Fast algorithms for computing high breakdown covariance matrices with missing data

Samuel Copt and Maria-Pia Victoria-Feser

No 2003.04

Cahiers du département d'économétrie
Faculté des sciences économiques et sociales
Université de Genève

Août 2003

Département d'économétrie
Université de Genève, 40 Boulevard du Pont-d'Arve, CH -1211 Genève 4
http://www.unige.ch/ses/metri/

Abstract

Robust estimation of covariance matrices when some of the data at hand are missing is an important problem. It has been studied by Little and Smith (1987) and more recently by Cheng and Victoria-Feser (2002). The latter propose the use of high breakdown estimators and so-called hybrid algorithms (see e.g. Woodruff and Rocke 1994). In particular, the minimum volume ellipsoid of Rousseeuw (1984) is adapted to the case of missing data. To compute it, they use (a modified version of) the forward search algorithm (see e.g. Atkinson 1994). In this paper, we propose to use instead a modification of the C-step algorithm proposed by Rousseeuw and Van Driessen (1999) which is actually a lot faster. We also adapt the orthogonalized Gnanadesikan-Kettering (OGK) estimator proposed by Maronna and Zamar (2002) to the case of missing data and use it as a starting point for an adapted S-estimator. Moreover, we conduct a simulation study to compare different robust estimators in terms of their efficiency and breakdown and use them to analyse real datasets.

Keywords:C-step algorithm, minimum volume ellipsoid, outliers, robust statistics, S-estimators, orthogonalized Gnanadesikan-Kettering robust estimator.

1 Introduction

Since the original works of Tukey (1960), Huber (1964) and Hampel (1971), robust statistics are nowadays widely used, and new or improved tools are continuously proposed. In this paper, we focus on the robust estimation of location and scatter of a multivariate normal distribution with missing data

The classical estimator of the covariance matrix, namely the maximum likelihood estimator (MLE) is very sensitive to model deviations. Indeed, one shouldn't forget that the common postulated models are only approximation of the reality. For example, there might be gross error in the data. Such errors appear as points lying very far from the core of the data and are extremely dangerous for classical statistical methods. It is therefore important to develop and use robust estimators for the mean and covariance of multivariate data since the latter can then be used in other analyses such as factor analysis. For example, Yuan and Bentler (1998) showed that the influence of such data on covariance structure analysis is limited if the covariance matrix is robustly estimated.

The aim of robust statistics is thus to provide tools not only to asses the robustness properties of classical procedures but also to produce estimators and tests that are robust to model deviations. In the case of robust estimation of multivariate location and scatter, robust covariances have been first investigated by Maronna (1976). In particular, he shows that robust estimators based on a weighting scheme that is not redescending (no weight of zero), fails to be robust in high dimensions. This happens because for such estimators (including the classical MLE), their breakdown point, i.e., the maximal amount of model misspecification they can withstand before they "breakdown" or their bias becomes arbitrarily large, is at most $1/(p+1)$, p being the dimension of the data. When working in high dimension it is therefore crucial to consider high breakdown estimators.

The statistical literature contains several proposals for high breakdown estimators of the mean and covariance in multivariate data when it is suspected that the data contain outliers or extreme observations. A well known one is the minimum covariance determinant (MCD) of Rousseeuw (1984). When they are missing data only Little and Smith (1987) and Cheng and Victoria-Feser (2002) propose different solutions. In this paper we actually concentrate on robust estimators with missing data, in particular we propose the use of faster algorithms for their computation and compare them through extensive simulations in terms of their robustness properties when data are contaminated and also in terms of the speed of two different algorithms used to compute the robust estimators. We also adapt the orthogonalized Gnanadesikan-Kettering (OGK) estimator proposed by Maronna and Zamar (2002) to the case of missing data and use it as a starting point for an adapted S-estimator. All our programs are readily available (upon request) in the form of an Splus library which has been used to produce the results and graphics presented in this paper. We will also consider real data to illustrate in another way the added value of robust estimators of mean and covariance, when the later is used for example as input to a principal component analysis.

The paper is organised as follows. In Section 2 we present a general class of estimators adapted to the case of missing data that includes as particular cases the MLE computed via the EM algorithm, its robust modification proposed by Little and Rubin (1987) and the adaptation of the S-estimator (Rousseeuw and Yohai 1984) proposed by Cheng and

Victoria-Feser (2002). In Section 3 we present the modified *MCD* proposed by Cheng and Victoria-Feser (2002) with the modification of a fast algorithm proposed by Rousseeuw and Van Driessen (1999), namely the *FAST-MCD*, to deal with missing data. We also present the adaptation of the *OGK* estimator to the case of missing data. In Section 4, an extensive simulation study is conducted to compare the speeds of the algorithms as well as the robustness properties of the different robust estimators. Finally, in Section 5 real datasets are analysed by means of a principal component analysis when classical and robust estimators are used as input.

2 A general class of estimators with missing data

The aim is to estimate the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, i.e., the mean and covariance of an underlying multivariate variable $Y = (Y_1, \ldots, Y_p)$ that has supposedly generated the sample $\mathbf{y}_i, i = 1, \ldots, n$ at hand. As it often happens in practice, we suppose that some of the observations might be missing in that some of the y_{ij} are observed for some $j \in \{1, \ldots, p\}$ and the others are not observed or missing for the other j's. In other terms, $\mathbf{y}_i = [\mathbf{y}_{[oi]}^T, \mathbf{y}_{[mi]}^T]^T$ so that a distinction is made between the observed (oi) and the missing (mi) data. We suppose that the data are missing at random (see Rubin 1976), a sufficient condition for correct likelihood-based inferences. Most known estimators of mean and covariance with missing data fall in the class proposed by Cheng and Victoria-Feser (2002), i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} w_i^{\mu}(\boldsymbol{\mu} - \hat{\mathbf{y}}_i) = 0 \tag{1}$$

