Applications of a Robust Dispersion Estimator

Jianfeng Zhang and David J. Olive*

Department of Mathematics, Southern Illinois University,

Mailcode 4408, Carbondale, IL 62901-4408, USA

September 6, 2009

Abstract

The RMVN estimator is an easily computed high breakdown robust $\sqrt{n}$ consistent estimator of multivariate location and dispersion, and the estimator is obtained by scaling the classical estimator applied to the "RMVN subset" that contains at least half of the cases. The applications for this estimator are numerous, and a simple method for performing robust principal component analysis, canonical correlation analysis and factor analysis is to apply the classical method to the "RMVN subset."

Keywords: Minimum covariance determinant estimator; Multivariate location and dispersion; Outliers

* *E-mail Address:* dolive@math.siu.edu

1. Introduction

Multivariate location and dispersion considers estimation of a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. 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 classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ of multivariate location and dispersion is the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \text{ and } \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})(\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})^{\mathrm{T}}. \tag{1.1}$$

An important model is the 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 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. For these distributions, the covariance matrix $\mathrm{Cov}(\boldsymbol{x}) = c_X \boldsymbol{\Sigma}$ if second moments exist where $c_X > 0$, and a dispersion estimator estimates $d\ \boldsymbol{\Sigma}$ for some $d > 0$. Many classical procedures originally meant for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution are semiparametric in that the procedures also perform well on a much larger class of elliptically contoured distributions.

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

$$D_i^2 = D_i^2(T(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X})) = (\boldsymbol{x}_i - T(\boldsymbol{X}))^T \boldsymbol{C}^{-1}(\boldsymbol{X})(\boldsymbol{x}_i - T(\boldsymbol{X})) \tag{1.2}$$

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

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \tag{1.3}$$

and for elliptically contoured distributions, U has density

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{1.4}$$

Olive and Hawkins (2009) provided the first practical estimators of multivariate location and dispersion that have been shown to be both high breakdown and $\sqrt{n}$ consistent. Competing estimators with theory have high computational complexity. The minimum covariance determinant (MCD) estimator has $O(n^v)$ complexity where $v = 1+p(p+3)/2$. The minimum volume ellipsoid (MVE) complexity is far higher, and there may be no known method for computing the S, τ, projection based, constrained M, MM, and Stahel-Donoho estimators described in Maronna, Martin and Yohai (2006, ch. 6).

In the literature, the above competing estimators are replaced by practical estimators, but none of the practical estimators have been shown to be both high breakdown and consistent. For example, the Rousseeuw and Van Driessen (1999) FAST-MCD (FMCD) estimator is used to replace the MCD estimator. Olive and Hawkins (2009) show that FAST-MCD is not a high breakdown estimator. Maronna and Zamar (2002, p. 309) claim, without proof, that their orthogonalized Gnanadesikan-Kettenring (OGK) estimator is consistent and high breakdown.

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.

Following Olive and Hawkins (2009), note that $(T_{t,j}, \boldsymbol{C}_{t,j}) = (\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ is the classical estimator applied to the "half set" of cases satisfying $\{\boldsymbol{x}_i : D_i^2(\overline{\boldsymbol{x}}_{t-1,j}, \boldsymbol{S}_{t-1,j}) \leq D^2_{(c_n)}(\overline{\boldsymbol{x}}_{t-1,j}, \boldsymbol{S}_{t-1,j})\}$ for $t \geq 0$. Hence $(T_{t,j}, \boldsymbol{C}_{t,j})$ is estimating $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, the population mean and covariance matrix of the truncated distribution covering half of the mass corresponding to $\{\boldsymbol{x} : (\boldsymbol{x}-\boldsymbol{\mu}_{t-1})^T \boldsymbol{\Sigma}_{t-1}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_{t-1}) \leq D^2_{(0.5)}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})\}$ where $D^2_{(0.5)}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ is the population median of the population squared distances $D^2(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$. Here $(\boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}_{-1})$ is the population analog of $(T_{-1,j}, \boldsymbol{C}_{-1,j})$.

