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

ABSTRACT

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 EM algorithm for computing the proposed estimators. This algorithm is initialized by an extension for missing data of the minimum volume ellipsoid. We asses the performance of our proposal by Monte Carlo simulation and give some real data examples. This article has Supplemental Material on line.

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, etc. See, for example, Salibian-Barrera et al (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 EM algorithm to maximize the likelihood for the observed data (see Dempster, Laird and Rubin, 1977 and Little and Rubin, 2002). To robustly estimate these parameters in the presence of outliers and missing data, Little and Smith (1987) proposed the 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 non–missing part of each observation and the current scatter. Little (1988) robustifies the Gaussian EM-algorithm using a multivariate Student's t-distribution or some other heavy tail distribution. It is well known, however, that such MLE estimators 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 loose 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 (MCD) 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 et al. (2011) proposed a general robust imputation method to deal with large datasets possessing outliers and missing data.

In this paper, we present two classes of robust estimators for missing data: *generalized S-estimators* (GSE) and *extended S-estimators* (ESE). Both classes coincide with the S-estimators introduced by Davies (1987) for complete data. Since GSE 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).

The rest of the paper 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 fix-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 extended S-estimates (ESE) and the extended minimum volume ellipsoid (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 $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$, $1 \leq i \leq n$, be p-dimensional i.i.d. random vectors with common density f belonging to the elliptical family

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

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 (1) gives a larger semiparametric family. Let $\mathbf{u}_i = (u_{i1}, ..., u_{ip})', 1 \leq i \leq n$, 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 $\mathbf{x} = (x_1, ..., x_p)'$ and $\mathbf{u} = (u_1, ..., u_p)'$ let $\mathbf{x}^{(\mathbf{u})}$ be the observed part of $\mathbf{x}$ and set

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

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 $1 \leq j < k \leq p$, there exists at least one $\mathbf{u}_i$, $1 \leq i \leq n$, with $u_{ij} = u_{ik} = 1$

Let $A_p = \{\mathbf{u} : (u_1, ..., u_p)',\ u_i \in \{0, 1\}\}$, then given a $p \times p$ positive definite matrix $\mathbf{\Sigma}$ and $\mathbf{u} \in A_p$, we denote by $\mathbf{\Sigma}^{(\mathbf{u})}$ the submatrix of $\mathbf{\Sigma}$ 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 $\left|\boldsymbol{\Sigma}^{*(\mathbf{u})}\right| = 1$.

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

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

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 non-decreasing 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

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

and where $\mathbf{X}$ has density given by (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

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

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

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

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

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

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

$$\widehat{\mathbf{\Sigma}}_n = \widehat{s}_n \widetilde{\mathbf{\Sigma}}_n, \tag{8}$$

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

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

3.2 Existence of GSE

Now we consider the existence of a solution $\left(\widehat{\mathbf{m}}_n, \widetilde{\Sigma}_n \right)$ to the minimization problem (7) with $\mathbf{m}$ ranging over R^p and Σ ranging over the set of $p \times p$ positive definite matrices. Consider the sample $(\mathbf{x}_i, \mathbf{u}_i)$, $i = 1, 2, ..., n$. 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 $\mathbf{u}_0 = (1, ..., 1)$ 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 Supplemental Material) gives a sufficient condition for the existence of the GES.

Assumption A. The function ρ is (i) non decreasing, (ii) strictly increasing at 0, (iii) continuous, (iv) $\rho(0) = 0$ and (v) $\lim_{v\to\infty} \rho(v) = 1$.

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

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

Then there exists at least one value $\left(\widehat{\mathbf{m}}_n, \widetilde{\Sigma}_n\right)$ *minimizing* $S_n^*(m, \mathbf{\Sigma})$ *with* $\widetilde{\Sigma}_n$ *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) conjecture that the S-estimator solution is unique with probability one 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” (i) $x + NA = NA$, for all x (NA means “non-available”), (ii) $x \times NA = NA$ for all $x \neq 0$ and (iii) $0 \times NA = 0$. Since GSE is defined using Mahalanobis distances, if $\mathbf{A}$ is an invertible matrix that preserves the missingness pattern [that is, for all $i = 1, 2, ..., n$, $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}$ 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

$$\mathbf{A} = \left(\begin{array}{cc} \mathbf{A}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{p-q} \end{array} \right)$$

preserves the missingness pattern.

3.4 Scaling the Mahalanobis Distances.

The partial Mahalanobis distances in (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 re-scale these distances using the "tuning constants" $\left|\widehat{\mathbf{\Omega}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}$. To fix ideas, suppose that $\mathbf{m} \approx \mathbf{m}_0$ and $\mathbf{\Sigma} \approx \widehat{\mathbf{\Omega}}_n \approx \mathbf{\Sigma}_0$ in (6). Then

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

where Y has density $f_0(||\mathbf{y}||^2)$ and so $||Y^{(\mathbf{u}_i)}||^2/c_{p(\mathbf{u}_i)}$ has M-scale equal to one 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 generalized S-estimator $\left(\widehat{\mathbf{m}}_n, \widehat{\mathbf{\Sigma}}_n\right)$ 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

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

where $S_n(\mathbf{m}, \mathbf{\Sigma})$ is defined in Section 2. Since the factor $\left|\widehat{\mathbf{\Omega}}_n\right|^{1/p}$ is constant, $\widetilde{\mathbf{\Sigma}}$ / $\left|\widetilde{\mathbf{\Sigma}}\right|^{1/p}$ 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 generalized S-estimators coincide in this case.

3.6 Fix-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 $\widehat{\mathbf{x}}\left(\mathbf{u}, \mathbf{x}^{(\mathbf{u})}, \mathbf{m}, \mathbf{\Sigma}\right)$ as the best linear predictor of $\mathbf{X}$ given $\mathbf{X}^{(\mathbf{u})} = \mathbf{x}^{(\mathbf{u})}$, when $E(\mathbf{X}) = \mathbf{m}$ and $Cov(\mathbf{X}) = \mathbf{\Sigma}$. Moreover $\mathbf{C}(\mathbf{u}, \mathbf{\Sigma})$ is the covariance matrix for the prediction error $\mathbf{X} - \widehat{\mathbf{x}}\left(\mathbf{u}, \mathbf{X}^{(\mathbf{u})}, \mathbf{m}, \mathbf{\Sigma}\right)$ 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 one and the remaining entries equal to zero, we have the following simple formulas. Let $\mathbf{v} = (v_1, ..., v_p)' \in \mathbf{A}_p$ such that $v_1 = \cdots = v_q = 0$ and $v_{q+1} = \cdots = v_p = 1$ and write

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

Then,

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

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

The following theorem (proved in Section 4.3 of the Supplemental Material) gives fix-point estimating equations for the GSE estimators of location and scatter shape component.

Theorem 2 *Let $\widehat{\mathbf{m}}_n$ and $\widetilde{\boldsymbol{\Sigma}}_n$ be defined by (7). Assume that ρ is a non-decreasing 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

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

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

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

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

$\widetilde{s}_n = S_n^*\left(\widehat{\mathbf{m}}_n, \widetilde{\mathbf{\Sigma}}_n\right)$ *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 fix-point equations to derive a computing algorithm in Section 5

4 Consistency of GSE

Theorem 3 below (proved in Section 4.4 of the Supplemental Material) shows that $(\widehat{\mathbf{m}}_n, \widetilde{\mathbf{\Sigma}}_n) \rightarrow (\mathbf{m}_0, t_0\mathbf{\Sigma}_0)$ a.s. for some $t_0 > 0$ even if the true f_0 is different from that used in (4). Therefore, GSE 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, e.g. principal component and canonical correlation analysis, only the scatter shape component is required (see Salibian-Barrera et al. and 2006; Taskinen et al., 2006).

Set $G\left(\mathbf{m}, \Sigma, F, c\right) = E_F\left(\rho\left(d\left(\mathbf{x}, \mathbf{m}, \mathbf{\Sigma}\right)/c\right)\right)$ 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 B. Given $1 \leq j < k \leq p$, there exists $\mathbf{u_i} = (u_{i1}, ..., u_{ip}) \in A_p$ such that $u_{ij} = u_{ik} = 1$.

Assumption C. 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 $\left(\mathbf{\Sigma}^{(\mathbf{u})}\right) = 1$ is $({\mathbf{m}_0}^{(\mathbf{u})}, \mathbf{\Sigma}_0^{*(\mathbf{u})})$.

Davies (1987) shows that if ρ satisfies Assumption A (in Theorem 1) and F is elliptical with f strictly decreasing (see equation (1)) then Assumption C holds.

In the case of complete data, Tatsuoka and Tyler (2000) show 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 $\mathbf{x} = A\mathbf{y} + \mathbf{m}$, $\mathbf{y} = (y_1, y_2, ..., y_p)'$, for i.i.d. Student's t-random variables $y_1, y_2, ..., y_p$ and invertible matrix A. Unfortunately, we cannot prove that Assumption C holds for these distributions. However, we believe based on Monte Carlo experiments with large samples (not presented here) that Assumption C, and therefore the consistency of GSE, hold for these distributions.

Theorem 3 *Suppose that* $\mathbf{x}_1, .., \mathbf{x}_n$ *is a random sample from* F_0, $(\widehat{\mathbf{m}}_n, \widehat{\Sigma}_n)$ *is defined by (6) and (8) with* $\widehat{\mathbf{\Omega}}_n \to \mathbf{\Omega}_0$ *a.s., where* $\mathbf{\Omega}_0$ *is positive definite, and assumptions A, B and C hold. Then, (i)* $\widehat{\mathbf{m}}_n \to \mathbf{m}_0$ *a.s. (ii)* $\widehat{\mathbf{\Sigma}}_n \to t_0 \mathbf{\Sigma_0}$ *a.s. where* t_0 *is defined by*

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

and (iii) 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 fix-point estimating equations (13) and (14) in Theorem 2.

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

Given $\left(\widehat{\mathbf{m}}_n^{(k)}, \widehat{\mathbf{\Sigma}}_n^{(k)}, \widetilde{s}_n^{(k)}\right)$ compute $\left(\widehat{\mathbf{m}}_n^{(k+1)}, \widehat{\mathbf{\Sigma}}_n^{(k+1)}, \widetilde{s}_n^{(k+1)}\right)$ 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 $\widehat{\mathbf{x}}_i^{(k)} = \widehat{\mathbf{x}}\left(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n^{(k)}, \widehat{\mathbf{\Sigma}}_n^{(k)}\right)$, $\mathbf{C}_i^{(k)} = \mathbf{C}\left(\mathbf{u}_i, \widehat{\mathbf{\Sigma}}_n^{(k)}\right)$, $w_i^{(k)} = w\left(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n^{(k)}, \widehat{\mathbf{\Sigma}}_n^{(k)}, \tilde{s}_n^{(k)}\right)$ and $w_i^* = w^*\left(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n^{(k)}, \widehat{\mathbf{\Sigma}}_n^{(k)}\right)$, 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 $\left|\tilde{s}_n^{(k+1)} / \tilde{s}_n^{(k)} - 1\right| < \delta$ 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) coincide, in the case of complete data, with the algorithm used to compute the S-estimator. In such case, Maronna et al (2006) Section 6.7.5 show 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 couldn't 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 $\left(\widehat{\mathbf{m}}_n^{(0)}, \widehat{\boldsymbol{\Sigma}}_n^{(0)}, \widetilde{s}_n^{(0)}\right)$ to compute GSE. We introduce now the class of extended S-estimators (ESE) 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 maximum likelihood estimator $(\widehat{\mathbf{m}}_n,\widehat{\boldsymbol{\Sigma}}_n)$ for the MCAR model is obtained as follows. Given $(\mathbf{m}, \boldsymbol{\Sigma})$, let $s_n\left(\mathbf{m}, \boldsymbol{\Sigma}\right)$ be the solution to

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

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

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

Finally

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

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

It is well known that this MLE estimator is not robust. To define a robust alternative Danilov (2010) considers a new scale $s_n\left(\mathbf{m}, \boldsymbol{\Sigma}\right)$ 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 $\left(\widehat{\mathbf{m}}_n, \widehat{\mathbf{\Sigma}}_n\right)$ 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 $s_n\left(\mathbf{m}, \mathbf{\Sigma}\right)$ subject to (21). Notice that ESE can be computed directly from the data using subsampling. Danilov (2010) showed that if ρ is non-decreasing, continuously differentiable and bounded, we obtain robustness and Fisher consistency at the multivariate normal model by taking c_j and k_j satisfying

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

where Z has a chi-squared distribution with j degrees of freedom. Notice that when $\rho\left(d\right) = d$ 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

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

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

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

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 $X_1, ..., X_j$ are i.i.d. standard normal random variables. More details on the derivation of the constants k_j and c_j are given in Section 3 of the Supplemental 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 $\widetilde{\boldsymbol{\Sigma}}_0$ be the sample covariance of the completed subsample multiplied by a scalar factor so that (21) holds. If $\widetilde{\boldsymbol{\Sigma}}_0$ is singular (or very badly conditioned) discard the subsample.