$$\frac{1}{n}\sum_{i=1}^{n} \left[w_i^{\delta}\boldsymbol{\Sigma} - w_i^{\eta}((\hat{\mathbf{y}}_i - \boldsymbol{\mu})(\hat{\mathbf{y}}_i - \boldsymbol{\mu})^T - \mathbf{C}_i)\right] = 0 \tag{2}$$

where

$$\begin{aligned}\hat{\mathbf{y}}_i &= \left[\mathbf{y}_{[oi]}^T, E\left[\mathbf{y}_{[mi]}\middle|\, \mathbf{y}_{[oi]}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right]^T\right]^T \\ &= \left[\mathbf{y}_{[oi]}^T, \boldsymbol{\mu}_{[mi]}^T + (\mathbf{y}_{[oi]} - \boldsymbol{\mu}_{[oi]})^T\boldsymbol{\Sigma}_{[ooi]}^{-1}\boldsymbol{\Sigma}_{[omi]}\right]^T \end{aligned} \tag{3}$$

and

$$\begin{aligned} C_{ijk} &= \operatorname{cov}\left[\begin{bmatrix}\mathbf{y}_{[oi]}\\ \mathbf{y}_{[mi]}\end{bmatrix}\left[\mathbf{y}_{[oi]}^T\ \mathbf{y}_{[mi]}^T\right]\middle|\, \mathbf{y}_{[oi]}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right] \\ &= \begin{bmatrix} 0 & 0 \\ 0 & \operatorname{cov}\left[\mathbf{y}_{[mi]}\mathbf{y}_{[mi]}^T\middle|\, \mathbf{y}_{[oi]}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right]\end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{[mmi]} - \boldsymbol{\Sigma}_{[moi]}\boldsymbol{\Sigma}_{[ooi]}^{-1}\boldsymbol{\Sigma}_{[omi]}\end{bmatrix}. \end{aligned} \tag{4}$$

where for example $\boldsymbol{\Sigma}_{[ooi]}$ denotes the partition of $\boldsymbol{\Sigma}$ corresponding to the observed part of $\mathbf{y}_i$, etc. The different estimators are actually defined through the data weighting system given

by w_i^μ, w_i^δ and w_i^η in (1) which in turn also depends on the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (see below). To compute the estimators, one can use an iterative procedure in which given current values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the $\hat{\mathbf{y}}_i$, $\mathbf{C}_i$ and the weights are first computed, and the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are then updated by

$$\boldsymbol{\mu}^* = \frac{1}{n}\sum_{i=1}^{n} w_i^\mu \hat{\mathbf{y}}_i \Bigg/ \frac{1}{n}\sum_{i=1}^{n} w_i^\mu \tag{5}$$

$$\boldsymbol{\Sigma}^* = \left[\frac{1}{n}\sum_{i=1}^{n} w_i^\eta ((\hat{\mathbf{y}}_i - \boldsymbol{\mu}^*)(\hat{\mathbf{y}}_i - \boldsymbol{\mu}^*)^T - \mathbf{C}_i)\right] \Bigg/ \left[\frac{1}{n}\sum_{i=1}^{n} w_i^\delta\right] \tag{6}$$

The classical MLE is obtained when $w_i^\mu = w_i^\eta = w_i^\delta = 1\ \forall i$, and (5) and (6) define the EM algorithm (see Dempster, Laird, and Rubin 1977). However, with complete data it is well known that the MLE of mean and covariance is not robust. When there are missing data, the situation doesn't change; see Cheng and Victoria-Feser (2002). Little and Rubin (1987) propose to base the M-step on a robust estimator belonging to the general class of M-estimator (Huber 1981). They call the resulting procedure the ER algorithm. Their estimator is defined by (5) and (6) with[1]

$$(w_i^\mu)^2 = w_i^\eta = w_i^\delta = w_i = \omega(d_{oi})/d_{oi}^2 \tag{7}$$

where

$$d_{oi}^2 = d_{oi}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\mathbf{y}_{[oi]} - \boldsymbol{\mu}_{[oi]}\right)^T \boldsymbol{\Sigma}_{[ooi]}^{-1} \left(\mathbf{y}_{[oi]} - \boldsymbol{\mu}_{[oi]}\right) \tag{8}$$

is the squared Mahalanobis distance corresponding to the observed part of $\mathbf{y}_i$. Here ω is a two-parameter weight function defined by

$$\omega(d_{oi}) = \begin{cases} d_{oi}^2 & \text{if } d_{io} \leq d_i^* \\ (d_i^*)^2 \exp\{-(d_{oi} - d_i^*)^2/b_2^2\} & \text{if } d_{io} > d_i^* \end{cases} \tag{9}$$

where $d_i^* = \sqrt{p_i} + b_1/2$, and p_i is the number of variables present for observation i. The quantities b_1 and b_2 are to be specified by the analyst and Little and Smith (1987) proposed $b_1 = 2$ and $b_2 = 1.25$. If the case i is uncontaminated, the data are normal and missing values are missing at random, then (8) is asymptotically $\chi_{p_i}^2$. The Wilson-Hilferty transformation of the chi-squared distribution yields $(d_{oi}^2/p_i)^{1/3} \sim N(1 - 2/(9p_i), 2/(9p_i))$. Following Little and Smith (1987), we also propose a probability plot of

$$Z_i = \frac{(d_{oi}^2/p_i)^{1/3} - 1 + 2/(9p_i)}{\sqrt{2/(9p_i)}} \tag{10}$$

versus standard normal order statistics, that should reveal atypical observation.

Little and Smith (1987) proposed as starting point of the ER algorithm, the MLE on the data where the missing ones have been replaced by the median of the corresponding

[1] The iteration step for the covariance matrix (6) doesn't exactly correspond to the same step in the ER algorithm in that the weights w_i^η are not applied to the correction matrix $\mathbf{C}_i$. We will however, in what follows consider this slight modification of the ER algorithm.

observations. Although the ER algorithm is relatively simple to implement, it suffer from an important drawback : its breakdown point is at most $1/(p+1)$ because it is based on a weighting scheme that is not redescending. This drawback will be highlighted by the simulation results. This means that if the proportion of outliers exceeds this value (or even is near it) the robust estimator is not robust anymore.