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. Thus $(\boldsymbol{\mu}_{-1,D}, \boldsymbol{\Sigma}_{-1,D})$ is the population mean and covariance matrix. For an elliptically contoured distribution with a nonsingular covariance matrix and for $t \geq 0$, $(\boldsymbol{\mu}_{t,D}, \boldsymbol{\Sigma}_{t,D})$ is the population mean and covariance matrix of the truncated distribution corresponding to the highest density region covering half the mass. Hence $\boldsymbol{\mu}_{t,D} = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{t,D} = c\boldsymbol{\Sigma}$ for some $c > 0$. Atkinson, Riani and Cerioli (2009) find the population mean and covariance matrices for such truncated multivariate normal distributions. We conjecture that the DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{k,D}, \boldsymbol{\Sigma}_{k,D})$ under mild conditions.

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 $\text{MED}(\boldsymbol{X})$ 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 $\text{MED}(\boldsymbol{X})$ in Euclidean distance while $(\boldsymbol{\mu}_{0,M}, \boldsymbol{\Sigma}_{0,M})$ is the population mean and covariance matrix of the truncated distribution corresponding to the hypersphere centered at the population median that contains half the mass. For a distribution that is spherical about $\boldsymbol{\mu}$ and for $t \geq 0$, $(\boldsymbol{\mu}_{t,M}, \boldsymbol{\Sigma}_{t,M}) = (\boldsymbol{\mu}, c\boldsymbol{I}_p)$ for some $c > 0$. For nonspherical elliptically contoured distributions, $\boldsymbol{\Sigma}_{t,M} \neq c\boldsymbol{\Sigma}$. However, the bias seems to be small even for $t = 0$, and to get smaller as k increases. If the median ball estimator is iterated to convergence, we do not know whether $\boldsymbol{\Sigma}_{\infty,M} = c\boldsymbol{\Sigma}$. We conjecture that the MB estimator is a high breakdown $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{k,M}, \boldsymbol{\Sigma}_{k,M})$ under mild conditions.

The FCH estimator uses the DGK estimator $(T_{k,D}, \boldsymbol{C}_{k,D})$ and the MB estimator $(T_{k,M}, \boldsymbol{C}_{k,M})$ as attractors. Let the "median ball" be the hypersphere containing the half set of data closest to $\text{MED}(\boldsymbol{X})$ in Euclidean distance. The FCH estimator uses the MB attractor if the DGK location estimator T_{DGK} is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{1.5}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom.

The RMVN estimator uses two reweighting steps. Let $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ be the classical estimator applied to the n_1 cases with $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$. 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 $(T_U, \boldsymbol{C}_U) \equiv (T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$ 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

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

Note that $(T_U, \boldsymbol{C}_U)$ is the unscaled RMVN estimator, and is the classical estimator applied to the "RMVN subset" of $n_2 \geq n/2$ cases. The RMVN subset tends to be a large clean subset even if certain types of outliers are present.

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 estimates $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ for $d > 1$. 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 and the $q = 5/6$ 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.

The following assumption (E1) gives a class of distributions where Olive (2008, ch. 10) and Olive and Hawkins (2009) showed that the FCH and RMVN estimators are high breakdown $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, d_R\boldsymbol{\Sigma})$ where R is the FCH or RMVN estimator, $d_R > 0$ and $d_R = 1$ for multivariate normal data. A similar result holds for $(T_U, \boldsymbol{C}_U)$, but $d_R \neq 1$ for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data. Lopuhaä (1999) showed that DGK is $\sqrt{n}$ consistent while Cator and Lopuhaä (2009) showed 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.

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 $\mathrm{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Since the Olive and Hawkins (2009) estimators such as $(T_U, \boldsymbol{C}_U)$, FCH and RMVN are the only practical estimators of multivariate location and dispersion that have been shown to be both high breakdown and $\sqrt{n}$ consistent, they are the default estimators for high breakdown inference. The OGK complexity is $O[p^3 + np^2 \log(n)]$ while that of the FCH and RMVN estimators is $O[p^3 + np^2 + np\log(n)]$. FCH and RMVN are roughly 100 times faster than FAST-MCD. Section 2 considers robust principal component analysis, Section 3 gives a diagnostic for the Hotelling's T^2 test, and Section 4 presents a small simulation study.

2. Robust Principal Component Analysis

Principal component analysis (PCA) is used to explain the dispersion structure with a few uncorrelated linear combinations of the original variables, called principal components. The analysis is used for data reduction and interpretation.

