(n^{-\min(\delta_1,\delta_2)})$.

The following proposition shows that it is easy to construct high breakdown concentration algorithms for MLR. Olive (2005) showed that OLS applied to the c_n cases with Y_i closest to the sample median of the Y_i provides a high breakdown start (that is affine equivariant but not regression equivariant).

Proposition 7 (See appendix for proof). The concentration algorithm estimator for CLTS, CLTA or CLMS is a high breakdown estimator if it includes a high breakdown start.

3 IMPROVING CONCENTRATION ALGORITHMS

This section shows that it is simple to modify existing concentration algorithms such that the resulting HB estimators have good statistical properties. For the MLD estimators, we will be interested in the attractor that minimizes the determinant $det(\boldsymbol{S}_{k,M})$ and in the attractor that minimizes the volume criterion

$$\sqrt{det(\boldsymbol{S}_{k,M})}[MED(D_i^2)]^p, \tag{3.1}$$

(see Rousseeuw and Leroy 1987, p. 259) which is proportional to the volume of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,M})^T \boldsymbol{S}_{k,M}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,M}) \le d^2\} \tag{3.2}$$

where $d^2 = \text{MED}(D_i^2(\overline{\boldsymbol{x}}_{k,M}, \boldsymbol{S}_{k,M}))$. The following theorem shows that the MBA estimator has good statistical properties.

Theorem 8 (See appendix for proof). Suppose (E1) holds.

i) If $(T_A, \boldsymbol{C}_A)$ is the attractor that minimizes the volume criterion (3.1), then $(T_A, \boldsymbol{C}_A)$ is a HB $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

ii) 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.

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

Theorem 9 (See appendix for proof). Suppose (E1) holds and that the CMCD algorithm uses $K_n \equiv K$ randomly selected elemental starts (e.g., K = 200), 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 spherically symmetric, then the CMCD estimator is asymptotically equivalent to the DGK estimator.

The following theorem shows that is simple to improve the CLTS and `lmsreg` estimators by adding two carefully chosen attractors. Hawkins and Olive (2002) suggested the CLTS estimator given in Theorem 10i and Maronna and Yohai (2002) claim that this CLTS estimator is consistent. Notice that the hybrid CLTS estimator has rate $\sqrt{n}$ while the rate of LTS is unknown. Also note that `lmsreg` is an inconsistent zero breakdown estimator but the modification to `lmsreg` is high breakdown and asymptotically equivalent to OLS. Hence the modified estimator has a $\sqrt{n}$ rate which is higher than the $n^{1/3}$ rate of the LMS estimator. Let $\hat{\boldsymbol{\beta}}_{k,OLS}$ denote the attractor that results when $\hat{\boldsymbol{\beta}}_{OLS}$ is the start. Let $\boldsymbol{b}_k$ be the attractor from the start consisting of OLS applied to the c_n cases 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 10 (See appendix for proof). i) Suppose that the CLTS algorithm uses $K_n \equiv K$ randomly selected elemental starts (e.g., K = 500) and the attractors $\hat{\boldsymbol{\beta}}_{k,OLS}$ and $\hat{\boldsymbol{\beta}}_{k,B}$. Then the resulting estimator is a HB $\sqrt{n}$ consistent estimator if $\hat{\boldsymbol{\beta}}_{OLS}$ is $\sqrt{n}$ consistent, and the estimator is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{k,OLS}$.
ii) Suppose a basic resampling algorithm is used for a HB criterion that is minimized by a consistent estimator for $\boldsymbol{\beta}$ (e.g., for LMS or LTS). Also assume that the algorithm uses $K_n \equiv K$ randomly selected elemental starts (e.g., K = 500), 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 $\boldsymbol{\beta}$.

Recall that the criterion is evaluated on the attractors. Then from the proof of the above theorem, it can be seen that the $\sqrt{n}$ consistent 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 $\hat{\boldsymbol{\beta}}_D$. Good choices for $\hat{\boldsymbol{\beta}}_D$ are OLS, L_1, the Wilcoxon rank estimator, $\hat{\boldsymbol{\beta}}_{k,OLS}$, the Mallows GM estimator and estimators that perform well when heteroscedasticity is present.

4 EXAMPLES

We examined several data sets from the archive (http://www.math.siu.edu/olive/olbookp.htm) to illustrate the DGK, MBA and FMCD 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*. For cases 61–65, the heights were about 0.75 inches with head lengths well over 5 feet. The DGK, FMCD 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 FMCD estimator to be singular. Hence these three variables were deleted and there were 6 outliers and 9 variables. The DGK, FMCD and MBA estimators 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, FMCD 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.