To construct a high breakdown estimator of mean and covariance matrix in multivariate data when some are missing, Cheng and Victoria-Feser (2002) propose two strategies. The first one is to provide an high breakdown estimator such as the MCD estimator as starting point for the ER algorithm and the second is to also adapt a high breakdown estimator such as an S-estimator (Rousseeuw and Yohai 1984) to incomplete data. The resulting estimator which is called the $ERTBS$ is then defined through (5) and (6) with

$$w_i^{\mu} = \psi(d_{oi}\left(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}\right)/k)\Big/(d_{oi}/k) \ , \tag{11}$$

$$w_i^{\eta} = pw_i^{\mu} \ , \tag{12}$$

$$w_i^{\delta} = (d_{oi}\left(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}\right)/k)^2 w_i^{\mu} \tag{13}$$

with $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ being the current values of the estimator, and

$$k = \frac{d_{[q]}}{\sqrt{(\chi^2_p)^{-1}(q/(n+1))}} \ , \tag{14}$$

where $d_{[q]}$ denotes the q-th ordered distance (based on the $d_{oi}\left(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}\right)$), $q = \lfloor (n+p+1)/2 \rfloor$ with $\lfloor x \rfloor$ denoting the integer part of x and

$$\psi(d;c,M) = \begin{cases} d & 0 \leq d < M \\ d\left(1-\left(\frac{d-M}{c}\right)^2\right)^2 & M \leq d \leq M+c \\ 0 & d > M+c \end{cases}$$

This ψ function defines the translated biweight S-estimator proposed by Rocke (1996) and is the derivative of the ρ function given for $M \leq d \leq M+c$ by

$$\begin{aligned} \rho_{M\leq d\leq M+c}(d;c,M) = \ & M^2/2 - M^2(M^4 - 5M^2c^2 + 15c^4)/30c^4 \\ & + d^2(0.5 + M^4/2c^4 - M^2/c^2) + d^3(4M/3c^2 - 4M^3/3c^4) \\ & + d^4(3M^2/2c^4 - 1/2c^2) - 4Md^5/5c^4 + d^6/6c^4 \end{aligned}$$

and for all d by

$$\rho(d;c,M) = \begin{cases} d^2/2 & 0 \leq d < M \\ \rho_{M\leq d\leq M+c}(d;c,M) & M \leq d \leq M+c \\ M^2/2 + c(5c+16M)/30 & d > M+c \end{cases}$$

The parameters M and c control the breakdown point ε^* and the asymptotic rejection probability ARP α of the $ERTBS$. The ARP can be interpreted as the probability for

an estimator, in large samples under a reference distribution, to give a null (or nearly null) weight. M and c are found implicitely by

$$\begin{aligned} \varepsilon^* \max_d \rho(d; c, M) &= E_{\chi^2_p}[\rho(d; c, M)] \ , \\ M + c &= \sqrt{(\chi^2_p)^{-1}(1-\alpha)} \ ; \end{aligned}$$

The choices for ε^* and α are to be made by the analyst. The former is the suspected maximal amount of contaminated data and for the latter Cheng and Victoria-Feser (2002) propose choices between 0.1% and 1%.

As Rocke (1996) noted, it is very important to choose a good starting point for any algorithm defining a high breakdown point estimator, otherwise the later can loose its high breakdown properties. For the $ERTBS$, Cheng and Victoria-Feser (2002) therefore propose an adaptation of the MCD estimator as a starting point as well as an algorithm to compute it. However, to compute the MCD one needs algorithms that are based on random starting subsamples. This can lead to situations in which the MCD is very long to compute, if not impossible. Therefore, in the following Section, we propose a fast algorithm to compute the MCD by adapting the *FAST-MCD* of Rousseeuw and Van Driessen (1999) and as an even faster alternative, we propose a modified version of the OGK estimator adapted to the case of missing data to be used as a starting point for the $ERTBS$.

3 Starting point robust estimators with missing data

3.1 The modified MCD

The objective of the MCD estimator is to find h observations (out of n) whose covariance matrix has the lowest determinant. The MCD mean estimator is then the sample mean of those h points, and the MCD covariance estimator is their sample covariance matrix. To compute the MCD, one needs an algorithm for finding the best subset of h points, which usually involves the repeated computation of the sample mean and covariance as well as Mahalanobis distances. When some observations are missing, Cheng and Victoria-Feser (2002) propose to use the EM algorithm to compute the sample means and covariances at all steps of the algorithm and to base the Mahalanobis distances on the observed part of the observation as in (8). The later are standardized by means of the Wilson-Hilferty transformation given in (10), so that one takes into account the non equal number of missing values for each observation.

A choice needs to be made on h and one way is to choose it such that the MCD has the highest breakdown. In this case, the minimal value of h is given by (Rousseeuw and Leroy 1987) :

$$h := \left\lfloor \frac{n+p+1}{2} \right\rfloor$$

But this is also the choice that give the largest efficiency loss. So when we suspect that the sample is not heavily contaminated we can reasonably choose a larger value for the proportion of points of say 75% or 80% so we can take $h := \lfloor 0.75n \rfloor$ or $h := \lfloor 0.80n \rfloor$.

The time needed to run the MCD can be quite large. That's why several authors focus on the development of algorithms able to deal with this problem. Hawkins (1994) presents a feasible solution algorithm for the MCD which involves taking random "trial solutions" and refining each ones to a local optimum satisfying the condition for the MCD criterion. Atkinson (1993,1994) proposes the forward search algorithm which also permits the detection of multiple outliers. This is the algorithm that is adapted by Cheng and Victoria-Feser (2002) to the case of missing data. More recently, Rousseeuw and Van Driessen (1999) present a new algorithm called *FAST-MCD* supposed to be even faster than the forward search algorithm and able to deal with very large data sets. In this paper, we propose to adapt it to compute the MCD when there are missing data.

