Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Robust Estimation of Multivariate Location and Scatter in the Presence of Missing Data

Mike DANILOV, Víctor J. YOHAI, and Ruben H. ZAMAR

Two main issues regarding data quality are data contamination (outliers) and data completion (missing data). These two problems have attracted much attention and research but surprisingly, they are seldom considered together. Popular robust methods such as S-estimators of multivariate location and scatter offer protection against outliers but cannot deal with missing data, except for the obviously inefficient approach of deleting all incomplete cases. We generalize the definition of S-estimators of multivariate location and scatter to simultaneously deal with missing data and outliers. We show that the proposed estimators are strongly consistent under elliptical models when data are *missing completely at random*. We derive an algorithm similar to the Expectation-Maximization algorithm for computing the proposed estimators. This algorithm is initialized by an extension for missing data of the minimum volume ellipsoid. We assess the performance of our proposal by Monte Carlo simulation and give some real data examples. This article has supplementary material online.

KEY WORDS: Consistent; Elliptical distribution; EM algorithm; Fixed point equation

1. INTRODUCTION

There are many problems that may affect the quality of data and the performance of an estimator. Two common problems are outliers and missing data. We address these two problems simultaneously, when the goal is to estimate multivariate location and scatter. The estimation of these parameters is a cornerstone for many robust multivariate analysis techniques such as principal components, canonical correlation, discriminant analysis, and so on. See, for example, Salibian-Barrera, Van Aelst, and Willems (2006), Taskinen et al. (2006), and Croux, Filzmoser, and Joossens (2008), and references therein.

We will assume that the data are *missing completely at random* (MCAR), that is, the probability that some components of a particular data point are missing does not depend on the values of this case.

Although outliers and missing data have been individually well studied, there are few works that address these two problems together. When there are no outliers, a common way to estimate multivariate location and scatter is to assume normality and to use the Expectation-Maximization (EM) algorithm to maximize the likelihood for the observed data (see Dempster, Laird, and Rubin 1977; Little and Rubin 2002). To robustly estimate these parameters in the presence of outliers and missing data, Little and Smith (1987) proposed the expectation-robust minimization (ER) algorithm which robustifies the EM algorithm using weights that penalize outliers. The weights are applied to Mahalanobis distances from the (possibly incomplete) data points to the current center, using the nonmissing part of each observation and the current scatter. Little (1988) robustified the Gaussian EM algorithm using a multivariate Student's t-distribution or some other heavy tail distribution. It is well known, however, that such maximum likelihood estimators (MLEs) have breakdown point equal to $1/(p+1)$ in the case of complete data (see Maronna 1976). Cheng and Victoria-Feser (2002) noticed that ER can lose its robustness when the fraction of contamination exceeds $1/(p+1)$. To remedy this problem, they proposed a procedure called ERTBS which replaces the Huber weights in ER by weights calculated using the translated biweight score function introduced by Rocke (1996). They also modified the way in which the weights are applied to the data. Since their procedure critically depends on an initial estimator, they introduced an extended minimum covariance determinant estimator for missing data as a possible starting value. Unfortunately, as evidenced by our simulation studies (see Section 7), ERTBS is not consistent for normal data and remains sensitive to clusters of outliers. Frahma and Jaekel (2010) extended the location and scatter M-estimators proposed by Tyler (1987) for the case of partially missing observations. However, since the score function of these estimators is monotone, their complete data breakdown point is $1/(p+1)$. Recently, Templ, Kowarik, and Filzmoser (2011) proposed a general robust imputation method to deal with large datasets possessing outliers and missing data.

In this article, we present two classes of robust estimators for missing data: *generalized S-estimators* (GSEs) and *extended S-estimators* (ESEs). Both classes coincide with the S-estimators introduced by Davies (1987) for complete data. Since GSEs require a robust initial estimator, we introduced the family of ESE to serve in this capacity. Following ideas in Section 6.7.5 of Maronna, Martin, and Yohai (2006), we propose, as initial estimator, a particular case of ESE that we call *extended minimum volume ellipsoid* (EMVE). This estimator generalizes the MVE estimator introduced by Rousseeuw (1985). EMVE is computed using subsampling followed by an appropriate concentration step as in Rousseeuw and Van Driessen (1999).

Mike Danilov is a Quantitative Analyst at Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA (E-mail: *mikedanilov@google.com*), Víctor J. Yohai is Emeritus Professor, Departamento de Matemáticas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 1, 1428 Buenos Aires, Argentina (E-mail: *vyohai@dm.uba.ar*), and Ruben H. Zamar is Professor, Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, BC V6T 1Z2, Canada (E-mail: *ruben@stat.ubc.ca*). This research was partially supported by grants X-018 and X-447 from the University of Buenos Aires, PIP 5505 from CONICET, PICT 00899 from ANPCyT, and Discovery grant from NSERC. We thank the Associate Editor and two referees for their comments and suggestions which resulted in several important improvements on the first version of this article.

© 2012 American Statistical Association
Journal of the American Statistical Association
September 2012, Vol. 107, No. 499, Theory and Methods
DOI: 10.1080/01621459.2012.699792

The rest of the article is organized as follows. In Section 2, we describe our setting. In Section 3, we define GSE, discuss some of its properties (including partial affine equivariance), and show that GSE satisfies a set of fixed point equations. In Section 4, we show that GSE is strongly consistent for the multivariate location and for the *scatter shape component*. That is, GSE converges a.s. to the scatter matrix except for a scalar factor under general elliptical distributions. GSE can also be scaled to be consistent for estimating the *scatter size component* for any particular elliptically symmetric family such as the multivariate normal family. ESE can only be made strongly consistent for a given single family of elliptical distributions. Fortunately, this does not affect the general consistency of the scatter shape component of the final GSE. In Section 5, we present an algorithm to compute GSE. In Section 6, we define ESEs and the EMVE. In Section 7, we conduct a Monte Carlo simulation study and some timing experiments. In Section 8, we give some real data examples.

2. NOTATION

Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, $1 \le i \le n$, be p-dimensional iid random vectors with common density f belonging to the elliptical family

$$f(\mathbf{x}, \mathbf{m}_0, \boldsymbol{\Sigma}_0) = |\boldsymbol{\Sigma}_0|^{-1} f_0\big((\mathbf{x} - \mathbf{m}_0)'\boldsymbol{\Sigma}_0^{-1}(\mathbf{x} - \mathbf{m}_0)\big), \qquad (1)$$

where, in general, $|A|$ denotes the determinant of the squared matrix A and $f_0(\|\mathbf{x}\|^2)$ is a density function in R^p. Notice that for each choice of f_0, we have a specific family of elliptical distributions. If f_0 is not specified, Equation (1) gives a larger semiparametric family. Let $\mathbf{u}_i = (u_{i1}, \ldots, u_{ip})'$, $1 \le i \le n$, be independent p-dimensional vectors of zeros and ones with common distribution G. The entries of $\mathbf{u}_i$ indicate which coordinates of $\mathbf{x}_i$ are actually observed: x_{ij} is observed when $u_{ij} = 1$. We also assume that $\mathbf{u}_i$ and $\mathbf{x}_i$ are independent (which corresponds to the MCAR assumption).

Given $\mathbf{x} = (x_1, \ldots, x_p)'$ and $\mathbf{u} = (u_1, \ldots, u_p)'$, let $\mathbf{x}^{(\mathbf{u})}$ be the observed part of $\mathbf{x}$ and set

$$p(\mathbf{u}) = \sum_{j=1}^{p} u_j. \qquad (2)$$

That is, $\mathbf{x}^{(\mathbf{u})}$ is a vector of dimension $p(\mathbf{u})$ formed with the available entries of $\mathbf{x}$. We assume the following identifiability condition: given $1 \le j < k \le p$, there exists at least one $\mathbf{u}_i$, $1 \le i \le n$, with $u_{ij} = u_{ik} = 1$.

Let $A_p = \{\mathbf{u} : (u_1, \ldots, u_p)',\ u_i \in \{0, 1\}\}$, then given a $p \times p$ positive definite matrix $\boldsymbol{\Sigma}$ and $\mathbf{u} \in A_p$, we denote by $\boldsymbol{\Sigma}^{(\mathbf{u})}$ the submatrix of $\boldsymbol{\Sigma}$ corresponding to the positive entries in $\mathbf{u}$. Similarly, given $\mathbf{m} \in R^p$, we denote by $\mathbf{m}^{(\mathbf{u})}$ the corresponding subvector of $\mathbf{m}$. Finally, we set $\boldsymbol{\Sigma}^{*(\mathbf{u})} = \boldsymbol{\Sigma}^{(\mathbf{u})}/|\boldsymbol{\Sigma}^{(\mathbf{u})}|^{1/p(\mathbf{u})}$ and note that $|\boldsymbol{\Sigma}^{*(\mathbf{u})}| = 1$.

