Robust Multivariate Location and Dispersion

David J. Olive and Douglas M. Hawkins*

Southern Illinois University and University of Minnesota

December 10, 2010

Abstract

Robust estimators for multivariate location and dispersion should be $\sqrt{n}$ consistent and highly outlier resistant, but estimators that have been shown to have these properties are impractical to compute. This paper gives easily computed $\sqrt{n}$ consistent outlier resistant estimators that can be used for inference. Applications are numerous, including outlier detection and a diagnostic for whether the data distribution is elliptically contoured.

KEY WORDS: minimum covariance determinant estimator, outliers.

*David J. Olive is Associate Professor (E-mail: *dolive@math.siu.edu*), Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. Douglas M. Hawkins is Professor (E-mail: *dhawkins@umn.edu*), School of Statistics, University of Minnesota, Minneapolis, MN 55455-0493, USA. Their work was supported by the National Science Foundation under grants DMS 0600933, DMS 0306304, DMS 9803622 and ACI 9619020.

1. INTRODUCTION

A long standing question in Statistics is whether high breakdown multivariate statistics is a viable field of study. Are there useful high breakdown estimators of multivariate location and dispersion that are practical to compute? Can high breakdown estimators be incorporated into a practical algorithm in such a way that the algorithm estimator is consistent? This paper provides practical $\sqrt{n}$ consistent estimators that incorporate a useful high breakdown estimator.

Let the ith case $\boldsymbol{x}_i$ be a $p \times 1$ random vector, and suppose the n cases are collected in an $n \times p$ matrix $\boldsymbol{X}$ with rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$. The fastest estimator of multivariate location and dispersion that has been shown to be both consistent and high breakdown is the minimum covariance determinant (MCD) estimator with $O(n^v)$ complexity where $v = 1 + p(p+3)/2$. See Bernholt and Fischer (2004). The minimum volume ellipsoid (MVE) complexity is far higher, and there may be no known method for computing S, τ, projection based, constrained M, MM, and Stahel-Donoho estimators. See Maronna, Martin and Yohai (2006, ch. 6) for descriptions and references.

Since the above estimators take too long to compute, they have been replaced by practical estimators. To our knowledge, no useful practical estimator of "high breakdown multivariate location and dispersion" has been shown to be consistent or high breakdown. When authors claim to have a method that uses a high breakdown estimator such as MCD, either the method takes too long to compute or the method actually uses a practical estimator that is not backed by theory. In particular, the "robust and computationally efficient multivariate techniques" (e.g., for principal component analysis, factor analysis and multivariate regression) that claim to use the impractical MCD estimator actually use the Rousseeuw and Van Driessen (1999) FAST-MCD (FMCD) estimator. These methods, reviewed by Hubert, Rousseeuw and Van Aelst (2008), should be classified as outlier diagnostics unless the FAST-MCD estimator can be shown to be consistent. See Huber and Ronchetti (2009, pp. xiii, 9, 196-198).

The classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ of multivariate location and dispersion is the sample mean and sample covariance matrix where

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

Some important joint distributions for $\boldsymbol{x}$ are completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. An important model is the elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with probability density function (pdf) $f(\boldsymbol{z}) = k_p|\boldsymbol{\Sigma}|^{-1/2}g[(\boldsymbol{z} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})]$ where $k_p > 0$ is some constant and g is some known function. The multivariate normal (MVN) $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution is a special case, and $\boldsymbol{x}$ is "spherical about $\boldsymbol{\mu}$" if $\boldsymbol{x}$ has an $EC_p(\boldsymbol{\mu}, c\boldsymbol{I}_p, g)$ distribution where $c > 0$ is some constant and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Let the $p \times 1$ column vector $T \equiv T(\boldsymbol{X})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C} \equiv \boldsymbol{C}(\boldsymbol{X})$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X})) = (\boldsymbol{x}_i - T(\boldsymbol{X}))^T\boldsymbol{C}^{-1}(\boldsymbol{X})(\boldsymbol{x}_i - T(\boldsymbol{X})) \tag{2}$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{X})$ is $D_i(T(\boldsymbol{X}), \boldsymbol{I}_p)$. The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

Following Johnson (1987, pp. 107-108), the population squared Mahalanobis distance

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \tag{3}$$

and for elliptically contoured distributions, U has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{4}$$

If d_n of the cases have been replaced by arbitrarily bad contaminated cases, then the contamination fraction is $\gamma_n = d_n/n$. Then the breakdown value of a multivariate location estimator is the smallest value of γ_n needed to make $\|T\|$ arbitrarily large. Let $0 \leq \lambda_p(\boldsymbol{C}) \leq \cdots \leq \lambda_1(\boldsymbol{C})$ denote the eigenvalues of $\boldsymbol{C}$. Then the breakdown value of $\boldsymbol{C}$ is the smallest value of γ_n needed to drive either λ_p to zero or λ_1 to ∞. High breakdown statistics have $\gamma_n \to 0.5$ as $n \to \infty$ if the (uncontaminated) clean data are in general position: no more than p points of the clean data lie on any $(p-1)$-dimensional hyperplane. *For the remainder of this paper, assume that the clean data are in general position.* Estimators are zero breakdown if $\gamma_n \to 0$ and positive breakdown if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

Many practical "robust estimators" generate a sequence of K trial fits called *attractors*: $(T_1, \boldsymbol{C}_1), ..., (T_K, \boldsymbol{C}_K)$. Then the attractor $(T_A, \boldsymbol{C}_A)$ that minimizes some criterion is used to obtain the final estimator. One way to obtain attractors is to generate trial fits called *starts*, and then use the *concentration* technique. Let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the jth start and compute all n Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j})$ 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 $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. Then $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the

jth attractor for $j = 1, ..., K$. Using $k = 10$ often works well, and the basic resampling algorithm is a special case $k = -1$ where the attractors are the starts.

