British Journal of Mathematical and Statistical Psychology (2002), 55, 317–335

www.bps.org.uk

High-breakdown estimation of multivariate mean and covariance with missing observations

Tsung-Chi Cheng[1] and Maria-Pia Victoria-Feser[2]*
[1]Department of Statistics, National Chengchi University, Taiwan
[2]Faculty of Psychology and Education, University of Geneva, Switzerland

We consider the problem of outliers in incomplete multivariate data when the aim is to estimate a measure of mean and covariance, as is the case, for example, in factor analysis. The ER algorithm of Little and Smith which combines the EM algorithm for missing data and a robust estimation step based on an M-estimator could be used in such a situation. However, the ER algorithm as originally proposed can fail to be robust in some cases, especially in high dimensions. We propose here two alternatives to avoid the problem. One is to combine a small modification of the ER algorithm with a so-called high-breakdown estimator as the starting point for the iterative procedure, and the other is to base the estimation step of the ER algorithm on a high-breakdown estimator. Among the high-breakdown estimators which are actually built to keep their robustness properties even if the number of variables is relatively large, we consider here the minimum covariance determinant estimator and the *t*-biweight S-estimator. Simulated and real data are used to compare and illustrate the different procedures.

1. Introduction

Many statistical procedures such as principal components analysis, factor analysis and covariance structure analysis require the estimation of a vector of means and a covariance matrix from the data at hand. A question that might arise when one performs these types of analysis is the extent of the influence of outliers or extreme data on the final results. Outliers or extreme data are taken here to mean observations of a subject that either do not behave like the majority (true extreme data) or that have not been recorded properly (false extreme data). Some might argue that in the first case, since there is no measurement error, the subject should be kept in the sample and the analysis carried out as usual. However, even in the fairest world, who wants a single subject to

*Requests for reprints should be addressed to Dr Maria-Pia Victoria-Feser, Faculty of Psychology and Education, University of Geneva, 40, Bd du Pont d'Ave, CH-1211 Geneva 4, Switzerland (e-mail: maria-pia.victoria-feser@hec.unige.ch).

dominate the outcome of the analysis? As will be shown below, such a situation can occur when classical sample means and covariances are used.

Robust statistics deal with the problems caused by outliers or extreme data which are a particular case of model misspecification. Robust theory provides tools not only to assess the robustness properties of statistical procedures, but also estimators and testing procedures that are resistant to model deviations in general and extreme data in particular. The general theory is given in Huber (1981) and Hampel, Ronchetti, Rousseeuw, and Stahel (1986), and a non-technical presentation of the subject can be found in Wilcox (1998). Robust covariances were first investigated by Devlin, Gnanadesikan, and Kettenring (1975), Maronna (1976), Huber (1977) and Campbell (1980), and robustness in the context of covariance structure analysis can be found in Yuan and Bentler (1998). The latter consider the case of complete data and show by means of the influence function *IF* (Hampel, 1974), a mathematical tool to assess the robustness properties of a statistical procedure, that classical estimators of structured parameters are not robust if the covariance matrix of the raw data is not estimated robustly. It is therefore crucial to have good robust procedures for the estimation of covariance matrices.

The statistical literature contains several proposals for estimators of the mean and covariance in multivariate data when it is suspected that the data contain outliers or extreme observations; see Stahel (1981), Donoho (1982), Tyler (1983, 1994), Rousseeuw (1984, 1985), Tamura and Boos (1986), Davies (1987), Lopuhaä (1991), Woodruff and Rocke (1994) and Kent and Tyler (1996). While the problems of high breakdown and efficient computation have been considered, one problem has been largely ignored: with real data one often encounters the problem of missing observations. There are several reasons why this problem is important, especially in the social and economics sciences, where missing values are the rule rather than the exception.

Let $\mathbf{y}_i$ be the ith of n observations on a p-variate distribution with mean μ and covariance Σ. It is often supposed that the distribution is multivariate normal or more generally an elliptical distribution. 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. In other words, $\mathbf{y}_i = [\mathbf{y}_{[oi]}^{\mathrm{T}}, \mathbf{y}_{[mi]}^{\mathrm{T}}]^{\mathrm{T}}$, so that a distinction is made between the observed (oi) and the missing (mi) data. According to Rubin (1976), missing values are usually assumed to be either missing at random (MAR), missing completely at random (MCAR), or neither MAR nor MCAR. An important condition for the missing data to be MAR is that their (missing) value is independent of the fact that they are missing. For example, they cannot be missing because they exceed a given threshold (see Little & Rubin, 1987). MCAR is a stronger hypothesis than MAR, but the latter is sufficient for correct likelihood-based inferences. In this paper, we therefore assume that data are at least MAR. One could ignore the missing values in that the vectors $\mathbf{y}_i$ containing missing observations are discarded and then proceed to apply the maximum likelihood (ML) or a robust estimator to estimate the parameters. However, this procedure has two important drawbacks: first, one could lose a lot of information by reducing the sample size considerably when only a few 'items' are missing, thus making the estimators less efficient; and second, the procedure could lead to a sample of size too small for any parameters to be estimable (if the size is smaller than p or even nil). In particular, when using robust M-estimators or S-estimators as proposed in Yuan and Bentler (1998), the sample size should be considerably larger than p.

Classically, to estimate the mean and covariance from a multivariate sample one uses the ML estimator by assuming independent and identically distributed observations

from the multivariate normal distribution. When there are missing data, these are replaced by their expected value in the ML estimating equations and the EM algorithm (Dempster, Laird, & Rubin, 1977) is used to compute the ML estimator. The EM algorithm is an iterative procedure which switches between an expectation (E) step in which the expected values of the missing data are computed and a maximization (M) step in which the ML estimating equations are solved.