Given a data point $(\mathbf{x}, \mathbf{u})$, a center $\mathbf{m} \in R^p$, and a $p \times p$ positive definite scatter matrix $\boldsymbol{\Sigma}$, the *partial square Mahalanobis distance* is given by

$$d(\mathbf{x}, \mathbf{u}, \mathbf{m}, \boldsymbol{\Sigma}) = \left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}^{(\mathbf{u})}\right)' \left(\boldsymbol{\Sigma}^{(\mathbf{u})}\right)^{-1} \left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}^{(\mathbf{u})}\right). \qquad (3)$$

3. GENERALIZED *S*-ESTIMATORS FOR MISSING DATA

3.1 Generalized *S*-Estimators (GSE)

We begin by recalling Davies (1987) definition of *S*-estimator for complete data. Suppose that $n > 2p$ and let $\rho : R_+ \to R_+$ (with $R_+ = [0, \infty)$) be a nondecreasing function such that $\max_t \rho(t) = 1$. Given $\mathbf{m} \in R^p$ and a $p \times p$ positive definite matrix $\boldsymbol{\Sigma}$, let $S_n(\mathbf{m}, \boldsymbol{\Sigma})$ be the solution in s to the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{d(\mathbf{x}_i, \mathbf{m}, \boldsymbol{\Sigma})}{c_p s}\right) = \frac{1}{2},$$

where c_p is such that

$$E\left(\rho\left(\frac{||\mathbf{X}||^2}{c_p}\right)\right) = 0.5, \qquad (4)$$

and where $\mathbf{X}$ has density given by Equation (1) with $\mathbf{m}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \mathbf{I}_p$. Usually, f_0 is chosen such that $f_0(||\mathbf{x}||^2)$ is the standard multivariate normal density. The *S*-estimator $(\widehat{\mathbf{m}}_n, \widehat{\boldsymbol{\Sigma}}_n)$ is then defined as

$$\begin{aligned}(\widehat{\mathbf{m}}_n, \widetilde{\boldsymbol{\Sigma}}_n) &= \arg \min_{\mathbf{m}, |\boldsymbol{\Sigma}|=1} S_n(\mathbf{m}, \boldsymbol{\Sigma}),\\ \hat{s}_n &= S_n(\widehat{\mathbf{m}}_n, \widetilde{\boldsymbol{\Sigma}}_n), \qquad (5)\\ \widehat{\boldsymbol{\Sigma}}_n &= \hat{s}_n \widetilde{\boldsymbol{\Sigma}}_n.\end{aligned}$$

We now generalize the definition of *S*-estimator for the case of incomplete data. Let $\widehat{\boldsymbol{\Omega}}_n$ be a $p \times p$ positive definite initial estimator for $\boldsymbol{\Sigma}_0$. Given $\mathbf{m} \in R^p$ and a $p \times p$ positive definite matrix $\boldsymbol{\Sigma}$, let $S_n^*(\mathbf{m}, \boldsymbol{\Sigma})$ be the solution in s to the equation

$$\sum_{i=1}^{n} c_{p(\mathbf{u}_i)} \rho\left(\frac{d\left(\mathbf{x}_i^{(\mathbf{u}_i)}, \mathbf{m}^{(\mathbf{u}_i)}, \boldsymbol{\Sigma}^{*(\mathbf{u}_i)}\right)}{s\left|\widehat{\boldsymbol{\Omega}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)} c_{p(\mathbf{u}_i)}}\right) = \frac{1}{2}\sum_{i=1}^{n} c_{p(\mathbf{u}_i)}, \qquad (6)$$

where c_p is defined in Equation (4) and $p(\mathbf{u})$ in Equation (2). We first define the multivariate location and scatter shape component estimators $(\widehat{\mathbf{m}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$ as

$$(\widehat{\mathbf{m}}_n, \widetilde{\boldsymbol{\Sigma}}_n) = \arg \min_{\mathbf{m}, \boldsymbol{\Sigma}} S_n^*(\mathbf{m}, \boldsymbol{\Sigma}). \qquad (7)$$

Note that S_n^* $(\mathbf{m}, t\boldsymbol{\Sigma}) = S_n^*$ $(\mathbf{m}, \boldsymbol{\Sigma})$ for all $t > 0$, and therefore $\widetilde{\boldsymbol{\Sigma}}_n$ is only determined up to a scalar factor, that is, $\widetilde{\boldsymbol{\Sigma}}_n$ only estimates the shape component of $\boldsymbol{\Sigma}_0$. Finally, we define the GSE of scatter for $\boldsymbol{\Sigma}_0$ (shape and size) as

$$\widehat{\boldsymbol{\Sigma}}_n = \widehat{s}_n \widetilde{\boldsymbol{\Sigma}}_n, \qquad (8)$$

where $\widehat{s}_n$ satisfies the equation

$$\sum_{i=1}^{n} c_{p(\mathbf{u}_i)} \rho\left(\frac{d\left(\mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n^{(\mathbf{u}_i)}, \widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right)}{c_{p(\mathbf{u}_i)} \widehat{s}_n}\right) = \frac{1}{2}\sum_{i=1}^{n} c_{p(\mathbf{u}_i)}. \qquad (9)$$

3.2 Existence of GSE

Now we consider the existence of a solution $(\widehat{\mathbf{m}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$ to the minimization problem (7) with $\mathbf{m}$ ranging over R^p and $\boldsymbol{\Sigma}$ ranging over the set of $p \times p$ positive definite matrices. Consider the sample $(\mathbf{x}_i, \mathbf{u}_i)$, $i = 1, 2, \ldots, n$. For a given configuration $\mathbf{u} \in A_p$, let $D_{\mathbf{u}} = \{i : \mathbf{u}_i = \mathbf{u}\}$ and $n_{\mathbf{u}} = \#D_{\mathbf{u}}$. Call

$\mathbf{u}_0 = (1, \dots, 1)$ the configuration corresponding to complete observations and let

$$\kappa_{\mathbf{0}} = \max_{\mathbf{c} \in R^{p(\mathbf{u})}, d \in R, \mathbf{c} \neq \mathbf{0}} \#\{i \in D_{\mathbf{u}_0} : \mathbf{c}'\mathbf{x}_i = d\}$$

be the maximum number of complete points $\mathbf{x}_i$ which lies in a p-dimensional subspace. Theorem 1 (proved in Section 4.2 of the supplementary material) gives a sufficient condition for the existence of the GES.

Assumption 1. The function ρ is (1) nondecreasing, (2) strictly increasing at 0, (3) continuous, (4) $\rho(0) = 0$, and (5) $\lim_{v \to \infty} \rho(v) = 1$.

Theorem 1. Suppose Assumption 1 holds. Consider a sample $(\mathbf{x}_i, \mathbf{u}_i)$, $i = 1, 2, \dots, n$ such that

$$n_{\mathbf{u}_0} > \frac{n}{2} + \kappa_0. \tag{10}$$

Then, there exists at least one value $(\widehat{\mathbf{m}}_n, \widetilde{\mathbf{\Sigma}}_n)$ minimizing $S_n^*(\mathbf{m}, \mathbf{\Sigma})$ with $\widetilde{\mathbf{\Sigma}}_n$ being positive definite.

Remark 1. If the complete observations are in general position, then $\kappa_0 = p + 1$. Condition (10) is by no means necessary for the existence of GSE. When Theorem 3 holds, then GES always exists for sufficiently large n, for any missing fraction. Moreover, the numerical results of Section 7 show that GSE is robust and efficient in situations where $n_{\mathbf{u}_0}$ is much smaller than $n/2$.

Regarding uniqueness of GSE, we notice that there are not such results for S-estimators for complete data. However, Tatsuoka and Tyler (2000) conjectured that the S-estimator solution is unique with probability 1 in the case of random samples from a continuous distribution. We believe that this may also be the case for the GSE.

3.3 Partial Equivariance of GSE

We use the "arithmetic rules" (1) $x + \mathrm{NA} = \mathrm{NA}$, for all x (NA means "nonavailable"), (2) $x \times \mathrm{NA} = \mathrm{NA}$ for all $x \neq 0$, and (3) $0 \times \mathrm{NA} = 0$. Since GSE is defined using Mahalanobis distances, if $\mathbf{A}$ is an invertible matrix that preserves the missingness pattern [that is, for all $i = 1, 2, \dots, n$, $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}$ has the same missing pattern as $\mathbf{x}_i$], then $\widehat{\mathbf{m}}_{y,n} = \mathbf{A}\widehat{\mathbf{m}}_{x,n} + \mathbf{b}$ and $\widehat{\mathbf{\Sigma}}_{y,n} = \mathbf{A}\widehat{\mathbf{\Sigma}}_{x,n}\mathbf{A}'$. Here, the x, y subscripts indicate whether the $\mathbf{x}_i$ or the $\mathbf{y}_i$ data are used to compute GES. In particular, GSE is location- and scale-equivariant because any invertible diagonal matrix $\mathbf{D}$ preserves the missingness pattern. As another example, suppose that $u_{i1} = u_{i2} = \cdots u_{iq} = 1$, for all $1 \le i \le n$. Let $\mathbf{A}_q$ be an invertible $q \times q$ matrix and $\mathbf{D}_{p-q}$ be a diagonal $(p-q) \times (p-q)$ matrix. Then

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{p-q} \end{pmatrix}$$

preserves the missingness pattern.