For classical principal component analysis, assume that the sample covariance matrix $\boldsymbol{S}$ has eigenvalue eigenvector pairs $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), (\hat{\lambda}_2, \hat{\boldsymbol{e}}_2), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Then the principal components corresponding to the jth case are $\hat{Y}_{1j} = \hat{\boldsymbol{e}}_1^T \boldsymbol{x}_j, ..., \hat{Y}_{pj} = \hat{\boldsymbol{e}}_p^T \boldsymbol{x}_j$. The estimated proportion of the total population variance due to the ith principal component is $\hat{\lambda}_i / \sum_{j=1}^p \hat{\lambda}_j$. The population analogs use the covariance matrix $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$ with eigenvalue eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$ for $i = 1, ..., p$. The analysis can also be based on the p eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ of the sample

correlation matrix $\boldsymbol{R}$. A robust "plug in" method uses an analysis based on the $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ computed from a robust dispersion estimator $\boldsymbol{C}$. See Croux and Haesbroeck (2000).

The RPCA method performs the classical principal component analysis on the RMVN subset, using either the sample covariance matrix $\boldsymbol{C}_U = \boldsymbol{S}_U$ or the sample correlation matrix $\boldsymbol{R}_U$. Under (E1), $\boldsymbol{C}_U$ and $\boldsymbol{R}_U$ are $\sqrt{n}$ consistent high breakdown estimators of $c\boldsymbol{\Sigma} = d\text{Cov}(\boldsymbol{X})$ and the population correlation matrix $\boldsymbol{D}\text{Cov}(\boldsymbol{X})\boldsymbol{D}$, respectively, where $\boldsymbol{D} = \text{diag}(1/\sqrt{\sigma}_{11}, ..., 1/\sqrt{\sigma}_{\text{pp}})$ and the σ_{ii} are the diagonal entries of $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$.

Theorem 1. Under (E1), the correlation of the eigenvalues computed from the classical PCA and RPCA converges to 1 in probability.

Proof: The eigenvalues are continuous functions of the dispersion estimator, hence consistent estimators of dispersion give consistent estimators of the population eigenvalues. See Eaton and Tyler (1991) and Bhatia, Elsner and Krause (1990). Under (E1), $\boldsymbol{S}$ estimates $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{C}_U$ estimates $d\ \boldsymbol{\Sigma}_{\boldsymbol{x}}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}}\ \boldsymbol{e} = \lambda \boldsymbol{e}$, then

$$d\ \boldsymbol{\Sigma}_{\boldsymbol{x}}\ \boldsymbol{e} = \frac{d}{c}\lambda c \boldsymbol{e}.$$

Hence the population eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $d\ \boldsymbol{\Sigma}_{\boldsymbol{x}}$ differ by some positive multiple d/c, and the population correlation is equal to one. The proof for $\boldsymbol{R}$ and $\boldsymbol{R}_U$ is similar. □

For principal components, a scree plot is a plot of component number versus eigenvalue, and often there is a sharp bend in the plot when the components are no longer important. See Cattell (1966). The above theorem suggests making the robust scree plot and the classical scree plot.

The eigenvectors are not continuous functions of the dispersion estimator, and the sample size may need to be massive before the robust and classical eigenvectors or prin-

cipal components have high absolute correlation. In the software, sign changes in the eigenvectors are common, since $\boldsymbol{\Sigma_x}\ \boldsymbol{e} = \lambda \boldsymbol{e}$ implies that $\boldsymbol{\Sigma_x}\ (-\boldsymbol{e}) = \lambda(-\boldsymbol{e})$.

The literature for robust PCA is large, but the "high breakdown" methods are impractical or not backed by theory. Some of these methods may be useful as outlier diagnostics. Spherical principal components is a bounded influence approach suggested by Locantore, Marron, Simpson, Tripoli, Zhang and Cohen (1999). Boente and Fraiman (1999) claim that basis of the eigenvectors is consistently estimated by spherical principal components for elliptically contoured distributions. Also see Maronna, Martin and Yohai (2006, pp. 212-213).

3. A diagnostic for the Hotelling's T_H^2 test

The Hotelling's T_H^2 test is used to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. The test rejects H_0 if

$$T_H^2 = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

if H_0 holds and the data are iid from a distribution with a nonsingular covariance matrix.