Little and Smith (1987) and Little (1988) propose basing the M-step on a robust estimator belonging to the general class of M-estimators (see Huber, 1981). This rather *ad hoc* procedure has been suggested because it is known that in general the ML estimator is not robust to small model departures or data contamination. Robust estimators are built to be resistant to model misspecifications in general and outliers in particular. In large dimensions, however, the choice of the M-estimator is important, since some of them are known to have a breakdown point of at most $1/(p + 1)$ (Maronna, 1976) which can be rather small in high dimensions. This means that if the proportion of outliers exceeds $1/(p + 1)$ (or even if it is near to this value), the robust estimator is no longer robust. This can happen because there are two types of robustness: infinitesimal and global. The first is concerned with the effect of infinitesimal model deviations as measured by *IF* and, therefore, estimators with a bounded *IF* are said to be robust in that sense. The second is concerned with the maximal amount of model misspecification (for example, proportion of extreme data) the estimators can withstand before they 'break down' or their bias becomes arbitrarily large (see also Hampel *et al.*, 1986). High-breakdown point estimators are robust in the latter sense (as well as in the former sense). Such estimators are desirable when robust estimators in the infinitesimal sense have low breakdown points. In this paper we propose high-breakdown estimators for the mean and covariance when there are missing data. This is achieved by adapting redescending M-estimators to the case of missing data.

In Section 2, we first highlight the robustness problems of the ML estimator and then present the ER algorithm and its limitations. Applications of high-breakdown point estimators to incomplete data are developed in Section 3. The estimators are then compared by means of simulated and real data sets in Section 4. Finally, in Section 5 we also provide details of where routines can be found to compute high-breakdown point estimators with missing data. These routines are in the form of an S-PLUS library which is easy to implement and easy to use for the non-specialist.

2. The ML estimator and the ER algorithm

We first describe the ML estimation of the mean and covariance matrix by means of the EM algorithm. Then, analytically and through a real data set, we show that outliers may spoil the estimates. The ER algorithm is then also presented and discussed.

2.1. Robustness properties of the ML estimator with missing data

In our case, we need to estimate the parameters μ and Σ, the mean and covariance of the underlying multivariate distribution. For notational convenience, let $\theta = [\mu^T, \text{vech}(\Sigma)^T]^T$, where the function vech stacks the non-duplicated elements of Σ into a $p/(p + 1)/2$ column vector. The objective function is

$$\max_{\theta} \sum_{i=1}^{n} \log f(\mathbf{y}_i, \theta),$$

where f is the density of the postulated distribution (here the multivariate normal distribution). For distributions of the exponential family, this is equivalent to solving for θ

$$E[\mathbf{t}(\mathbf{y}) \mid \theta] - E[\mathbf{t}(\mathbf{y}) \mid \mathbf{y}_{[o]}, \theta] = \mathbf{0}, \tag{1}$$

where $\mathbf{t}(\mathbf{y})$ are sufficient statistics. Equation (1) defines the ML estimator with MAR values. Note that (1) may have multiple solutions when there are missing values. If the multivariate normal distribution is postulated, (1) becomes

$$\sum_{i=1}^{n} s(\mathbf{y}_i; \theta) = \sum_{i=1}^{n} \begin{bmatrix} \mu - \hat{\mathbf{y}}_i \\ \text{vech}(\Sigma) - \text{vech}((\hat{\mathbf{y}}_i - \mu)(\hat{\mathbf{y}}_i - \mu)^{\mathrm{T}}) - \text{vech}(\mathbf{C}_i) \end{bmatrix} = \mathbf{0}, \tag{2}$$

where s is the score function,

$$\begin{aligned} \hat{y}_{ij} &= \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed,} \\ E[y_{ij} \mid \mathbf{y}_{[oi]}, \theta], & \text{if } y_{ij} \text{ is missing,} \end{cases} \\ &= \begin{cases} y_{ij} & \text{if } y_{ij} \text{ is observed,} \\ \mu_{[mi]} + \Sigma_{[moi]}\Sigma_{[ooi]}^{-1}(\mathbf{y}_{[oi]} - \mu_{[oi]}) & \text{if } y_{ij} \text{ is missing,} \end{cases} \end{aligned} \tag{3}$$

and

$$\begin{aligned} C_{ijk} &= \begin{cases} 0, & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed,} \\ \text{Cov}[y_{ij}, y_{ik} \mid \mathbf{y}_{[oi]}, \theta], & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing,} \end{cases} \\ &= \begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed,} \\ \Sigma_{[mmi]} - \Sigma_{[moi]}\Sigma_{[ooi]}^{-1}\Sigma_{[omi]} & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing,} \end{cases} \end{aligned} \tag{4}$$

where, for example, $\Sigma_{[ooi]}$ denotes the partition of Σ corresponding to the observed part of $\mathbf{y}_i$, etc.

There is no analytical solution to (2) and therefore one can use the EM algorithm (Dempster *et al.*, 1977) to solve the equations. The EM algorithm is an iterative computational method to find ML estimates of parameters when the data are not fully observed. The special case of estimating the mean and the covariance matrix from incomplete multivariate data has also been discussed, among others, by Beale and Little (1975) and Little and Rubin (1987). The EM algorithm switches between an E-step in which the $\hat{\mathbf{y}}_i$ and $\mathbf{C}_i$ are computed given values of μ and Σ, and an M-step in which (2) is solved using $\hat{\mathbf{y}}_i$ and $\mathbf{C}_i$ computed in the E-step. One can also use the sweep operator of Beale and Little (1975) to ease the programming.