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. We found that estimators based on concentration estimators were much more effective on such data sets than estimators based on the basic resampling algorithm.

5 APPLICATIONS AND CONCLUSIONS

The MBA estimator is robust and the hybrid CMCD estimator of Theorem 9 that uses the 2 MBA starts as well as 200 randomly chosen elemental starts will be a HB $\sqrt{n}$ consistent estimator that is asymptotically equivalent to the DGK estimator. This CMCD estimator is also about twice as fast as the current zero breakdown inconsistent FMCD estimator. The CLTS estimator of Theorem 10i should use at least $K = 500$ elemental starts and should outperform the estimator of Theorem 10ii that uses $K = 10000$ starts.

Although the estimators in Section 3 are very attractive, there is still room for improvement. The MBA estimator is useful for data mining where speed is crucial. HB

MLR estimators can be made using HB MLD estimators (and vice verca), and the HB MLR estimators that use the MBA or hybrid CMCD estimator may have greater outlier resistance than the estimators from Theorem 10. These estimators can also be used to correctly implement some two stage estimators, including the cross checking 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, or of Mahalanobis distances from several estimators including starts and attractors. Keep track of the best starts and attractors that have desirable properties including i) rate $n^{1/2}$, ii) high breakdown and iii) affine equivariance combined with a low value of the criterion.

Many papers have been written that need a HB consistent estimator of MLD. Since no practical HB estimator was available, inconsistent zero breakdown estimators were often used in implementations, resulting in zero breakdown estimators that were often inconsistent (although perhaps useful as diagnostics).

Applications of the robust $\sqrt{n}$ consistent CLTS and CMCD estimators are numerous. For example, robustify the ideas in the following papers by using the CMCD estimator instead of the FMCD, MCD or MVE estimator. *Binary regression:* see Croux and Haesbroeck (2003). *Canonical correlation analysis:* see Branco, Croux, Filzmoser, and Oliviera (2005). *Discriminant analysis:* see He and Fung (2000). *Factor analysis:* see Pison, Rousseeuw, Filzmoser, and Croux (2003). *Analogs of Hotelling's T^2 test:* see

Willems, Pison, Rousseeuw, and Van Aelst (2002). *Longitudinal data analysis:* see He, Cui and Simpson (2004). *Multiple linear regression:* see He, Simpson and Wang (2000). Robust asymptotically efficient MLR estimators can be made by using this modified t-type estimator to create a cross checking estimator. See He (1991) and Davies (1993). *Resistant regression:* see Olive (2005). *Multivariate analysis diagnostics:* the Rousseeuw and Van Driessen (1999) DD plot of classical Mahalanobis distances versus CMCD distances should be used for multivariate analysis much as Cook's distances are used for MLR. Olive (2002) shows that the plotted points in the DD plot will follow the identity line with zero intercept and unit slope if the data distribution is multivariate normal (MVN), and will follow a line with zero intercept but non–unit slope if the data distribution is elliptically contoured but not MVN. *Multivariate regression:* see Rousseeuw, Van Aelst, Van Driessen and Agulló (2004). *Principal components:* see Hubert, Rousseeuw, and Vanden Branden (2005). *Asymptotically efficient estimators of MLD:* see He and Wang (1996).

Regression via Dimension Reduction: Regression is the study of the conditional distribution of the response Y given the vector of predictors $\boldsymbol{x} = (1, \boldsymbol{w}^T)^T$ where $\boldsymbol{w}$ is the vector of nontrivial predictors. Make a DD plot of the classical Mahalanobis distances versus the robust distances computed from $\boldsymbol{w}$. If $\boldsymbol{w}$ comes from an elliptically contoured distribution, then the plotted points in the DD plot should follow a straight line through the origin. Give zero weight to cases in the DD plot that do not cluster tightly about "the best straight line" through the origin (often the identity line with unit slope), and run a weighted regression procedure. This technique can increase the resistance of regression procedures such as sliced inverse regression (SIR, see Li, 1991) and MAVE (Xia, Tong, Li, and Zhu, 2002). Also see Cook and Nachtsheim (1994) and Li, Cook and Nachtsheim (2004). Gather, Hilker and Becker (2001, 2002) also develop a robust version of SIR.