3.4 Scaling the Mahalanobis Distances

The partial Mahalanobis distances in Equation (6) are not properly scaled because they are based on the normalized matrices $\mathbf{\Sigma}^*$. Although this causes no problem regarding the consistency of the estimator, to achieve robustness it is necessary to rescale these distances using the "tuning constants" $|\widehat{\mathbf{\Omega}}_n^{(\mathbf{u}_i)}|^{1/p(\mathbf{u}_i)}$. To fix ideas, suppose that $\mathbf{m} \approx \mathbf{m}_0$ and $\mathbf{\Sigma} \approx \widehat{\mathbf{\Omega}}_n \approx \mathbf{\Sigma}_0$ in Equation (6). Then

$$\frac{d\left(\mathbf{x}_i^{(\mathbf{u}_i)}, \mathbf{m}^{(\mathbf{u}_i)}, \mathbf{\Sigma}^{*(\mathbf{u}_i)}\right)}{c_{p(\mathbf{u}_i)}\left|\widehat{\mathbf{\Omega}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}} \approx \frac{d\left(\mathbf{x}_i^{(\mathbf{u}_i)}, \mathbf{m}_0^{(\mathbf{u}_i)}, \mathbf{\Sigma}_0^{*(\mathbf{u}_i)}\right)}{c_{p(\mathbf{u}_i)}\left|\mathbf{\Sigma}_0^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}} \sim \frac{||\mathbf{Y}^{(\mathbf{u}_i)}||^2}{c_{p(\mathbf{u}_i)}},$$

where $\mathbf{Y}$ has density $f_0(||\mathbf{y}||^2)$ and so $||\mathbf{Y}^{(\mathbf{u}_i)}||^2/c_{p(\mathbf{u}_i)}$ has M-scale equal to 1 for the given ρ function. Hence, with this scaling, large Mahalanobis distances are downweighted and do not upset the estimator. A discussion of a possible choice for the initial scatter estimator $\widehat{\mathbf{\Omega}}_n$ and an algorithm to compute the final estimators are given in Section 5.

3.5 GSE on Complete Data

When the data are complete, for any $\widehat{\mathbf{\Omega}}_n$, the GSE $(\widehat{\mathbf{m}}_n, \widehat{\mathbf{\Sigma}}_n)$ reduces to the regular S-estimator given by (5). In fact, in this case, Equation (6) becomes

$$\sum_{i=1}^{n} c_p \rho\left(\frac{d(\mathbf{x}_i, \mathbf{m}, \mathbf{\Sigma}^*)}{\left|\widehat{\mathbf{\Omega}}_n\right|^{1/p} c_p s}\right) = \frac{1}{2}\sum_{i=1}^{n} c_p = \frac{n c_p}{2},$$

that is,

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{d(\mathbf{x}_i, \mathbf{m}, \mathbf{\Sigma}^*)}{\left|\widehat{\mathbf{\Omega}}_n\right|^{1/p} c_p s}\right) = \frac{1}{2}.$$

Therefore

$$|\widehat{\mathbf{\Omega}}_n|^{1/p} S_n^*(\mathbf{m}, \mathbf{\Sigma}) = S_n(\mathbf{m}, \mathbf{\Sigma}^*),$$

where $S_n(\mathbf{m}, \mathbf{\Sigma})$ is defined in Section 2. Since the factor $|\widehat{\mathbf{\Omega}}_n|^{1/p}$ is constant, $\widetilde{\mathbf{\Sigma}}/|\widetilde{\mathbf{\Sigma}}|^{1/p}$ minimizes $S_n(\mathbf{m}, \mathbf{\Sigma}^*)$ if and only if $\widetilde{\mathbf{\Sigma}}$ minimizes $S_n^*(\mathbf{m}, \mathbf{\Sigma})$. In other words, the classical and GSEs coincide in this case.

3.6 Fixed Point Estimating Equations for GSE

For $\mathbf{u} \in A_p$ (see Section 2), $\mathbf{x}^{(\mathbf{u})} \in R^{p(\mathbf{u})}$, $\mathbf{m} \in R^p$, and a $p \times p$ positive definite matrix $\mathbf{\Sigma}$, we define $\widehat{\mathbf{x}}(\mathbf{u}, \mathbf{x}^{(\mathbf{u})}, \mathbf{m}, \mathbf{\Sigma})$ as the best linear predictor of $\mathbf{X}$ given $\mathbf{X}^{(\mathbf{u})} = \mathbf{x}^{(\mathbf{u})}$, when $E(\mathbf{X}) = \mathbf{m}$ and $\mathrm{cov}(\mathbf{X}) = \mathbf{\Sigma}$. Moreover, $\mathbf{C}(\mathbf{u}, \Sigma)$ is the covariance matrix for the prediction error $\mathbf{X} - \widehat{\mathbf{x}}(\mathbf{u}, \mathbf{X}^{(\mathbf{u})}, \mathbf{m}, \mathbf{\Sigma})$ when $\mathbf{X}$ has expectation $\mathbf{m}$ and covariance $\mathbf{\Sigma}$. In particular, if $\mathbf{u}$ has the first $q = p\ (\mathbf{u})$ entries equal to 1 and the remaining entries equal to 0, we have the following simple formulas. Let $\mathbf{v} = (v_1, \dots, v_p)' \in \mathbf{A}_p$ such that $v_1 = \cdots = v_q = 0$ and $v_{q+1} = \cdots = v_p = 1$ and write

$$\mathbf{m} = \begin{pmatrix} \mathbf{m}^{(\mathbf{u})} \\ \mathbf{m}^{(\mathbf{v})} \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{uu}} & \mathbf{\Sigma}_{\mathbf{uv}} \\ \mathbf{\Sigma}_{\mathbf{vu}} & \mathbf{\Sigma}_{\mathbf{vv}} \end{pmatrix}.$$

Then

$$\widehat{\mathbf{x}}\left(\mathbf{u}, \mathbf{x}^{(\mathbf{u})}, \mathbf{m}, \Sigma\right) = \begin{pmatrix} \mathbf{x}^{(\mathbf{u})} \\ \mathbf{m}^{(\mathbf{v})} + \mathbf{\Sigma}_{\mathbf{vu}}\mathbf{\Sigma}_{\mathbf{uu}}^{-1}\left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}^{(\mathbf{u})}\right) \end{pmatrix}, \tag{11}$$

$$\mathbf{C}\left(\mathbf{u}, \mathbf{\Sigma}\right) = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{\mathbf{vv}} - \mathbf{\Sigma}_{\mathbf{vu}}\mathbf{\Sigma}_{\mathbf{uu}}^{-1}\mathbf{\Sigma}_{\mathbf{uv}} \end{pmatrix}. \tag{12}$$

The following theorem (proved in Section 4.3 of the supplementary material) gives fixed point estimating equations for the GSE estimators of location and scatter shape component.

Theorem 2. Let $\widehat{\mathbf{m}}_n$ and $\widetilde{\mathbf{\Sigma}}_n$ be defined by Equation (7). Assume that ρ is a nondecreasing and continuously differentiable function. Then, we have

$$\widehat{\mathbf{m}}_n = \frac{\sum_{i=1}^n w_i \widehat{\mathbf{x}}_i}{\sum_{i=1}^n w_i} \tag{13}$$

and

$$\widetilde{\mathbf{\Sigma}}_n = \frac{\sum_{i=1}^n [w_i\,(\widehat{\mathbf{x}}_i - \widehat{\mathbf{m}}_n)\,(\widehat{\mathbf{x}}_i - \widehat{\mathbf{m}}_n)' + w_i w_i^* \mathbf{C}_i]}{\sum_{i=1}^n w_i w_i^*}, \tag{14}$$

where $\widehat{\mathbf{x}}_i = \widehat{\mathbf{x}}(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n, \widehat{\mathbf{\Sigma}}_n)$, $\mathbf{C}_i = \mathbf{C}(\mathbf{u}_i, \widehat{\mathbf{\Sigma}}_n)$, $w_i = w(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n, \widehat{\mathbf{\Sigma}}_n, \widetilde{s}_n), w_i^* = w^*(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n, \widehat{\mathbf{\Sigma}}_n)$ with

$$w(\mathbf{u}, \mathbf{x}^{(\mathbf{u})}, \mathbf{m}, \Sigma, s) = \frac{|\mathbf{\Sigma}^{(\mathbf{u})}|^{1/p(\mathbf{u})}}{|\widehat{\mathbf{\Omega}}_n^{(\mathbf{u})}|^{1/p(\mathbf{u})}} \rho' \left(\frac{d(\mathbf{x}^{(\mathbf{u})}, \mathbf{m}^{(\mathbf{u})}, \mathbf{\Sigma}^{(\mathbf{u})})}{c_{p(\mathbf{u})}\ s} \frac{|\mathbf{\Sigma}^{(\mathbf{u})}|^{1/p(\mathbf{u})}}{|\widehat{\mathbf{\Omega}}_n^{(\mathbf{u})}|^{1/p(\mathbf{u})}} \right), \tag{15}$$

$$w^*\left(\mathbf{u}, \mathbf{x}^{(\mathbf{u})}, \mathbf{m}, \mathbf{\Sigma}\right) = \frac{d\left(\mathbf{x}^{(\mathbf{u}_i)}, \mathbf{m}^{(\mathbf{u})}, \mathbf{\Sigma}^{(\mathbf{u})}\right)}{p(\mathbf{u})}, \tag{16}$$