To assess the robustness properties of any statistic, for example an estimator, one can use the influence function IF. For M-estimators (the ML belongs to this class) it is known to be proportional to the score function (see Hampel *et al.*, 1986). Let $\mathbf{z} = [\mathbf{z}_{[o]}^{\mathrm{T}} \mathbf{z}_{[m]}^{\mathrm{T}}]^{\mathrm{T}}$ be any point in the p-dimensional space from which we observe only a part ($\mathbf{z}_{[o]}$). The influence function of the ML with missing values is proportional to

$$[IF_{\mu_o}^{\mathrm{T}}\, IF_{\mu_m}^{\mathrm{T}}\, IF_{\Sigma_{oo}}^{\mathrm{T}}\, IF_{\Sigma_{om}}^{\mathrm{T}}\, IF_{\Sigma_{mm}}^{\mathrm{T}}]^{\mathrm{T}},$$

where

$$\begin{aligned} IF_{\mu_o} &= [\mu_{[o]} - \mathbf{z}_{[o]}], \\ IF_{\mu_m} &= [\Sigma_{[mo]}\Sigma_{[oo]}^{-1}(\mu_{[o]} - \mathbf{z}_{[o]})], \\ IF_{\Sigma_{oo}} &= [\text{vech}(\Sigma_{[ooi]}) - \text{vech}((\mu_{[o]} - \mathbf{z}_{[o]})(\mu_{[o]} - \mathbf{z}_{[o]})^{\mathrm{T}})] \end{aligned}$$

$$IF_{\Sigma_{om}} = [\text{vech}(\Sigma_{[om]}) - \text{vech}((\mu_{[o]} - \mathbf{z}_{[o]})(\mu_{[o]} - \mathbf{z}_{[o]})^{\mathrm{T}}\Sigma_{[oo]}^{-1}\Sigma_{[om]})],$$

$$IF_{\Sigma_{mm}} = [\text{vech}(\Sigma_{[mo]}\Sigma_{[oo]}^{-1}\Sigma_{[om]}) - \text{vech}(\Sigma_{[mo]}\Sigma_{[oo]}^{-1}(\mu_{[o]} - \mathbf{z}_{[o]})(\mu_{[o]} - \mathbf{z}_{[o]})^{\mathrm{T}}\Sigma_{[oo]}^{-1}\Sigma_{[om]})].$$

This shows that the ML is not robust since its *IF* is unbounded for arbitrary values of $\mathbf{z}$. Actually, the influence of outliers when there are missing values is even worse than in the complete-data case. Indeed, an extreme value in the observed part of $\mathbf{z}$ not only influences the corresponding part in the mean vector and covariance matrix, but also the non-observed part of the latter. In other words, IF_{μ_m} depends on $\mathbf{z}_{[o]}$, and so also does $IF_{\Sigma_{om}}$ and $IF_{\Sigma_{mm}}$.

2.2. Example: Working memory data

We illustrate here the non-robustness of the ML estimator with missing data by means of data belonging to a data set collected for the study of age differences in working memory (see Ribaupierre and Ludwig, 2000). The data were collected on a group of 98 men and women aged 56 and over who performed a set of different tasks: box crossing (Baddeley, Della Sala, Gray, Papagno, & Spinnler, 1997); logical memory, which is a subtest taken from the Wechsler Memory Scale—Revised (Wechsler, 1987); and the continuous monitoring task (Kray, Frensch, & Lindenberger, 1996). The box crossing task is a combination of a verbal memory span and the crossing of boxes on a sheet of paper. The scores considered here are the number of crosses made on the single (BCXS) and dual (BCXD) conditions of the task. The logical memory task is either an immediate (ML1TOT) or a postponed (ML2TOT) story recall task. Finally, in the continuous monitoring task (CMT), which is computerized, participants adjust a half disc to a model which changes in either size, colour or both. The scores we consider are time needed for adjusting the half disc in size along (CMTTMSS), colour alone (CMTTMCS) and in size and colour in dual condition (CMTTMSD and CMTTMCD respectively). The data are incomplete in that for 22 subjects not all the scores have been recorded. We can suppose that the missing data are MAR.

In theory, one would expect a relatively strong within-task correlation, and a relatively strong negative correlation between the box crossing and the CMT scores because they are both connected to a processing speed factor (Salthouse, 1996). On the other hand, the correlation between these variables and the logical memory scores should be weaker. One can look at the scatter plots of the data (see Fig. 1) to see if these expectations are observed. On the whole, we indeed can see a relatively strong within-class correlation, except that for the logical memory task (ML1TOT and ML2TOT) there seem to be a few participants whose score on ML1TOT is weak whereas their score on ML2TOT is high, and others for whom the relationship is the other way round. For the majority, however, the correlation looks strong and positive. The correlation between the scores on the box crossing and the CMT does indeed look negative and probably also relatively strong, but for some participants the relationship between these scores seems to deviate from that of the majority.

Assuming that the complete data are from a multivariate normal distribution, we apply the EM algorithm to find the ML estimator of $\theta = [\mu^{\mathrm{T}}, \text{vech}(\Sigma)^{\mathrm{T}}]^{\mathrm{T}}$ with starting value the ML estimator computed on the data where the missing values have been replaced by the median value of the corresponding observed variables. The resulting estimates will be identified by the subscript EM. For the sake of clarity, we present here