Three important starts will be examined by this paper. Hawkins and Olive (1999) and Rousseeuw and Van Driessen (1999) use elemental starts: $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ is the classical estimator applied to a randomly selected "elemental set" of $p + 1$ cases. The Devlin, Gnanadesikan and Kettenring (1981) DGK estimator $(T_{k,D}, \boldsymbol{C}_{k,D})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start. The Olive (2004) median ball (MB) estimator $(T_{k,M}, \boldsymbol{C}_{k,M})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{X}), \boldsymbol{I}_p)$ as the only start where MED($\boldsymbol{X}$) is the coordinatewise median. Hence $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to MED($\boldsymbol{X}$) in Euclidean distance. Section 3 will show that the MB estimator is a high breakdown estimator and that the DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$, the same quantity estimated by the MCD estimator. For nonspherical elliptically contoured distributions, the MB estimator is a biased estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$, but the bias seems to be small even for $k = 0$, and to get smaller as k increases.

Rousseeuw (1984) suggested the following criteria for screening attractors. Suppose the attractor is $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j})$ computed from a subset of c_n cases. The MCD(c_n) criterion is the determinant $det(\boldsymbol{S}_{k,j})$. The volume of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,j})^T \boldsymbol{S}_{k,j}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,j}) \leq h^2\} \text{ is equal to } \frac{2\pi^{\mathrm{p}/2}}{\mathrm{p}\Gamma(\mathrm{p}/2)} \mathrm{h}^{\mathrm{p}} \sqrt{\det(\boldsymbol{S}_{\mathrm{k,j}})}, \qquad (5)$$

see Johnson and Wichern (1988, pp. 103-104). Let $h = D_{(c_n)}(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j})$. Then the "MVE($c_n$)" criterion is $h^p \sqrt{\det(\boldsymbol{S}_{\mathrm{k,j}})}$ (but does not actually correspond to the MVE estimator).

In the following sections, three promising algorithms for robust multivariate location and dispersion are examined. The first algorithm uses concentration with a few consistent and outlier resistant starts. See Hubert, Rousseeuw and Verdonck (2010), Olive (2004) and the estimators developed in Section 3. The second algorithm uses concentration with randomly selected elemental starts. The third algorithm is the orthogonalized Gnanadesikan-Kettenring (OGK) estimator which Maronna and Zamar (2002, p. 309) claim, without proof, is consistent and high breakdown.

Reyen, Miller, and Wegman (2009) simulate the OGK and the Olive (2004) median ball algorithm (MBA) estimators for $p = 100$ and n up to 50000, and note that the OGK complexity is $O[p^3 + np^2 \log(n)]$ while that of MBA (and FMCD) is $O[p^3 + np^2 + np \log(n)]$.

Section 2 shows algorithms that use many attractors may not be trustworthy. Section 3 develops $\sqrt{n}$ consistent outlier resistant estimators that use the high breakdown MB estimator. Section 4 considers outlier resistance with a small simulation study.

2. THEORY FOR SOME PRACTICAL ESTIMATORS

Suppose the algorithm estimator uses some criterion to choose an attractor as the final estimator where there are K attractors and K is fixed, e.g. $K = 500$, so K does not depend on n. The main point of this section is that the theory of the algorithm estimator depends on the theory of the attractors, not on the estimator corresponding to the criterion.

Hawkins and Olive (2002) noted that if K randomly selected elemental starts are used with concentration to produce the attractors, then the resulting estimator is inconsistent and zero breakdown if K and k are fixed and free of n. Note that each elemental start can be made to breakdown by changing one case. Hence the breakdown value of the final estimator is bounded by $K/n \to 0$ as $n \to \infty$. The classical estimator applied to a randomly drawn elemental set is an inconsistent estimator, so the K starts and the K attractors are inconsistent. Note that if the $\boldsymbol{x}_i$ are iid and $P(\boldsymbol{x}_i = \boldsymbol{\mu}) < 1$, then $\overline{\boldsymbol{x}}_{-1,j}$ is the sample mean applied to $p+1$ iid cases. Thus there exists $\epsilon > 0$ such that $P(\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) \equiv \delta_\epsilon > 0$, and $P(\min_j \|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = P(\text{all} \ \ \|\overline{\boldsymbol{x}}_{-1,\mathrm{j}} - \boldsymbol{\mu}\| > \epsilon) \to \delta_\epsilon^{\mathrm{K}} > 0$ as $n \to \infty$ where equality would hold if the $\overline{\boldsymbol{x}}_{-1,j}$ were iid. Hence the "best start" that minimizes $\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\|$ is inconsistent, and the algorithm needs $K_n \to \infty$ as $n \to \infty$ to produce a consistent estimator.

This theory shows that the Maronna, Martin and Yohai (2006, pp. 198-199) estimators that use $K = 500$ and one concentration step ($k = 0$) are inconsistent and zero breakdown. The following theorem is useful because it does not depend on the criterion used to choose the attractor. If the algorithm needs to use many attractors to achieve outlier resistance, then the individual attractors have little outlier resistance. Such estimators include elemental concentration algorithms, heuristic and genetic algorithms and projection algorithms. Algorithms such as elemental concentration algorithms where all K of the attractors are inconsistent are especially untrustworthy. For example, Stahel Donoho algorithms, discussed in Maronna, Martin and Yohai (2006, pp. 193-194), use randomly chosen projections and the attractor is a weighted mean and covariance matrix computed for each projection. If randomly chosen projections result in inconsistent attractors, then the Stahel Donoho algorithm is likely inconsistent.

Suppose there are K consistent estimators $(T_j, \boldsymbol{C}_j)$ of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ for some constant $a > 0$, 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 by Pratt (1959).

Theorem 1. Suppose the algorithm estimator chooses an attractor as the final estimator where there are K attractors and K is fixed.

i) If all of the attractors are consistent, then the algorithm estimator is consistent.

ii) If all of the attractors are consistent with the same rate, e.g., n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is consistent with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

iv) The elemental concentration algorithm is zero breakdown.