A key idea of the *FAST-MCD* algorithm is the fact that starting from any approximation to the MCD, it is possible to find an approximation with a lower determinant. Indeed Rousseeuw and Van Driessen (1999) observed that from a subset H_k of size h in which μ, Σ and the Mahalanobis distances are computed, one can create a subset H_{k+1} by taking among the n observations the h ones with the smallest Mahalanobis distances with the property that the determinant of Σ based on $H_{k+1}$is smaller. Each step is called a C-step. The initial subset is created by choosing randomly $p+1$ observations on which the Mahalanobis distances are computed to order the n observations. The first h ones define the initial subset H_1. If the determinant of Σ based on the randomly chosen $p+1$ observations is nil, one adds one randomly chosen observation at the time until the determinant becomes positive. If for any subset H_k there are missing values, we compute μ_k and Σ_k with the EM algorithm. The Mahalanobis distances are also changed as in (8) and standardized using the Wilson-Hilferty transformation. The absolute value of the later is used to order the observations. The initial subset is created choosing randomly $p+1$ observations among the fully observed ones.

For each subset H_k, one must compute a covariance matrix, a determinant and the Mahalanobis distances. This can be rather heavy if the data set is large. Therefore Rousseeuw and Van Driessen (1999) suggest a simplification: they show empirically that it is possible to make a distinction between good (robust) estimations and bad ones after only two or three steps. This means that the C-step doesn't need to be iterated until the covariance matrix with minimal determinant is found, the algorithm can switch to another initial subset. We found the same feature with our simulations. Finally, another particularity of the *FAST-MCD* algorithm is that it can be split into a nested system of subsets to improve the speed of convergence in large datasets (see Rousseeuw and Van Driessen 1999).

Through extensive simulations we compare in Section 4 the forward search algorithm and the *FAST-MCD* algorithm for the computation of the MCD with missing data.

3.2 The modified OGK

Maronna and Zamar (2002) base their OGK on the robust estimator for covariances σ_{jk} proposed by Gnanadesikan and Kettenring (1972) which is very simple to compute. Indeed the later is defined for a pair of random variables (i.e. $p=2$) as

$$\frac{1}{4}\left(\sigma(Y_j+Y_k)^2-\sigma(Y_j-Y_k)^2\right)$$

where $\sigma()$ is a standard deviation function applied on its argument. A robust estimator for σ_{jk} is obtained when $\sigma()$ is a robust function. When $p > 2$, the covariance matrix $\mathbf{\Sigma}$ is estimated by replacing all its elements by all pairwise estimates. It is well known that such an estimator may produce non positive definite matrices and the estimator is not affine equivariant. To overcome the lack of positive definitness, Maronna and Zamar (2002) propose an estimator defined by the following four steps:

1. Let $\mathbf{D} = \text{diag}\left(\sigma\left(Y_j\right)\right)\big|_{j=1,\ldots,p}$ and define $\mathbf{x}_i = \mathbf{D}^{-1}\mathbf{y}_i, i = 1, \ldots, n$, i.e., realizations from $X = (X_1, \ldots, X_p)$

2. Compute the matrix $\mathbf{U} = (u_{jk})$ with

$$u_{jk} = \begin{cases} \frac{1}{4}\left(\sigma(X_j + X_k)^2 - \sigma(X_j - X_k)^2\right) & j \neq k \\ 1 & j = k \end{cases} \tag{15}$$

3. Decompose $\mathbf{U}$ as $\mathbf{U} = \mathbf{E}\Lambda\mathbf{E}^T$ with $\Lambda =$diag$(\lambda_1, \ldots, \lambda_p)$

4. Define $\mathbf{z}_i = \mathbf{E}^T\mathbf{x}_i$, i.e., realizations from $Z = (Z_1, \ldots, Z_p)$ and $\mathbf{A} = \mathbf{DE}$. The estimator of $\mathbf{\Sigma}$ is $\mathbf{A}\Gamma\mathbf{A}^T$ with $\Gamma = \text{diag}\left(\sigma\left(Z_j\right)^2\right)\big|_{j=1,\ldots,p}$.

A location estimator for $\boldsymbol{\mu}$ is given by $\mathbf{A}\nu$ with $\nu = \left(m\left(Z_j\right)\right)\big|_{j=1,\ldots,p}$, $m()$ being a (robust) mean function. The procedure can be iterated by replacing $\mathbf{U}$ in step 2 by $\mathbf{E}\Gamma\mathbf{E}^T$ until convergence. For $\sigma()$ and $m()$, Maronna and Zamar (2002) propose the following functions

$$m(Y) = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \tag{16}$$

and

$$\sigma\left(Y\right)^2 = \frac{MAD\left(Y\right)}{n}\sum_{i=1}^n \rho_{c2}\left(\frac{y_i - m(Y)}{MAD\left(Y\right)}\right) \tag{17}$$

with

$$w_i = W_{c1}\left(\frac{y_i - m(Y)}{\sigma_0\left(Y\right)}\right)$$

$$W_c(x) = \begin{cases} \left(1 - \left(\frac{x}{c}\right)^2\right)^2 & |x| \leq c \\ 0 & \text{otherwise} \end{cases}$$

and

$$\rho_c\left(x\right) = \min\left(x^2, c^2\right)$$

Maronna and Zamar (2002) propose to use the values of $c1 = 4.5$ and $c2 = 3$. Moreover, they argue that to improve the efficiency of the OGK, one could use it as a hard rejection tool in that a reweighted estimator as in (1) is used in which $\hat{\mathbf{y}}_i = \mathbf{y}_i \forall i$ and $w_i^\mu = w_i^\eta = w_i^\delta = w_i$ with

$$w_i = \begin{cases} 1 & \left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_{\text{OGK}}\right)^T \widehat{\mathbf{\Sigma}}_{\text{OGK}}^{-1}\left(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_{\text{OGK}}\right) \leq \chi_p^2(.9) \\ 0 & \text{otherwise} \end{cases}$$

The resulting estimator will be called the reweighted OGK ($rOGK$). Note that this strategy is also used most of the times with the MCD but with the quatile 0.975 (instead of 0.9) of the χ^2_p. We will call the resulting estimator the $rMCD$.