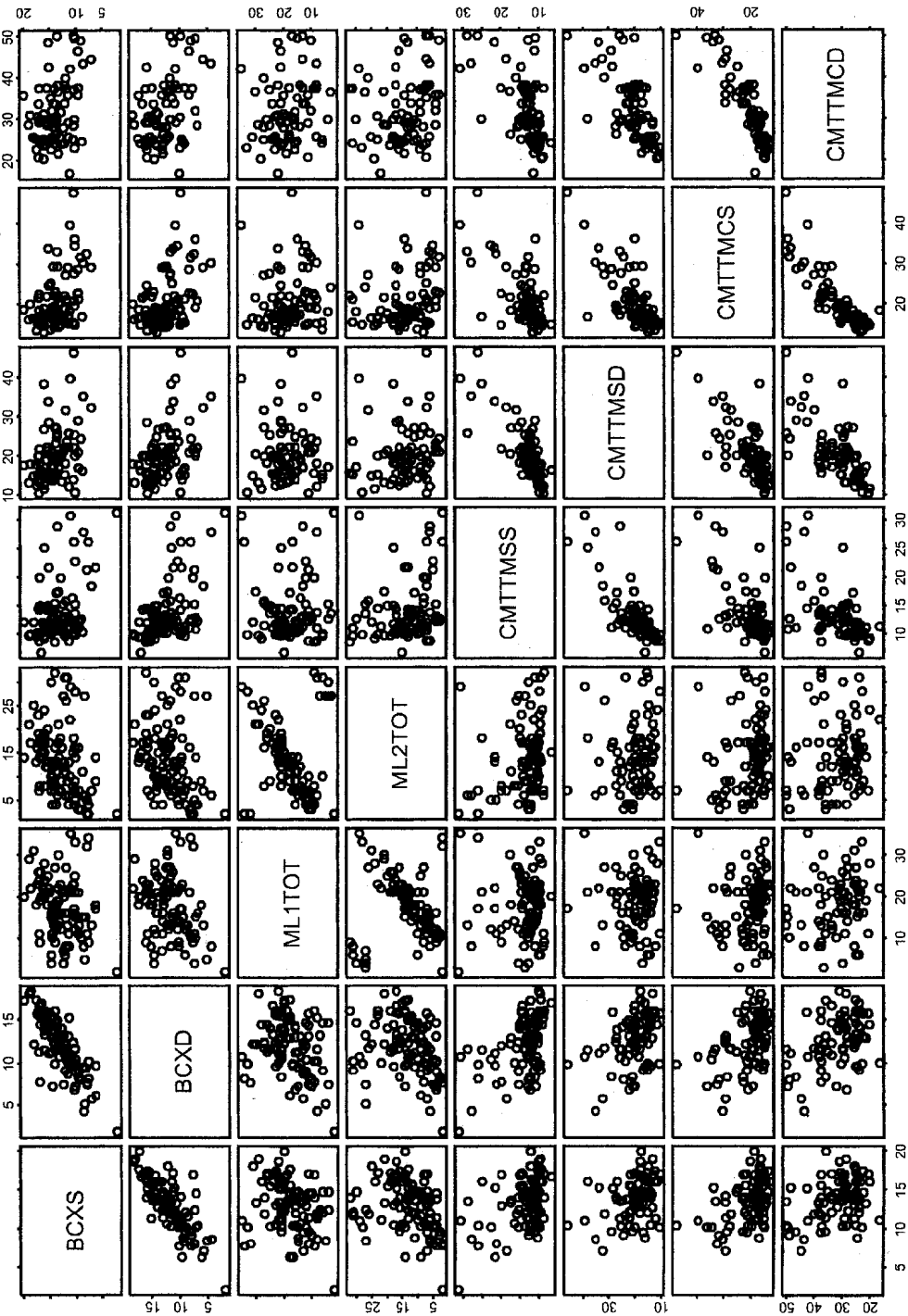

Figure 1. Scatter plots of the working memory data.

the results for the correlation matrix Φ

$$\mu_{EM} = [12.90 \quad 11.84 \quad 17.57 \quad 13.56 \quad 14.24 \quad 21.93 \quad 21.94 \quad 33.05],$$

$$\Phi_{EM} = \begin{bmatrix} & \text{BCXS} & \text{BCXD} & \text{ML1TOT} & \text{ML2TOT} & \text{CMTTMSS} & \text{CMTTMSD} & \text{CMTTMCS} & \text{CMTTMCD} \\ \text{BCXS} & 1.0 & 0.80 & 0.28 & 0.31 & -0.43 & -0.54 & -0.59 & -0.52 \\ \text{BCXD} & & 1.0 & 0.29 & 0.26 & -0.51 & -0.53 & -0.55 & -0.53 \\ \text{ML1TOT} & & & 1.0 & 0.20 & -0.01 & -0.02 & -0.13 & -0.23 \\ \text{ML2TOT} & & & & 1.0 & -0.25 & -0.27 & -0.28 & -0.26 \\ \text{CMTTMSS} & & & & & 1.0 & 0.84 & 0.72 & 0.65 \\ \text{CMTTMSD} & & & & & & 1.0 & 0.78 & 0.72 \\ \text{CMTTMCS} & & & & & & & 1.0 & 0.89 \\ \text{CMTTMCD} & & & & & & & & 1.0 \end{bmatrix}.$$

A robust procedure (which will be explained later) gives the following estimation results

$$\mu_{ERTBS} = [13.34 \quad 12.49 \quad 18.49 \quad 12.32 \quad 12.92 \quad 20.16 \quad 20.00 \quad 31.18], \tag{5}$$

$$\Phi_{ERTBS} = \begin{bmatrix} & \text{BCXS} & \text{BCXD} & \text{ML1TOT} & \text{ML2TOT} & \text{CMTTMSS} & \text{CMTTMSD} & \text{CMTTMCS} & \text{CMTTMCD} \\ \text{BCXS} & 1.0 & 0.81 & 0.40 & 0.46 & -0.42 & -0.49 & -0.50 & -0.44 \\ \text{BCXD} & & 1.0 & 0.29 & 0.41 & -0.45 & -0.45 & -0.48 & -0.43 \\ \text{ML1TOT} & & & 1.0 & 0.84 & -0.15 & -0.24 & -0.31 & -0.34 \\ \text{ML2TOT} & & & & 1.0 & -0.25 & -0.41 & -0.43 & -0.44 \\ \text{CMTTMSS} & & & & & 1.0 & 0.84 & 0.67 & 0.71 \\ \text{CMTTMSD} & & & & & & 1.0 & 0.73 & 0.77 \\ \text{CMTTMCS} & & & & & & & 1.0 & 0.88 \\ \text{CMTTMCD} & & & & & & & & 1.0 \end{bmatrix}. \tag{6}$$

It is interesting to note that the correlation matrices are on the whole not that different, except for the correlation between the two logical memory tasks (ML1TOT and ML2TOT). This correlation is very small when estimated by means of the classical ML estimator, whereas it appears very strong when a robust estimator is used. If one recalls the scatter plot of the data (see Fig. 1), this result is not surprising since we have already seen that the correlation between ML1TOT and ML2TOT seems strong for the majority of the participants, but a few of them do not seem to follow the same pattern. This example shows that outliers can bias the ML estimator, whereas the robust estimator is not so influenced by a few outlying subjects.