Since the FMCD estimator is a zero breakdown elemental concentration algorithm, the Hubert, Rousseeuw and Van Aelst (2008) claim that "MCD can be efficiently computed with the FAST-MCD estimator" is false. Suppose K is fixed, but each randomly drawn start is iterated to convergence so that k is not fixed. Then it is not known whether the attractors are inconsistent or consistent estimators, so it is not known whether FMCD is consistent. Let γ_o be the highest percentage of large outliers that FMCD can detect reliably. Following Hawkins and Olive (2002), if n is large then for many data sets

$$\gamma_o \approx \min(0.5, 1 - [1 - (0.2)^{1/K}]^{1/(p+1)})100\%. \tag{6}$$

3. PRACTICAL CONSISTENT ROBUST ESTIMATORS

This section shows that the MB estimator is high breakdown and that the DGK estimator is $\sqrt{n}$ consistent. The new FCH estimator and the Olive (2004) MBA estimator are defined after Theorem 4. Theorem 5 shows that MBA and FCH are $\sqrt{n}$ consistent. Then new RFCH and RMVN estimators are defined and shown to be $\sqrt{n}$ consistent.

If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ is the original data set, let $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ be the contaminated data collected into an $n \times p$ matrix $\boldsymbol{W}$ after d_n of the $\boldsymbol{x}_i$ have been replaced by arbitrary outliers. If a high breakdown estimator $(T, \boldsymbol{C}) \equiv (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ is evaluated on the contaminated data $\boldsymbol{W}$, then the location estimator T is contained in some ball about the origin of radius r, and $0 < a < \lambda_n \leq \lambda_1 < b$ where the constants a, r and b depend on the clean data but not on $\boldsymbol{W}$ if the number of outliers d_n satisfies $0 \leq d_n < n\gamma_n < n/2$ where the breakdown value $\gamma_n \to 0.5$ as $n \to \infty$.

The following theorem is closely related to a result in Olive (2004) and will be used to show that if the classical estimator $(\overline{\boldsymbol{x}}_B, \boldsymbol{S}_B)$ is applied to $c_n \approx n/2$ cases contained in a ball about the origin of radius r where r depends on the clean data but not on $\boldsymbol{W}$, then $(\overline{\boldsymbol{x}}_B, \boldsymbol{S}_B)$ is a high breakdown estimator.

Theorem 2. If the classical estimator $(\overline{\boldsymbol{x}}_B, \boldsymbol{S}_B)$ is applied to c_n cases that are contained in some bounded region where $p + 1 \leq c_n \leq n$, then the maximum eigenvalue λ_1 of $\boldsymbol{S}_B$ is bounded.

The proof of the following theorem implies that a high breakdown estimator $(T, \boldsymbol{C})$ has $\text{MED}(D_i^2) \leq V$ and that the hyperellipsoid $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq D^2_{(c_n)}\}$ that contains c_n of the cases is in some ball about the origin of radius r, where V and r do not depend on the outliers even if the number of outliers is close to $n/2$. Also the attractor of a high breakdown estimator is a high breakdown estimator if the number of concentration steps k is fixed, e.g., $k = 10$. The theorem implies that the MB estimator $(T_{k,M}, \boldsymbol{C}_{k,M})$ is high breakdown. Olive (2004) proved this result for $k = 0$ and Arcones (1995) showed that $T_{0,M}$ is a $\sqrt{n}$ consistent high breakdown estimator of $\boldsymbol{\mu}$.

Theorem 3. Suppose $(T, \boldsymbol{C})$ is a high breakdown estimator where $\boldsymbol{C}$ is a symmetric, positive definite $p \times p$ matrix if the contamination proportion d_n/n is less than the breakdown value. Then the concentration attractor $(T_k, \boldsymbol{C}_k)$ is a high breakdown estimator if the coverage $c_n \approx n/2$ and the data are in general position.

Lopuhaä (1999) and Pratt (1959) will be used to provide simple proofs for the theory of the new FCH, RFCH and RMVN estimators. Lopuhaä (1999) shows that if a start $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$, then the attractor is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a, s > 0$ are some constants. Also, the attractor and the start have the same rate. If the start is inconsistent, then so is the attractor. The constant a depends on s, p, and on the elliptically contoured distribution, but does not otherwise depend on the consistent start. The constant a also depends on h^2 in the weight function $I(D_i^2(T, \boldsymbol{C}) \leq h^2)$ where h^2 is a positive constant and the indicator is 1 if $D_i^2(T, \boldsymbol{C}) \leq h^2$ and 0 otherwise.

To see that the Lopuhaä (1999) theory extends to concentration where the weight function uses $h^2 = D^2_{(c_n)}(T, \boldsymbol{C})$, note that $(T, \tilde{\boldsymbol{C}}) \equiv (T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$ where $b > 0$ is derived in (10), and weight function $I(D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1)$ is equivalent to the concentration weight function $I(D_i^2(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C}))$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^δ where $0 < \delta \leq 0.5$, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= s^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}). \tag{7}$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $s^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose $c_n/n \to \xi \in (0,1)$ as $n \to \infty$, and let $D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the ξth percentile of the population squared distances. Then $D_{(c_n)}^2(T, \boldsymbol{C}) \xrightarrow{P} s^{-1} D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $b\boldsymbol{\Sigma} = s^{-1} D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) s\boldsymbol{\Sigma} = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}$. Thus

$$b = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{8}$$

does not depend on $s > 0$ or $\delta \in (0, 0.5]$. Theorem 4 shows that $a = a_{MCD}$ where $\xi = 0.5$. Hence concentration with a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ as a start results in a consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with rate n^δ. This result can be applied iteratively for a finite number of concentration steps. Hence DGK is a $\sqrt{n}$ consistent estimator of the same quantity that MCD is estimating. It is not known if the results hold if concentration is iterated to convergence. For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi_p^2$.

The following assumption (E1) gives a class of distributions where we can prove that the new robust estimators are $\sqrt{n}$ consistent. Cator and Lopuhaä (2009, 2010) show that MCD is consistent provided that the MCD functional is unique. Distributions where the functional is unique are called “unimodal,” and rule out, for example, a spherically symmetric uniform distribution. Theorem 4 shows that under (E1), both MCD and DGK are estimating $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

Assumption (E1): The $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a “unimodal” $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Theorem 4. Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where the constants $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ computed from the $c_n \approx n/2$ of cases with the smallest distances $D_i(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with the same rate n^δ.