Visualizing 1D Regression: In a 1D regression model the response Y is independent of the predictors $\boldsymbol{x}$ given $\boldsymbol{\beta}^T\boldsymbol{x}$. Generalized linear models and single index models are important special cases. Resistant methods that use trimming for visualizing 1D regression are given in Olive (2002, 2004b).

APPENDIX: MATHEMATICAL PROOFS

Proof of Proposition 1. 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 $a_{MCD}\boldsymbol{\Sigma}$. QED

Proof of Proposition 2. To prove i) and ii), notice that each start is inconsistent. Hence each attractor is inconsistent by He and Portnoy (1992) for the CLTS and Lopuhaä (1999) for CMCD. Choosing from K inconsistent estimators still results in an inconsistent estimator. To prove iii) for MLR, replace one observation in each start by a high leverage case (with y tending to ∞). For multivariate data with $h \geq p+1$, replace $h-p$ cases so that the start is singular and the covariance matrix can not be computed. QED

Proof of Lemma 3. If W_i is the rank of the ith start, then $W_1, ..., W_K$ are iid discrete uniform on $\{1, ..., B\}$ and $R = \min(W_1, ..., W_K)$. If r is an integer in $[1, B]$, then

$$P(R \leq r) = 1 - (\frac{B-r}{B})^K.$$

Solve the above equation $\alpha = P(R \leq R_\alpha)$ for R_α. QED

Proof of Proposition 4. i) The $h_n = g(n)$ cases are randomly sampled without replacement. Hence the classical estimator applied to these $g(n)$ cases has rate $[g(n)]^\delta$. Thus all K starts have rate $[g(n)]^\delta$, and the result follows by Pratt (1959). ii) The result follows by Pratt (1959). iii) and iv) By He and Portnoy (1992) for CLTS and by Lopuhaä (1999) for CMCD, all K attractors have $[g(n)]^\delta$ rate, and the result follows by Pratt (1959). v) The DGK estimator uses $K = 1$ and $h_n = n$, and the k concentration steps are performed after using the classical estimator as a start. Hence the result follows by Lopuhaä (1999). vi) Each of the K starts in the MBA algorithm is $\sqrt{n}$ consistent (if $M > 0$ then the $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = (T_{-1}, \boldsymbol{C}_{-1})$ can be regarded as the start). Hence the result follows by Proposition 1 and Pratt (1959). QED

Proof of Proposition 5. 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 $0 \le \det(\boldsymbol{C}_{MCD}) \le \det(\boldsymbol{C}_A)$. QED

Proof of Proposition 6. $\|\hat{\boldsymbol{\beta}}_{Q,n} - \hat{\boldsymbol{\beta}}_{A,n}\| \le \|\hat{\boldsymbol{\beta}}_{Q,n} - \boldsymbol{\beta}\| + \|\hat{\boldsymbol{\beta}}_{A,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta_1}) + O_P(n^{-\delta_2}) = O_P(n^{-\min(\delta_1,\delta_2)})$. QED

Proof of Proposition 7. Olive (2005) showed that an MLR estimator is high breakdown if the median absolute residual stays bounded under high contamination. Concentration insures that the criterion function of the $c_n \approx n/2$ absolute residuals gets smaller. QED

Proof of Theorem 8. i) The estimator is HB since $(\overline{\boldsymbol{x}}_{0,50}, \boldsymbol{S}_{0,50})$ is a high breakdown estimator and hence has a bounded volume if up to nearly 50% of the cases are outliers. If the distribution is spherically symmetric then the result follows by Proposition 4vi. Otherwise, the hyperellipsoid corresponding to the highest density region has at least one major axis and at least one minor axis. The estimators with $M > 0$ trim too much data in the direction of the major axis and hence the resulting attractor is not estimating the highest density region. But the DGK estimator ($M = 0$) is estimating the highest density region. Thus the probability that the DGK estimator is the attractor that minimizes the volume goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0})$.

ii) The estimator is HB since $0 < det(\boldsymbol{S}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{S}_{0,50}) < \infty$ if up to nearly 50% of the cases are outliers. If the distribution is spherically symmetric then the result follows by Proposition 4vi. Otherwise, the estimators with $M > 0$ trim too much data in the direction of the major axis and hence the resulting attractor is not estimating the highest density region. Hence $\boldsymbol{S}_{k,M}$ 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 $\|\boldsymbol{S}_{MCD} - \boldsymbol{S}_{k,0}\| = O_P(n^{-1/2})$. 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 $(T_{k,0}, \boldsymbol{C}_{k,0})$. QED