2.3. The ER algorithm

In the presence of contaminated multivariate data with missing values, Little and Smith (1987) proposed the expectation–robust (ER) algorithm which modifies in an *ad hoc* manner the EM algorithm so that extreme observations are downweighted. The estimated weights are based on the Mahalanobis distance. The algorithm is defined by

combining the usual E-step with the following robust modification (R-step):

$$\mu^{(t+1)} = \frac{\sum_{i=1}^{n} w_i \hat{\mathbf{y}}_i^{(t)}}{\sum_{i=1}^{n} w_i},$$

$$\text{vech}(\Sigma)^{(t+1)} = \frac{\sum_{i=1}^{n} w_i^2 \, \text{vech}(\hat{\mathbf{y}}_i^{(t)} - \mu^{(t+1)})(\hat{\mathbf{y}}_i^{(t)} - \mu^{(t+1)}) + \text{vech}(\mathbf{C}_i^{(t)})}{\sum_{i=1}^{n} w_i^2 - 1},$$

where $w_i = w(d_i^*)/d_i^*$, and

$$(d_i^*)^2 = (\hat{\mathbf{y}}_i - \mu)^{\mathrm{T}} (\Sigma)^{-1} (\hat{\mathbf{y}}_i - \mu) \tag{7}$$

is the squared Mahalanobis distance at iteration t (i.e. with $\hat{\mathbf{y}}_i^{(t)}$, $\mu^{(t)}$ and $\Sigma^{(t)}$). The vector of filled-in values $\hat{\mathbf{y}}_i$ and $\mathbf{C}_i$ are defined in (3) and (4). Here w denotes a two-parameter bounded-influence function (Hampel, 1974) defined by

$$w(d_i^*) = \begin{cases} d_i^*, & \text{if } d_i^* \leq d_{oi}, \\ d_{oi} \exp\{-(d_i^* - d_{oi})^2 / 2 b_2^2, & \text{if } d_i^* > d_{oi}, \end{cases} \tag{8}$$

where $d_{oi} = \sqrt{p_i} + b_1/2$ 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 data analyst. The choice of b_1 determines the cut-off, and b_2 specifies how rapidly the weights decrease. Based on Hampel (1973), Little and Smith (1987) suggested $b_1 = 2$ and $b_2 = 1.25$. If case i is uncontaminated, the data are normal and missing values are MAR, then (7) is asymptotically $\chi^2_{p_i}$. The Wilson–Hilferty transformation of the chi-square distribution (see Kendall & Stuart, 1969, Chapter 16) yields

$$((d_i^*)^2/p_i)^{1/3} \dot{\sim} N(1 - 2/(9p_i), 2/(9p_i)). \tag{9}$$

In order to detect atypical observations, Little and Smith (1987) therefore proposed a probability plot of

$$Z_i = \frac{((d_i^*)^2/p_i)^{1/3} - 1 + 2/(9p_i)}{\sqrt{[2/(9p_i)]}} \tag{10}$$

versus standard normal order statistics in which d_i^* are computed using the ML estimates of μ and Σ obtained by the ER algorithm. Little and Smith (1987) propose as starting value the ML estimator computed on the data where the missing values have been replaced by the median values of the corresponding observed variables. The resulting estimates will be identified by the subscript ER.

2.4. Robustness properties of the ER algorithm

Although the ER algorithm is relatively simple to implement, it suffers from an important drawback: its breakdown point is low. This is essentially due to the starting point of the algorithm and also to the form of the weights. We now illustrate this problem with simulated data. Since the estimators we study here are all affine equivariant, the choice of covariance matrix is arbitrary and therefore we chose to generate 50 data points from a multivariate standard normal distribution $MN(\mathbf{0}, \mathbf{I})$. We constructed so-called shift outliers (see Rocke & Woodruff, 1996) which are well known to be the hardest to detect. They are built by adding the quantity $r\sqrt{0.999 p^{-1} (\chi^2_p)^{-1}}$ to all components of some of the data. r roughly represents the importance of the shift added to the data, and we chose $r = 2$ for the first 10 observations and $r = 0$ for the others. Thus 20% of the data are outliers. We also randomly removed 25 elements of the data matrix. It should be stressed that we tried smaller amounts of contamination as well as $r = 1.5$ and $r = 4$, and we found similar results to the case we present here.

The ML estimator computed using the EM algorithm gives the following results:

$$\mu_{EM} = [0.631 \quad 1.03 \quad 0.977 \quad 0.804 \quad 1.04],$$

$$\Sigma_{EM} = \begin{bmatrix} 4.35 & 3.31 & 3.35 & 2.98 & 2.98 \\ & 3.81 & 3.52 & 2.58 & 2.94 \\ & & 4.19 & 2.86 & 2.86 \\ & & & 3.50 & 2.71 \\ & & & & 3.42 \end{bmatrix}.$$

We can see that the outlying observations have a large influence on the estimates. Indeed, the variances are overestimated, as are the covariances. Actually a correlation is found between the variables which are supposed to be independent. By using the ER algorithm it is hoped, however, that the outlying observations do not have such an effect. The estimation results are

$$\mu_{ER} = [0.647 \quad 1.05 \quad 0.989 \quad 0.792 \quad 1.04],$$

$$\Sigma_{ER} = \begin{bmatrix} 4.37 & 3.34 & 3.40 & 3.13 & 3.03 \\ & 3.89 & 3.59 & 2.69 & 3.02 \\ & & 4.28 & 2.96 & 2.94 \\ & & & 3.59 & 2.80 \\ & & & & 3.52 \end{bmatrix}.$$

EM and ER lead to similar estimated values for μ and Σ. These estimators are clearly influenced by the outlying observations. This shows that ER may fail to be robust when the proportion of outliers is relatively large.