Next we define the new easily computed robust $\sqrt{n}$ consistent FCH estimator, so named since it is fast, consistent and uses a high breakdown attractor. The FCH and MBA estimators use the $\sqrt{n}$ consistent DGK estimator $(T_{k,D}, \boldsymbol{C}_{k,D})$ and the high breakdown MB estimator $(T_{k,M}, \boldsymbol{C}_{k,M})$ as attractors. The MBA estimator uses the attractor with the smallest determinant. The difference between the FCH and MBA estimators is that the FCH estimator also uses a location criterion to choose the attractor: if the DGK location estimator $T_{k,D}$ has a greater Euclidean distance from MED$(\boldsymbol{X})$ than half the data, then FCH uses the MB attractor. The FCH estimator only uses the attractor with the smallest determinant if $\|T_{k,D} - \text{MED}(\boldsymbol{X})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{X}), \boldsymbol{I}_p))$. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_F, \boldsymbol{C}_F)$ takes $T_F = T_A$ and

$$\boldsymbol{C}_F = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{9}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom and F is the MBA or FCH estimator. We conjecture that FCH is high breakdown.

Theorem 5. T_{FCH} is high breakdown. Suppose (E1) holds. If $(T_A, \boldsymbol{C}_A)$ is the DGK or MB attractor with the smallest determinant, then $(T_A, \boldsymbol{C}_A)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence the MBA and FCH estimators are outlier resistant $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c = 1$ for multivariate normal data.

Many variants of the FCH and MBA estimators can be given where the algorithm gives a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$. One such variant uses K starts $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ that are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, s_j\boldsymbol{\Sigma})$ where $s_j > 0$. The MCD criteria is used to choose the final attractor, and scaling is done as in (11). A second variant is the same as the first, but the Kth attractor is replaced by the MB estimator, and for $j < K$ the jth attractor $(T_{k,j}, \boldsymbol{C}_{k,j})$ is not used if $T_{k,j}$ has a greater Euclidean distance from MED$(\boldsymbol{X})$ than half the data. Then the location estimator of the algorithm is high breakdown.

We also considered several estimators that use the MB and DGK estimators as attractors. CMVE is a concentration algorithm like FCH, but the "MVE" criterion is used in place of the MCD criterion. A standard method of reweighting can be used to produce the RMBA, RFCH and RCMVE estimators. RMVN uses a slightly modified method of reweighting so that RMVN gives good estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate normal data, even when certain types of outliers are present.

The RFCH estimator uses two standard reweighting steps. Let $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ be the classical estimator applied to the n_1 cases with $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$, and let

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \leq \chi^2_{p,0.975}$, and let

$$\boldsymbol{C}_{RFCH} = \frac{\text{MED}(D_i^2(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_2.$$

RMBA and RFCH are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$, but the two estimators use nearly 97.5% of the cases if the data is multivariate normal. We conjecture CMVE and RMVE are also $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$.

The RMVN estimator uses $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ and n_1 as above. Let $q_1 = \min\{0.5(0.975)n/n_1, 0.995\}$, and

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,q_1}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the n_2 cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)) \leq \chi^2_{p,0.975}$. Let $q_2 = \min\{0.5(0.975)n/n_2, 0.995\}$, and

$$\boldsymbol{C}_{RMVN} = \frac{\text{MED}(D_i^2(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,q_2}} \tilde{\boldsymbol{\Sigma}}_2.$$

The RMVN estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$ and $d = u_{0.5}/\chi^2_{p,q}$ where $q_2 \to q$ in probability as $n \to \infty$. Here $0.5 \leq q < 1$ depends on the elliptically contoured distribution, but $q = 0.5$ and $d = 1$ for multivariate normal data.

If the bulk of the data is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the RMVN estimator can give useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for certain types of outliers where FCH and RFCH estimate $(\boldsymbol{\mu}, d_E\boldsymbol{\Sigma})$ for $d_E > 1$. To see this claim, let $0 \leq \gamma < 0.5$ be the outlier proportion. If $\gamma = 0$, then $n_i/n \xrightarrow{P} 0.975$ and $q_i \xrightarrow{P} 0.5$. If $\gamma > 0$, suppose the outlier configuration is such that the $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH})$ are roughly χ^2_p for the clean cases, and the outliers have larger D_i^2 than the clean cases. Then $\text{MED}(D_i^2) \approx \chi^2_{p,q}$ where $q = 0.5/(1-\gamma)$. For example, if $n = 100$ and $\gamma = 0.4$, then there are 60 clean cases, $q = 5/6$, and the quantile $\chi^2_{p,q}$ is being estimated instead of $\chi^2_{p,0.5}$. Now $n_i \approx n(1-\gamma)0.975$, and q_i estimates q. Thus $\boldsymbol{C}_{RMVN} \approx \boldsymbol{\Sigma}$. Of course consistency cannot generally be claimed when outliers are present.

Simulations suggested $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ gives useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a variety of outlier configurations. Using 20 runs and $n = 1000$, the averages of the dispersion matrices were computed when the bulk of the data are iid $N_2(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2)$.

Table 1: Average Dispersion Matrices for Near Point Mass Outliers

RMVN	FMCD	OGK	MB
$\begin{bmatrix} 1.002 & -0.014 \\ -0.014 & 2.024 \end{bmatrix}$	$\begin{bmatrix} 0.055 & 0.685 \\ 0.685 & 122.46 \end{bmatrix}$	$\begin{bmatrix} 0.185 & 0.089 \\ 0.089 & 36.244 \end{bmatrix}$	$\begin{bmatrix} 2.570 & -0.082 \\ -0.082 & 5.241 \end{bmatrix}$

Table 2: Average Dispersion Matrices for Mean Shift Outliers

RMVN	FMCD	OGK	MB
$\begin{bmatrix} 0.990 & 0.004 \\ 0.004 & 2.014 \end{bmatrix}$	$\begin{bmatrix} 2.530 & 0.003 \\ 0.003 & 5.146 \end{bmatrix}$	$\begin{bmatrix} 19.671 & 12.875 \\ 12.875 & 39.724 \end{bmatrix}$	$\begin{bmatrix} 2.552 & 0.003 \\ 0.003 & 5.118 \end{bmatrix}$