To extend the OGK or $rOGK$ to the case of missing data, we propose to impute the missing values by means of the $\widehat{y}$ in (3) obtained by the EM algorithm, i.e., with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ estimated by (1) where all weights are equal to 1. The reason is that the EM algorithm is very fast, and although it leads to baised estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and therefore of the imputed values $\widehat{y}$, this shouldn't affect the resulting OGK. Indeed, the OGK downweigths extreme observations in (16) and (17), and these observations can be extreme because of either the observed values or the imputed ones. Through extensive simulations, we will study this adapted OGK in Section 4.

4 Simulation study

The aim of our simulation study is first to compare the behaviors under different situations of the different estimators proposed by Cheng and Victoria-Feser (2002) as well as the modified OGK for missing values as such or as a starting point for the $ERTBS$. Second,we also compare the speed of the two algorithms for the MCD with missing data in different settings, i.e., the (modified) forward search algorithm and our adaptation of the *FAST-MCD*, as well as with the modified OGK. We will see that the *FAST-MCD* outperforms the forward search in all situations but that the OGK is the fastest of all.

4.1 The design

The model is the multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For the affine equivariant estimators (i.e. the MCD and the $ERTBS$), their performance is supposed to be independent of the choice for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ so that one could choose $\mathbf{0}$ and $\mathbf{I}$. This is however not the case for the OGK, and in fact the more the variables are correlated, the larger the potential bias of the estimator. In order to control for the strenght of the correlation, Maronna and Zamar (2002) suggest to transform the $N(\mathbf{0}, \mathbf{I})$ simulated data $\mathbf{y}_i$ by $\mathbf{x}_i = \mathbf{R}(\rho)\mathbf{y}_i$ where $\mathbf{R}(\rho) = (r_{jk})|_{j,k=1,\ldots,p}$ and

$$r_{jk} = \begin{cases} 1 & j = k \\ \rho & j \neq k \end{cases}$$

so that the resulting covariance matrix is $\boldsymbol{\Sigma} = \mathbf{R}(\rho)^2$. Maronna and Zamar (2002) suggest to use the value of $\rho = 0.2$ which yields a $\boldsymbol{\Sigma}$ with the largest eigen value more than 12 times the second one, thus indicating a relatively strong correlation. In our simulations, we will use the $N(\mathbf{0}, \mathbf{I})$ and the $N(\mathbf{0}, \mathbf{R}(.2)^2)$ models for all estimators.

For the problem of how to contaminate the data, we follow the proposition of Woodruff and Rocke (1993). To generate an ϵ-proportion of so-called shift-outliers, i.e., the ones which are the hardest to find, we put the center of the contaminated data at a distances of $\sqrt{p}+\beta/\sqrt{2}$ from the mean where β parametrizes the distances of the contamination from the main body of the data. The missing data, if any, are chosen randomly among the mixture distribution between the good data ($N(\mathbf{0}, \boldsymbol{\Sigma})$) and the bad data ($N(\left(\sqrt{p} + \beta/\sqrt{2}\right)\mathbf{e}_p, \boldsymbol{\Sigma})$, $\mathbf{e}_p$

being a p-dimensional vector of ones). Table 1 summarizes the combinations of quantities for ϵ, β and the proportion of missing data ($miss$) that we have considered. Table 2 shows the different values for n and p used in the simulations.

$miss =$	0.1	0.2	0.3	
$\epsilon =$	0	0.02	0.05	0.1
$\beta =$	1.6			

Table 1: values for $miss$, ϵ and β

	$p = 10$	$p = 20$	$p = 50$
$n =$	50	100	200
$n =$	100	200	400
$n =$	500	500	600

Table 2: values for n and p

Each robust estimator requires a decision on its initialization parameters. For the MCD estimator, $h = [0.6n]$.was chosen. For the OGK, $c1 = 4.5$ and $c2 = 3$ were chosen. For the $ERTBS$ estimator we chose for our simulations the breakdown point $\varepsilon^* = 0.3$ and the ARP $\alpha = 0.001$. All computational experiments were done on a Athlon 1900Mhz with 512 MB of memory. The core of the program was written in Fortran 77 and Splus was used as a front-end (to produce the various graphics). For all combinations of parameters, 1000 samples were generated.

4.2 Computational times

We describe now the time needed to compute the $rOGK$ or the $rMCD$, when the later is computed using the adapted *FAST-MCD* algorithm ($rMCD/FAST$) or the forward search algorithm ($rMCD/FWD$). We chose the reweighted version of the two starting point estimators, because as we will see later, the non-reweighted versions can lead to biased estimates. For each of the parameters given in Table 2 and for different sample sizes, a time in second has been computed. Figure 1 shows the results (in a log-scale) for the datasets with $\epsilon = 10\%$ and $miss = 30\%$ (for other combinations the results are comparatively similar).

We notice the following features. The speed for the $rMCD/FWD$ as expected is slower than the speed of the $rMCD/FAST$, with an increasing difference as the sample size increases. The $rMCD/FAST$ can be up to 150 times faster than the $rMCD/FWD$. However, when the $rOGK$ is used as a starting point, the computational times decrease drastically, with sometimes a ratio of 18 compared with the $rMCD/FAST$. However, the speed of the $rMCD/FAST$ doesn't depend very much on the sample size \$n\$, whereas the $rOGK$ does quite substantially.