Figure 2 presents the transformed distances (10) on the simulated data when using several estimators. We can see that when one uses the ER algorithm, none of the contaminated observations is revealed as an outlier when we know that there are ten of them (the value of 1.96 is taken as a benchmark for detecting outliers). The ER algorithm is therefore not satisfactory in situations of this kind. We will return to these data when we discuss high-breakdown estimators which actually are able to detect outliers when they are relatively numerous.

3. High-breakdown estimators in incomplete data

We propose two strategies for constructing a high-breakdown point estimator of the mean and covariance in multivariate data with missing values. The first is to provide a high-breakdown point estimator as starting value for the ER algorithm, and the second is to adapt a high-breakdown estimator to incomplete data. For the latter we also need a high-breakdown estimator as a starting point for the algorithm. We propose here to adapt the minimum covariance determinant (MCD) estimator to the case of missing values. But why not just consider the MCD estimator alone as a high-breakdown estimator for the mean and covariance? The problem is that it is known to be very inefficient, so that usually it is used as a starting point for more efficient estimators such as M-estimators.

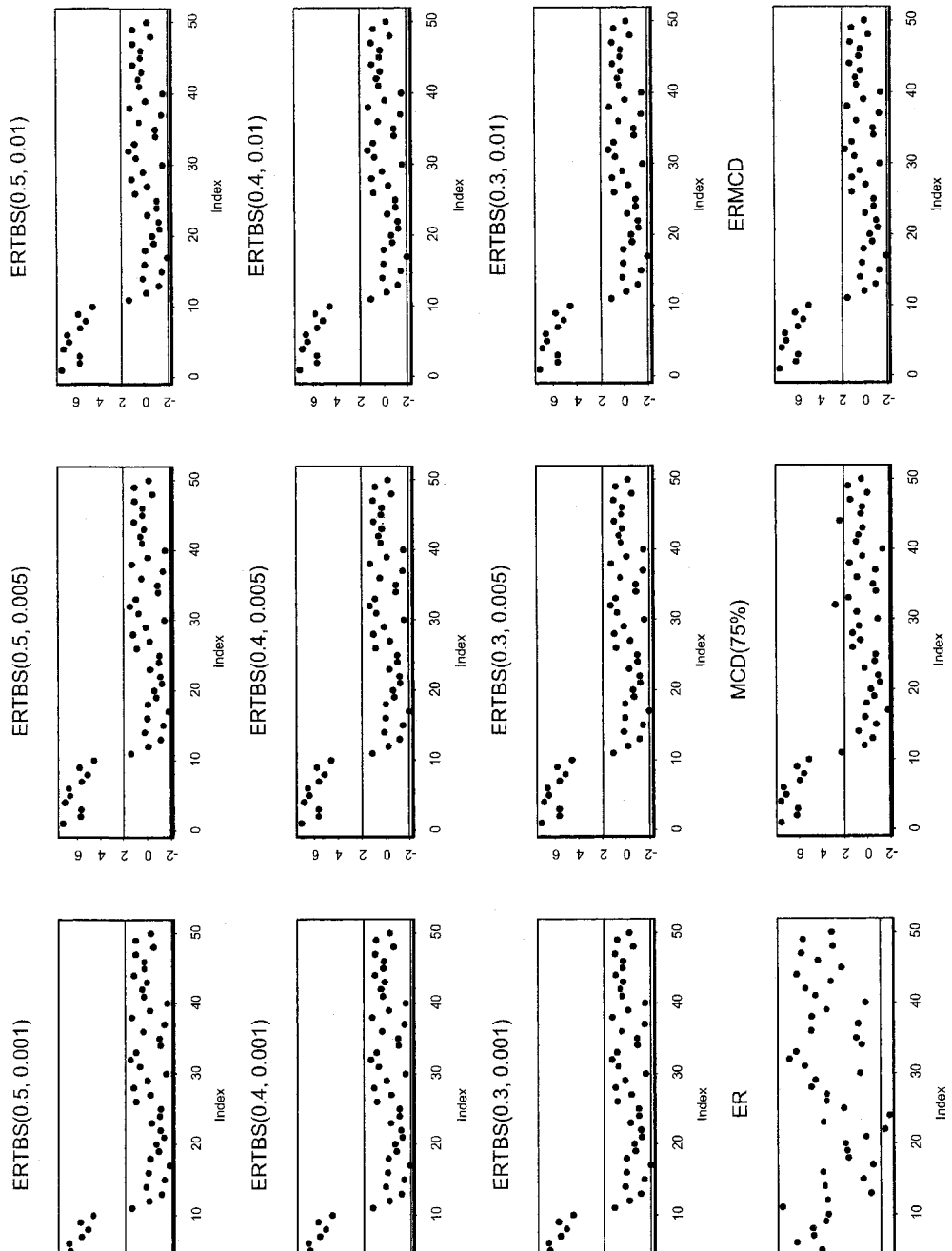

Figure 2. Index plots of transformed distances for the simulated data.

3.1. The MCD estimator in incomplete data

We now present the MCD estimator and an algorithm to compute it when the data are incomplete.

3.1.1. The MCD estimator

The minimum covariance determinant estimator is given by the sample mean and covariance of the subset of h observations for which the determinant of their covariance matrix is minimal. The MCD mean estimator is then the sample mean of those h points, and the MCD covariance estimator is their sample covariance matrix. The usual value of h, that which achieves the highest breakdown point, is $h = \lfloor (n + p + 1)/2 \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of x. Such a choice gives a breakdown point of nearly 50%, but also the larger efficiency loss with respect to the ML estimator at the true model (i.e. with no data contamination). We can reasonably consider a smaller value for the breakdown point of, say, 25% or 20% and therefore choose $h = \lfloor 0.75n \rfloor$ or $h = \lfloor 0.8n \rfloor$ to increase the efficiency of the MCD when the sample is not suspected to be heavily contaminated. For multivariate data sets, it takes too much time to find the exact estimate, so an approximation is computed. We present here the forward search algorithm to compute (an approximation to) the MCD.