4. Compute the EMVE scale $s_0 = s_n(\widehat{\mathbf{m}}_0, \widetilde{\boldsymbol{\Sigma}}_0)$ defined by (23) with $\rho(t) = I_{[1,\infty)}(t)$ and set $\widehat{\boldsymbol{\Sigma}}_0 = s(\widehat{\mathbf{m}}_0, \widetilde{\boldsymbol{\Sigma}}_0)\widetilde{\boldsymbol{\Sigma}}_0$.

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 can not compare the d_i directly. Then, for comparison purposes we compute $\pi_i = F_{p(\mathbf{u}_i)}\left(d_i\right),\ 1 \leq i \leq n$, 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 $\left(\widehat{\mathbf{m}}_1, \widetilde{\mathbf{\Sigma}}_1\right)$ 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 $s_1 = s_n(\widehat{\mathbf{m}}_1, \widetilde{\mathbf{\Sigma}}_1)$ and put $\widehat{\mathbf{\Sigma}}_1 = s_n(\widehat{\mathbf{m}}_1, \widetilde{\mathbf{\Sigma}}_1)\widetilde{\mathbf{\Sigma}}_1$.

9. If $s_1 < s_0$, we set $\left(\widehat{\mathbf{m}}_0, \widehat{\mathbf{\Sigma}}_0, s_0\right) = \left(\widehat{\mathbf{m}}_1, \widehat{\mathbf{\Sigma}}_1, s_1\right)$.

Finally we choose as the EMVE, the pair $\left(\widehat{\mathbf{m}}_0, \widehat{\mathbf{\Sigma}}_0\right)$ 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 zero 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 $\mathbf{\Sigma}$ all equal to r, with $r = 0.5, 0.6, ..., 0.9$.

We introduce 10% point mass contaminations of different sizes, located at k Mahalanobis distances away from $\mathbf{0}$ ($k = 1, 2, ..., 12$), in the direction of the eigenvector of $\mathbf{\Sigma}$ 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 $N = 100$.

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

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 et al. (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.

TABLE 1 ABOUT HERE

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 on Figure 1.

FIGURE 1 ABOUT HERE

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 (i) a pairwise S-estimators applied to all pair of variables using the corresponding complete cases and (ii) fast S-estimator with Tukey's bisquare function applied to the completed data after NA's are replaced by the coordinatewise median on the available data. Details are given in Section 1 of the Supplemental 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 Supplemental Material

8 Examples

Example 1 ***Boston Housing Data:*** *Our first example uses Harrison and Rubinfeld (1978) "Boston Housing Dataset" 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.*

FIGURE 2 ABOUT HERE

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 RAD =24. Note that the median and mad of RAD are equal to 5 and 2.96, respectively. Deletion of these outliers and re-calculation 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 outliers deletion followed by re-calculation 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 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 i^{th} case. The maximum likelihood approach (EM algorithm in this case) only finds 8 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 5 non-identified points are cases 159, 171, 392, 394 and 464. Cases 159 and 171 have a large number of missing entries (5 and 4, 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 estimated before and after random missingness is introduced in the data. The averages over 20 replicates are displayed in Table 2.*

TABLE 2 ABOUT HERE

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 due to the fact that these data are highly correlated (the inverse condition number for FS and MLE scatter estimates computed on complete data are 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 7 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 *ABOUT HERE*

Figure 3 shows that the three estimates roughly agree regarding the multivariate location for the 9 variables with the exception of Race (variable number 7) where *the ERTBS and EM estimates are somewhat larger.*

FIGURE 4 ABOUT HERE

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 non-outlying cases using GSE and ERTBS follow an approximate Chi-square distribution with 9 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.

9 Supplemental Material

The Supplemental 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, 2 and 3.

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.

Copt, S., and Victoria-Feser, M. P. (2003). "Fast Algorithms for Computing High Breakdown Covariance matrices with Missing Data",Tech. Rep. 2003.04, Université de Geneve.

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

Danilov, M. (2010), "Robust Estimation or Multivariate Scatter under Non-Affine Equivariant Scenarios". Ph.D. thesis, Department of Statistics, University of British Columbia.

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

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.

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.

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.

Kenward, M. G. and Molenberghs, G. (1998). "Likelihood Based Frequentist Inference when Data are Missing at Random". *Statistical Science*, 13 236–247.

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.

Little, R. J. A. (1988), "Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values". *Journal of the Royal Statist. Society. Series C (Applied Statistics)*, 37, 23- 38.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, Wiley, New York.

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

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

Petersen, K. B., and Pedersen, M. S. (2008), *The Matrix Cookbook*, Version 20080216. Available at
http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf.

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

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

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

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.

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

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

Taskinen, S., C. Croux, A. Kankainen, E. Ollila, and, H. Oja (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.

Tatsuoka, K., and Tyler, D. (2000). “The Uniqueness of S and M-Functionals Under Non-elliptical Distributions”. *Annals of Statistics*, 28, 1219-1243.

Templ, M., Kowarik, A., and Filzmoser, P. (2011). “Iterative Stepwise Regression Imputation Using Standard and Robust Methods”. *Computational Statistics & Data Analysis*, 55, 2793 - 2806.

Tyler, D. (1987). “A Distribution-Free M-Estimator of Multivariate Scatter”. *Annals of Statistics*, 15, 234–251.

TABLES

ρ	Estimates				
	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

Table 1. Monte Carlo Study. 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.

Estimator	Percentage of missing data		
	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) $**$

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). (*) Average obtained from 19 replicates because ERTBS crashed in one occasion. (**) Average obtained from 7 replicates.

FIGURES

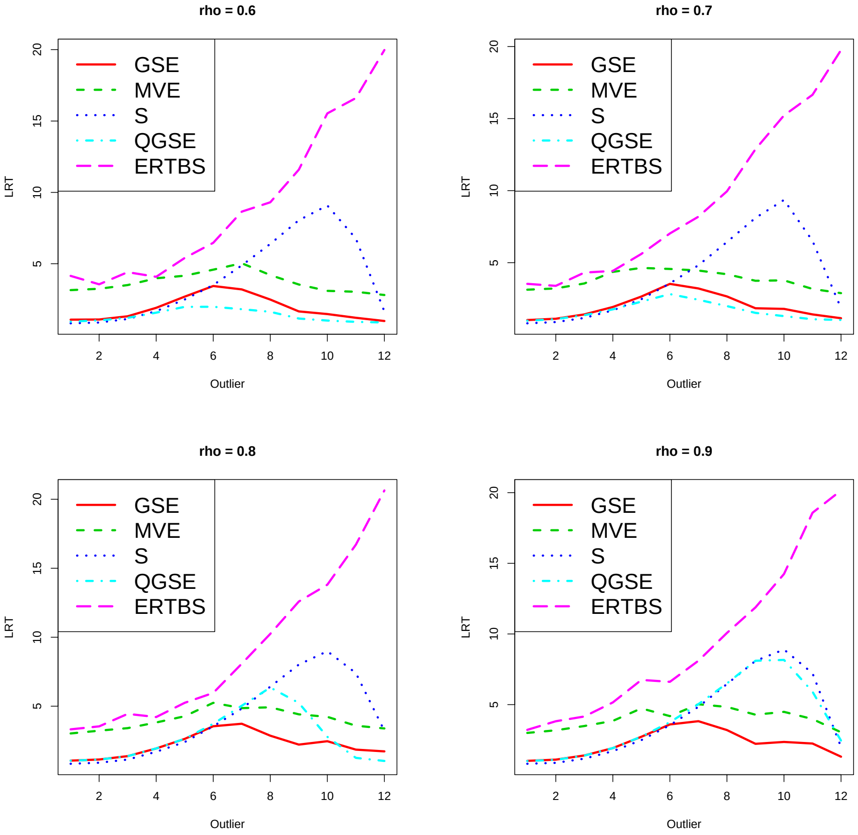

Figure 1: Monte Carlo Study. We cosider samples of size 100, of 10-dimensional observations. We plot the average LRT distances as a function of the outlier size, for different correlation structures and 10% of missing data.

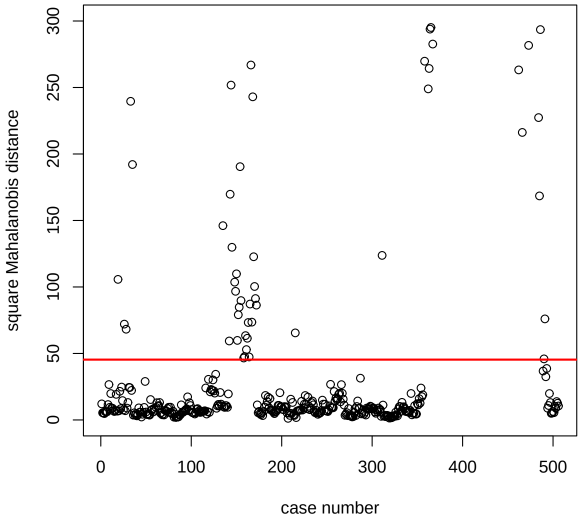

Figure 2: Boston Housing Data. Squared Mahalanobis distances using the fast S estimate of scatter matrix and multivariate location with all the data. There are in total 174 outliers. One hundred and twenty five distances exceeding the value 300 have been excluded from the plot.

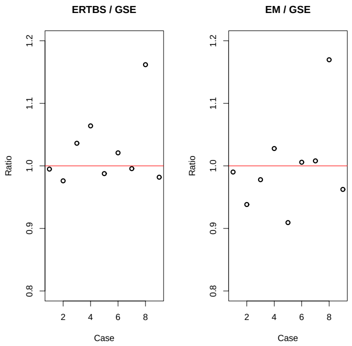

Figure 3: Wages and Hours Data. Ratio comparison of the location estimates for each single variable, taking GES as baseline. All the ratios are between 0.9 and 1.1 except for variable 7 (Race) where the ratios are a bit larger

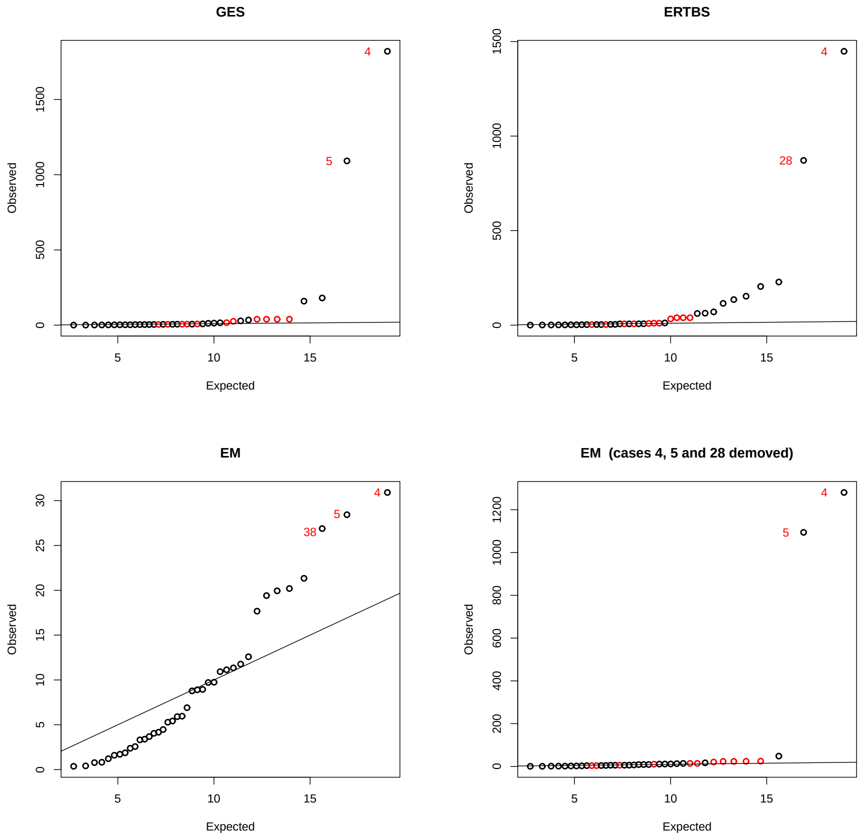

Figure 4: Wages and Hours Data. QQ plots for the Mahalnobis distances using GSE, ERTBS, EM and EM applied to the "clean data" (removing the big outliers detected by GSE and ERTBS). Only the GSE big outliers (cases 4 and 5) remain as such when EM is applied to the "clean" data. The partial Mahalanobis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

Supplemental Material for:
Robust Estimation of Multivariate Location and Scatter in the Presence of Missing Data

1 Discussion on Possible Initial Estimators

In Section 7 of the paper we use the so called quadrant correlation as initial estimate to compute a fast version of GSE which we call *quadrant correlation GSE* and denote by *QGSE*. Following the suggestion of an anonymous referee we consider the possible use of other fast initial estimators. For that purpose, we conduct a Monte Carlo experiment to compare the performances of the final GSE under 10% of contamination when the initial estimators are:

(i) Quadrant correlation, already considered in Section 7 of the paper. The final estimate in this case is called *quadrant correlation GSE* and denoted QGSE.