For clean data, FCH, RFCH and RMVN give $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while FMCD and OGK seem to be approximately unbiased for $\boldsymbol{\Sigma}$. The median ball estimator was scaled using (11) and estimated $diag(1.13, 1.85)$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((0, 15)^T, 0.0001\boldsymbol{I}_2)$, a near point mass at the major axis. FCH, MB and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. FMCD and OGK failed to estimate $d\ \boldsymbol{\Sigma}$. Note that $\chi^2_{2,5/6}/\chi^2_{2,0.5} = 2.585$. See Table 1.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((20, 20)^T, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Rocke and Woodruff (1996) suggest that outliers with mean shift are hard to detect. FCH, FMCD, MB and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$, and OGK failed. See Table 2.

Example 1. Tremearne (1911) recorded *height* $= x_1$ and *height while kneeling* $= x_2$ of 112 people. Figure 1 shows a scatterplot of the data. Case 3 has the largest Euclidean distance of 214.767 from $\text{MED}(\boldsymbol{X}) = (1680, 1240)^T$, but if the distances correspond to the

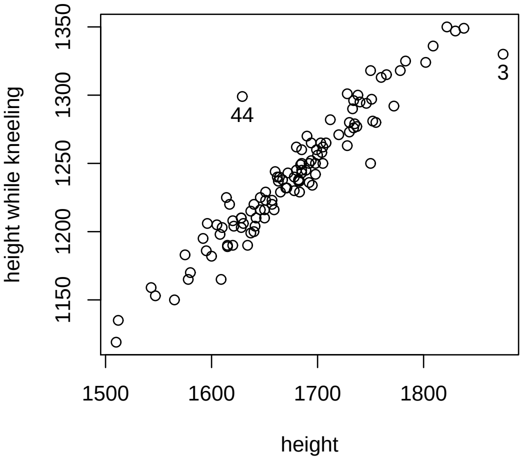

Figure 1: Scatterplot for Tremearne (1911) Data

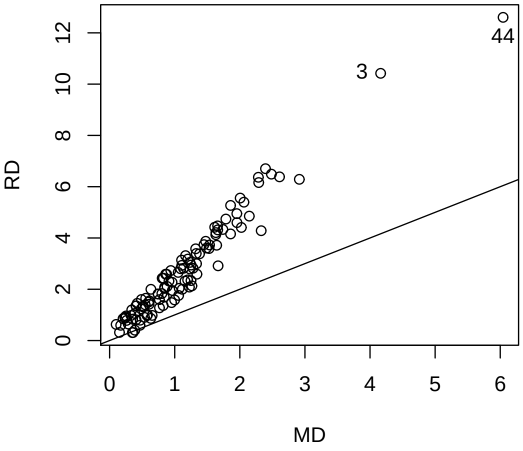

Figure 2: DD Plot for Tremearne (1911) Data

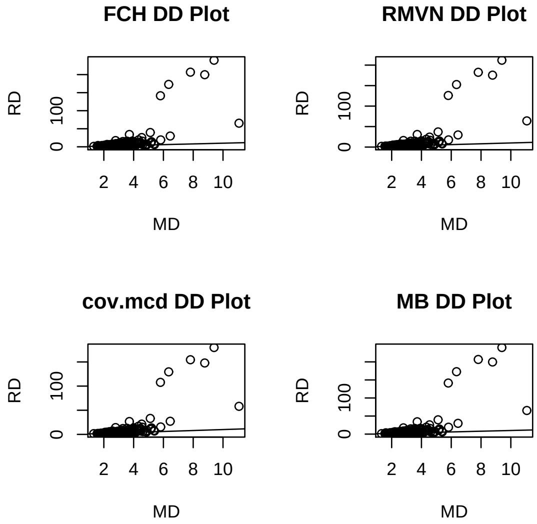

Figure 3: DD Plots for Gladstone Data

contours of a covering ellipsoid, then case 44 has the largest distance. The hypersphere (circle) centered at MED($\boldsymbol{X}$) that covers half the data is small because the data density is high near MED($\boldsymbol{X}$). The median Euclidean distance is 59.661 and case 44 has Euclidean distance 77.987. Hence the intersection of the sphere and the data is a highly correlated clean ellipsoidal region. The Rousseeuw and Van Driessen (1999) DD plot is a plot of classical distances (MD) versus "robust" distances (RD). Figure 2 shows the DD plot using the MB estimator. Notice that both the classical and MB estimators give the largest distances to cases 3 and 44. As the dimension p gets larger, outliers that can not be detected by marginal methods (case 44 in Example 1) become harder to detect.

Example 2. The estimators can be useful when the data is not elliptically contoured. The Gladstone (1905-6) data has 11 variables on 267 persons after death. Head measurements were *breadth, circumference, head height, length* and *size* as well as *cephalic index* and *brain weight. Age*, *height* and two categorical variables *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. The OGK and FAST-MCD estimators were singular. Figure 3 shows the DD plots for the FCH, RMVN, `cov.mcd` (from R version 2.4.1) and MB estimators. The DD plots from the DGK, MBA, CMVE, RCMVE and RFCH estimators were similar, and the six outliers in Figure 3 correspond to the six infants in the data set.

Olive (2002) showed that if a consistent robust estimator is scaled as in (11), then the plotted points in the DD plot will cluster about the identity line with unit slope and zero intercept if the data is multivariate normal, and about some other line through the origin if the data is from some other elliptically contoured distribution with a nonsingular covariance matrix. Since multivariate procedures tend to perform well for elliptically contoured data, the DD plot is useful even if outliers are not present.