The MCD estimator is affine equivariant but is not the only high-breakdown point affine-equivariant estimator. The minimum volume ellipsoid (Rousseeuw, 1984) is also a high-breakdown point affine-equivariant estimator. However, Butler, Davies, and Jhun (1993) show the consistency and asymptotic normality of the MCD estimator of the multivariate mean, and the consistency of that of the covariance, with a rate of convergence of $n^{-1/2}$ compared to $n^{-1/3}$ for the minimum volume ellipsoid.

Several numerical algorithms have been proposed for computing the MCD. Atkinson and Cheng (2000) show that by using the forward search algorithm (see below), the resulting procedure is fast—in particular, it is much faster than the feasible solution algorithm of Hawkins (1994). They also provide a procedure for choosing the right value for h. In the following subsection, we adapt the procedure to the problem of missing data, although the same ideas could in principle be applied to any procedure such as the one proposed by Rousseeuw and Van Driessen (1999), which is suspected to be even faster than the forward search algorithm in large-sample problems.

For complete data, the forward search algorithm as presented by Atkinson (1994) can be summarized for the MCD estimator by the following pseudo-code. Given values for s (to be discussed later) and a subset Q_k of q_k observations:

1. Compute the sample mean and covariance, $\mathbf{y}(Q_k)$ and $\mathbf{S}(Q_k)$.
2. Order all n observations according to increasing Mahalanobis distances computed using $\mathbf{y}(Q_k)$ and $\mathbf{S}(Q_k)$.
3. Choose the first h observations, compute their sample mean $\mathbf{y}_h(Q_k)$ and covariance matrix $\mathbf{S}_h(Q_k)$ and its determinant D_k.
4. If $D_k < D_{k-1}$, replace the previous values of μ_{MCD}, Σ_{MCD} and D by $\bar{\mathbf{y}}_h(Q_k)$, $\mathbf{S}_h(Q_k)$ and D_k, respectively.
5. If $q_k = n$ stop, else choose the first $q_{k+1} = q_k + s \leq n$ observations of the ordered sample in step 2 which define a new subset Q_{k+1} and go to step 1 by replacing Q_k by Q_{k+1}.

The algorithm starts with a randomly chosen subset of size $q_1 = p + 1$. One forward search finds an MCD estimator μ_{MCD} and Σ_{MCD} with minimum determinant D. The forward search procedure is in fact repeated for several randomly chosen initial subsets.

In our experience 100 initial samples is a good choice, but there is no theoretical result on the optimal choice. The final MCD estimator is then given by the MCD of the search with minimum determinant D. If for one search $\mathbf{S}(Q_k)$ becomes singular, the search is cancelled and replaced by another one. Finally, the choice of the increment s is usually $s = 1$, although it is suspected that larger values would increase the speed of the algorithm without jeopardizing too much the probability of finding the MCD. To our knowledge, no research has yet been done on the choice of s.

3.1.2. MCD with incomplete data

The MCD estimator with missing data can easily be defined by computing the sample mean and covariance via the EM algorithm of the subset of h observations for which the determinant of their covariance matrix is minimal. The forward search algorithm is then adapted here to the case of missing data. However, although with complete data the usual choice is $h = \lfloor (n + p + 1)/2 \rfloor$, we found that with incomplete data this value is too small, probably because of the loss of information due to the missing data. We have not studied the relationship between the value of h and the percentage of missing data. We can only recommend taking larger values of h; in our experience $h = \lfloor 0.75n \rfloor$ or $h = \lfloor 0.8n \rfloor$ are reasonable values when the proportion of contaminated data is not expected to be greater than 20–25%

Adapting the procedure to the case of missing data is straightforward. Basically, we use the EM algorithm to compute the sample mean and covariance in the process of the forward searches when some of the observations in the subset Q_k are missing. Thus, the only difference between this and the complete data case lies in the calculation of the Mahalanobis distances used to order the observations. If for the ith observation there are missing values, the Mahalanobis distances are based on the observed values, leading to the distances

$$d^2_{[oi]}(Q_k) = (\mathbf{y}_{[oi]} - \bar{\mathbf{y}}(Q_k)_{[oi]})^{\mathrm{T}} \mathbf{S}(Q_k)^{-1}_{[ooi]} (\mathbf{y}_{[oi]} - \bar{\mathbf{y}}(Q_k)_{[oi]}). \tag{11}$$

To order the distances and take into account the non-equal number of missing values for each observations, we need to standardize the distances. We selected the Wilson–Hilferty transformation of the chi-squared distribution given in (10) to order the observations in step 2. The reason for this choice is that we suspect that using (7) with imputed values instead of (11) with the Wilson–Hilferty transformation would give an advantage to observations with missing values. The reason is that the imputed values in (7) are nearer to the estimated vector of the mean and therefore have smaller Mahalanobis distance than full observations with similar values for the observed part. With (11), the non-observed part is not taken into account in the computation of the Mahalanobis distance, and the latter is standardized for the number of observed values by means of the Wilson–Hilferty transformation. We have, however, no proof that our statement is correct. Simulations have actually shown no significant differences between the two possible approaches in that the forward search algorithm led to the same or similar estimates. In what follows, the resulting estimates will be identified by the subscript ERMCD when the MCD is used as a starting point for the ER algorithm.

3.2. The TBS estimator

The best-known high-breakdown point estimators are actually S-estimators, first proposed by Rousseeuw and Leroy (1987, p. 263). In particular, the translated-biweight

S-estimator (TBS), proposed by Rocke (1996), belongs to this class. In what follows, we will show how the estimating equations for the TBS estimator and the TBS estimator for missing data can be seen as special cases of

$$\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} w_i^{\mu}(\mu - \hat{\mathbf{y}}_i) \\ w_i^{\delta}\,\mathrm{vech}(\Sigma) - w_i^{\eta}(\mathrm{vech}((\hat{\mathbf{y}}_i - \mu)(\hat{\mathbf{y}}_i - \mu)^{\mathrm{T}}) - \mathrm{vech}(\mathbf{C}_i)) \end{bmatrix} = \mathbf{0}. \tag{12}$$

Equation (12) actually also defines an M-estimator which generalizes (2) by incorporating weights. With missing data $\hat{\mathbf{y}}_i$ and $\mathbf{C}_i$ are given by (3) and (4), whereas with complete data $\hat{\mathbf{y}}_i = \mathbf{y}_i$ and $\mathbf{C}_i = \mathbf{0}$.