(ii) Fast S-estimator based on Tukey's bisquare function applied to the completed data after NA's are replaced by coordinate-wise medians of the available data. The final estimate in this case is called *median filled GSE* and denoted GSE-MF.

(iii) Pairwise S-estimator based on Tukey's bisquare function applied to all pairs of variables using the complete cases for each pair. To enforce the required positive definiteness in this case we replace the negative eigenvalues by a small positive value. The final estimate in this case is called *pairwise-S GSE* and denoted GSE-PW.

(iv) EMVE, our main proposal, naturally denoted by GSE.

Note: We included EMVE for comparison purposes. Moreover, we notice that pairwise S-estimators are not particularly fast and there are several truly fast alternatives (see e.g. Ma and Genton, 2001; Maronna and Zamar, 2002; Serneels et al., 2005; Khan et al., 2007). However, from the performance point of view, we can expect that pairwise S-estimator are superior to other fast pairwise alternatives (S-estimates are highly robust, affine equivariant, consistent and asymptotically normal). Therefore, we used the pairwise S-estimators to represent the performance of pairwise approaches.

Table 1 gives the maximum LRT distances to the true covariance matrices averaged over 100 replicates (the maximum is taken over all the considered contamination sizes). As expected, GSE has the overall best performance as we range over different correlation settings. The other alternatives perform well in some cases but poorly in other cases. Among them QGSE seems more stable as we range over different correlation structures. Based on these results and the timing results reported in the next section, we recommend the use of EMVE as default initial estimator for GSE.

%missing	ρ	Estimates			
		GSE	GSE-PW	GSE-MF	QGSE
10	0.50	5.60	60.63	5.19	2.14
	0.70	7.49	15.54	6.05	3.82
	0.90	7.95	23.93	8.53	11.26
	0.95	7.67	5.23	15.91	14.65
	0.99	6.73	14.72	40.10	31.73
20	0.50	7.35	106.52	5.69	2.21
	0.70	7.68	55.42	7.05	3.20
	0.90	7.60	18.25	12.23	10.78
	0.95	7.48	5.02	25.01	14.21
	0.99	9.68	12.40	55.70	31.06

Table 1: Maximum average LRT performance under 10% of contamination (averages taken over 100 replicates) for the final GSE estimate using our default initial estimator EMVE and other robust covariance estimates as initial estimates.

2 Timing Experiment

Here we report the *mean time* needed to compute GSE and ERTBS under normal samples of size $n = 250$, using an R code in a PC computer with an Intel Core i5-540M Processor, 2.53 GHz. We averaged over 90 realizations to include different correlation setups: 30 cases with low correlation, 30 with medium correlation and 30 with high correlation. We considered different dimensions, correlation structures and fractions of

%miss.	Estimates				
	GSE		QGSE		ERTBS
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$
10	5.35	11.01	1.05	3.53	3.08 (89)
20	5.37	12.83	1.55	3.56	13.34 (78)
30	5.73	14.08	2.98	5.22	23.63 (72)
40	6.16	15.17	5.19	14.86	46.12 (49)
50	6.69	16.09	4.89	12.85	52.47 (2)
60	7.70	17.42	5.41	13.57	NaN (0)
70	8.12	22.53	6.02	14.96	NaN (0)

Table 2. Average computing time - in seconds - evaluated using the command *system.time* in R ("User Time"), for normal samples of size 250.

As expected, the computing time increases with the dimension and percentage of missing data. The results in Table 2 were obtained with the R command *system.time ("User Time")* for $p = 10$ and 20. We were not able to evaluate ERTBS in the case of $p = 20$ because the program repeatedly crashed (itself and R) for percentage of missing larger than 10. ERTBS also often crashed when $p = 10$. The number in brackets in the ERTBS column give the number of successful evaluations of the estimates for the given sampling situations.

3 Derivation of the Consistency Constants k_j for EMVE.

It is immediate that

$$c_j = \lim_{\varepsilon \to 0} c_j\left(\varepsilon\right) = \text{Median}\left(Y\right),$$

where Y has chi-squared distribution with j degrees of freedom. It can also be shown that

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

Moreover, since

$$\rho_\varepsilon'\left(t\right) = \left\{ \begin{array}{ll} 0 & t < (1-\varepsilon) \\ \frac{1}{2\varepsilon} & (1-\varepsilon) < t < (1+\varepsilon) \\ 0 & t > (1+\varepsilon) \end{array} \right. ,$$

we have

$$\begin{aligned} k_j &= \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} E\left\{ X_1^2 I_{((1-\varepsilon)c_j,\ (1+\varepsilon)c_j)} \left(\sum_{h=1}^{j} X_h^2 \right) \right\} \\ &= \lim_{\varepsilon \to 0} \frac{\left[E\left\{ X_1^2 I_{(0\ ,\ (1+\varepsilon)\ c_j)} \left(\sum_{h=1}^{j} X_h^2 \right) \right\} - E\left\{ X_1^2 I_{(0\ ,\ (1-\varepsilon)\ c_j)} \left(\sum_{h=1}^{j} X_h^2 \right) \right\} \right]}{2\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{g\left(\varepsilon\right) - g\left(-\varepsilon\right)}{2\varepsilon} = g'\left(0\right), \qquad (1) \end{aligned}$$

where

$$\begin{aligned} g\left(\varepsilon\right) &= E\left\{ X_1^2 I_{(0\ ,\ (1+\varepsilon)\ c_j)} \left(\sum_{h=1}^{j} X_h^2 \right) \right\} \\ &= \frac{1}{j} E\left\{ \left(\sum_{h=1}^{j} X_h^2 \right) I_{(0\ ,\ (1+\varepsilon)\ c_j)} \left(\sum_{h=1}^{j} X_h^2 \right) \right\} \\ &= \frac{1}{j} E\left\{ Y_j I_{(0\ ,\ (1+\varepsilon)\ c_j)} \left(Y_j\right) \right\}, \end{aligned}$$

where Y_j has chi-square distribution with j degrees of freedom. We also have

$$\begin{aligned} g'(\varepsilon) &= \frac{1}{j} E\left\{ Y_j I_{(0\ ,\ (1+\varepsilon)\ c_j)}\left(Y_j\right)\right\} \\ &= \frac{1}{j} \frac{1}{2^{j/2}\Gamma\left(j/2\right)} \frac{d}{d\varepsilon}\left(\int_0^{c_j(1+\varepsilon)} t \times t^{(j/2)-1} \exp\left(-t/2\right) dt \right)\Bigg|_{\varepsilon=0} \\ &= \frac{1}{j 2^{j/2}\Gamma\left(j/2\right)} c_j^{(1+j/2)} \exp\left(-c_j/2\right). \end{aligned}$$

4 Proof of Theorems

4.1 Some Lemmas Needed to Prove the Theorems

Notation. In what follows $\mathbf{x}$ and $\mathbf{m}$ are random vectors of dimension p, $\mathbf{\Sigma}$ belongs to $R_+^{p\times p}$, the set of $p\times p$ positive definite matrices. Moreover, $||\mathbf{x}||$ and $||\mathbf{\Sigma}||$ will denote the L$_2$ norm of the vector $\mathbf{x}$ and of the matrix $\mathbf{\Sigma}$ respectively. Given $\mathbf{m}_0 \in R^p$, $\mathbf{\Sigma}_0 \in R_+^{p\times p}$ and $\varepsilon > 0$, $S(\mathbf{m}_0, \mathbf{\Sigma}_0, \varepsilon)$ is defined by

$$S(\mathbf{m}_0, \mathbf{\Sigma}_0, \varepsilon) = \left\{(\mathbf{m}, \mathbf{\Sigma}) : ||\mathbf{m} - \mathbf{m}_0|| \leq \varepsilon, ||\mathbf{\Sigma} - \mathbf{\Sigma}_0|| \leq \varepsilon\right\}.$$

Given $\mathbf{\Sigma} \in R_+^{p\times p}$, its eigenvalues are denoted by $\lambda_1(\mathbf{\Sigma}) \leq \lambda_2(\mathbf{\Sigma}) \leq ... \leq \lambda_p(\mathbf{\Sigma})$. Given a distribution function F on R^p and a Borel set $\mathbf{c} \subset R^p$ we denote by $P_F(A)$ the probability of the set A under F and by $E_F(g(\mathbf{x}))$ the expectation of $g(\mathbf{x})$ when $\mathbf{x}$ has distribution F. Finally, given a set A, we denote by A^C its complement.

Lemma 1 *Suppose that* $P_F(\mathbf{x}'\mathbf{c} + d = 0) < t < 1$ *for all* $\mathbf{c} \neq \mathbf{0}$ *and* $d \in R$. *Then*
(i) There exists $\delta_1 > 0$ *such that*

$$\inf_{||\mathbf{c}||=1, d\in R} P(|\mathbf{x}'\mathbf{c} + d| > \delta_1) > 1 - t + \delta_1.$$

(ii) Suppose that $\mathbf{x}_1, ..., \mathbf{x}_n$ *be i.i.d. observations with distribution* F. *Then there exists* δ_1 *such that*

$$\underline{\lim}_{n\to\infty} \inf_{||\mathbf{c}||=1, d\in R} \frac{1}{n}\sum_{i=1}^{n} I\left(|\mathbf{x}_i'\mathbf{c} + d| > \delta_1\right) > 1 - t + \delta_1 \ \ a.s..$$

Proof
(i) Let K_1 be so that

$$P_F(||\mathbf{x}_1|| \geq K_1) < t - 0.1 \tag{2}$$

and let $K = K_1 + 0.1$. Then for all $\mathbf{c}$ with $||\mathbf{c}|| = \mathbf{1}$ and $d \geq K$

$$\{|\mathbf{x}_i'\mathbf{c} + d| > 0.1\} \supset \{||\mathbf{x}|| \leq \mathbf{k}_1\}$$

and then

$$\begin{aligned} \inf_{||\mathbf{c}||=1, d>K} P(|\mathbf{x}_i'\mathbf{c} + d| > 0.1) &\geq P(||\mathbf{x}|| \leq \mathbf{k}_1) \\ &> 1 - t + 0.1. \end{aligned}$$

On the other hand according to the first part of the proof of Lemma A1.4 in the Supplementary Material of Marazzi, Vilar and Yohai (2009), there exists δ_0 such that

$$\inf_{||\mathbf{c}||=1, d>K} P(|\mathbf{x}_i'\mathbf{c} + d| > \delta_0) > 1 - t + \delta_0.$$

Then, part (i) follows taking $\delta_1 = \min(\delta_0, 0.1)$.

(ii) Let K_1 and K as in part (i). We have

$$\begin{aligned} \inf_{||\mathbf{c}||=1, |d|>K} \frac{1}{n} \sum_{i=1}^{n} I\left(|\mathbf{x}_i'\mathbf{c} + d| > 0.1\right) &\geq \frac{1}{n} \sum_{i=1}^{n} \inf_{||\mathbf{c}||=1, |d|>K} I\left(|\mathbf{x}_i'\mathbf{c} + d| > 0.1\right) \\ &\geq \frac{1}{n} \sum_{i=1}^{n} I\left(||\mathbf{x}_i|| \geq K_1\right), \end{aligned}$$

then from (2) and the Law of Large Numbers (LLN) we get

$$\underline{\lim}_{n\to\infty} \inf_{||\mathbf{c}||=1, |d|>K} \frac{1}{n} \sum_{i=1}^{n} I\left(|\mathbf{x}_i'\mathbf{c} + d| > 0.1\right) > 1 - t + 0.1 \text{ a.s..} \quad (3)$$

Lemma A1.4 in the Supplemental Material of Marazzi et al. (2009) implies that there exists δ_0 such that

$$\underline{\lim}_{n\to\infty} \inf_{||\mathbf{c}||=1, d|\leq K} \frac{1}{n} \sum_{i=1}^{n} I\left(|\mathbf{x}_i'\mathbf{c} + d| > \delta_0\right) > 1 - t + \delta_0 \text{ a.s..} \quad (4)$$

Then from (3) and (4) part (ii) follows taking $\delta_1 = \min(\delta_0, 1)$.