If a location estimator T satisfies

$$\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{D}),$$

then a competing test rejects H_0 if

$$T_C^2 = n(T - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{D}}^{-1}(T - \boldsymbol{\mu}_0) > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

if H_0 holds and $\hat{\boldsymbol{D}}$ is a consistent estimator of $\boldsymbol{D}$. The F cutoff can be used since

$$\frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha} \to \chi^2_{p,1-\alpha}$$

as $n \to \infty$. This idea is used for small p by Srivastava and Mudholkar (2001) where T is the coordinatewise trimmed mean.

Now the RMVN estimator is asymptotically equivalent to a scaled DGK estimator that uses $k = 5$ concentration steps and two "reweight for efficiency" steps. Lopuhaä (1999, pp. 1651-1652) shows that if (E1) holds, then for $k = 0$, the DGK estimator $(T_{0,D}, \boldsymbol{C}_{0,D})$ is asymptotically normal with

$$\sqrt{n}(T_{0,D} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \kappa_p \boldsymbol{\Sigma}).$$

We conjecture that a similar result holds after concentration:

$$\sqrt{n}(T_{RMVN} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \tau_p \boldsymbol{\Sigma})$$

for a wide variety of elliptically contoured distributions where τ_p depends on both p and the underlying distribution. Since the test is based on a conjecture, it is ad hoc, and should be used as an outlier diagnostic rather than for inference. Willems, Pison, Rousseeuw, and Van Aelst (2002) use similar reasoning to present a diagnostic based on the FMCD estimator.

For MVN data, simulations suggest that τ_p is close to 1. The ad hoc test that rejects H_0 if

$$T_R^2/f_{n,p} = n(T_{RMVN} - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{C}}_{RMVN}^{-1}(T_{RMVN} - \boldsymbol{\mu}_0)/f_{n,p} > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

where $f_{n,p} = 1.04 + 0.12/p + (40+p)/n$ gave fair results in the simulations described in the following section for $n \geq 15p$ and $2 \leq p \leq 100$.

4. Example and simulations

A simulation was done to check that RMVN estimates $\mathbf{\Sigma}$ if γ is the percentage of outliers. The clean cases were MVN: $\boldsymbol{x} \sim N_p(\mathbf{0}, diag(1, 2, ..., p))$. Outlier types were $\boldsymbol{x} \sim N_p((0, ..., 0, pm)^T, 0.0001\boldsymbol{I}_p)$, a near point mass at the major axis, and the mean shift $\boldsymbol{x} \sim N_p(pm\mathbf{1}, diag(1, 2, ..., p))$ where $\mathbf{1} = (1, ..., 1)^T$. On clean MVN data, $n \geq 20p$ gave good results for $2 \leq p \leq 100$. For the contaminated MVN data, the first $n\gamma$ cases were outliers, and the classical estimator $\boldsymbol{S}_c$ was computed on the clean cases. The diagonal elements of $\boldsymbol{S}_c$ and $\hat{\mathbf{\Sigma}}_{RMVN}$ should both be estimating $(1, 2, ..., p)^T$. The average diagonal elements of both matrices were computed for 20 runs, and the criterion Q was the sum of the absolute differences of the p diagonal elements from the two averaged matrices. Since $\gamma = 0.4$ and the initial subsets for the RMVN estimator are half sets, the simulations used $n = 35p$. The values of Q shown in Table 1 correspond to good estimation of the diagonal elements. Values of pm slightly smaller than the tabled values led to poor estimation of the diagonal elements.

Example. Buxton (1920) gives various measurements on 87 men including *height*, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five *heights* were recorded to be about 19mm with the true heights recorded under head length. Performing a classical principal components analysis on these five variables using the covariance matrix resulted in a first principal component that was created by the outliers. See Figure 1 where the second principal component is plotted versus the first. The robust PCA, or the classical PCA performed after the outliers are removed, resulted in a first principal component that was approximately $-$ *height* with $\hat{\boldsymbol{e}}_1 \approx (-1.000, 0.002, -0.023, -0.002, -0.009)^T$ while the second robust principal component was based on the eigenvector $\hat{\boldsymbol{e}}_2 \approx (-0.005, 0.848, -0.054, -0.048, 0.525)^T$. The plot of the first two robust principal components, with the outliers

deleted, is shown in Figure 2. These two components explain about 86% of the variance.