$\widetilde{s}_n = S_n^*(\widehat{\mathbf{m}}_n, \widetilde{\mathbf{\Sigma}}_n)$, and $\widehat{\mathbf{\Sigma}}_n = \widehat{s}_n \widetilde{\mathbf{\Sigma}}_n$, where $\widehat{s}_n$ satisfies (9).

These equations show that the GSE estimators of location and scatter shape component are a weighted mean and a weighted/corrected sample covariance matrix. We will use the above fixed point equations to derive a computing algorithm in Section 5.

4. CONSISTENCY OF GSE

Theorem 3 (proved in Section 4.4 of the supplementary material) shows that $(\widehat{\mathbf{m}}_n, \widetilde{\mathbf{\Sigma}}_n) \to (\mathbf{m}_0, t_0\mathbf{\Sigma}_0)$ a.s. for some $t_0 > 0$ even if the true f_0 is different from that used in (4). Therefore, GSEs are strongly consistent for the scatter shape component under the semiparametric elliptical model (1). Davies (1987) proved similar results for *S*-estimators in the case of complete data. Note that for many applications, for example, principal component analysis and canonical correlation analysis, only the scatter shape component is required (see Salibian-Barrera, Van Aelst, and Willems 2006; Taskinen et al. 2006).

Set $G\,(\mathbf{m}, \Sigma, F, c) = E_F\,(\rho\,(d\,(\mathbf{x}, \mathbf{m}, \mathbf{\Sigma})\,/c))$ and let $F^{(\mathbf{u})}$ be the marginal distribution of $\mathbf{x}^{(\mathbf{u})}$ when $\mathbf{x}$ has distribution F. We consider the following assumptions:

Assumption 2. Given $1 \le j < k \le p$, there exists $\mathbf{u_i} = (u_{i1}, \ldots, u_{ip}) \in A_p$ such that $u_{ij} = u_{ik} = 1$.

Assumption 3. There exists $(\mathbf{m}_0, \mathbf{\Sigma}_0)$ such that for any $c > 0$ and any $\mathbf{u} \in A_p$, the only minimizer of $G(\mathbf{m}^{(\mathbf{u})}, \mathbf{\Sigma}^{(\mathbf{u})}, F^{(\mathbf{u})}, c)$ subject to the constraint $\det(\mathbf{\Sigma}^{(\mathbf{u})}) = 1$ is $(\mathbf{m}_0^{(\mathbf{u})}, \mathbf{\Sigma}_0^{*(\mathbf{u})})$.

Davies (1987) showed that if ρ satisfies Assumption 1 (in Theorem 1) and F is elliptical with f strictly decreasing (see Equation (1)), then Assumption 3 holds.

In the case of complete data, Tatsuoka and Tyler (2000) showed in Section 4 that the consistency of *S*-estimators holds for a more general family of distributions. This family includes the distribution function corresponding to $\mathbf{x} = A\mathbf{y} + \mathbf{m}$, $\mathbf{y} = (y_1, y_2, \ldots, y_p)'$, for iid Student's t random variables $y_1, y_2, \ldots, y_p$ and invertible matrix A. Unfortunately, we cannot prove that Assumption 3 holds for these distributions. However, we believe based on Monte Carlo experiments with large samples (not presented here) that Assumption 3, and therefore the consistency of GSE, holds for these distributions.

Theorem 3. Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is a random sample from F_0, $(\widehat{\mathbf{m}}_n, \widehat{\Sigma}_n)$ is defined by Equations (6) and (8) with $\widehat{\mathbf{\Omega}}_n \to \mathbf{\Omega}_0$ a.s., where $\mathbf{\Omega}_0$ is positive definite, and Assumptions 1–3 hold. Then (1) $\widehat{\mathbf{m}}_n \to \mathbf{m}_0$ a.s., (2) $\widehat{\mathbf{\Sigma}}_n \to t_0\mathbf{\Sigma}_0$ a.s., where t_0 is defined by

$$\sum_{\mathbf{u}\in A_p^*} \lambda_{\mathbf{u}} c_{\mathbf{u}} E_{F_0} \left\{ \rho \left(\frac{\left(\mathbf{x}_i - \mathbf{m}_0^{(\mathbf{u})}\right)' \left(\mathbf{\Sigma}_0^{(\mathbf{u})}\right)^{-1} \left(\mathbf{x}_i - \mathbf{m_0}^{(\mathbf{u})}\right)}{t_0 c_{p(\mathbf{u})})} \right) \right\} = \frac{1}{2} \sum_{\mathbf{u}\in A_p^*} \lambda_{\mathbf{u}} c_{p(\mathbf{u})}, \tag{17}$$

and (3) if F_0 is $N(\mathbf{m}_0, \mathbf{\Sigma}_0)$, then $t_0 = 1$.

5. COMPUTING ALGORITHM FOR GSE

In this section, we describe an iterative algorithm for computing GSE based on the fixed point estimating Equations (13) and (14) in Theorem 2.

Given initial estimates $(\widehat{\mathbf{m}}_n^{(0)}, \widehat{\mathbf{\Sigma}}_n^{(0)}, \widetilde{s}_n^{(0)})$, put $\widehat{\mathbf{\Omega}}_n = \widehat{\mathbf{\Sigma}}_n^{(0)}$ and define the sequence $(\widehat{\mathbf{m}}_n^{(k)}, \widehat{\mathbf{\Sigma}}_n^{(k)}, \widetilde{s}_n^{(k)})$, $k \ge 0$, using the recursion below. A procedure to compute the initial estimates $(\widehat{\mathbf{m}}_n^{(0)}, \widehat{\mathbf{\Sigma}}_n^{(0)}, \widetilde{s}_n^{(0)})$ is given in the next section.

Given $(\widehat{\mathbf{m}}_n^{(k)}, \widehat{\mathbf{\Sigma}}_n^{(k)}, \widetilde{s}_n^{(k)})$, compute $(\widehat{\mathbf{m}}_n^{(k+1)}, \widehat{\mathbf{\Sigma}}_n^{(k+1)}, \widetilde{s}_n^{(k+1)})$ as follows:

$$\widehat{\mathbf{m}}_n^{(k+1)} = \frac{\sum_{i=1}^n w_i^{(k)} \widehat{\mathbf{x}}_i^{(k)}}{\sum_{i=1}^n w_i^{(k)}} \tag{18}$$

and

$$\widetilde{\mathbf{\Sigma}}_n^{(k+1)} = \frac{\sum_{i=1}^n \left[w_i^{(k)} \left(\widehat{\mathbf{x}}_i^{(k)} - \widehat{\mathbf{m}}_n^{(k)}\right) \left(\widehat{\mathbf{x}}_i^{(k)} - \widehat{\mathbf{m}}_n^{(k)}\right)' + w_i^{(k)} w_i^{*(k)} \mathbf{C}_i^{(k)}\right]}{\sum_{i=1}^n w_i^{(k)} w_i^{*(k)}}, \tag{19}$$

where $\widehat{\mathbf{x}}_i^{(k)} = \widehat{\mathbf{x}}(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n^{(k)}, \widehat{\mathbf{\Sigma}}_n^{(k)})$, $\mathbf{C}_i^{(k)} = \mathbf{C}(\mathbf{u}_i, \widehat{\mathbf{\Sigma}}_n^{(k)})$, $w_i^{(k)} = w(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n^{(k)}, \widehat{\mathbf{\Sigma}}_n^{(k)}, \tilde{s}_n^{(k)})$, and $w_i^* = w^*(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n^{(k)}, \widehat{\mathbf{\Sigma}}_n^{(k)})$, where w and w^* are defined in (15) and (16), respectively. Set $\tilde{s}_n^{(k+1)} = S_n^*(\widehat{\mathbf{m}}_n^{(k+1)}, \widetilde{\mathbf{\Sigma}}_n^{(k+1)})$ and $\widehat{\mathbf{\Sigma}}_n^{(k+1)} = \widehat{s}_n^{(k+1)} \widetilde{\mathbf{\Sigma}}_n^{(k+1)}$, where $\widehat{s}_n^{(k+1)}$ is the solution to (9) with $\widehat{\mathbf{m}}_n = \widehat{\mathbf{m}}_n^{(k+1)}$ and $\widetilde{\mathbf{\Sigma}}_n = \widetilde{\mathbf{\Sigma}}_n^{(k+1)}$. The iteration stops when $|\tilde{s}_n^{(k+1)}/\tilde{s}_n^{(k)} - 1| < \delta$ for some appropriately chosen $\delta > 0$.