Lemma 2 *Suppose that $P_F(\mathbf{x}'\mathbf{c} + d = 0) < t$ for all $\mathbf{c} \neq \mathbf{0}$ and $d \in R$, and ρ satisfies Assumption A.*
(i) Then there exists $\delta_2 > 0$ such that

$$\inf_{\lambda_1(\mathbf{\Sigma})<\delta_2, \mathbf{m}\in R^p} E_F\left(\rho(\mathbf{x} - \mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{m})\right) \geq 1 - t + \delta_2.$$

(ii) Suppose that $\mathbf{x}_1, ..., \mathbf{x}_n$ be i.i.d. observations with distribution F. Then there exists δ_2 such that

$$\underline{\lim}_{n\to\infty} \inf_{\lambda_1(\mathbf{\Sigma})<\delta_2, \mathbf{m}\in R^p} \frac{1}{n} \sum_{i=1}^{n} \rho\left((\mathbf{x}_i - \mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{m})\right) \geq 1 - t + \delta_2 \textit{ a.s..}$$

Proof (i) Let $\mathbf{\Sigma} \in R_+^{p\times p}$, with eigenvalues $\lambda_1(\mathbf{\Sigma}) \leq ... \leq \lambda_p(\mathbf{\Sigma})$ and corresponding orthonormal eigenvectors $\mathbf{u}_1, ..., \mathbf{u}_p$. Then

$$\begin{aligned}(\mathbf{x}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}) &= \sum_{i=1}^{p} \frac{(\mathbf{u}_i'(\mathbf{x}-\mathbf{m}))^2}{\lambda_i(\mathbf{\Sigma})} \\ &\geq \frac{(\mathbf{u}_1'(\mathbf{x}-\mathbf{m}))^2}{\lambda_1(\mathbf{\Sigma})} \qquad (5)\end{aligned}$$

Then for any $\mathbf{m} \in R^p$ and $\mathbf{\Sigma} \in R_+^{p\times p}$ we have

$$\begin{aligned}E_F(\ \rho(\mathbf{x}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})) &\geq \rho\left(\frac{\delta_1^2}{\lambda_1(\mathbf{\Sigma})}\right) P(|\mathbf{u}_1'(\mathbf{x}-\mathbf{m})| \geq \delta_1) \\ &\geq \rho\left(\frac{\delta_1^2}{\lambda_1(\mathbf{\Sigma})}\right)(1-t+\delta_1).\end{aligned}$$

We can choose δ so that

$$\rho\left(\delta_1^2/\delta\right)(1-t+\delta_1) \geq (1-t+\delta_1/2) \qquad (6)$$

and put $\delta_2 = \min(\delta, \delta_1/2)$. Then

$$\inf_{\lambda_1(\mathbf{\Sigma})<\delta_2, \mathbf{m}\in R^p} E_F\left(\rho(\mathbf{x}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right) \geq 1-t+\delta_2,$$

and this proves part (i).

(ii) Let δ_1 as in Lemma 1, then for any $\mathbf{m} \in R^p$ and $\mathbf{\Sigma} \in R_+^{p\times p}$ (5) implies

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left((\mathbf{x}_i-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m})\right) \geq \rho\left(\frac{\delta_1^2}{\lambda_1(\mathbf{\Sigma})}\right)\frac{1}{n}\sum_{i=1}^{n} I(|\mathbf{u}_1'(\mathbf{x}_i-\mathbf{m})| \geq \delta_1).$$

Take $\delta_2 = \min(\delta, \delta_1/2)$ where δ is as in (6). Then part (ii) follows from Lemma 1 (ii).

Lemma 3 *Suppose that $P_F(\mathbf{x}'\mathbf{c}+d=0) < t$, for all $\mathbf{c} \neq \mathbf{0}$ and $d \in R$, and let ρ satisfying Assumption A. Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be i.i.d. random variables with distribution F. Then there exists $\delta_3 > 0$ such that*

$$\underline{\lim}_{n\to\infty} \inf_{\lambda_p(\mathbf{\Sigma})>1/\delta_3, |\mathbf{\Sigma}|=1, \mathbf{m}\in R^p} \frac{1}{n}\sum_{i=1}^{n} \rho\left((\mathbf{x}_i-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m})\right) \geq 1-t+\delta_3 \ \textit{a.s.}. \qquad (7)$$

Proof. If $|\mathbf{\Sigma}| = 1$ we have

$$\{\lambda_p(\mathbf{\Sigma}) > 1/\delta_3\} \subset \left\{\lambda_1(\mathbf{\Sigma}) < \delta_3^{1/(p-1)}\right\}.$$

Then (7) follows from Lemma 2 (ii) taking $\delta_3 = \delta_2^{(p-1)}$.

Lemma 4 *Given $K_1 > 0$ and $\eta > 0$, then*
(i) There exists $K_2 > 0$ such that

$$\inf_{\lambda_p(\mathbf{\Sigma})\leq K_1, ||\mathbf{m}||>K_2} E_F\left(\rho(\mathbf{x}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right) \geq 1-\eta.$$

(ii) Suppose that $\mathbf{x}_1, ..., \mathbf{x}_n$ are i.i.d. random variables with distribution F. Then, given $K_1 > 0$, there exists $K_2 > 0$ such that

$$\underline{\lim}_{n\rightarrow\infty} \inf_{\lambda_p(\mathbf{\Sigma})\leq K_1, ||\mathbf{m}||>K_2} \frac{1}{n}\sum_{i=1}^{n} \rho\left((\mathbf{x}_i-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m})\right) \geq 1-\eta \textit{ a.s..}$$

Proof. (i) We have

$$\inf_{\lambda_p(\mathbf{\Sigma})\leq K_1, ||\mathbf{m}||>K_2} (\mathbf{x}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}) \geq \inf_{||\mathbf{m}||>K_2} \frac{||\mathbf{x}-\mathbf{m}||^2}{K_1} \geq \frac{|I(||\mathbf{x}|| \leq \mathbf{K}_2/2)K_2}{2K_1}. \tag{8}$$

Let K_2 be such that $P(||\mathbf{x}|| \leq \mathbf{K}_2/2) \geq 1-(\eta/2)$ and such that $\rho(K_2/2K_1) > 1-t$ where $(1-t)(1-(\eta/2)) > 1-\eta$. Then from (8) we get

$$\begin{aligned} E_F\left(\inf_{\lambda_p(\mathbf{\Sigma})\leq K_1, ||\mathbf{m}||>K} \rho\left((\mathbf{x}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right)\right) &\geq \rho(K_2/2K_1)P_F(||\mathbf{x}|| \leq \mathbf{K}_2/2) \\ &\geq (1-t)(1-\eta/2) \\ &> 1-\eta. \end{aligned}$$

(ii) Follows immediately from (i) and the LLN.

The following Lemma is proved in Muler and Yohai (2002).

Lemma 5 *Suppose that the function $h(x,\theta)$ is defined in a subset of $C = C_1 \times C_2$, where $C_1 \subset R^h$ and $C_2 \subset R^k$ is compact. Assume that h is continuos in θ and that $E_F(h(\mathbf{x},\theta)) > a$ for all $\theta \in C_2$. If $\mathbf{x}_1, ..., \mathbf{x}_n$ is a random sample from F then*

$$\underline{\lim}_{n\rightarrow\infty} \inf_{\theta\in C} \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{x}_i,\theta) > a \textit{ a.s..}$$

Lemma 6 *Suppose that $(\mathbf{m}_0, \mathbf{\Sigma}_0)$ is the unique minimum of*

$$E_F\left(\rho((\mathbf{x}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}))\right)$$

subject to $|\mathbf{\Sigma}| = 1$ is. Let $\alpha = E\left(\rho((\mathbf{x}-\mathbf{m}_0)'\mathbf{\Sigma}_0(\mathbf{x}-\mathbf{m}_0))\right)$. Assume that ρ satisfies Assumption A and

$$P_F(\mathbf{c}'\mathbf{x}+d=0) < 1-\alpha, \tag{9}$$

for all $\mathbf{c} \neq 0$ *in* R^p *and* $d \in R$. *Let* $\mathbf{x}_1, ..., \mathbf{x}_n$ *be a random sample of* F, *then*
(i) Given $\varepsilon > 0$, *there exists* $\delta > 0$ *such that*

$$\underline{\lim}_{n\to\infty} \inf_{(\mathbf{m},\mathbf{\Sigma})\in S(\mathbf{m}_0,\mathbf{\Sigma}_0,\varepsilon)^C \cap \{|\mathbf{\Sigma}|=1\}} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{(\mathbf{x}_i-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m})}{1+\delta}\right) > \alpha+\delta \quad a.s..$$

(ii) Given $\varepsilon > 0$ *and* $\eta > 0$ *there exists* $\delta > 0$ *such that*

$$\underline{\lim}_{n\to\infty} \inf_{\mathbf{m}\in R^p, |\mathbf{\Sigma}|=1} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{(\mathbf{x}_i-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m})}{1+\delta}\right) > \alpha-\eta \quad a.s..$$

(iii) For any $\delta > 0$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{(\mathbf{x}_i-\mathbf{m}_0)'\mathbf{\Sigma}_0^{-1}(\mathbf{x}_i-\mathbf{m}_0)}{1+\delta}\right) < \alpha \quad a.s..$$

Proof. (i) Since for all $\mathbf{m} \in R^p$ and $\Sigma \in R_+^{p\times p}$ we have

$$E\left(\rho\left(\frac{(\mathbf{x}_i-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m})}{2}\right)\right) > \alpha,$$

Lemmas 2 (ii), 3 and 4 (ii) applied to $\rho(u/2)$ and (9) imply that it is possible to find a compact set $D \subset R^p \times R_+^{p\times p}$ and $\delta_1 > 0$ such that

$$\underline{\lim}_{n\to\infty} \inf_{(\mathbf{m},\mathbf{\Sigma})\in S(\mathbf{m}_0,\mathbf{\Sigma}_0,\varepsilon)^C \cap \mathbf{D}^{C\cdot} \cap \{|\mathbf{\Sigma}|=1\}} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{(\mathbf{x}_i-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m})}{2}\right) > \alpha+\delta_1 \text{ a.s..} \tag{10}$$

Let

$$\gamma(\mathbf{m},\mathbf{\Sigma},\delta) = E\left(\rho\left(\frac{\mathbf{x}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})}{1+\delta}\right)\right),$$

then, by the Dominated Convergence Theorem (DCT) $\gamma(m,\mathbf{\Sigma},\delta)$ is continuous in $m, \mathbf{\Sigma}$ and δ. Moreover

$$\gamma(\mathbf{m},\mathbf{\Sigma},0) > a \tag{11}$$

for all $(m,\mathbf{\Sigma}) \in E, S$ where E is the closure of $S(m_0,\mathbf{\Sigma}_0,\varepsilon)^{C\cdot} \cap \mathbf{D} \cap \{|\mathbf{\Sigma}|=1\}$. We will show that there exists $\delta_2 > 0$ such that

$$\gamma(\mathbf{m},\mathbf{\Sigma},\delta_2) > a+\delta_2 \tag{12}$$

for all $(m,\mathbf{\Sigma}) \in E$. Suppose that (12) does not hold. Then, there exists a sequence $(m_n,\mathbf{\Sigma}_n,\delta_n)$ with $(m_n,\mathbf{\Sigma}_n) \in E$, $\delta_n \to 0$ and $\lim \gamma(m_n,\mathbf{\Sigma}_n,\delta_n) \leq a+\delta_n$. Since E is compact, without loss of generality we can suppose that $(m_n,\mathbf{\Sigma}_n) \to (m^*,\mathbf{\Sigma}^*) \in E$. Therefore

$$\gamma(\mathbf{m}^*,\mathbf{\Sigma}^*,0) \leq a,$$

contradicting (11). Applying Lemma 5 to

$$h(\mathbf{m},\mathbf{\Sigma}) = \rho\left(\frac{\mathbf{x}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})}{1+\delta_2}\right),$$

we have

$$\underline{\lim}_{n\to\infty} \inf_{(\mathbf{m},\mathbf{\Sigma})\in S(\mathbf{m}_0,\mathbf{\Sigma}_0,\varepsilon)^C\cap\mathbf{D}\cap\{|\mathbf{\Sigma}|=1\}} \frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{(\mathbf{x}_i-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m})}{1+\delta_2}\right) > \alpha+\delta_2. \tag{13}$$

Then, part (i) follows from (10) and (13) taking $\delta=\min(1,\delta_1,\delta_2)$.