If $W_{in} \sim N(0, \tau^2/n)$ for $i = 1, ..., r$ and if S^2_W is the sample variance of the W_{in}, then $E(nS^2_W) = \tau^2$ and $V(nS^2_W) = 2\tau^4/(r-1)$. So $nS^2_W \pm \sqrt{5}SE(nS^2_W) \approx \tau^2 \pm \sqrt{10}\tau^2/\sqrt{r-1}$. So for $r = 1000$ runs, expect nS^2_W to be between $\tau^2 - 0.1\tau^2$ and $\tau^2 + 0.1\tau^2$ with high confidence. Similar results hold for many estimators if W_{in} is $\sqrt{n}$ consistent and asymptotically normal and if n is large enough. If W_{in} has less than $\sqrt{n}$ rate, e.g. $n^{1/3}$ rate, then the scaled sample variance $nS^2_W \to \infty$ as $n \to \infty$.

Table 3 considers $W = T_p$ and $W = C_{p,p}$ for eight estimators, $p = 5$ and 10 and $n = 10p$ and 5000 when $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, ..., p))$. For the classical estimator, denoted by CLAS, $T_p = \overline{x}_p \sim N(0, p/n)$, and $nS^2(T_p) \approx p$ while $C_{p,p}$ is the sample variance of n iid $N(0, p)$ random variables. Hence $nS^2(C_{p,p}) \approx 2p^2$. RFCH, RMVN, FMCD and possibly OGK use a "reweight for efficiency" concentration step that uses a random number of cases with percentage close to 97.5%. These four estimators had similar behavior. DGK,

Table 3: Scaled Variance $nS^2(T_p)$ and $nS^2(C_{p,p})$

p	n	V	FCH	RFCH	RMVN	DGK	OGK	CLAS	FMCD	MB
5	50	C	216.0	72.4	75.1	209.3	55.8	47.12	153.9	145.8
5	50	T	12.14	6.50	6.88	10.56	6.70	4.83	8.38	13.23
5	5000	C	307.6	64.1	68.6	325.7	59.3	48.5	60.4	309.5
5	5000	T	18.6	5.34	5.33	19.33	6.61	4.98	5.40	20.20
10	100	C	817.3	276.4	286.0	725.4	229.5	198.9	459.6	610.4
10	100	T	21.40	11.42	11.68	20.13	12.75	9.69	14.05	24.13
10	5000	C	955.5	237.9	243.8	966.2	235.8	202.4	233.6	975.0
10	5000	T	29.12	10.08	10.09	29.35	12.81	9.48	10.06	30.20

FCH and MB used about 50% of the cases and had similar behavior. By Lopuhaä (1999), estimators with less than $\sqrt{n}$ rate still have zero efficiency after the reweighting. Although FMCD, MB and OGK have not been proven to be $\sqrt{n}$ consistent, their values did not blow up even for $n = 5000$.

4. OUTLIER RESISTANCE

Geometrical arguments suggest that the MB estimator has considerable outlier resistance. Suppose the outliers are far from the bulk of the data. Let the “median ball” correspond to the half set of data closest to MED($\boldsymbol{X}$) in Euclidean distance. If the outliers are outside of the median ball, then the initial half set in the iteration leading to the MB estimator will be clean. Thus the MB estimator will tend to give the outliers the largest MB distances unless the initial clean half set has very high correlation in a direction about which the outliers lie. This property holds for very general outlier configurations. The FCH estimator tries to use the DGK attractor if the $det(\boldsymbol{C}_{DGK})$ is small and the DGK location estimator T_{DGK} is in the median ball. Distant outliers that make $det(\boldsymbol{C}_{DGK})$ small also drag T_{DGK} outside of the median ball. Then FCH uses the MB attractor.

Compared to OGK and FMCD, the MB estimator is vulnerable to outliers that lie within the median ball. If the bulk of the data is highly correlated with the major axis of an ellipsoidal region, then the distances based on the clean data can be very large for outliers that fall within the median ball. The outlier resistance of the MB estimator decreases as p increases since the volume of the median ball rapidly increases with p.

A simple simulation for outlier resistance is to count the number of times the minimum distance of the outliers is larger than the maximum distance of the clean cases. The simulation used 100 runs. If the count was 97, then in 97 data sets the outliers can be separated from the clean cases with a horizontal line in the DD plot, but in 3 data sets the robust distances did not achieve complete separation.

The clean cases had $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, 2, ..., p))$. Outlier types were the mean shift $\boldsymbol{x} \sim N_p(pm\boldsymbol{1}, diag(1, 2, ..., p))$ where $\boldsymbol{1} = (1, ..., 1)^T$, and $\boldsymbol{x} \sim N_p((0, ..., 0, pm)^T, 0.0001$ $\boldsymbol{I}_p)$, a near point mass at the major axis. Notice that the clean data can be transformed to a $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ distribution by multiplying $\boldsymbol{x}_i$ by $\text{diag}(1, 1/\sqrt{2}, ..., 1/\sqrt{p})$, and this transformation changes the location of the near point mass to $(0, ..., 0, pm/\sqrt{p})^T$.

For near point mass outliers, an ellipsoid with very small volume can cover half of the data if the outliers are at one end of the ellipsoid and some of the clean data are at the other end. This half set will produce a classical estimator with very small determinant

Table 4: Number of Times Mean Shift Outliers had the Largest Distances

p	γ	n	pm	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	4	49	49	85	84	38	76	57
10	.1	100	5	91	91	99	99	93	98	91
10	.4	100	7	90	90	90	90	0	48	100
40	.1	100	5	3	3	3	3	76	3	17
40	.1	100	8	36	36	37	37	100	49	86
40	.25	100	20	62	62	62	62	100	0	100
40	.4	100	20	20	20	20	20	0	0	100
40	.4	100	35	44	98	98	98	95	0	100
60	.1	200	10	49	49	49	52	100	30	100
60	.1	200	20	97	97	97	97	100	35	100
60	.25	200	25	60	60	60	60	100	0	100
60	.4	200	30	11	21	21	21	17	0	100
60	.4	200	40	21	100	100	100	100	0	100

by (5). In the simulations for large γ, as the near point mass is moved very far away from the bulk of the data, only the classical, MB and OGK estimators did not have numerical difficulties. Since the MCD estimator has smaller determinant than DGK while MVE has smaller volume than DGK, estimators like FAST-MCD and MBA that use the MVE or MCD criterion without using location information will be vulnerable to these outliers. FAST-MCD is also vulnerable to outliers if γ is slightly larger than γ_o given by (6).