Proof of Theorem 9. The estimator is HB since $0 < det(\boldsymbol{S}_{MCD}) \leq det(\boldsymbol{C}_{CMCD}) \leq det(\boldsymbol{S}_{0,50}) < \infty$ 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 spherically symmetric 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 spherically symmetric, 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 $(T_{k,0}, \boldsymbol{C}_{k,0})$. QED

Proof of Theorem 10. Proposition 7 shows that concentration and basic resampling algorithms that use a HB start are HB, and $\hat{\boldsymbol{\beta}}_{k,B}$ is a HB estimator.

i) By He and Portnoy (1992), the OLS attractor $\hat{\boldsymbol{\beta}}_{k,OLS}$ is $\sqrt{n}$ consistent estimator. As $n \to \infty$, the estimator that minimizes the LTS criterion gets arbitrarily close to $\boldsymbol{\beta}$ since the LTS estimator is consistent by Mašîček (2004). Since $\hat{\boldsymbol{\beta}}_{k,B}$ is a biased estimator of $\boldsymbol{\beta}$, with probability tending to one, the OLS attractor will have a smaller criterion value. With probability tending to one, the OLS attractor will also have a smaller criterion value than the criterion value of the attractor from a randomly drawn elemental set (by Lemma 3 and He and Portnoy 1992, also see Remark 4 in Hawkins and Olive 2002). Since K randomly elemental sets are used, the CLTS estimator is asymptotically equivalent to the OLS attractor.

ii) As in the proof of i), the OLS estimator will minimize the criterion value with probability tending to one as $n \to \infty$. QED

6 References

Arcones, M.A. (1995), "Asymptotic Normality of Multivariate Trimmed Means," *Statistics and Probability Letters,* 25, 43-53.

Bernholt, T., and Fischer, P. (2004), "The Complexity of Computing the MCD-Estimator," *Theoretical Computer Science*, 326, 383-398.

Branco, J.A., Croux, C., Filzmoser, P., and Oliviera, M.R. (2005), "Robust Canonical Correlations: a Comparative Study," *Computational Statistics*, 20, 203-229.

Butler, R.W., Davies, P.L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics,* 21, 1385-1400.

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Cook, R.D., and Nachtsheim, C.J. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *Journal of the American Statistical Association,* 89, 592-599.

Croux, C., and Haesbroeck, G. (2003), "Implementing the Bianco and Yohai Estimator for Logistic Regression," *Computational Statistics and Data Analysis,* 44, 273-295.

Croux, C., and 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," *Journal of the American Statistical Association,* 97, 1006-1009.

Davies, P.L. (1993), "Aspects of Robust Linear Regression," *The Annals of Statistics,* 21, 1843-1899.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1975), "Robust Estimation and Outlier Detection with Correlation Coefficients," *Biometrika,* 62, 531-545.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association,* 76, 354-362.

Dollinger, M.B., and Staudte, R.G. (1991), "Influence Functions of Iteratively Reweighted

Least Squares Estimators," *Journal of the American Statistical Association,* 86, 709-716.

Feller, W. (1957), *An Introduction to Probability Theory and Its Applications,* Vol. 1, 2nd ed., New York: Wiley.

Gather, U., Hilker, T., and Becker, C. (2001), "A Robustified Version of Sliced Inverse Regression," in *Statistics in Genetics and in the Environmental Sciences*, eds. Fernholtz, T.L., Morgenthaler, S., and Stahel, W., Basel, Switzerland: Birkhäuser,145-157.

Gather, U., Hilker, T., and Becker, C. (2002), "A Note on Outlier Sensitivity of Sliced Inverse Regression," *Statistics,* 36, 271-281.

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., and Kettenring, J.R. (1972), "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," *Biometrics,* 28, 81-124.

Hawkins, D.M., and Olive, D.J. (1999), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics and Data Analysis,* 30, 1-11.

Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," (with discussion), *Journal of the American Statistical Association,* 97, 136-159.

He, X. (1991), "A Local Breakdown Property of Robust Tests in Linear Regression," *Journal of Multivariate Analysis,* 38, 294-305.

He, X., Cui, H., and Simpson, D.G. (2004), "Longitudinal Data Analysis Using t-type Regression," *Journal of Statistical Planning and Inference,* 122, 253-269.

He, X., and Fung, W.K. (2000), "High Breakdown Estimation for Multiple Populations with Applications to Discriminant Analysis," *Journal of Multivariate Analysis,* 72, 151-162.