(ii) By part (i), it will be enough to show that there exists $\delta>0$ such that

$$\underline{\lim}_{n\to\infty} \inf_{(\mathbf{m},\mathbf{\Sigma})\in S(\mathbf{m}_0,\mathbf{\Sigma}_0,\varepsilon)\cap\{|\mathbf{\Sigma}|=1\}} \frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{(\mathbf{x}_i-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{m})}{1+\delta}\right) > \alpha-\eta \text{ a.s..} \tag{14}$$

The set $D=S(m_0,\mathbf{\Sigma}_0,\varepsilon)\cap\{|\mathbf{\Sigma}|=1\}$ is compact and $\gamma(m,\mathbf{\Sigma},0)>a-\eta$ for all $(m,\mathbf{\Sigma})\in D$. Then, by an argument similar to the one used to prove (12), it can be shown that there exist $\delta_1>0$ such that $\gamma(m,\mathbf{\Sigma},\delta_1)>\alpha-\eta$ for all $(m,\mathbf{\Sigma})\in D$. Put now

$$h(\mathbf{m},\mathbf{\Sigma}) = \rho\left(\frac{\mathbf{x}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})}{1+\delta_1}\right).$$

Then applying Lemma 5 to h we obtain (14) with $\delta=\delta_1$.

(iii) This part of the lemma follows from the fact that for any $\delta>0$ we have that $\gamma(m_0,\mathbf{\Sigma}_0,\delta)<a$ and the LLN.

We introduce the following notation. Let $\mathbf{u}=(u_1,...,u_p)'\in A_p$, $q=p(\mathbf{u})$, $\mathbf{A}$ be a $q\times q$ matrix and $1\le h_1<\cdots<h_q\le p$ such that $u_{h_i}=1$, $1\le i\le q$. Then we denote by $A_{(\mathbf{u})}$ the $p\times p$ matrix such that

$$\left(\mathbf{A}_{(\mathbf{u})}\right)_{ij} = \begin{cases} A_{\alpha\beta} & \text{if} \quad \alpha=h_i,\beta=h_j \\ 0 & \text{if} \quad \text{otherwise} \end{cases}.$$

In other words we expanded the $q\times q$ matrix $\mathbf{A}$ to a $p\times p$ matrix $\mathbf{A}_{(\mathbf{u})}$ completing with zeros the rows and columns corresponding to $u_i=0$. Similarly, given $\mathbf{u}=(u_1,...,u_p)'\in A_p$, $q=p(\mathbf{u})$, $\mathbf{z}\in R^q$ and $1\le h_1<\cdots<h_q\le p$ such that $u_{h_i}=1$, $1\le i\le q$, we define the expanded vector $\mathbf{z}_{(\mathbf{u})}\in R^p$ with zeros in the places where $u_i=0$, $1\le i\le p$.

The following lemma is required for the proof of Theorem 2

Lemma 7 *Let $\mathbf{\Sigma}$ be a $p\times p$ positive definite matrix, $\mathbf{m}\in R^p$ and $\mathbf{u}\in A_p$. Moreover, let $\widehat{\mathbf{x}}\left(\mathbf{u},\mathbf{x}^{(\mathbf{u})},\mathbf{m},\mathbf{\Sigma}\right)$, and $\mathbf{C}\left(\mathbf{u},\mathbf{\Sigma}\right)$ be as defined in equations (11) and (12) of the paper respectively. Then*

(a) $\mathbf{\Sigma}\left[\left(\mathbf{\Sigma}^{(\mathbf{u})}\right)^{-1}\right]_{(\mathbf{u})}\mathbf{\Sigma}=\mathbf{\Sigma}-\mathbf{C}\left(\mathbf{u},\mathbf{\Sigma}\right),$

(b) $\mathbf{\Sigma}\left[\left(\mathbf{\Sigma}^{(\mathbf{u})}\right)^{-1}\right]_{(\mathbf{u})}(\mathbf{x}-\mathbf{m})=\widehat{\mathbf{x}}\left(\mathbf{u},\mathbf{x}^{(\mathbf{u})},\mathbf{m},\mathbf{\Sigma}\right)-\mathbf{m}.$

Proof: For simplicity we can assume without loss of generality that the missingness pattern $\mathbf{u}$ is such that the first q components are observed and the last $(p-q)$ components are missing. Using equation (12) of the paper we can write

$$\begin{aligned}\boldsymbol{\Sigma}\left[\left(\boldsymbol{\Sigma}^{(\mathbf{u})}\right)^{-1}\right]_{(\mathbf{u})}\boldsymbol{\Sigma} &= \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{uu}} & \boldsymbol{\Sigma}_{\mathbf{uv}} \\ \boldsymbol{\Sigma}_{\mathbf{vu}} & \boldsymbol{\Sigma}_{\mathbf{vv}} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{uu}}^{-1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{uu}} & \boldsymbol{\Sigma}_{\mathbf{uv}} \\ \boldsymbol{\Sigma}_{\mathbf{vu}} & \boldsymbol{\Sigma}_{\mathbf{vv}} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{uu}} & \boldsymbol{\Sigma}_{\mathbf{uv}} \\ \boldsymbol{\Sigma}_{\mathbf{vu}} & \boldsymbol{\Sigma}_{\mathbf{vu}}\boldsymbol{\Sigma}_{\mathbf{uu}}^{-1}\boldsymbol{\Sigma}_{\mathbf{uv}} \end{pmatrix} \\ &= \boldsymbol{\Sigma} - \mathbf{C}\left(\mathbf{u},\boldsymbol{\Sigma}\right),\end{aligned}$$

and (a) is proved.

Note that

$$\begin{aligned}\left[\left(\boldsymbol{\Sigma}^{(\mathbf{u})}\right)^{-1}\right]_{(\mathbf{u})}\boldsymbol{\Sigma} &= \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{uu}}^{-1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{uu}} & \boldsymbol{\Sigma}_{\mathbf{uv}} \\ \boldsymbol{\Sigma}_{\mathbf{vu}} & \boldsymbol{\Sigma}_{\mathbf{vv}} \end{pmatrix} \\ &= \begin{pmatrix} I & \boldsymbol{\Sigma}_{\mathbf{uu}}^{-1}\boldsymbol{\Sigma}_{\mathbf{uv}} \\ 0 & 0 \end{pmatrix}.\end{aligned}$$

Then. using equation (11) of the paper we can write

$$\begin{aligned}(\mathbf{x}-\mathbf{m})'\left[\left(\boldsymbol{\Sigma}^{(\mathbf{u})}\right)^{-1}\right]_{(\mathbf{u})}\boldsymbol{\Sigma} &= \left(\left(\mathbf{x}^{(\mathbf{u})}-\mathbf{m}^{(\mathbf{u})}\right)',\left(\mathbf{x}^{(\mathbf{v})}-\mathbf{m}^{(\mathbf{v})}\right)'\right)\begin{pmatrix} I & \boldsymbol{\Sigma}_{\mathbf{uu}}^{-1}\boldsymbol{\Sigma}_{\mathbf{uv}} \\ 0 & 0 \end{pmatrix} \\ &= \left(\left(\mathbf{x}^{(\mathbf{u})}-\mathbf{m}^{(\mathbf{u})}\right)',\left(\mathbf{x}^{(\mathbf{u})}-\mathbf{m}^{(\mathbf{u})}\right)'\boldsymbol{\Sigma}_{\mathbf{uu}}^{-1}\boldsymbol{\Sigma}_{\mathbf{uv}}\right) \\ &= \left(\widehat{\mathbf{x}}\left(\mathbf{u},\mathbf{x}^{(\mathbf{u})},\mathbf{m},\boldsymbol{\Sigma}\right)-\mathbf{m}\right)'\end{aligned}$$

proving (b).

4.2 Proof of Theorem 1 in the Paper (Existence of GSE)

Consider a sequence $(\mathbf{m}_j,\boldsymbol{\Sigma}_j)$ with $\mathbf{m}_j\in R^p$ and $\boldsymbol{\Sigma}_j\in R_+^{p\times p}$ so that

$$\lim_{n\to\infty}S_n^*(\mathbf{m}_j,\boldsymbol{\Sigma}_j)=\inf_{\mathbf{m},\boldsymbol{\Sigma}}S_n^*(\mathbf{m},\boldsymbol{\Sigma}).$$

Without loss of generality we can assume that $\lambda_p(\boldsymbol{\Sigma}_n)=1$ and that $\boldsymbol{\Sigma}_n\to\boldsymbol{\Sigma}_0$ where $\boldsymbol{\Sigma}_0$ is positive semidefinite. We will show that $\boldsymbol{\Sigma}_0$ is nonsingular.

Take

$$t_n>c_p|\widehat{\boldsymbol{\Omega}}_n|^{1/p}\sup_{j\geq 1}S_n^*(\mathbf{m}_j,\boldsymbol{\Sigma}_j)\ <\infty.$$

Then, applying Lemma 2 (i) to the empirical distribution of $\mathbf{x}_i$, $i\in D_{\mathbf{u}_0}$ we can find δ_2 such that

$$\inf_{\lambda_1(\boldsymbol{\Sigma})<\delta_2,\mathbf{m}\in R^p}\sum_{i\in D_{\mathbf{u}_0}}\rho\left(\frac{(\mathbf{x}_i-\mathbf{m})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}_i-\mathbf{m})}{t_n}\right)\geq\ n_{\mathbf{u}_0}\left(1-\frac{\kappa_0}{n_{\mathbf{u}_0}}\right). \tag{15}$$

We will show that $\lambda_1(\boldsymbol{\Sigma}_j)>\delta_0$, where $\delta_0=\delta_2^{2-(1/p)}$ for all j. Suppose that this does not hold for j_0. Then, $\lambda_1(\boldsymbol{\Sigma}_{j_0}^*)\leq\delta_2^{2-(1/p)}/\delta_2^{(p-1)/p}=\delta_2$. Therefore using (15) we get

$$\begin{aligned}
&\sum_{i=1}^n c_{p(\mathbf{u}_i)}\rho\left(\frac{\left(\mathbf{x}_i^{(\mathbf{u}_i)}-\mathbf{m}_{j_0}^{(\mathbf{u}_i)}\right)'\left(\boldsymbol{\Sigma}_{j_0}^{*(\mathbf{u}_i)}\right)^{-1}\left(\mathbf{x}_i^{(\mathbf{u}_i)}-\mathbf{m}_{j_0}^{(\mathbf{u}_i)}\right)}{S_n^*(\mathbf{m},\boldsymbol{\Sigma})c_{p(\mathbf{u}_i)}|\widehat{\boldsymbol{\Omega}}_n|^{1/p(\mathbf{u}_i)}}\right)\\
&=\inf_{\lambda_1(\boldsymbol{\Sigma})\leq\delta_0,\mathbf{m}\in R^p}\sum_{i=1}^n c_{p(\mathbf{u}_i)}\rho\left(\frac{\left(\mathbf{x}_i^{(\mathbf{u}_i)}-\mathbf{m}^{(\mathbf{u}_i)}\right)'\left(\boldsymbol{\Sigma}^{*(\mathbf{u}_i)}\right)^{-1}\left(\mathbf{x}_i^{(\mathbf{u}_i)}-\mathbf{m}^{(\mathbf{u}_i)}\right)}{S_n^*(\mathbf{m},\boldsymbol{\Sigma})c_{p(\mathbf{u}_i)}|\widehat{\boldsymbol{\Omega}}_n|^{1/p(\mathbf{u}_i)}}\right)\\
&>c_p\inf_{\lambda_1(\boldsymbol{\Sigma})<\delta_2,\mathbf{m}\in R^p}\sum_{i\in D_{\mathbf{u}_0}}\rho\left(\frac{(\mathbf{x}_i-\mathbf{m})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}_i-\mathbf{m})}{t_n}\right)\\
&\geq n_{\mathbf{u}_0}c_p(1-\frac{\kappa_0}{n_{\mathbf{u}_0}}).
\end{aligned}$$

Then since c_k is increasing with k, according to the definition of $S_n^*(\mathbf{m}_j,\boldsymbol{\Sigma}_j)$ we get

$$\begin{aligned}
n_{\mathbf{u}_0}c_p(1-\kappa_0/n_{\mathbf{u}_0})&<\frac{1}{2}\sum_{i=1}^n c_{p(\mathbf{u}_i)}\\
&\leq\frac{nc_p}{2}
\end{aligned}$$

and then $n_{u_0}<(n/2)+\kappa_0$contradicting the assumption of the Theorem. Then $\lambda_1(\boldsymbol{\Sigma}_0)\geq\delta_0$.