4.3 Comparing estimators

The aim of this subsection is to study the robustness properties (bias versus efficiency) of the different estimators proposed with incomplete data by means of simulations. For the MCD,

all calculations were made using the modified *FAST-MCD* for missing data. It should be stressed that this exercise has not been done in Cheng and Victoria-Feser (2002). The estimators we consider here are those presented in Section 2 namely, the MLE computed via the EM algorithm (which is taken as a benchmark), the ER algorithm with the MLE as starting point (ER/MLE), the ER algorithm with the MCD, $rMCD$, OGK and $rOGK$ as starting point (ER/MCD, $ER/rMCD$, ER/OGK and $ER/rOGK$), the $ERTBS$ algorithm with the MCD, $rMCD$, OGK and $rOGK$ as starting point ($ERTBS/MCD$, $ERTBS/rMCD$, etc.). The data were generated using the designs presented in Section 4.1. The percentage of missing observations and the sizes n and p do not seem to have an influence on the behaviour of the different estimators. The influencial factors are the covariance structure and the percentage of contamination. Indeed, when the data are correlated ($N(\mathbf{0}, \mathbf{R}(.2))$) the OGK can be biased when there is data contamination which is not the case when the data are uncorrelated ($N(\mathbf{0}, \mathbf{I})$). The consequence is that the ER and the $ERTBS$ become also biased. We use boxplots to compare the estimators. They are built on the estimated biases of one of the element of the mean vector, one of the diagonal elements of the covariance matrix and one of the off-diagonal elements of the covariance matrix. Only the results for μ_1, σ_{11}, and σ_{12} are represented, since for other parameters, the same pattern is found. In Figure 2 are presented the boxplots of the sampling distributions of the MCD, $rMCD$, OGK and $rOGK$ with $miss = 0.1$ and $n = 50$ and $p = 10$. One can see that for the variance and covariance the OGK is biased when there is 5% or more data contamination. Fortunately, the $rOGK$ doesn't show the same pattern and therefore we propose to use the later one as a starting point for the ER or the $ERTBS$. In Figure 3 are presented the sampling distributions of the final estimators when the $rMCD$ and the $rOGK$ are used as starting points for the robust ones. The EM (i.e. MLE) is taken as a benchmark. The MLE clearly fails even if the contamination is small. However it is the most efficient with no contamination but the efficiency loss for the robust estimators seems to be quite small. The ER/MLE breaksdown at (at most) 10% of data contamination. Finally, the $ER/rMCD$, $ER/rOGK$, $ETBBS/rMCD$ or $ERTBS/rOGK$ are very robust and can withstand at least 10% of data contamination.

If we want to see a difference between the ER and $ERTBS$ with the same high breakdown strating point, we have to push the percentage of contamination up to 30%. We haven't done a full coverage of such situation since its very unlikely that someone will want to study data sets with such a percentage of contamination. We show here an example based on one simulated dataset of size $n = 100$ and $p = 10$ with 30% of contamination (the first 30 observations) and 10% of missing values. We plot the transformed Mahalanobis distances to see if the estimators can detect all the contaminated values. The results are displayed in Figure 4 for the $ER/rOGK$ and $ERTBS/rOGK$ but we found the same result with the $rMCD$ as starting point. Clearly the $ER/rOGK$ breaks down in such a case but the $ERTBS/rOGK$ does not since it is able to detect the 30 outliers.

5 Conclusion

In this paper we have considered high breakdown estimation of the mean and covariance of a multivariate normal distribution with missing data. We have proposed to use a modification of the *FAST-MCD* algorithm to compute the MCD which is used as a starting point for the ER of the $ERTBS$. We found through simulations that the computational speed is much more improved when on uses the C-step instead of the forward search. We have also conducted a simulation study to compare the different high breakdown estimators computed in different ways. First we found that the results are independent of the chosen method to compute the MCD. As expected, the MLE breaks down at very low levels of contamination (2%). The ER breaks down at at least 10% of contamination if the starting point is not the MCD and breaks down at 30% anyway. The $ERTBS$ is the most robust overall and its variance is comparable to the one of the other estimators (including the MLE) so that the efficiency loss in using this high breakdown estimator is very small. Finally, the program to compute the $ERTBS$ by means of the *FAST-MCD* is available as an Splus library from the authors.

References

Atkinson, A. C. (1993). Stalactite plots and robust estimation for the detection of multivariate outliers. In S. Morgenthaler, E. Ronchetti, and W. A. Stahel (Eds.), *New Directions in Statistical Data Analysis and Robustness*. Basel: Birkhäuser.

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association 89*, 1329–1339.

Cheng, T.-C. and M. Victoria-Feser (2002). High breakdown estimation of multivariate location and scale with missing observations. *British Journal of Mathematical and Statistical Psychology*. To appear.

Dempster, A. P., M. N. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Serie B 39*, 1–22.

Gnanadesikan, R. and J. R. Kettenring (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics 29*, 81–124.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics 42*, 1887–1896.

Hawkins, D. M. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics & Data Analysis 17*, 197–210.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 35*, 73–101.

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley.

Little, R. J. A. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R. J. A. and P. J. Smith (1987). Editing and imputing for quantitative survey data. *Journal of the American Statistical Association 82*, 58–68.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics 4*, 51–67.

Maronna, R. A. and R. H. Zamar (2002). Robust multivariate estimates for high-dimensional datasets. *Technometrics 44*, 307–317.

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in hish dimension. *The Annals of Statistics 24*, 1327–1345.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association 79*, 871–880.

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

Rousseeuw, P. J. and K. Van Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*, 212–223.

Rousseeuw, P. J. and V. J. Yohai (1984). Robust regression by means of S-estimators. In J. W. Franke, Hardle, and R. D. Martin (Eds.), *Robust and Nonlinear Time Series Analysis*, pp. 256–272. New York: Springer-Verlag.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin (Ed.), *Contributions to Probability and Statistics*, pp. 448–485. Stanford (CA): Stanford University Press.

Woodruff, D. L. and D. M. Rocke (1993). Heuristic search algorithm for the minimum volume ellipsoid. *Journal of Computational and Graphical Statisitcs 2*, 69–95.

Woodruff, D. L. and D. M. Rocke (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association 89*, 888–896.

Yuan, K.-H. and P. M. Bentler (1998). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology 51*, 63–88.