Note that the recursion equations for the classical EM algorithm are obtained from (18) and (19) setting $w_i^{(k)} = w_i^{*(k)} = 1$ for all i.

Remark 2. The recursive algorithm determined by Equations (18) and (19) coincides, in the case of complete data, with the algorithm used to compute the S-estimator. In such a case Maronna, Martin, and Yohai (2006), Section 6.7.5, showed that the target scale function $S_n^*(\mathbf{m}, \mathbf{\Sigma})$ is decreased by the recursion, that is, in the complete data case we have

$$S_n^*\left(\widehat{\mathbf{m}}_n^{(k+1)}, \widetilde{\mathbf{\Sigma}}_n^{(k+1)}\right) \leq S_n^*\left(\widehat{\mathbf{m}}_n^{(k)}, \widetilde{\mathbf{\Sigma}}_n^{(k)}\right)$$

for all k. We could not prove this property for GES. However, we verified it numerically in our Monte Carlo study in Section 7.

6. EXTENDED *S*-ESTIMATORS

As mentioned before, we need an initial robust estimator $(\widehat{\mathbf{m}}_n^{(0)}, \widehat{\mathbf{\Sigma}}_n^{(0)}, \widetilde{s}_n^{(0)})$ to compute GSE. We introduce now the class of ESEs which can be computed from scratch. The extended MVE described in Section 6.2 is a particularly robust member of this family which we use as default initial estimate in our GSE implementation.

6.1 Definition of ESE

The Gaussian MLE $(\widehat{\mathbf{m}}_n, \widehat{\mathbf{\Sigma}}_n)$ for the MCAR model is obtained as follows. Given $(\mathbf{m}, \mathbf{\Sigma})$, let $s_n(\mathbf{m}, \mathbf{\Sigma})$ be the solution to

$$\sum_{i=1}^n \frac{d\left(\mathbf{x}_i^{(\mathbf{u}_i)}, \mathbf{m}^{(\mathbf{u}_i)}, \mathbf{\Sigma}^{(\mathbf{u}_i)}\right)}{s} = \sum_{i=1}^n p(\mathbf{u}_i). \tag{20}$$

Now, let $(\widehat{\mathbf{m}}_n, \widetilde{\mathbf{\Sigma}}_n)$ be the minimizers of $s_n(\mathbf{m}, \mathbf{\Sigma})$ subject to the constraint

$$\sum_{i=1}^n \log(\det(\mathbf{\Sigma}^{(\mathbf{u}_i)})) = 0. \tag{21}$$

Finally,

$$\widehat{\mathbf{\Sigma}}_n = s_n(\widehat{\mathbf{m}}_n, \widetilde{\mathbf{\Sigma}}_n)\widetilde{\mathbf{\Sigma}}_n. \tag{22}$$

Notice that Equation (21) is a scaling constraint. It is easy to show that, for any $p \times p$ positive definite $\mathbf{\Sigma}$, there exists $a > 0$ such that $\sum_{i=1}^n \log(\det(a\mathbf{\Sigma}^{(\mathbf{u}_i)})) = 0$. In fact, it suffices to take $a = \exp\{-\sum \log(\det(\mathbf{\Sigma}^{(\mathbf{u}_i)}))/\sum p(\mathbf{u}_i)\}$. Good references for the Gaussian MLE include Tanner (1993), Schafer (1997), Kenward and Molenberghs (1998), and Little and Rubin (2002) among many others.

It is well known that this MLE is not robust. To define a robust alternative, Danilov (2010) considered a new scale $s_n(\mathbf{m}, \mathbf{\Sigma})$ defined as the solution to

$$\sum_{i=1}^n k_{p(\mathbf{u}_i)} c_{p(\mathbf{u}_i)} \rho\left(\frac{d_i\left(\mathbf{x}_i, \mathbf{m}^{(\mathbf{u}_i)}, \mathbf{\Sigma}^{(\mathbf{u}_i)}\right)}{s c_{p(\mathbf{u}_i)}}\right) = \frac{1}{2}\sum_{i=1}^n c_{p(\mathbf{u}_i)} k_{p(\mathbf{u}_i)}. \tag{23}$$

As in the Gaussian MLE case, the robust ESE estimator is defined as $(\widehat{\mathbf{m}}_n, \widehat{\mathbf{\Sigma}}_n)$, where $\widehat{\mathbf{\Sigma}}_n = s_n(\widehat{\mathbf{m}}_n, \widetilde{\mathbf{\Sigma}}_n)\widetilde{\mathbf{\Sigma}}_n$ and $(\widehat{\mathbf{m}}_n, \widetilde{\mathbf{\Sigma}}_n)$ minimizes $s_n(\mathbf{m}, \mathbf{\Sigma})$ subject to (21). Notice that ESE can be computed directly from the data using subsampling. Danilov (2010) showed that if ρ is nondecreasing, continuously differentiable, and bounded, we obtain robustness and Fisher consistency at the multivariate normal model by taking c_j and k_j satisfying

$$\frac{1}{2} = E\left(\rho\left(\frac{Z}{c_j}\right)\right), \quad k_j = \frac{1}{j}E\left(\rho'\left(\frac{Z}{c_j}\right)Z\right),$$

where Z has a chi-squared distribution with j degrees of freedom. Notice that when $\rho(d) = d$, we have $c_j = 2j$ and $k_j = 1$ for all j and (23) reduces to (20). Unfortunately, unlike for GSE, the consistency of ESE for the scatter shape component cannot be established for general elliptical distributions.

6.2 MVE for Incomplete Data

To achieve maximum robustness—specially in the case of large p—we wish to compute the ESE version of the minimum volume ellipsoid (EMVE) which has the discontinuous loss function

$$\rho_0(t) = I_{[1,\infty)}(t). \tag{24}$$

To obtain the EMVE consistency correction constants k_j, we consider the approximating loss functions

$$\rho_\varepsilon(t) = \begin{cases} 0, & t \leq (1-\varepsilon), \\ \dfrac{1}{2\varepsilon}(t+\varepsilon-1), & (1-\varepsilon) < t < (1+\varepsilon), \\ 1, & t \geq (1+\varepsilon), \end{cases}$$

and calculate $c_j = \lim_{\varepsilon \to 0} c_j(\varepsilon)$, where $c_j(\varepsilon)$ satisfies

$$E\left\{\rho_\varepsilon\left(\frac{Y}{c_j(\varepsilon)}\right)\right\} = \frac{1}{2}, \tag{25}$$

where Y has a chi-squared distribution with j degrees of freedom. Moreover, k_j is computed by

$$k_j = \lim_{\varepsilon \to 0} k_j(\varepsilon) = \lim_{\varepsilon \to 0} E\left\{X_1^2 \rho_\varepsilon'\left(\frac{X_1^2 + X_2^2 + \cdots + X_j^2}{c_j(\varepsilon)}\right)\right\}, \tag{26}$$

where $X_1, \ldots, X_j$ are iid standard normal random variables. More details on the derivation of the constants k_j and c_j are given in Section 3 of the supplementary material.

6.3 Resampling Algorithm for EMVE

We take N subsamples of size $n_0 = p/(1-\alpha)$, where α is the fraction of missing data (α = number of missing entries/np). As usual for algorithms based on subsampling, N can be taken so that we get an outlier-free subsample with a desired probability. The subsample size n_0 is taken larger than p to avoid singularity.