Now we will prove that $\mathbf{m}_n$ is bounded. Note that since $\lambda_1(\boldsymbol{\Sigma}_j)\geq\delta_0$ and $\lambda_p(\boldsymbol{\Sigma}_j)=1$ we have $\lambda_p(\boldsymbol{\Sigma}_j^*)\leq 1/\lambda_0^{(p-1)/p}=K_1$. Then, applying Lemma 4 (i) to the empirical distribution of $\mathbf{x}_i$, $i\in D_{\mathbf{u}_0}$ we can find K_2 such that

$$\inf_{\lambda_1(\boldsymbol{\Sigma})<\delta_2,||\mathbf{m}||\geq K_2}\sum_{i\in D_{\mathbf{u}_0}}\rho\left(\frac{(\mathbf{x}_i-\mathbf{m})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}_i-\mathbf{m})}{t_n}\right)\geq .6n_{\mathbf{u}_0}. \tag{16}$$

We will show that $||\mathbf{m}_j||<K_2$ for all j. Suppose that this does not hold for j_0, then, $\lambda_1(\boldsymbol{\Sigma}_{j_0}^*)\leq\delta_2^{2-(1/p)}/\delta_2^{(p-1)/p}=\delta_2$. Therefore using (16) we get

$$\sum_{i=1}^{n} c_{p(\mathbf{u}_i)} \rho \left(\frac{\left(\mathbf{x}_i^{(\mathbf{u}_i)} - \mathbf{m}_{j_0}^{(\mathbf{u}_i)}\right)' \left(\boldsymbol{\Sigma}_{j_0}^{*(\mathbf{u}_i)}\right)^{-1} \left(\mathbf{x}_i^{(\mathbf{u}_i)} - \mathbf{m}_{j_0}^{(\mathbf{u}_i)}\right)}{S_n^*(\mathbf{m}, \boldsymbol{\Sigma}) c_{p(\mathbf{u}_i)} |\widehat{\boldsymbol{\Omega}}_n|^{1/p(\mathbf{u}_i)}} \right)$$

$$= \inf_{\lambda_1(\boldsymbol{\Sigma}) \leq \delta_0, ||\mathbf{m}|| \geq K_2} \sum_{i=1}^{n} c_{p(\mathbf{u}_i)} \rho \left(\frac{\left(\mathbf{x}_i^{(\mathbf{u}_i)} - \mathbf{m}^{(\mathbf{u}_i)}\right)' \left(\boldsymbol{\Sigma}^{*(\mathbf{u}_i)}\right)^{-1} \left(\mathbf{x}_i^{(\mathbf{u}_i)} - \mathbf{m}^{(\mathbf{u}_i)}\right)}{S_n^*(\mathbf{m}_j, \boldsymbol{\Sigma}_j) c_{p(\mathbf{u}_i)} |\widehat{\boldsymbol{\Omega}}_n|^{1/p(\mathbf{u}_i)}} \right)$$

$$> c_p \inf_{\lambda_1(\boldsymbol{\Sigma}) < \delta_2, ||\mathbf{m}|| \geq K_2} \sum_{i \in D_{\mathbf{u}_0}} \rho \left(\frac{(\mathbf{x}_i - \mathbf{m})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{m})}{t_n} \right)$$

$$\geq 0.6 n_{\mathbf{u}_0} c_p$$

and then

$$\begin{aligned} 0.6 n_{\mathbf{u}_0} c_p &< \frac{1}{2} \sum_{\mathbf{u}} n_{\mathbf{u}} c_{p(\mathbf{u})} \\ &\leq \frac{n c_p}{2} \end{aligned}$$

and this is absurd. Then $||\mathbf{m}_j|| < K_2$ for all j and this implies that there exists a subsequence of $\mathbf{m}_j$ converging to $\mathbf{m}_0$. Using the continuity of S_n^* we obtain that $S_n^*(m_0, \boldsymbol{\Sigma}_0) = \inf_{\mathbf{m}, \boldsymbol{\Sigma}} S_n^*(\mathbf{m}, \boldsymbol{\Sigma})$.

4.3 Proof of Theorem 2 in the Paper (Fix Point Estimating Equations)

The unscaled generalized S-estimator can be seen as an M-estimator if we fix the scale s at $\tilde{s}_n$. Let hand side $G(\mathbf{m}, \boldsymbol{\Sigma})$, be defined by

$$G(\mathbf{m}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} c_{p(\mathbf{u}_i)} \rho \left(\frac{d(\mathbf{x}_i^{(\mathbf{u}_i)}, \mathbf{m}^{(\mathbf{u}_i)}, \boldsymbol{\Sigma}^{(\mathbf{u}_i)})}{\tilde{s}_n c_{p(\mathbf{u}_i)}} \frac{\left|\boldsymbol{\Sigma}^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}}{\left|\widehat{\boldsymbol{\Omega}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}} \right). \tag{17}$$

Then, the unscaled generalized S-estimator $\left(\widehat{\mathbf{m}}_n, \widetilde{\boldsymbol{\Sigma}}_n\right)$ minimizes $G(\mathbf{m}, \boldsymbol{\Sigma})$ with respect to $\mathbf{m}$ and $\boldsymbol{\Sigma}$. Therefore,

$$\frac{\partial}{\partial \mathbf{m}} G(\mathbf{m}, \boldsymbol{\Sigma}) \bigg|_{\mathbf{m} = \widehat{\mathbf{m}}_n, \boldsymbol{\Sigma} = \widetilde{\boldsymbol{\Sigma}}_n} = \mathbf{0}, \quad \frac{\partial}{\partial \boldsymbol{\Sigma}} G(\mathbf{m}, \boldsymbol{\Sigma}) \bigg|_{\mathbf{m} = \widehat{\mathbf{m}}_n, \boldsymbol{\Sigma} = \widetilde{\boldsymbol{\Sigma}}_n} = 0. \tag{18}$$

To simplify the notations we set

$$Q_i(\mathbf{m}, \boldsymbol{\Sigma}) = \frac{d(\mathbf{x}_i^{(\mathbf{u}_i)}, \mathbf{m}^{(\mathbf{u}_i)}, \boldsymbol{\Sigma}^{(\mathbf{u}_i)})}{\widetilde{s}_n c_{p(\mathbf{u}_i)}} \frac{\left|\boldsymbol{\Sigma}^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}}{\left|\widehat{\boldsymbol{\Omega}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}}$$

and $\widehat{Q}_i$=$Q_i(\widehat{\mathbf{m}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$.

Consider the derivative of G with respect to $\mathbf{m}$. Using

$$\frac{\partial}{\partial \mathbf{v}} \left(\mathbf{x}-\mathbf{v}\right)' \mathbf{W} \left(\mathbf{x}-\mathbf{v}\right) = -2\mathbf{W}\left(\mathbf{x}-\mathbf{v}\right)$$

(see expression (78) from Petersen and Pedersen, 2008) we obtain

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{m}} G(\mathbf{m}, \boldsymbol{\Sigma})\bigg|_{\mathbf{m}=\widehat{\mathbf{m}}_n, \boldsymbol{\Sigma}=\widetilde{\boldsymbol{\Sigma}}_n} &= \sum_{i=1}^{n} c_{p(\mathbf{u}_i)} \rho'\left(\widehat{Q}_i\right) \frac{\partial Q_i(\mathbf{m}, \boldsymbol{\Sigma})}{\partial \mathbf{m}}\bigg|_{\mathbf{m}=\widehat{\mathbf{m}}_n, \boldsymbol{\Sigma}=\widetilde{\boldsymbol{\Sigma}}_n} \\
= \sum_{i=1}^{n} c_{p(\mathbf{u}_i)} \rho'\left(\widehat{Q}_i\right) & \frac{\left|\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}}{\tilde{s}_n c_{p(\mathbf{u}_i)} \left|\widehat{\boldsymbol{\Omega}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}} \frac{\partial d(\mathbf{x}_i^{(\mathbf{u}_i)}, \mathbf{m}^{(\mathbf{u}_i)}, \boldsymbol{\Sigma}^{(\mathbf{u}_i)})}{\partial \mathbf{m}}\Bigg|_{\mathbf{m}=\widehat{\mathbf{m}}_n, \boldsymbol{\Sigma}=\widetilde{\boldsymbol{\Sigma}}_n} \\
= \frac{2}{\tilde{s}_n} \sum_{i=1}^{n} \rho'\left(\widehat{Q}_i\right) & \frac{\left|\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}}{\left|\widehat{\boldsymbol{\Omega}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}} \left[\left(\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right)^{-1} \left(\mathbf{x}_i^{(\mathbf{u}_i)} - \widehat{\mathbf{m}}_n{}^{(\mathbf{u}_i)}\right)\right]_{(\mathbf{u}_i)} = \mathbf{0}. \qquad (19)
\end{aligned}$$

This yields the following equation

$$\sum_{i=1}^{n} w_i \left[\left(\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right)^{-1}\right]_{(\mathbf{u})} \left(\mathbf{x}_i - \widehat{\mathbf{m}}_n\right) = 0, \qquad (20)$$

where

$$\begin{aligned}
w_i &= \frac{\left|\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}}{\left|\widehat{\boldsymbol{\Omega}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}} \rho'\left(\hat{Q}_i\right) \\
&= w\left(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n, \widetilde{\boldsymbol{\Sigma}}_n\right)
\end{aligned}$$

and w is defined in equation (15) of the paper. Pre-multiplying both sides of (20) by $\widetilde{\boldsymbol{\Sigma}}_n$ and using Lemma 7 (b), we get

$$\sum_{i=1}^{n} w_i \left(\widehat{\mathbf{x}}_i - \widehat{\mathbf{m}}_n\right) = 0,$$

where $\widehat{\mathbf{x}}_i = \widehat{\mathbf{x}}\left(\mathbf{u}_i, \mathbf{x}_i^{(\mathbf{u})}, \widehat{\mathbf{m}}_n, \widetilde{\boldsymbol{\Sigma}}_n\right)$ .Finally we can write

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

proving equation (13) of the paper.

We now consider the derivative with respect to $\boldsymbol{\Sigma}$. Using that

$$\frac{\partial \mathbf{a}' \mathbf{W}^{-1} \mathbf{b}}{\partial \mathbf{W}} = -\left(\mathbf{W}'\right)^{-1} \mathbf{ab}^{'} \left(\mathbf{W}'\right)^{-1}$$

(see expression (55) from Petersen and Pedersen, 2008) and

$$\frac{\partial \log |\mathbf{W}|}{\partial \mathbf{W}} = (\mathbf{W}')^{-1}$$

we obtain

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} G(\mathbf{m}, \boldsymbol{\Sigma})\Big|_{\mathbf{m}=\widehat{\mathbf{m}}_n, \boldsymbol{\Sigma}=\widetilde{\boldsymbol{\Sigma}}_n} = \sum_{i=1}^{n} c_{p(\mathbf{u}_i)} \rho' \left(\widehat{Q}_i\right) \frac{\partial Q_i(\mathbf{m}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}}\Big|_{\mathbf{m}=\widehat{\mathbf{m}}_n, \boldsymbol{\Sigma}=\widetilde{\boldsymbol{\Sigma}}_n}$$

$$= \sum_{i=1}^{n} \frac{c_{p(\mathbf{u}_i)} \rho' \left(\widehat{Q}_i\right)}{\hat{s}_n c_{p(\mathbf{u}_i)} \left|\widehat{\boldsymbol{\Omega}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}} \left[\frac{\partial d(\mathbf{x}_i^{(\mathbf{u}_i)}, \mathbf{m}^{(\mathbf{u}_i)}, \boldsymbol{\Sigma}^{(\mathbf{u}_i)})}{\partial \boldsymbol{\Sigma}}\Big|_{\mathbf{m}=\widehat{\mathbf{m}}_n, \boldsymbol{\Sigma}=\widetilde{\boldsymbol{\Sigma}}_n} \left|\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)} \right.$$

$$\left. + \, d(\mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n{}^{(\mathbf{u}_i)}, \widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}) \times \frac{\partial \left|\boldsymbol{\Sigma}^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}}{\partial \boldsymbol{\Sigma}}\Big|_{\mathbf{m}=\widehat{\mathbf{m}}_n, \boldsymbol{\Sigma}=\widetilde{\boldsymbol{\Sigma}}_n} \right]$$

$$= \frac{1}{\hat{s}_n} \sum_{i=1}^{n} w_i \left[- \left[\left(\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right)^{-1}\right]_{(\mathbf{u}_i)} (\mathbf{x}_i - \widehat{\mathbf{m}}_n)(\mathbf{x}_i - \widehat{\mathbf{m}}_n)' \left[\left(\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right)^{-1}\right]_{(\mathbf{u}_i)} \right.$$

$$\left. + \, w_i^* \left[\left(\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right)^{-1}\right]_{(\mathbf{u}_i)} \right] = 0,$$

where

$$w_i^* = \frac{d\left(\mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n{}^{(\mathbf{u}_i)}, \widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right)}{p(\mathbf{u}_i)}$$
$$= w^*(\mathbf{u}, \mathbf{x}_i^{(\mathbf{u}_i)}, \widehat{\mathbf{m}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$$

and w^* is defined in equation (16) of the paper. Note that w_i^* can be thought of as a secondary weight for the i^{th} case. Equivalently, we can write

$$\sum_{i=1}^{n} w_i \left[\left(\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right)^{-1}\right]_{(\mathbf{u}_i)} (\mathbf{x}_i - \widehat{\mathbf{m}}_n)(\mathbf{x}_i - \widehat{\mathbf{m}}_n)' \left[\left(\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right)^{-1}\right]_{(\mathbf{u}_i)}$$
$$= \sum_{i=1}^{n} w_i w_i^* \left[\left(\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u}_i)}\right)^{-1}\right]_{(\mathbf{u}_i)}. \qquad (21)$$

Pre and post multiplying both sides of (21) by $\widetilde{\boldsymbol{\Sigma}}_n$ and using Lemma 7 we get

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

where $\mathbf{C}_i = \mathbf{C}\left(\mathbf{u}_i, \widehat{\boldsymbol{\Sigma}}_n\right)$ or equivalently

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

proving equation (14) of the paper.