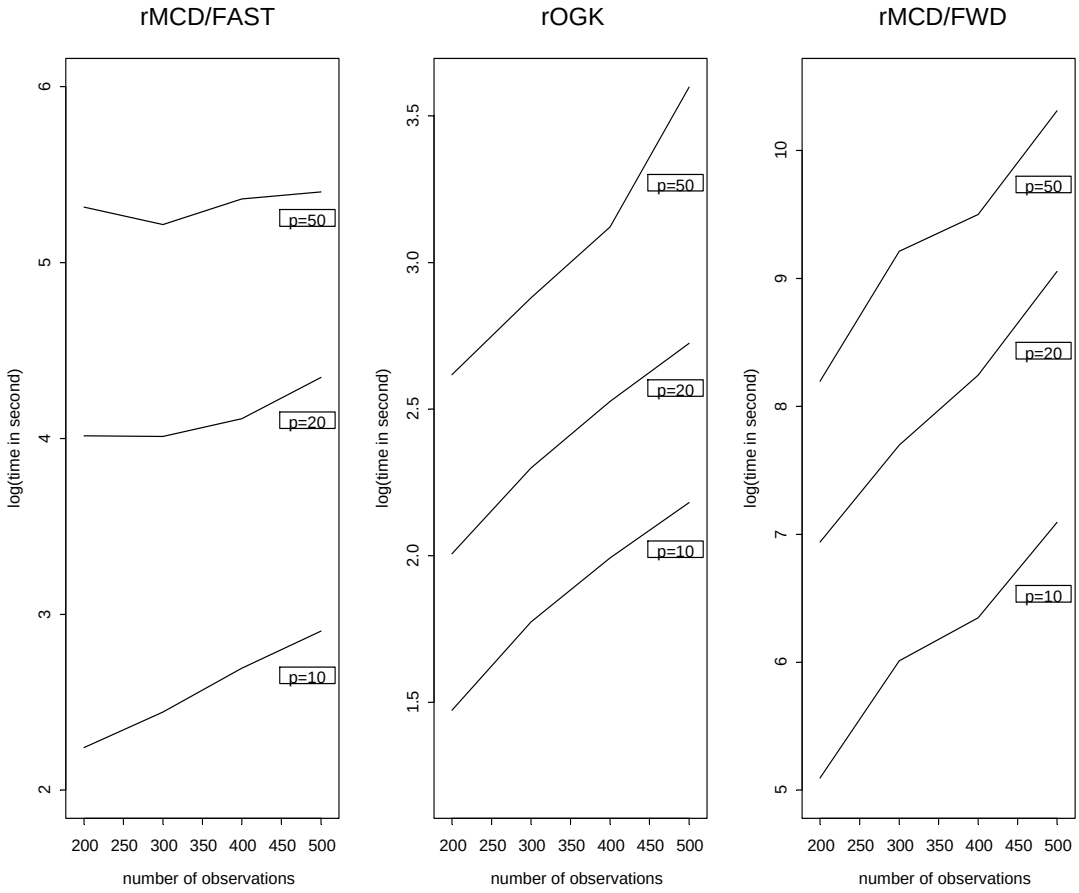

Figure 1: Log of the mean time in seconds needed to compute the $rOGK$ and the $rMCD$ by means of the forward search (FWD) algorithm and FAST-MCD algorithm as a function of the sample size and the data dimension p.

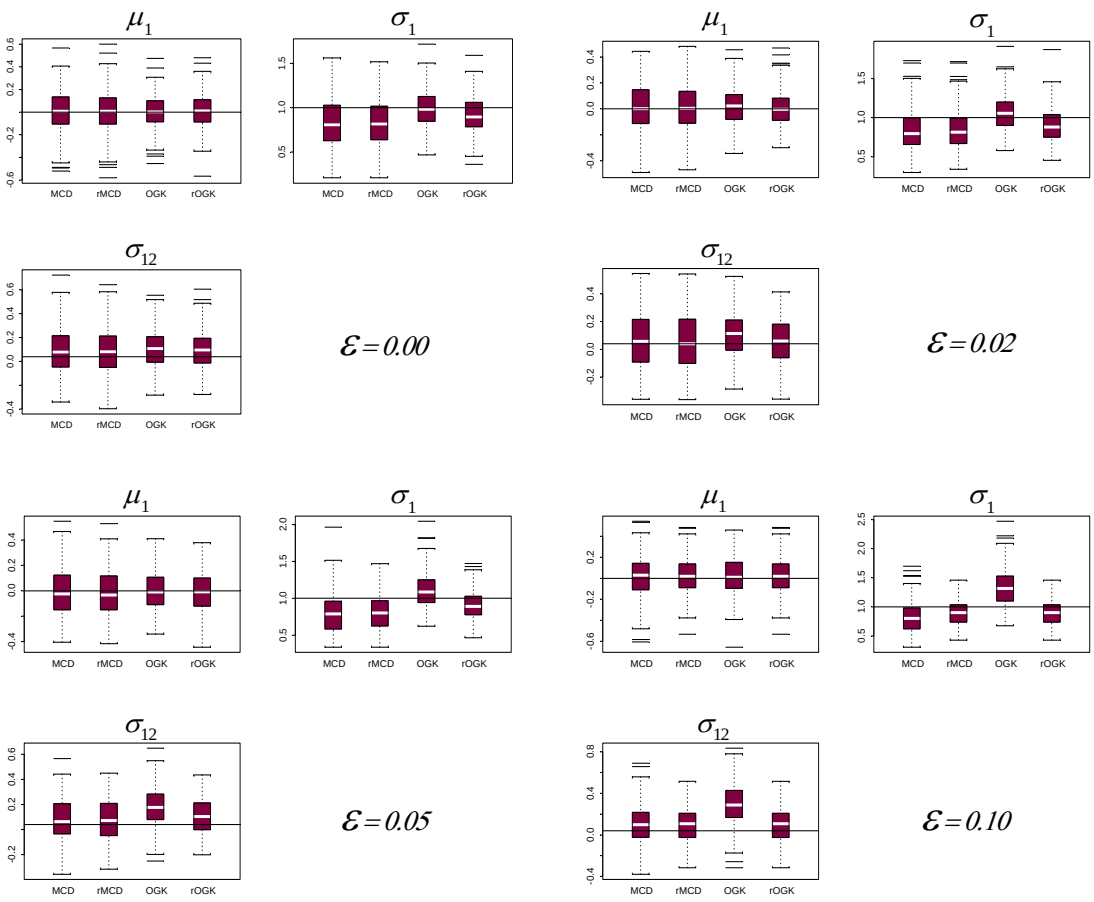

Figure 2: Sampling distribution of starting point robust estimators with missing data for different amounts of data contamination.