In simulations for principal component analysis, FCH, RMVN, OGK and FAST-MCD seem to estimate $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$ where $\boldsymbol{z} = (z_1, ..., z_p)^T$ and the z_i are iid from a continuous distribution with variance σ^2. Here $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T$. The bias for the MB estimator seemed to be small. It is known that affine equivariant estimators give unbiased estimators of $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if the distribution of z_i is also symmetric. DGK and FAST-MCD are affine equivariant. FCH and RMVN are asymptotically equivalent to a scaled DGK estimator. But in the simulations the results also held for skewed distributions.

The simulations used 1000 runs where $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}$ and $\boldsymbol{z} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{z} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{z} \sim MVT_p(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy(0,1). The choice $\boldsymbol{A} = diag(\sqrt{1}, ..., \sqrt{p})$ results in $\boldsymbol{\Sigma} = diag(1, ..., p)$. Note that the population eigenvalues will be proportional to $(p, p-1, ..., 1)^T$ and the population "variance explained" by the ith principal component is $\lambda_i / \sum_{j=1}^{p} \lambda_j = 2(p+1-i)/[p(p+1)]$. For $p = 4$, these numbers are 0.4, 0.3 and 0.2 for the first three principal components. If the "correlation" option is used, then the population "correlation matrix" is the identity matrix $\boldsymbol{I}_p$, the ith population eigenvalue is proportional to $1/p$ and the population "variance explained" by the ith principal component is $1/p$.

Table 2 shows the mean "variance explained" along with the standard deviations for the first three principal components. Also a_i and p_i are the average absolute value of the correlation between the ith eigenvectors or the ith principal components of the classical and robust methods. Two rows were used for each "n–data type" combination. The a_i are shown in the top row while the p_i are in the lower row. The values of a_i and p_i were

similar. The standard deviations were slightly smaller for the classical PCA for normal data. The classical method failed to estimate (0.4,0.3,0.2) for the Cauchy data. For the lognormal data, RPCA gave better estimates, and the p_i were not high except for $n = 10000$.

To compare affine equivariant and non-equivariant estimators, Maronna and Zamar (2002) suggest using $\boldsymbol{A}_{i,i} = 1$ and $\boldsymbol{A}_{i,j} = \rho$ for $i \neq j$ and $\rho = 0, 0.5, 0.7, 0.9$, and 0.99. Then $\boldsymbol{\Sigma} = \boldsymbol{A}^2$. If ρ is high, or if p is high and $\rho \geq 0.5$, then the data are concentrated about the line with direction $\mathbf{1} = (1, ..., 1)^T$. For $p = 50$ and $\rho = 0.99$, the population variance explained by the first principal component is 0.999998. If the "correlation" option is used, then there is still one extremely dominant principal component unless both p and ρ are small.

Table 3 shows the mean "variance explained" along with the standard deviations multiplied by 10^7 for the first principal component. The a_1 value is given but p_1 was always 1.0 to many decimal places even with Cauchy data. Hence the eigenvectors from the robust and classical methods could have low absolute correlation, but the data was so tightly clustered that the first principal components from the robust and classical methods had absolute correlation near 1.

For the Hotelling's T_H^2 simulation, the data is $N_p(\delta\mathbf{1}, diag(1, 2, ..., p))$ where $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is being tested with 5000 runs at a nominal level of 0.05. In Table 4, $\delta = 0$ so H_0 is true, while hcv and rhcv are the proportion of rejections by the T_H^2 test and by the ad hoc robust test. Sample sizes are $n = 15p, 20p$ and $30p$. The robust test is not recommended for $n < 15p$ and appears to be conservative except when $n = 15p$ and $75 \leq p \leq 100$.

If $\delta > 0$, then H_0 is false and the proportion of rejections estimates the power of the

test. Table 5 shows that T_H^2 has more power than the robust test, but suggests that the power of both tests rapidly increases to one as δ increases.

5. Conclusions

The RMVN subset is often a large clean subset of the data even when certain types of outliers are present. The classical estimator $(T_U, \boldsymbol{C}_U)$ applied to this subset is a $\sqrt{n}$ consistent high breakdown estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ on a large class distributions. This result suggests applying classical methods for principal components, canonical correlation analysis and factor analysis on the RMVN subset. This method is very simple since it uses the software available for the classical method.