4.4 Proof of Theorem 3 in the Paper (Consistency of GSE)

(i) Let $A_p = \{\mathbf{u} : (u_1, ..., u_p)', u_i \in \{0,1\}\}$ and for any $\mathbf{u} \in A_p$, let $\lambda_{\mathbf{u}} = P(\mathbf{u}_i = \mathbf{u})$ and $A_p^* = \{\mathbf{u} \in A_p : \lambda_{\mathbf{u}} > 0\}$. For $\mathbf{u} \in A_p^*$ let $D_{n,\mathbf{u}} = \{i : \mathbf{u}_i = \mathbf{u}, 1 \leq i \leq n\}$. We start showing that

$$(\widehat{\mathbf{m}}_n^{(\mathbf{u})}, \widehat{\mathbf{\Sigma}}_n^{*(\mathbf{u})}) \to (\mathbf{m}_0^{(\mathbf{u})}, \mathbf{\Sigma}_0^{*(\mathbf{u})}) \text{ a.s..} \tag{22}$$

for all $\mathbf{u} \in A_p^*$.

Define γ

$$\gamma = \frac{\min_{\mathbf{u} \in A_p^*} \lambda_{\mathbf{u}} c_{\mathbf{u}}}{2 \sum_{\mathbf{u} \in A_p^*} \lambda_{\mathbf{u}} c_{\mathbf{u}}},$$

then

$$\min_{\mathbf{u} \in A_p^*} \lambda_{\mathbf{u}} c_{\mathbf{u}} - \gamma \sum_{\mathbf{u} \in A_p^*} \lambda_{\mathbf{u}} c_{\mathbf{u}} \geq \frac{\gamma}{2}. \tag{23}$$

Define s_0 by

$$\sum_{\mathbf{u} \in A_p^*} \lambda_{\mathbf{u}} c_{\mathbf{u}} E\rho \left(\frac{(\mathbf{x}_i - \mathbf{m}_0^{(\mathbf{u})})' \left(\mathbf{\Sigma}_0^{*(\mathbf{u})} \right)^{-1} (\mathbf{x}_i - \mathbf{m_0}^{(\mathbf{u})})}{s_0 c_{p(\mathbf{u})})} \right) = \frac{1}{2} \sum_{\mathbf{u} \in A_p^*} \lambda_{\mathbf{u}} c_{p(\mathbf{u})} \tag{24}$$

and $a_{\mathbf{u}}$ by

$$a_{\mathbf{u}} = E_F \left(\rho \left(\frac{(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}_0)' \left(\mathbf{\Sigma}_0^{*(\mathbf{u})} \right)^{-1} (\mathbf{x}^{(\mathbf{u})} - \mathbf{m}_0)}{s_0 c_u |\mathbf{\Omega}_0^{(\mathbf{u})}|^{1/p(\mathbf{u})}} \right) \right).$$

Note that according to (24) we have

$$\sum_{\mathbf{u} \in A_p^*} \lambda_{\mathbf{u}} c_{\mathbf{u}} a_{\mathbf{u}} = \frac{1}{2} \sum_{\mathbf{u} \in A_p^*} \lambda_{\mathbf{u}} c_{\mathbf{u}}. \tag{25}$$

Since A_p^* is finite and Assumption C holds, applying part (i) of Lemma 6 to the function $\rho \left(v / \left(s_0 c_{p(\mathbf{u})} |\mathbf{\Omega}_0^{(\mathbf{u})}|^{1/p} \right) \right)$ we can find δ_1 such that for all $\mathbf{u} \in A_p^*$ we have

$$\underline{\lim} \inf_{\left\{ (\mathbf{m}, \mathbf{\Sigma}) : (\mathbf{m}^{(\mathbf{u})}, \mathbf{\Sigma}^{*(\mathbf{u})}) \notin S(\mathbf{m}_0^{(\mathbf{u})}, \mathbf{\Sigma}_0^{*(\mathbf{u})}, \varepsilon) \right\}} V_{n,\mathbf{u}} \geq a_{\mathbf{u}} + \delta_1, \tag{26}$$

where

$$V_{n,\mathbf{u}} = \frac{1}{n_{\mathbf{u}}} \sum_{i \in D_{n,\mathbf{u}}} \rho \left(\frac{(\mathbf{x}_i^{(\mathbf{u})} - \mathbf{m}^{(\mathbf{u})})' \left(\mathbf{\Sigma}^{*(\mathbf{u})} \right)^{-1} (\mathbf{x}_i^{(\mathbf{u})} - \mathbf{m}^{(\mathbf{u})})}{(1 + \delta_1) s_0 c_{p(\mathbf{u}_0)} |\mathbf{\Omega}_0^{(\mathbf{u})}|^{1/p}} \right).$$

Then, using part (ii) of Lemma 6 we can find $\delta_2 \leq \delta_1$ such that for all $\mathbf{u} \in A_p^*$ we have

$$\begin{aligned}
&\underline{\lim} \inf_{\mathbf{m}\in R^p, \boldsymbol{\Sigma}\in R_+^{p\times p}} \frac{1}{n_{\mathbf{u}}} \sum_{i\in D_{n,\mathbf{u}}} \rho\left(\frac{(\mathbf{x}_i^{(\mathbf{u})} - \mathbf{m}^{(\mathbf{u})})' \left(\boldsymbol{\Sigma}^{*(\mathbf{u})}\right)^{-1} (\mathbf{x}_i^{(\mathbf{u})} - \mathbf{m}^{(\mathbf{u})})}{(1+\delta_2)s_0 c_{\mathbf{u}} |\boldsymbol{\Omega}_0^{(\mathbf{u})}|^{1/p(\mathbf{u})}} \right) && (27)\\
&\geq a_{\mathbf{u}} - \delta_1 \gamma c_{p(\mathbf{u})} \lambda_{\mathbf{u}}. && (28)
\end{aligned}$$

Take any $\mathbf{u}_0 \in A_p^*$. We will prove that $(\widehat{\mathbf{m}}_n^{(\mathbf{u}_0)}, \widehat{\boldsymbol{\Sigma}}_n^{*(\mathbf{u})}) \to (m_0^{(\mathbf{u}_0)}, \boldsymbol{\Sigma}_0^{*(\mathbf{u}_0)})$. Let $\varepsilon > 0$ and

$$M_n = \left\{ (\mathbf{m}, \boldsymbol{\Sigma}) : (\mathbf{m}^{(\mathbf{u}_0)}, \boldsymbol{\Sigma}^{*(\mathbf{u}_0)}) \notin S(\mathbf{m}_0^{(\mathbf{u}_0)}, \boldsymbol{\Sigma}_0^{*(\mathbf{u}_0)}, \varepsilon) \right\}.$$

Since $|\widehat{\boldsymbol{\Omega}}_n^{(\mathbf{u})}| \to |\boldsymbol{\Omega}_0^{(\mathbf{u})}|$ a.s. for any $\mathbf{u} \in A_p$, using (25),(26), (27), (23) and the fact that $\delta_2 \leq \delta_1$we obtain

$$\begin{aligned}
&\underline{\lim}_{n\to\infty} \inf_{(\mathbf{m},\boldsymbol{\Sigma})\in M_n} \frac{1}{n} \sum_{i=1}^{n} c_{p(\mathbf{u}_i)} \rho\left(\frac{\left(\mathbf{x}_i^{(\mathbf{u}_i)} - \mathbf{m}^{(\mathbf{u}_i)}\right)' \left(\boldsymbol{\Sigma}^{*(\mathbf{u}_i)}\right)^{-1} \left(\mathbf{x}_i^{(\mathbf{u}_i)} - \mathbf{m}^{(\mathbf{u}_i)}\right)}{(1+\delta_2/2)s_0 c_{p(\mathbf{u}_i)} |\widehat{\boldsymbol{\Omega}}_n^{(\mathbf{u}_i)}|^{1/p(\mathbf{u}_i)}} \right)\\
&\geq \lambda_{\mathbf{u}_0} c_{p(\mathbf{u}_0)} \underline{\lim}_{n\to\infty} \inf_{(\mathbf{m},\boldsymbol{\Sigma})\in M_n} \frac{1}{n_{\mathbf{u}_0}} \sum_{i\in D_{n,\mathbf{u}_0}} \rho\left(\frac{\left(\mathbf{x}_i^{(\mathbf{u}_0)} - \mathbf{m}^{(\mathbf{u}_0)}\right)' \left(\boldsymbol{\Sigma}^{*(\mathbf{u}_0)}\right)^{-1} \left(\mathbf{x}_i^{(\mathbf{u}_0)} - \mathbf{m}^{(\mathbf{u}_0)}\right)}{(1+\delta_2)s_0 c_{p(\mathbf{u}_0)} |\boldsymbol{\Omega}_0^{(\mathbf{u}_0)}|^{1/p(\mathbf{u}_0)}} \right)\\
&+ \sum_{\mathbf{u}\in A_p^*-\mathbf{u}_0} \lambda_{\mathbf{u}} c_{p(\mathbf{u})} \underline{\lim}_{n\to\infty} \inf_{(\mathbf{m},\boldsymbol{\Sigma})} \frac{1}{n_{\mathbf{u}}} \sum_{i\in D_{n,\mathbf{u}}} \rho\left(\frac{\left(\mathbf{x}_i^{(\mathbf{u})} - \mathbf{m}^{(\mathbf{u})}\right)' \left(\boldsymbol{\Sigma}^{*(\mathbf{u})}\right)^{-1} \left(\mathbf{x}_i^{(\mathbf{u})} - \mathbf{m}^{(\mathbf{u})}\right)}{(1+\delta_2)s_0 c_{p(\mathbf{u})} |\boldsymbol{\Omega}_0^{(\mathbf{u})}|^{1/p(\mathbf{u})}} \right)\\
&\geq \lambda_{\mathbf{u}_0} c_{p(\mathbf{u}_0)} (a_{\mathbf{u}_0} + \delta_1) + \sum_{\mathbf{u}\in A_p-\mathbf{u}_0} \lambda_{\mathbf{u}} c_{p(\mathbf{u})} (a_{\mathbf{u}} - \delta_1 \gamma c_{p(\mathbf{u})} \lambda_{\mathbf{u}})\\
&\geq \sum_{\mathbf{u}\in A_p^*} a_{\mathbf{u}} \lambda_{\mathbf{u}} c_{p(\mathbf{u})} + \delta_1 \left(\lambda_{\mathbf{u}_0} c_{p(\mathbf{u}_0)} - \gamma \sum_{\mathbf{u}\in A_p^*} c_{p(\mathbf{u})} \lambda_{\mathbf{u}}) \right)\\
&= \sum_{\mathbf{u}\in A_p^*} a_{\mathbf{u}} \lambda_{\mathbf{u}} c_{p(\mathbf{u})} + \frac{\delta_1 \gamma}{2}\\
&= \frac{1}{2} \sum_{\mathbf{u}\in A_p^*} \lambda_{\mathbf{u}} c_{p(\mathbf{u})} + \frac{\delta_1 \gamma}{2}
\end{aligned}$$

Then

$$\underline{\lim}_{n\to\infty} \inf_{\left\{(\mathbf{m},\boldsymbol{\Sigma}):(\mathbf{m}^{(\mathbf{u}_0)}, \boldsymbol{\Sigma}^{*(\mathbf{u}_0)}) \notin S(\mathbf{m}_0^{(\mathbf{u}_0)}, \boldsymbol{\Sigma}_0^{*(\mathbf{u}_0)}, \varepsilon)\right\}} S_n^*(\mathbf{m}, \boldsymbol{\Sigma}) \geq s_0 + \frac{\delta_2}{2} \qquad (29)$$

Applying part (iii) of Lemma 6 to the function $\rho\left(v/\left(s_0 c_{p(\mathbf{u})} |\boldsymbol{\Omega}_0^{(\mathbf{u})}|^{1/p}\right)\right)$ we obtain