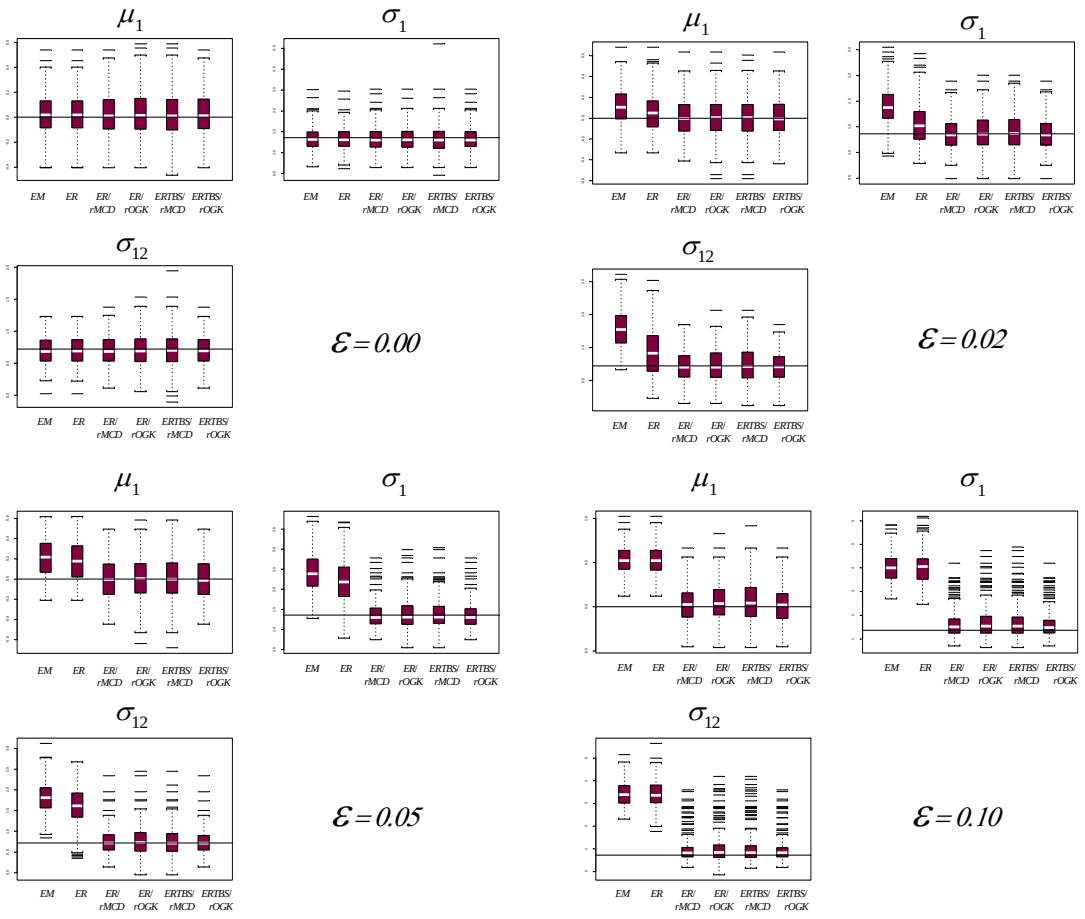

Figure 3: Sampling distribution of robust estimators with missing data for different amounts of data contamination

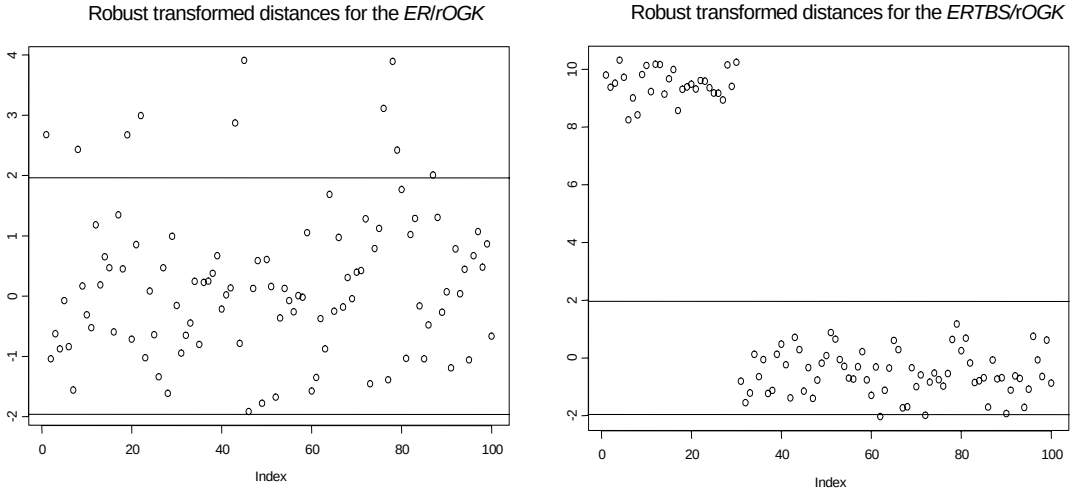

Figure 4: Transformed Mahalanobis distances using the $ERTBS$ or ER with the $rOGK$ start to detect outlying observations