There are many other applications of $(T_U, \boldsymbol{C}_U)$, FCH and RMVN. The DD plot a plot of the classical versus robust Mahalanobis distances, can be used to check for outliers. MVN data scatters about the identity line while elliptically contoured data satisfying (E1) scatters about some line through the origin. The three robust estimators can be used as plug in estimators replacing $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ to make robust analogs for many multivariate procedures. RMVN is useful if an estimator of $\boldsymbol{\Sigma}$ is needed instead of an estimator of $d\,\boldsymbol{\Sigma}$. Since the MB estimator does not make the expensive $O(p^3)$ determinant calculation, it may be useful if the main concern is outlier detection.

FCH and RMVN need $n > 2p$ to be computed, and $n \geq 10p$ to produce DD plots where the plotted points cluster tightly about the identity line for MVN data. RMVN needs $n \geq 20p$ before it gives good estimates of $\boldsymbol{\Sigma}$ for MVN data. The estimators can be modified so that the initial estimator covers more than half of the cases (e.g. 75% of the cases) with the price of decreased outlier resistance.

The previous sections illustrated PCA and a robust analog of Hotelling's T^2 test as

applications. The robust T_R^2 statistic tends not to be as inflated as T_H^2 when outliers are present, as can be demonstrated with the `rhotsim` program referenced below. Ideally software users would make a DD plot and other checks on the model, but users of statistical software too often fail to make such checks. Since both statistics are easily computed, if $n \geq 15n$ software could produce a warning if the two statistics differ.

Simulations were done in *R*. The `MASS` library was used to compute FMCD and the `robustbase` library was used to compute OGK. Programs are in the collection of functions *rpack.txt* at (www.math.siu.edu/olive/ol-bookp.htm). Function `covrmvn` computes the FCH, RMVN and MB estimators while `covfch` computes the FCH, RFCH and MB estimators. The following functions were used in the three simulations and have more outlier configurations than the two described in the paper. Function `covesim` was used to produce Table 1, `pcasim` for Tables 2 and 3 and `rhotsim` for Tables 4 and 5.

References

Atkinson, A.C., Riani, M., and Cerioli, A. (2009), "Finding an Unknown Number of Outliers," *Jounrnal of the Royal Statistical Society, B*, 71, 447-466.

Bhatia, R., Elsner, L., and Krause, G. (1990), "Bounds for the Variation of the Roots of a Polynomial and the Eigenvalues of a Matrix," *Linear Algebra and Its Applications*, 142, 195-209.

Boente, G. and Fraiman, R. (1999) "Discussion of 'Robust Principal Component Analysis for Functional Data' by Locantore et al," *Test*, 28-35.

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

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).

Cattell, R.B. (1966), "The Scree Test for the Number of Factors," *Multivariate Behavioral Research,* 1, 245-276.

Croux, C., and Haesbroeck, G. (2000), "Principal component analysis based on robust estimators of the covariance or correlation," *Biometrika*, 87, 603618.

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

Eaton, M.L., and Tyler, D.E. (1991), "On Wielands's Inequality and its Application to te Asymptotic Distribution of the Eigenvalues of a Random Symmetric Matrix," *The Annals of Statistics*, 19, 260-271.

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

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., and Cohen, K.L. (1999), "Robust Principal Component Analysis for Functional Data," (with discussion), *Test*, 8, 1-73.

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

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

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

Olive, D.J., 2004. A resistant estimator of multivariate location and dispersion. Com-

putat. Statist. Data Analys. 46, 99-102.

Olive, D.J., 2008. Applied Robust Statistics, unpublished online text available from (www.math.siu.edu/olive/ol-bookp.htm).

Olive, D.J., Hawkins, D.M., 2009. Robust multivariate location and dispersion. Preprint, see (www.math.siu.edu/olive/pphbmld.pdf).

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

Srivastava, D.K., Mudholkar, G.S., 2001. Trimmed $\tilde{T}^2$: a robust analog of Hotelling's T^2. J. Statist. Plann. Inference 97, 343-358.

Willems, G., Pison, G., Rousseeuw, P.J., Van Aelst, S., 2002. A robust Hotelling test. Metrika 55, 125-138.