Tables 4 and 5 help illustrate the results for the simulation. Large counts and small pm for fixed γ suggest greater ability to detect outliers. Values of p were 5, 10, 15, ..., 60. First consider the mean shift outliers and Table 4. For $\gamma = 0.25$ and 0.4, MB usually had the highest counts. For $5 \leq p \leq 20$ and the mean shift, the OGK estimator often had the smallest counts, although FMCD could not handle 40% outliers for $p = 20$. For $25 \leq p \leq 60$, OGK usually had the highest counts for $\gamma = 0.05$ and 0.1. For $p \geq 30$, FMCD could not handle 25% outliers even for enormous values of pm.

In Table 5, FCH greatly outperformed MBA although the only difference between the two estimators is that FCH uses a location criterion as well as the MCD criterion. OGK performed well for $\gamma = 0.05$ and $20 \leq p \leq 60$ (not tabled). For large γ, OGK often has large bias for $c\boldsymbol{\Sigma}$. Then the outliers may need to be enormous before OGK can detect them. Also see Table 2, where OGK gave the outliers the largest distances for all runs, but $\boldsymbol{C}_{OGK}$ does not give a good estimate of $c\boldsymbol{\Sigma} = c \;\; diag(1,2)$.

5. CONCLUDING REMARKS

Now that practical outlier resistant $\sqrt{n}$ consistent estimators have been shown to exist, they can be used for outlier detection and inference. We recommend using FCH

Table 5: Number of Times Near Point Mass Outliers had the Largest Distances

p	γ	n	pm	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	40	73	92	92	92	100	95	100
10	.25	100	25	0	99	99	90	0	0	99
10	.4	100	25	0	100	100	100	0	0	100
40	.1	100	80	0	0	0	0	79	0	80
40	.1	100	150	0	65	65	65	100	0	99
40	.25	100	90	0	88	87	87	0	0	88
40	.4	100	90	0	91	91	91	0	0	91
60	.1	200	100	0	0	0	0	13	0	91
60	.25	200	150	0	100	100	100	0	0	100
60	.4	200	150	0	100	100	100	0	0	100
60	.4	200	20000	0	100	100	100	64	0	100

instead of MBA, and RFCH or RMVN instead of RMBA. RMVN is useful for inference if the parametric $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ model is reasonable, and estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are needed. If estimates of $\boldsymbol{\mu}$ and $c\boldsymbol{\Sigma}$ are needed, then RMVN, RFCH and FCH can be used for inference on the large class of elliptically contoured distributions that satisfy (E1).

Note that large sample inference is often immediate. For example, since RMVN is a consistent estimator of $c \operatorname{Cov}(\boldsymbol{x})$ under (E1), the correlation of the eigenvalues computed from the classical estimator and from RMVN converges to 1 in probability. RMVN, RFCH and FCH can also be used as the plug in estimators, replacing estimators such as MBA, RMBA, FAST-MCD and OGK. There are many applications including standard multivariate methods such as canonical analysis, discrimination, factor analysis, principal components and regression. See Hubert, Rousseeuw and Van Aelst (2008), Maronna, Martin and Yohai (2006), Reyen, Miller and Wegman (2009), and Wilcox (2008ab, 2009, 2010). Applications for dimension reduction methods such as 1D regression and sliced inverse regression include Chang and Olive (2010), Cook and Nachtsheim (1994) and Olive (2002).

The new estimators can also be used to improve outlier diagnostics. Making a scatterplot matrix of the classical, DGK, MB, OGK and FAST-MCD distances is useful.

Simulations were done in *R*. Programs are in the collection of functions *rpack.txt* at (www.math.siu.edu/olive/ol-bookp.htm). The `robustbase` library was downloaded from (www.r-project.org/#doc) to compute OGK and FAST-MCD. The `rpack` function *mldsim* was used for Tables 1 to 5. The function *cmve* computes CMVE and RCMVE, function *covfch* computes FCH and RFCH while *covrmvn* computes the RMVN and MB estimators. The function *covrmb* computes MB and RMB where RMB is like RMVN except the MB estimator is reweighted instead of FCH.

APPENDIX

Proof of Theorem 1. i) Choosing from K consistent estimators results in an consistent estimator, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$. iv) The classical estimator with breakdown $1/n$ is applied to each elemental start. Hence $\gamma_n \leq K/n \to 0$ as $n \to \infty$.

Proof of Theorem 2. The largest eigenvalue of a $p \times p$ matrix $\boldsymbol{A}$ is bounded above by $p \max |a_{i,j}|$ where $a_{i,j}$ is the (i,j) entry of $\boldsymbol{A}$. See Datta (1995, p. 403). Denote the c_n cases by $\boldsymbol{z}_1, ..., \boldsymbol{z}_{c_n}$. Then the (i,j)th element $a_{i,j}$ of $\boldsymbol{A} = \boldsymbol{S}_B$ is

$$a_{i,j} = \frac{1}{c_n - 1} \sum_{m=1}^{c_n} (z_{i,m} - \overline{z}_i)(z_{j,m} - \overline{z}_j).$$

Hence the maximum eigenvalue λ_1 is bounded.

Proof of Theorem 3. Following Leon (1986, p. 280), if $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalues $\tau_1 \geq \cdots \geq \tau_n$, then for any nonzero vector $\boldsymbol{x}$,

$$0 < \|\boldsymbol{x}\|^2 \, \tau_n \leq \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|^2 \, \tau_1. \tag{10}$$

Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of $\boldsymbol{C}$. By (7),

$$\frac{1}{\lambda_1}\|\boldsymbol{x} - T\|^2 \leq (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq \frac{1}{\lambda_n}\|\boldsymbol{x} - T\|^2. \tag{11}$$

By (8), if the $D^2_{(i)}$ are the order statistics of the $D^2_i(T, \boldsymbol{C})$, then $D^2_{(i)} < V$ for some constant V that depends on the clean data but not on the outliers even if i and d_n are near $n/2$. (Note that $1/\lambda_n$ and MED($\|\boldsymbol{x}_i - T\|^2$) are both bounded for high breakdown estimators even for d_n near $n/2$.)