He, X., and Portnoy, S. (1992), "Reweighted LS Estimators Converge at the Same Rate as the Initial Estimator," *The Annals of Statistics,* 20, 2161-2167.

He, X., Simpson, D.G., and Wang, G.Y. (2000), "Breakdown Points of t-type Regression Estimators," *Biometrika,* 87, 675-687.

He, X., and Wang, G. (1996), "Cross-Checking Using the Minimum Volume Ellipsoid Estimator," *Statistica Sinica,* 6, 367-374.

He, X., and Wang, G. (1997), "A Qualitative Robustness of S*- Estimators of Multivariate Location and Dispersion," *Statistica Neerlandica*, 51, 257-268.

Hubert, M., Rousseeuw, P.J., and Vanden Branden, K. (2005), "ROBPCA: a New Approach to Robust Principal Component Analysis," *Technometrics,* 47, 64-79.

Johnson, M.E. (1987), *Multivariate Statistical Simulation,* New York: Wiley.

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Englewood Cliffs, NJ: Prentice Hall.

Kim, J. (2000), "Rate of Convergence of Depth Contours: with Application to a Multivariate Metrically Trimmed Mean," *Statistics and Probability Letters,* 49, 393-400.

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

Li, K.C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association,* 86, 316-342.

Li, L., Cook, R.D, and Nachtsheim, C.J. (2004), "Cluster-based Estimation for Sufficient Dimension Reduction," *Computational Statistics and Data Analysis,* 47, 175-193.

Lopuhaä, H.P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 27, 1638-1665.

Maronna, R.A., and 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," *Journal of the American Statistical Association,* 97, 154-155.

Maronna, R.A., and Zamar, R.H. (2002), "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," *Technometrics,* 44, 307-317.

Mašíček, L. (2004), "Consistency of the Least Weighted Squares Regression Estimator," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst, S., Series: Statistics for Industry and Technology, Basel, Switzerland: Birkhäuser, 183-194.

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," *Technometrics,* 44, 64-71.

Olive, D.J. (2004a), "A Resistant Estimator of Multivariate Location and Dispersion," *Computational Statistics and Data Analysis*, 46, 99-102.

Olive, D.J. (2004b), "Visualizing 1D Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst, S., Series: Statistics for Industry and Technology, Basel, Switzerland: Birkhäuser, 221-233.

Olive, D.J. (2005), "Two Simple Resistant Regression Estimators," *Computational Statistics and Data Analysis*, 49, 809-819.

Pison, G., Rousseeuw, P.J., Filzmoser, P., and Croux, C. (2003), "Robust Factor Analysis," *Journal of Multivariate Analysis,* 84, 145-172.

Pratt, J.W. (1959), "On a General Concept of 'in Probability'," *The Annals of Mathematical Statistics,* 30, 549-558.

Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association,* 91, 1047-1061.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

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

Rousseeuw, P.J., Van Aelst, S., Van Driessen, K., and Agulló, J. (2004), "Robust Multivariate Regression," *Technometrics*, 46, 293-305.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rousseeuw, P.J., and Van Driessen, K. (2000), "An Algorithm for Positive-Breakdown Regression Based on Concentration Steps," in *Data Analysis: Modeling and Practical Application,* eds. W. Gaul, O. Opitz, and M. Schader, New York: Springer-Verlag.

Rousseeuw, P.J., and Van Driessen, K. (2002), "Computing LTS Regression for Large Data Sets," *Estadistica*, 54, 163-190.

Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics,* 1, 253-270.

Ruppert, D., and Carroll, R. J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association,* 75, 828-838.

Schaaffhausen, H. (1878), "Die Anthropologische Sammlung Des Anatomischen Der Universitat Bonn," *Archiv fur Anthropologie,* 10, 1-65, Appendix.

Víšek, J.Á. (1996), "On High Breakdown Point Estimation," *Computational Statistics,* 11, 137-146.

Welsh, A.H., and Ronchetti, E. (2002), "A Journey in Single Steps: Robust One-Step M-estimation in Linear Regression," *Journal of Statistical Planning and Inference,* 103, 287-310.

Willems, G., Pison, G., Rousseeuw, P.J., and Van Aelst, S. (2002), "A Robust Hotelling Test," *Metrika*, 55, 125-138.

Xia, Y., Tong, H., Li, W.K., and Zhu, L.-X. (2002), "An Adaptive Estimation of Dimension Reduction Space," (with discussion), *Journal of the Royal Statistical Society, B*, 64, 363-410.