Table 1: Estimation of $\mathbf{\Sigma}$ with $\gamma = 0.4$, $n = 35p$

p	type	n	pm	Q
5	1	135	16	0.153
5	2	135	6	0.213
10	1	350	21	0.326
10	2	350	6	0.326
15	1	525	26	0.856
15	2	525	7	0.675
20	1	700	33	0.798
20	2	700	8	0.792
25	1	875	39	1.014
25	2	875	10	1.867

Table 2: Variance Explained by PCA and RPCA, $p = 4$

n	type	M/S	vexpl	rvexpl	a_1/p_1	a_2/p_2	a_3/p_3
40	N	M	0.445,0.289,0.178	0.472,0.286,0.166	0.895	0.821	0.825
		S	0.050,0.037,0.032	0.062,0.043,0.037	0.912	0.813	0.804
100	N	M	0.419,0.295,0.191	0.425,0.293,0.189	0.952	0.926	0.963
		S	0.033,0.030,0.024	0.040,0.032,0.027	0.956	0.923	0.953
400	N	M	0.404,0.298,0.198	0.406,0.298,0.198	0.994	0.991	0.996
		S	0.019,0.017,0.014	0.021,0.019,0.015	0.995	0.990	0.994
40	C	M	0.765,0.159,0.056	0.514,0.275,0.147	0.563	0.519	0.511
		S	0.165,0.112,0.051	0.078,0.055,0.040	0.776	0.383	0.239
100	C	M	0.762,0.156,0.060	0.455,0.286,0.173	0.585	0.527	0.528
		S	0.173,0.112,0.055	0.054,0.041,0.034	0.797	0.377	0.269
400	C	M	0.756,0.162,0.060	0.413,0.296,0.194	0.608	0.562	0.575
		S	0.172,0.113,0.054	0.030,0.025,0.022	0.796	0.397	0.308
40	L	M	0.539,0.256,0.139	0.521,0.268,0.146	0.610	0.509	0.530
		S	0.127,0.075,0.054	0.099,0.061,0.047	0.643	0.439	0.398
100	L	M	0.482,0.270,0.165	0.459,0.279,0.172	0.647	0.555	0.566
		S	0.180,0.063,0.052	0.077,0.047,0.041	0.654	0.492	0.474
400	L	M	0.437,0.282,0.185	0.416,0.290,0.194	0.748	0.639	0.739
		S	0.080,0.048,0.044	0.049,0.035,0.033	0.727	0.594	0.690
10000	L	M	0.400,0.301,0.200	0.402,0.300,0.199	0.982	0.967	0.991
		S	0.027,0.023,0.018	0.013,0.011,0.009	0.976	0.967	0.989

Table 3: Variance Explained by PCA and RPCA, SSD = 10^7 SD, $p = 50$

n	type	vexpl	SSD	rvexpl	SSD	a_1
200	N	0.999998	1.958	0.999998	2.867	0.687
1000	N	0.999998	0.917	0.999998	0.971	0.944
1000	C	0.999996	161.3	0.999998	1.482	0.112
1000	L	0.999998	0.919	0.999998	1.508	0.175