Following Johnson and Wichern (1988, pp. 50, 103), the boundary of the set $\{\boldsymbol{x}|(\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) \leq h^2\} = \{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq h^2\}$ is a hyperellipsoid centered at T with axes of length $2h\sqrt{\lambda_i}$. Hence $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq D^2_{(c_n)}\}$ is contained in some ball about the origin of radius r where r does not depend on the number of outliers even for d_n near $n/2$. This is the set containing the cases used to compute $(T_0, \boldsymbol{C}_0)$. Since the set is bounded, T_0 is bounded and the largest eigenvalue $\lambda_{1,0}$ of $\boldsymbol{C}_0$ is bounded by Theorem 2. Since $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_0)$, the smallest eigenvalue $\lambda_{n,0}$ is bounded away from 0. Since these bounds do not depend on the outliers even for d_n near $n/2$, $(T_0, \boldsymbol{C}_0)$ is a high breakdown estimator. Now repeat the argument with $(T_0, \boldsymbol{C}_0)$ in place of $(T, \boldsymbol{C})$ and $(T_1, \boldsymbol{C}_1)$ in place of $(T_0, \boldsymbol{C}_0)$. Then $(T_1, \boldsymbol{C}_1)$ is high breakdown. Repeating the argument iteratively shows $(T_k, \boldsymbol{C}_k)$ is high breakdown.

Proof of Theorem 4. By Lopuhaä (1999) the estimator is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate n^δ. By the remarks above, a will be the same for any consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ and a does not depend on $s > 0$ or $\delta \in (0, 0.5]$. Hence the result follows if $a = a_{MCD}$. The MCD estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ by Butler, Davies and Jhun (1993) and Cator and Lopuhaä (2009, 2010). If the MCD estimator is the start, then it is also the attractor by Rousseeuw and Van Driessen (1999) who show that concentration does not increase the MCD criterion. Hence $a = a_{MCD}$.

Proof of Theorem 5. T_{FCH} is high breakdown since it is a bounded distance from MED($\boldsymbol{X}$) even if the number of outliers is close to $n/2$. Under (E1) the FCH and MBA estimators are asymptotically equivalent since $\|T_{k,D} - \text{MED}(\boldsymbol{X})\| \to 0$ in probability. The estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{S}_{0,M}) < \infty$ by Theorem 3 if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$, then the result follows from Pratt (1959) and Theorem 4 since both starts are $\sqrt{n}$ consistent. Otherwise, the MB estimator $\boldsymbol{S}_{k,M}$ is a biased estimator of $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{S}_{k,D}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ by Theorem 4 and $\|\boldsymbol{C}_{MCD} - \boldsymbol{S}_{k,D}\| = O_P(n^{-1/2})$. 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 $(\overline{\boldsymbol{x}}_{k,D}, \boldsymbol{S}_{k,D})$.

Let $P(U \leq u_\alpha) = \alpha$ where U is given by (3). Then the scaling in (11) makes $\boldsymbol{C}_F$ a consistent estimator of $c\boldsymbol{\Sigma}$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data.

REFERENCES

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

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

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

Cator, E. A., and Lopuhaä, H. P. (2009), "Central Limit Theorem and Influence Function for the MCD Estimators at General Multivariate Distributions," preprint. See (http://arxiv.org/abs/0907.0079).

Cator, E.A., and Lopuhaä, H.P. (2010), "Asymptotic Expansion of the Minimum Covariance Determinant Estimators," *Journal of Multivariate Analysis,* 101, 2372-2388.

Chang, J., and Olive, D. J. (2010), "OLS for 1D Regression Models," *Communications in Statistics: Theory and Methods*, 39, 1869-1882.

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.

Datta, B. N. (1995), *Numerical Linear Algebra and Applications,* Pacific Grove, CA: Brooks/Cole.

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.

Gladstone, R. J. (1905-6), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

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.

Huber, P. J., and Ronchetti, E. M. (2009), *Robust Statistics,* 2nd ed., Hoboken, NJ: Wiley.

Hubert, M., Rousseeuw, P. J., and Van Aelst, S. (2008), "High Breakdown Multivariate Methods," *Statistical Science*, 23, 92-119.

Hubert, M., Rousseeuw, P.J., and Verdonck, T. (2010), "A Deterministic Algorithm for the MCD," technical report, (http://wis.kuleuven.be/stat/robust/publications.html).

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

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

Leon, S. J. (1986), *Linear Algebra with Applications*, 2nd ed., New York, NY: Macmillan Publishing Company.

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

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, Hoboken, NJ: Wiley.

Maronna, R. A., and Zamar, R. H. (2002), "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," *Technometrics,* 50, 295-304.

Olive, D. J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

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

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

Reyen, S. S., Miller, J. J., and Wegman, E. J. (2009), "Separating a Mixture of Two Normals with Proportional Covariances," *Metrika*, 70, 297-314.

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 Van Driessen, K. (1999), “A Fast Algorithm for the Minimum Covariance Determinant Estimator,” *Technometrics,* 41, 212-223.

Tremearne, A. J. N. (1911), “Notes on Some Nigerian Tribal Marks,” *Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 41, 162-178.

Wilcox, R. R. (2008a), “Some Small–Sample Properties of Some Recently Proposed Outlier Detection Techniques,” *Journal of Statistical Computation and Simulation*, 78, 701-712.

Wilcox, R. R. (2008b), “Robust Principal Components: a Generalized Variance Perspective,” *Behavior Research Methods*, 40, 102-108.

Wilcox, R. R. (2009), “Robust Multivariate Regression When There is Heteroscedasticity,” *Communications in Statistics-Simulation and Computation*, 38, 1-13.

Wilcox, R. R. (2010), “Regression: Comparing Predictors and Groups of Predictors Based on a Robust Measure of Association,” *Journal of Data Science*, 8, 429-441.