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} c_{p(\mathbf{u}_i)}\rho\left(\frac{(\mathbf{x}_i^{(\mathbf{u}_i)}-\mathbf{m}_0^{(\mathbf{u}_i)})'\left(\boldsymbol{\Sigma}_0^{*(\mathbf{u}_i)}\right)^{-1}(\mathbf{x}_i^{(\mathbf{u}_i)}-\mathbf{m}_0^{(\mathbf{u}_i)})}{(s_0+\delta_2/2)c_{p(\mathbf{u}_i)}|\widehat{\boldsymbol{\Omega}}_n^{(\mathbf{u}_i)}|^{1/p(\mathbf{u}_i)}}\right)\\
&\le \sum_{\mathbf{u}\in A_p^*}\lambda_{\mathbf{u}}c_{p(\mathbf{u})}\lim_{n\to\infty}\frac{1}{n_{\mathbf{u}}}\sum_{i\in D_{n,\mathbf{u}}}\rho\left(\frac{(\mathbf{x}_i-\mathbf{m_0}^{(\mathbf{u})})'\left(\boldsymbol{\Sigma}_0^{*(\mathbf{u})}\right)^{-1}(\mathbf{x}_i-\mathbf{m_0}^{(\mathbf{u})})}{(s_{0+}\delta_2/4)c_{p(\mathbf{u})}|\boldsymbol{\Omega}_0^{(\mathbf{u})}|^{1/p(\mathbf{u})}}\right)\\
&< \sum_{\mathbf{u}\in A_p^*}\lambda_{\mathbf{u}}c_{p(\mathbf{u})}a_{\mathbf{u}},
\end{aligned}$$

and therefore,

$$\lim_{n\to\infty} S_n^*(\mathbf{m}_0,\boldsymbol{\Sigma}_0) < s_0+\frac{\delta_2}{2}. \tag{30}$$

Using (29) and (30) we obtain (22). Then, Assumption B implies that $\widehat{\mathbf{m}}_n \to m_0$ a.s. proving (i)

(ii) Put $\phi_n^{(\mathbf{u})} = 1/|\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u})}|^{1/p(\mathbf{u})}$ and $\phi_0^{(\mathbf{u})} = 1/|\boldsymbol{\Sigma}_0^{(\mathbf{u})}|^{1/p(\mathbf{u})}$. Then (22) implies

$$\phi_n^{(\mathbf{u})}\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u})} \to \phi_0^{(\mathbf{u})}\boldsymbol{\Sigma}_0^{(\mathbf{u})} \text{ a.s.}$$

and putting $\eta_n^{(\mathbf{u})} = \phi_n^{(\mathbf{u})}/\phi_0^{(\mathbf{u})}$ we get

$$\eta_n^{(\mathbf{u})}\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u})} \to \boldsymbol{\Sigma}_0^{(\mathbf{u})}\text{a.s..}$$

We will show that for all $\mathbf{u}_1 \in A_p^*$ and $\mathbf{u}_2 \in A_p^*$ we have

$$\eta_n^{(\mathbf{u}_1)}/\eta_n^{(\mathbf{u}_2)} \to 1 \text{ a.s..}$$

Take i such that $u_{1i} = 1$ and j such that $u_{2j} = 1$. By Assumption B there exists $\mathbf{u}_3$ such that $u_{3i} = 1$ and $u_{3j} = 1$. Put $\widetilde{\boldsymbol{\Sigma}}_n = (\widetilde{\sigma}_{n,ij})$ and $\boldsymbol{\Sigma}_0 = (\sigma_{0,ij})$. Then

$$\begin{aligned}
\eta_n^{(\mathbf{u}_1)}\widetilde{\sigma}_{n,.ii} &\to \sigma_{0,ii} \text{ a.s.,}\\
\eta_n^{(\mathbf{u}_2)}\widetilde{\sigma}_{n.jj} &\to \sigma_{0,jj} \text{ a.s.,}\\
\eta_n^{(\mathbf{u}_3)}\widetilde{\sigma}_{n,ii} &\to \sigma_{0,ii} \text{ a.s.,}\\
\eta_n^{(\mathbf{u}_3)}\widetilde{\sigma}_{n,jj} &\to \sigma_{0,jj} \text{ a.s.}
\end{aligned}$$

Then $\eta_n^{(\mathbf{u}_1)}/\eta_n^{(\mathbf{u}_3)} \to 1$ a.s. and $\eta_n^{(\mathbf{u}_2)}/\eta_n^{(\mathbf{u}_3)} \to 1$ a.s., and this implies $\eta_n^{(\mathbf{u}_1)}/\eta_n^{(\mathbf{u}_2)} \to 1$ a.s. Then, if we take a fixed $\mathbf{u}_0 \in A_p^*$, for all $\mathbf{u} \in A_p^*$ we have $\eta_n^{(\mathbf{u}_0)}\widetilde{\boldsymbol{\Sigma}}_n^{(\mathbf{u})} \to \boldsymbol{\Sigma}_0^{(\mathbf{u})}$ a.s.. Since Assumption B is satisfied, this implies

$$\eta_n^{(\mathbf{u}_0)}\widetilde{\boldsymbol{\Sigma}}_n \to \boldsymbol{\Sigma}_0 \text{ a.s..} \tag{31}$$

We will show now that $\widehat{\boldsymbol{\Sigma}}_n \to t_0\boldsymbol{\Sigma}_0$ a.s.. The scale $\widehat{s}_n$ is defined by

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

Put $\overline{\boldsymbol{\Sigma}}_n = \widetilde{\boldsymbol{\Sigma}}_n/\eta_n^{(\mathbf{u}_0)}$ and $\widehat{t}_n = \widehat{s}_n/\eta_n^{(\mathbf{u}_0)}$. Then we can write

$$\sum_{i=1}^{n} c_{p(\mathbf{u}_i)}\rho\left(\frac{(\mathbf{x}_i^{(\mathbf{u}_i)}-\widehat{\mathbf{m}}_n^{(\mathbf{u}_i)})'\left(\overline{\boldsymbol{\Sigma}}^{(\mathbf{u}_i)}\right)^{-1}(\mathbf{x}_i^{(\mathbf{u}_i)}-\widehat{\mathbf{m}}_n^{(\mathbf{u}_i)})}{\widehat{t}_n c_{p(\mathbf{u}_i})}\right)=\frac{1}{2n}\sum_{i=1}^{n} c_{p(\mathbf{u}_i)},$$

or equivalently

$$\sum_{\mathbf{u}\in A_p^*}\lambda_{n,\mathbf{u}}c_{p(\mathbf{u})}\frac{1}{n_{\mathbf{u}}}\sum_{\mathbf{i}\in D_{n,\mathbf{u}}}\rho\left(\frac{\left(\mathbf{x}_i^{(\mathbf{u})}-\widehat{\mathbf{m}}_n^{(\mathbf{u})}\right)'\left(\overline{\boldsymbol{\Sigma}}^{(\mathbf{u})}\right)^{-1}\left(\mathbf{x}_i^{(\mathbf{u})}-\widehat{\mathbf{m}}_n^{(\mathbf{u})}\right)}{\widehat{t}_n c_{p(\mathbf{u})})}\right)=\frac{1}{2}\sum_{\mathbf{u}\in A_p^*}\lambda_{n,\mathbf{u}}c_{p(\mathbf{u})}. \tag{32}$$

Note that (31) implies

$$\lim_{n\to\infty}\overline{\boldsymbol{\Sigma}}_n=\boldsymbol{\Sigma}_0 \text{ a.s..}$$

In order to prove (ii) it is enough to show that

$$\lim_{n\to\infty}\widehat{t}_n\to t_0 \text{ a.s.,} \tag{33}$$

where t_0 is defined in equation (17) of the paper.

Using Lemma A3.1 of the Supplemental Material of Marazzi et al. (2009) we have that for any $\varepsilon>0$ and $\mathbf{u}\in A_p^*$

$$\lim_{n\to\infty}\frac{1}{n_{\mathbf{u}}}\sum_{\mathbf{i}\in D_{n,\mathbf{u}}}\rho\left(\frac{\left(\mathbf{x}_i^{(\mathbf{u})}-\widehat{\mathbf{m}}_n^{(\mathbf{u})}\right)'\left(\overline{\boldsymbol{\Sigma}}^{(\mathbf{u})}\right)^{-1}\left(\mathbf{x}_i-\widehat{\mathbf{m}}_n^{(\mathbf{u})}\right)}{(t_0+\varepsilon)c_{p(\mathbf{u})})}\right)$$
$$=E\rho\left(\frac{\left(\mathbf{x}^{(\mathbf{u})}-\mathbf{m}_0^{(\mathbf{u})}\right)'\left(\boldsymbol{\Sigma}_0^{(\mathbf{u})}\right)^{-1}\left(\mathbf{x}^{(\mathbf{u})}-\mathbf{m}_0^{(\mathbf{u})}\right)}{(t_0+\varepsilon)c_{p(\mathbf{u})})}\right),$$

and putting in left hand side of (32) $\widehat{t}_n = t_0 + \varepsilon$ we get

$$\lim_{n\to\infty} \sum_{\mathbf{u}\in A_p^*} \lambda_{n,\mathbf{u}} c_{p(\mathbf{u})} \frac{1}{n_{\mathbf{u}}} \sum_{\mathbf{i}\in D_{n,\mathbf{u}}} \rho \left(\frac{\left(\mathbf{x}_i^{(\mathbf{u})} - \widehat{\mathbf{m}}_n^{(\mathbf{u})}\right)' \left(\overline{\boldsymbol{\Sigma}}^{(\mathbf{u})}\right)^{-1} \left(\mathbf{x}_i^{(\mathbf{u})} - \widehat{\mathbf{m}}_n^{(\mathbf{u})}\right)}{(t_0+\varepsilon) c_{p(\mathbf{u})}} \right)$$

$$= \frac{1}{2} \sum_{\mathbf{u}\in A_p^*} \lambda_{\mathbf{u}} c_{p(\mathbf{u})} E\rho \left(\frac{\left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}_0^{(\mathbf{u})}\right)' \left(\boldsymbol{\Sigma}_0^{(\mathbf{u})}\right)^{-1} \left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}_0^{(\mathbf{u})}\right)}{(t_0+\varepsilon) c_{p(\mathbf{u})}} \right)$$

$$< \frac{1}{2} \sum_{\mathbf{u}\in A_p^*} \lambda_{\mathbf{u}} c_{p(\mathbf{u})} E\rho \left(\frac{\left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}_0^{(\mathbf{u})}\right)' \left(\boldsymbol{\Sigma}_0^{(\mathbf{u})}\right)^{-1} \left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}_0^{(\mathbf{u})}\right)}{t_0 c_{p(\mathbf{u})}} \right).$$

This shows that $\overline{\lim}\, \widehat{t}_n \leq t_0 + \varepsilon$ a.s. Similarly it can be proved that $\underline{\lim}_{n\to\infty} \widehat{t}_n \geq t_0 - \varepsilon$ a.s. This proves (33) and therefore

$$\lim_{n\to\infty} \widehat{\boldsymbol{\Sigma}}_n = \lim_{n\to\infty} \widehat{s}_n \widetilde{\boldsymbol{\Sigma}}_n = \lim_{n\to\infty} \widehat{t}_n \lim_{n\to\infty} \overline{\boldsymbol{\Sigma}}_n = t_0 \boldsymbol{\Sigma}_0$$

proving (ii).

(iii) In the case that the $\mathbf{x}_i$'s have normal distribution, we have

$$E\rho \left(\frac{\left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}_0^{(\mathbf{u})}\right)' \left(\boldsymbol{\Sigma}_0^{(\mathbf{u})}\right)^{-1} \left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}_0^{(\mathbf{u})}\right)}{c_{p(\mathbf{u})}} \right) = \frac{1}{2},$$

and therefore

$$\sum_{\mathbf{u}\in A_p^*} \lambda_{\mathbf{u}} E\rho \left(\frac{\left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}_0^{(\mathbf{u})}\right)' \left(\boldsymbol{\Sigma}_0^{(\mathbf{u})}\right)^{-1} \left(\mathbf{x}^{(\mathbf{u})} - \mathbf{m}_0^{(\mathbf{u})}\right)}{c_{p(\mathbf{u})}} \right) = \frac{1}{2} \sum_{\mathbf{u}\in A_p^*} \lambda_{\mathbf{u}} c_{p(\mathbf{u})}$$

proving that $t_0 = 1$.

5 References

Khan, J. A., Van Aelst, S., and Zamar, R. H. (2007). "Robust Linear Model Selection Based on Least Angle Regression," *Journal of the American Statistical Association*, 102, 1289-1299.

Ma, Y., and Genton, M. G. (2001). "Highly Robust Estimation of Dispersion Matrices," *Journal of Multivariate Analysis*, 78, 11-36.

Marazzi, A., Villar, A. J., and Yohai, V. J. (2009). "Robust Response Transformations Based on Optimal Prediction, " *Journal of the American Statistical Association,* 104, 360-370.

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

Muler, N., and Yohai, V. J. (2002). "Robust Estimates for ARCH Processes," *Journal of Time Series,* 23, 341-375.

Serneels, S., Croux, C., Filzmoser, P., and Van Espen, P. J. (2005). "Partial Robust M-Regression " *Chemometrics and Intelligent Laboratory Systems*, 79, 55-64.