The following steps are performed for each subsample:

1. Compute $\widehat{\mathbf{m}}_0$, the coordinate-wise median (for the given subsample).
2. Complete the subsample replacing each missing entry by the overall median for that variable (calculated on the entire dataset).
3. Let $\widetilde{\mathbf{\Sigma}}_0$ be the sample covariance of the completed subsample multiplied by a scalar factor so that Equation (21) holds. If $\widetilde{\mathbf{\Sigma}}_0$ is singular (or very badly conditioned), discard the subsample.
4. Compute the EMVE scale $s_0 = s_n(\widehat{\mathbf{m}}_0, \widetilde{\mathbf{\Sigma}}_0)$ defined by Equation (23) with $\rho(t) = I_{[1,\infty)}(t)$ and set $\widehat{\mathbf{\Sigma}}_0 = s(\widehat{\mathbf{m}}_0, \widetilde{\mathbf{\Sigma}}_0)\widetilde{\mathbf{\Sigma}}_0$.
5. Compute the partial squared Mahalanobis distances $d_i = d(\mathbf{x}_i, \mathbf{u}_i, \widehat{\mathbf{m}}_0, \widehat{\mathbf{\Sigma}}_0)$, $1 \leq i \leq n$.
6. Since the $p(\mathbf{u}_i)$ may be different for each case, we cannot compare the d_i directly. Then, for comparison purposes, we compute $\pi_i = F_{p(\mathbf{u}_i)}(d_i)$, $1 \leq i \leq n$, where F_j is the

chi-squared distribution function with j degrees of freedom.

7. Concentration step: choose 50% of the points with the smallest π_i and compute $(\widehat{\mathbf{m}}_1, \widetilde{\boldsymbol{\Sigma}}_1)$ as the Gaussian MLE for this half-sample using the classical EM algorithm multiplied by a scalar factor so that (21) holds.
8. Again, compute the EMVE scale $s_1 = s_n(\widehat{\mathbf{m}}_1, \widetilde{\boldsymbol{\Sigma}}_1)$ and put $\widehat{\boldsymbol{\Sigma}}_1 = s_n(\widehat{\mathbf{m}}_1, \widetilde{\boldsymbol{\Sigma}}_1)\widetilde{\boldsymbol{\Sigma}}_1$.
9. If $s_1 < s_0$, we set $(\widehat{\mathbf{m}}_0, \widehat{\boldsymbol{\Sigma}}_0, s_0) = (\widehat{\mathbf{m}}_1, \widehat{\boldsymbol{\Sigma}}_1, s_1)$.

Finally, we choose as the EMVE, the pair $(\widehat{\mathbf{m}}_0, \widehat{\boldsymbol{\Sigma}}_0)$ with smallest MVE scale s_0.

7. MONTE CARLO SIMULATION STUDY

We conduct a simulation study to investigate the performance of the proposed estimators. We consider samples of size $n = 100$ from uncontaminated and contaminated normal distributions of dimension $p = 10$. Since the estimators are scale- and location-equivariant, we assume without loss of generality that the means are equal to 0 and the variances are equal to 1. Since the model and estimators are not affine-equivariant, we consider several correlation structures by taking the off-diagonal entries of the covariance matrix $\boldsymbol{\Sigma}$ all equal to r, with $r = 0.5, 0.6, \ldots, 0.9$. We introduce 10% point mass contaminations of different sizes, located at k Mahalanobis distances away from $\mathbf{0}$ ($k = 1, 2, \ldots, 12$), in the direction of the eigenvector of $\boldsymbol{\Sigma}$ associated with the smallest eigenvalue. It has been empirically observed that this is the least favorable placement for the outliers. The percentage of missing values is fixed at 10%. Results for other fractions of contaminations and missing values (not reported here) present similar patterns. The number of replicates is $N = 100$.

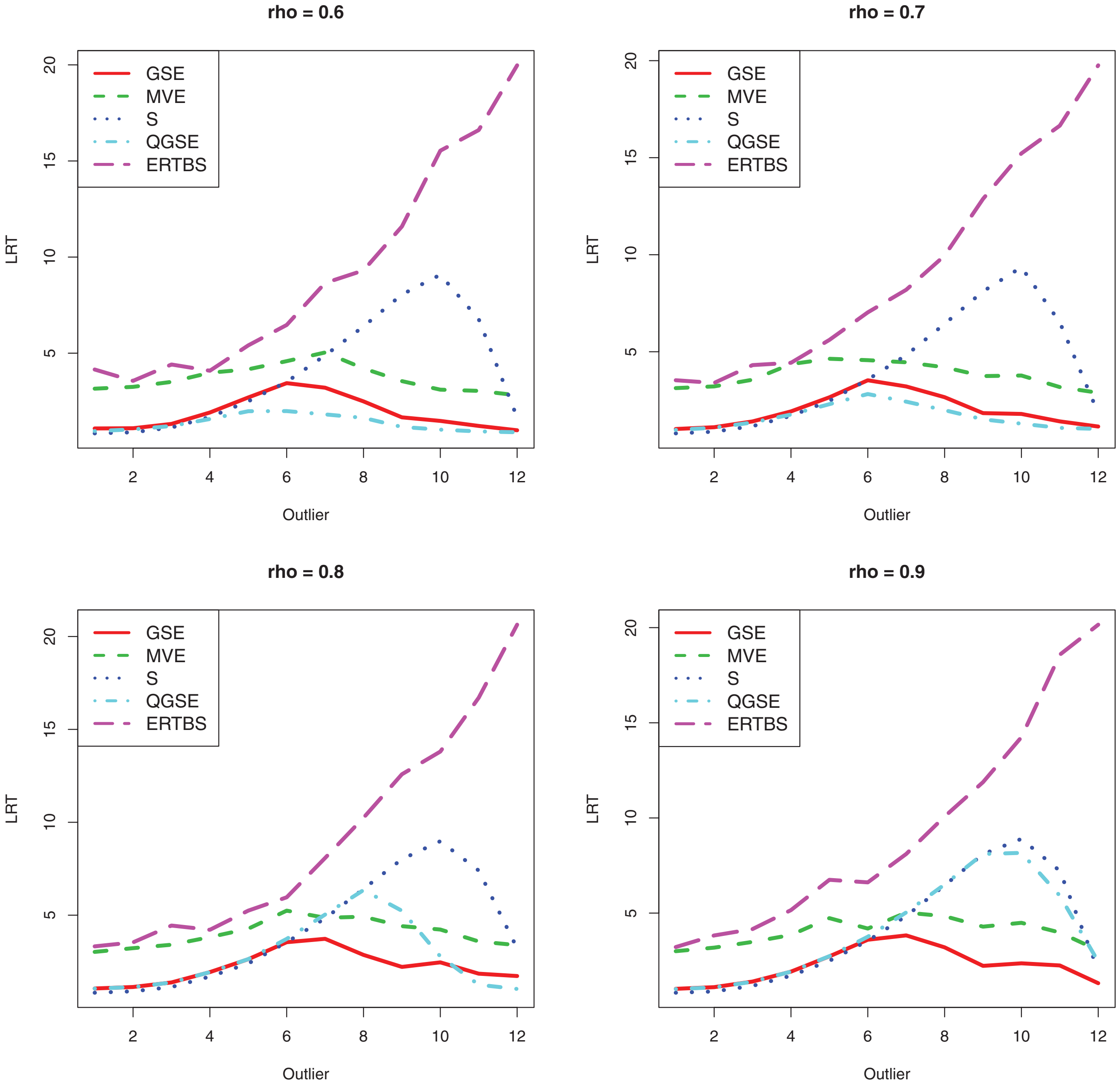

Figure 1. Monte Carlo Study. We consider samples of size 100, of 10-dimensional observations. We plot the average LRT distances as a function of the outlier size, for different correlation structures and 10% of missing data. The online version of this figure is in color.

Table 1. Monte Carlo study

	Estimates				
ρ	ERTBS	EMVE	GSE	QGSE	EM
0.50	0.27	0.29	0.87	0.88	1.00
0.60	0.24	0.30	0.88	0.91	1.00
0.70	0.29	0.31	0.88	0.89	1.00
0.80	0.26	0.30	0.89	0.89	1.00
0.90	0.25	0.29	0.87	0.87	1.00
0.99	0.24	0.31	0.87	0.87	1.00

NOTES. Gaussian LRT efficiency (relative to EM) for some robust scatter estimates. We consider clean 10-dimensional samples of size 100 with 10% of missing values.

Performance measure: The performance of a given scatter estimator $\widehat{\boldsymbol{\Sigma}}_n$ is measured by $E(\mathrm{LRT}(\widehat{\boldsymbol{\Sigma}}_n, \boldsymbol{\Sigma}_0))$, where $\mathrm{LRT}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0)$ is the likelihood radio test distance $\mathrm{LRT}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0) = \mathrm{trace}(\boldsymbol{\Sigma}\Sigma_0^{-1}) - \log(\det(\boldsymbol{\Sigma}\Sigma_0^{-1})) - p$. This distance appears naturally in the context of the Gaussian likelihood ratio test statistic to test the hypothesis that the population covariance matrix equals $\boldsymbol{\Sigma}_0$.