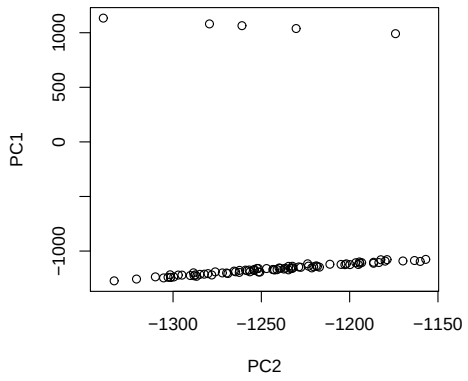

Figure 1: First Two Principal Components for Buxton data

Table 4: Hotelling simulation

p	n=15p	hcv	rhcv	n=20p	hcv	rhcv	n=30p	hcv	rhcv
10	150	0.0476	0.0300	200	0.0516	0.0304	300	0.0498	0.0286
15	225	0.0474	0.0318	300	0.0506	0.0308	450	0.0492	0.0320
20	300	0.0540	0.0368	400	0.0548	0.0314	600	0.0520	0.0354
25	375	0.0444	0.0334	500	0.0462	0.0296	750	0.0456	0.0288
30	450	0.0472	0.0324	600	0.0516	0.0358	900	0.0484	0.0342
35	525	0.0490	0.0384	700	0.0522	0.0358	1050	0.0502	0.0374
40	600	0.0534	0.0440	800	0.0486	0.0354	1200	0.0526	0.0336
45	675	0.0406	0.0390	900	0.0544	0.0390	1350	0.0512	0.0366
50	750	0.0498	0.0430	1000	0.0522	0.0394	1500	0.0512	0.0364
55	825	0.0504	0.0502	1100	0.0496	0.0392	1650	0.0510	0.0374
60	900	0.0482	0.0514	1200	0.0488	0.0404	1800	0.0474	0.0376
65	975	0.0568	0.0602	1300	0.0524	0.0414	1950	0.0548	0.0410
70	1050	0.0462	0.0530	1400	0.0558	0.0432	2100	0.0522	0.0424
75	1125	0.0474	0.0632	1500	0.0502	0.0486	2250	0.0490	0.0370
80	1200	0.0524	0.0620	1600	0.0524	0.0432	2400	0.0468	0.0356
85	1275	0.0482	0.0758	1700	0.0496	0.0456	2550	0.0520	0.0404
90	1350	0.0504	0.0746	1800	0.0484	0.0454	2700	0.0484	0.0398
95	1425	0.0524	0.0892	1900	0.0472	0.0506	2850	0.0538	0.0424
100	1500	0.0554	0.0808	2000	0.0452	0.0506	3000	0.0488	0.0392

Table 5: Hotelling power simulation

p	n	hcv	rhcv	δ	n	hcv	rhcv	δ	n	hcv	rhcv	δ
5	75	0.459	0.245	0.20	100	0.366	0.184	0.15	150	0.333	0.208	0.12
5	75	0.682	0.416	0.25	100	0.599	0.368	0.20	150	0.577	0.394	0.16
5	75	0.840	0.588	0.30	100	0.816	0.587	0.30	150	0.860	0.708	0.40
10	150	0.221	0.113	0.10	200	0.312	0.182	0.10	300	0.469	0.340	0.10
10	150	0.621	0.400	0.17	200	0.655	0.467	0.15	300	0.647	0.504	0.12
10	150	0.888	0.729	0.22	200	0.848	0.692	0.18	300	0.872	0.767	0.15
15	225	0.314	0.188	0.10	300	0.442	0.294	0.10	450	0.317	0.228	0.07
15	225	0.714	0.543	0.15	300	0.623	0.449	0.12	450	0.648	0.522	0.10
15	225	0.881	0.738	0.18	300	0.858	0.755	0.15	450	0.853	0.762	0.12
20	300	0.408	0.276	0.10	400	0.341	0.230	0.08	600	0.291	0.216	0.06
20	300	0.691	0.525	0.13	400	0.674	0.534	0.11	600	0.554	0.433	0.08
20	300	0.935	0.852	0.17	400	0.858	0.742	0.13	600	0.790	0.701	0.10
25	375	0.304	0.214	0.08	500	0.434	0.319	0.08	750	0.354	0.266	0.06
25	375	0.728	0.580	0.12	500	0.676	0.531	0.10	750	0.660	0.556	0.08
25	375	0.926	0.837	0.15	500	0.868	0.771	0.12	750	0.887	0.815	0.10
30	450	0.374	0.264	0.08	600	0.395	0.290	0.07	900	0.290	0.217	0.05
30	450	0.602	0.467	0.10	600	0.639	0.517	0.09	900	0.743	0.642	0.08
30	450	0.883	0.763	0.13	600	0.867	0.770	0.11	900	0.876	0.808	0.09

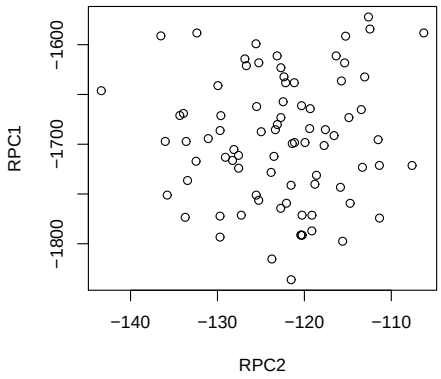

Figure 2: First Two Robust Principal Components with Outliers Omitted