We compare the following estimators:

(a) EMVE, the extended *S*-estimate described in Section 6.2;
(b) GSE, the generalized *S*-estimate with function $\rho(u) = \rho_B(\sqrt{u})$, where $\rho_B(u) = \min(1, 1 - (1 - u^2)^3)$ is the Tukey's bisquare rho function, and using the EMVE as initial estimator;
(c) QGSE, a fast version of GSE with the pairwise *quadrant correlation* as initial estimator;
(d) ERTBS, the estimator proposed by Copt and Victoria-Feser (2003), evaluated using the R-code kindly provided by the authors; and
(e) FS, the fast *S*-estimator proposed in Section 6.7.5 of Maronna, Martin, and Yohai (2006). FS was computed using the function covSest (method = bisquare) from the R-package rrcov, evaluated on the complete data. FS is not an estimator for incomplete data, however it is included for comparison purposes.

Table 1 shows the finite sample relative efficiency of the robust estimates with respect to the classical EM estimator based on the average LRT distances over the Monte Carlo replicates, when there are 10% of missing data and no outliers.

We note that ERTBS and EMVE are quite inefficient while GSE and QGSE have efficiencies close to 0.9. The average LRT distances when we have 10% of outlier contamination of different sizes and 10% of missing data are reported in Figure 1.

Notice the stable and good performance of GSE, comparable to FS computed on the complete dataset. QGSE also performs pretty well, specially for small *r*. EMVE is a robust and stable initial estimator responsible for the excellent performance of GSE. However, EMVE has a relatively weak performance in this simulation due to its low efficiency.

As suggested by an anonymous referee, we also investigated the possible use of other fast initial estimators besides the quadrant correlation, such as (1) a pairwise *S*-estimator applied to all pairs of variables using the corresponding complete cases and (2) fast *S*-estimator with Tukey's bisquare function applied to the completed data after NAs are replaced by the coordinatewise median on the available data. Details are given in Section 1 of the supplementary material. Unlike GSE, which has a stable good performance for all the considered correlation structures, all the considered fast versions worked well for some correlation structures but poorly for others. The most stable among the fast versions is the GSE using the quadrant correlation as initial estimator.

Finally, computing times for the different estimates can be found in the supplementary material.

8. EXAMPLES

Example 1. Boston Housing Data: Our first example uses the Harrison and Rubinfeld (1978) "Boston Housing Dataset" downloaded from the R-package "spdep," with 506 cases and 12 variables. The Mahalanobis distances for the complete data using FS estimates as center and scatter matrix are given in Figure 2.

There are 174 outliers accounting for 34% of the cases. The outliers correspond mostly to cases 142–172 and 357–492 with 132 of them having variable index of accessibility to radial highways (RAD) = 24. Note that the median and median absolute deviation (MAD) of RAD are equal to 5 and 2.96, respectively. Deletion of these outliers and recalculation of the robust estimator and Mahalanobis distances reveal no further outlying cases. The need for robust analysis is justified by the fact that the MLE approach identifies only 10 outliers (cases 366, 369, 381, 399, 405, 406, 411, 415, 419, and 428) after several iterations of outlier deletion followed by recalculation of the mean, covariance matrix, and Mahalanobis distances. We now set a randomly chosen 10% of the entries equal to NA and use the partial Mahalanobis distances $\tilde{d}_i$ to identify outliers. The

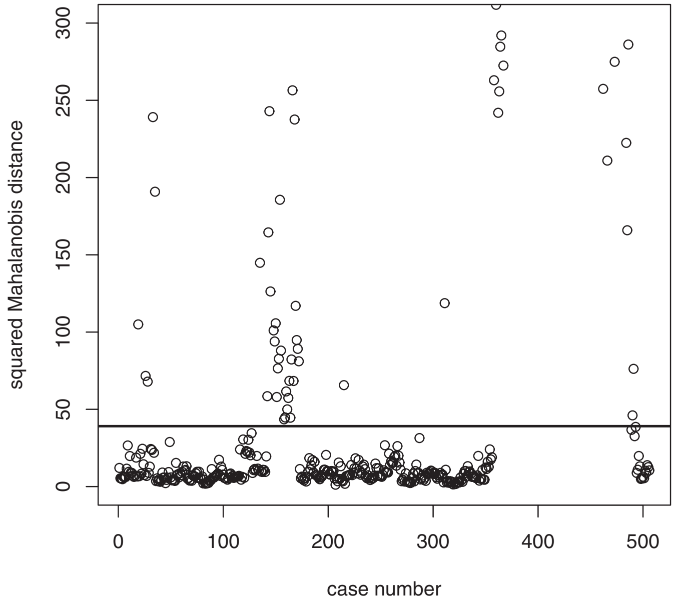

Figure 2. Boston Housing Data. Squared Mahalanobis distances using the fast *S*-estimate of scatter matrix and multivariate location with all the data. There are in total 174 outliers. One hundred and twenty-five distances exceeding the value 300 have been excluded from the plot.

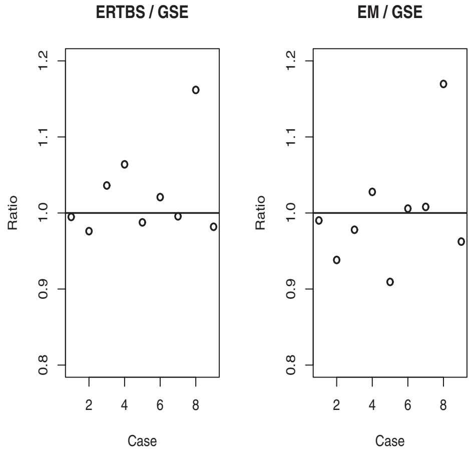

Figure 3. Wages and Hours Data. Ratio comparison of the location estimates for each single variable, taking GES as baseline. All the ratios are between 0.9 and 1.1 except for variable 7 (Race) where the ratios are a bit larger.

partial distances are adjusted using the formula $d_i = F_p^{-1}(F_{p_i}(\tilde{d}_i))$, where $p = 12$ and p_i is the number of observed variables for the ith case. The maximum likelihood approach (EM algorithm in this case) only finds eight outliers (cases 366, 381, 399, 405, 406, 411, 415, and 419) after a few iterations. On the other hand, GSE identifies 169 outliers which are a subset of the 174 outliers found in the complete case analysis. The five nonidentified points are cases 159, 171, 392, 394, and 464. Cases 159 and 171 have a large number of missing entries (five and four, respectively). The number of missing entries (per case) has mean = 1.3 and standard deviation = 1. Moreover, cases 392, 394, and 464 have RAD = NA while RAD = 24 in the complete data. This may have been useful to identify these three cases as outliers in the complete data analysis. We also conduct an experiment to illustrate the estimators' ability (or lack of) to cope with missing data. In this experiment, we do not evaluate the robustness of the estimators but their ability to emulate their complete data values using the incomplete data. Hence, we compute the LRT distance between the scatter matrices

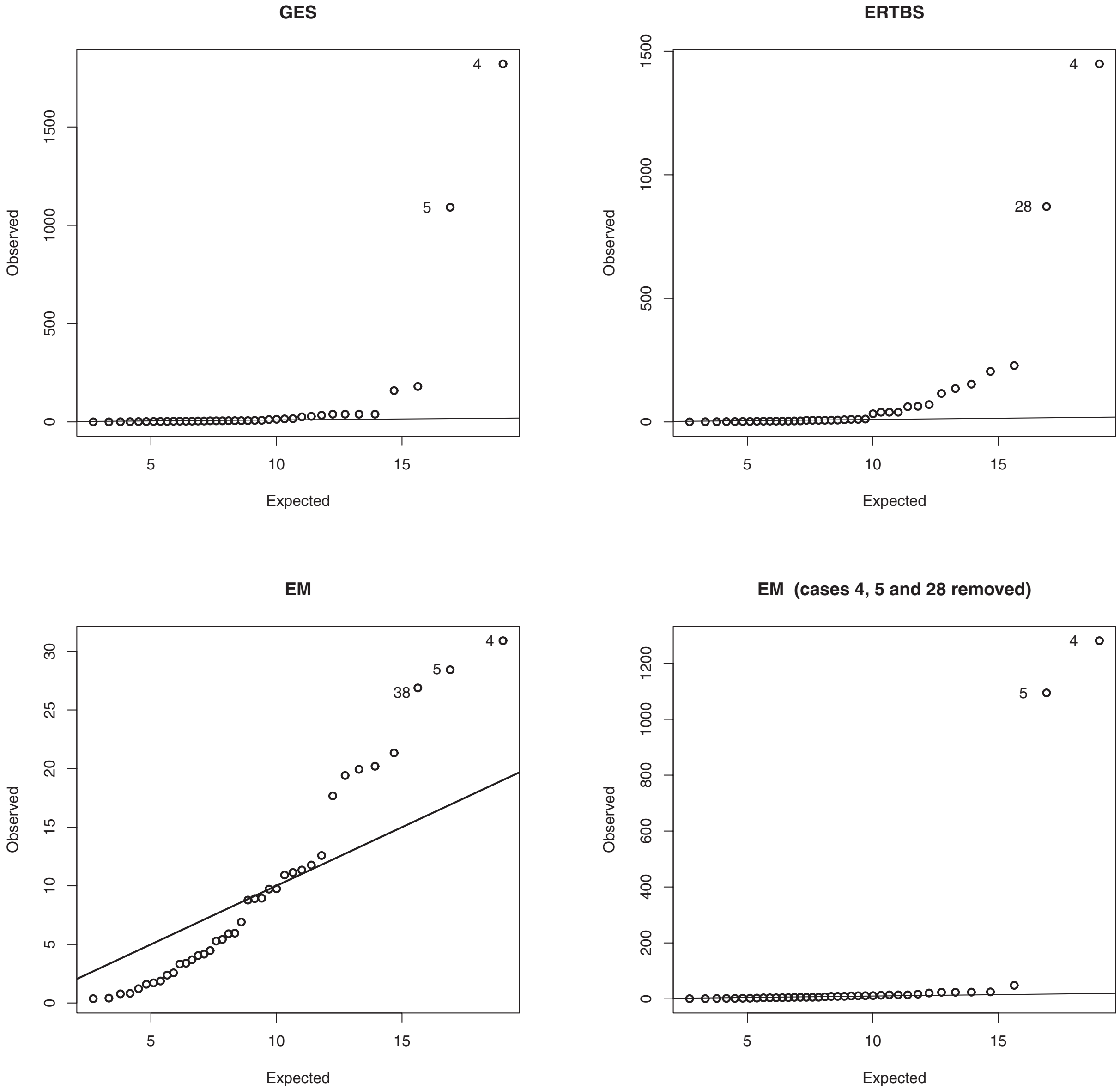

Figure 4. Wages and Hours Data. QQ plots for the Mahalanobis distances using GSE, ERTBS, EM, and EM applied to the "clean data" (removing the big outliers detected by GSE and ERTBS). Only the GSE big outliers (cases 4 and 5) remain as such when EM is applied to the "clean" data. The partial Mahalanobis distances are adjusted (see the text) to make them comparable.

Table 2. Average effect of missing data on the estimators' "intended results" (that is, those that would be obtained if the complete data were available)

	Percentage of missing data		
Estimator	10%	20%	30%
EM	0.07(0.03)	0.15(0.06)	0.32(0.13)
GSE	0.10(0.03)	0.26(0.08)	0.56(0.15)
QGSE	1.14(0.42)	2.68(0.60)	4.73(0.88)
EMVE	1.91(0.62)	2.55(0.85)	2.96(0.96)
ERTBS	0.39(0.16)	3.58(5.3)*	25.76(11)**

NOTES. (*) Average obtained from 19 replicates because ERTBS crashed on one occasion. (**) Average obtained from seven replicates.

estimated before and after random missingness is introduced in the data. The averages over 20 replicates are displayed in Table 2.

Not surprisingly, EM shows the best performance closely followed by GSE. The other three robust estimators are considerably worse. The poor performance of QGSE may be because these data are highly correlated (the inverse condition number for FS and MLE scatter estimates computed on complete data is 2e-07 and 6e-08, respectively).

Example 2. Wages and Hours: In this example, we have 39 cases and 9 variables. A national sample of 6000 households with earnings below $15,000 was obtained in 1966. The 6000 households were divided into 39 demographic subgroups and the averages over these groups were used to investigate the relation between "average hours worked during the year" and "average hourly wages" adjusting for other seven variables. Overall, 4.3% of the data are missing and 28% of the cases have at least one missing value. We computed GSE, ERTBS, and EM for these data.

Figure 3 shows that the three estimates roughly agree regarding the multivariate location for the nine variables with the exception of Race (variable number 7) where the ERTBS and EM estimates are somewhat larger.

The first three panels of Figure 4 display the chi-squared qq-plots for the adjusted partial square Mahalanobis distances for the three estimates. The adjusted square distances for nonoutlying cases using GSE and ERTBS follow an approximate chi-square distribution with nine degrees of freedom. The EM-adjusted Mahalanobis distances do not seem to follow an approximate chi-square distribution and do not highlight any clear big outlier. GSE finds two big outliers—cases 4 and 5—and two marginal outliers. ERTBS finds two big outliers—cases 4 and 28—and seven marginal outliers. Notice that case 5 is not an ERTBS outlier while case 28 is not a GSE outlier. Finally, we remove the large outliers found by GSE and ERTBS (cases 4, 5, and 28) and apply EM to the remaining data. In this case, only cases 4 and 5 are identified as outliers and the adjusted partial square Mahalanobis distances are very similar to those produced by the original GSE fit.

SUPPLEMENTARY MATERIAL

The supplementary material (available online) has four sections. Section 1 contains simulation results (performance) for other initial estimates. Section 2 includes a table showing the computing times for the different estimates. Section 3 derives the consistency constants k_j for defining EMVE. Finally, Section 4 gives detailed proofs for Theorems 1–3.

[Received January 2011. Revised February 2012.]

REFERENCES

Cheng, T. C., and Victoria-Feser, M. P. (2002), "High-Breakdown Estimation of Multivariate Mean and Covariance With Missing Observations," *British Journal of Mathematical and Statistical Psychology*, 55, 317–335. [1178]

Copt, S., and Victoria-Feser, M. P. (2003), "Fast Algorithms for Computing High Breakdown Covariance Matrices With Missing Data," Technical Report 2003.04, Université de Geneve. [1184]

Croux, C., Filzmoser, P., and Joossens, K. (2008), "Classification Efficiencies for Robust Discriminant Analysis," *Statistica Sinica*, 18, 588–599. [1178]

Danilov, M. (2010), "Robust Estimation or Multivariate Scatter Under Non-Affine Equivariant Scenarios," Ph.D. dissertation, Department of Statistics, University of British Columbia. [1182]

Davies, P. (1987), "Asymptotic Behaviour of S-Estimates of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics*, 15, 1269–1292. [1178,1179,1181]

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [1178]

Frahma, G., and Jaekel, U. (2010), "A Generalization of Tyler's M-Estimators to the Case of Incomplete Data," *Computational Statistics and Data Analysis*, 54, 374–393. [1178]

Harrison, D., and Rubinfeld, D. L. (1978), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81–102. [1184]

Kenward, M. G., and Molenberghs, G. (1998), "Likelihood Based Frequentist Inference When Data Are Missing at Random," *Statistical Science*, 13, 236–247. [1182]

Little, R. J. A. (1988), "Robust Estimation of the Mean and Covariance Matrix From Data With Missing Values," *Journal of the Royal Statistical Society*, Series C, 37, 23–38. [1178]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley. [1178,1182]

Little, R. J. A., and Smith, P. J. (1987), "Editing and Imputing for Quantitative Survey Data," *Journal of the American Statistical Association*, 82, 58–68. [1178]

Maronna, R. A. (1976), "Robust M-Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51–67. [1178]

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, Chichister: Wiley. [1178,1182,1184]

Rocke, D. M. (1996), "Robustness Properties of S-Estimators of Multivariate Location and Shape in High Dimension," *The Annals of Statistics*, 24, 1327–1345. [1178]

Rousseeuw, P. (1985), "Multivariate Estimation With High Breakdown Point," *Mathematical Statistics and Applications*, 8, 283–297. [1178]

Rousseeuw, P. J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223. [1178]

Salibian-Barrera, M., Van Aelst, S., and Willems, G. (2006), "PCA Based on Multivariate MM-Estimators With Fast and Robust Bootstrap," *Journal of the American Statistical Association*, 101, 1198–1211. [1178,1181]

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall. [1182]

Tanner, M. A. (1993), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (2nd ed.), New York: Springer. [1182]

Taskinen, S., Croux, C., Kankainen, A., Ollila, E., and Oja, H. (2006), "Influence Functions and Efficiencies of the Canonical Correlation and Vector Estimates Based on Scatter and Shape Matrices," *Journal of Multivariate Analysis*, 97, 359–384. [1178,1181]

Tatsuoka, K., and Tyler, D. (2000), "The Uniqueness of S and M-Functionals Under Non-Elliptical Distributions," *The Annals of Statistics*, 28, 1219–1243. [1180,1181]

Templ, M., Kowarik, A., and Filzmoser, P. (2011), "Iterative Stepwise Regression Imputation Using Standard and Robust Methods," *Computational Statistics & Data Analysis*, 55, 2793–2806. [1178]

Tyler, D. (1987), "A Distribution-Free M-Estimator of Multivariate Scatter," *The Annals of Statistics*, 15, 234–251. [1178]