An S-estimator of multivariate mean and covariance is defined as the solution in μ and Σ which minimizes $|\Sigma|$ subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho(((\mathbf{y}_i - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{y}_i - \mu))^{1/2}) = \frac{1}{n}\sum_{i=1}^{n}\rho(d_i) = b_0, \tag{13}$$

where ρ is a non-decreasing function which usually satisfies $E_{\chi_p^2}[\rho(d)] = b_0$. The breakdown point is given by the ratio of b_0 to the maximum of ρ (see Lopuhaä & Rousseeuw, 1991). Therefore, b_0 is usually computed for a chosen breakdown ε^* and a ρ-function by means of

$$b_0 = \varepsilon^* \max_d \rho(d). \tag{14}$$

It is known that such an S-estimator also satisfies the equations of an M-estimator of mean and covariance defined by (12) in which

$$w_i^{\mu} = \nu_1(d_i) = k\psi(d_i/k)/d_i, \tag{15}$$

$$w_i^{\eta} = \nu_2(d_i) = p\,\nu_1(d_i/k), \tag{16}$$

$$w_i^{\delta} = \nu_3(d_i) = \psi(d_i/k)\,d_i/k, \tag{17}$$

with $\hat{\mathbf{y}}_i = \mathbf{y}_i$, $\mathbf{C}_i = \mathbf{0}$, $\psi(d) = \partial/\partial d\,\rho(d)$, d_i are the Mahalanobis distances

$$d_i^2 = (\mathbf{y}_i - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{y}_i - \mu), \tag{18}$$

and k is such that

$$\frac{1}{n}\sum_{i=1}^{n}\rho(d_i/k) = b_0. \tag{19}$$

Rocke (1996) showed that an S-estimator can be found iteratively once b_0 has been set in (14), by first computing the scaling factor k for the Mahalanobis distances (computed using current values of μ and Σ) in (19), then the weights in (15), (16) and (17). The estimates of μ and Σ are then updated in (12) given these weights. Rocke (1996) also proposes replacing the standardization step (19) with one that consists of equating the median of $\rho(d_i)$ with the median under normality. In this case, the scaling factor k would be

$$k = \frac{d_{(q)}}{\sqrt{\chi_p^2(q/(n+1))}}, \tag{20}$$

where $d_{(q)}$ denotes the qth ordered distance, and $q = \lfloor (n + p + 1)/2 \rfloor$.

Equation (13) defines general S-estimators which depend on the choice of the ρ-function or its derivative, the ψ-function. A common choice for the function ψ is Tukey's biweight. However, as Rocke (1996) argues, in high dimensions it fails to

downweight outliers with large distances. This is measured by using the concept of *asymptotic rejection probability* (ARP) which can be interpreted as the probability of an estimator, in large samples under a reference distribution, giving a null (or nearly null) weight. Although the ARP should be small for the sake of efficiency, it is useful to be able to downweight points that are very improbable under the null model. Rocke (1996) shows that the ARP of the S-estimator based on the biweight function tends to 0 as the dimension p rises. This means that points lying far away from the centre of the data are not downweighted when p is large. Therefore he proposes a modified biweight estimator, namely the TBS estimator defined through

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

The corresponding ρ function is given, for $M \le d \le M + c$, by

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

and for all d by

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

The parameters c and M can be chosen to give the desired breakdown point ε^* and ARP α, i.e.

$$\varepsilon^* \max_d \rho(d; c, M) = E_{\chi^2_p}[\rho(d; c, M)] = b_0,$$

$$M + c = \sqrt{(\chi^2_p)^{-1}(1 - \alpha)}.$$

The choices of ε^* and α are to be made by the analyst. The former is the suspected maximal amount of contaminated data, and for the latter we propose choices between 0.1% and 1%

Rocke (1996) discusses several choices for the function ρ defining the S-estimator. The striking feature is that whatever the choice, what remains very important is the starting point of the algorithms. Indeed, (12) admits several solutions which depend on the starting point of the algorithms. Even the TBS estimator can loose its high-breakdown properties if the starting point is not a high-breakdown point estimator, as would be the case if one chooses the sample mean and covariance on the whole data set. We therefore recommend using, for example, the MCD estimator computed by means of the forward search algorithm as a starting point for the TBS estimator.

3.3. The TBS estimator with incomplete data

When the data are incomplete and an S-estimator is preferred to a monotone M-estimator, then weights in (12) can be chosen accordingly and $\hat{\mathbf{y}}_i$ and $\mathbf{C}_i$ are given by (3) and (4). As for the MCD estimator, one has to choose how to define the Mahalanobis distances (18) used in the computation of the weights. We propose basing the weights on Mahalanobis distances computed on the observed values, i.e.

$$d^2_{[oi]} = (\mathbf{y}_{[oi]} - \mu_{[oi]})^{\mathrm{T}} \Sigma^{-1}_{[ooi]} (\mathbf{y}_{[oi]} - \mu_{[oi]}). \tag{21}$$

We then obtain the following system defining the S-estimator of μ and Σ with missing data

$$\frac{1}{n} \sum_{i=1}^{n} \nu_1(d_{[oi]}) \begin{bmatrix} \mathbf{y}_{[oi]} - \mu_{[oi]} \\ \Sigma_{[moi]} \Sigma^{-1}_{[ooi]} (\mathbf{y}_{[oi]} - \mu_{[oi]}) \end{bmatrix} = \mathbf{0}, \tag{22}$$

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \nu_2(d_{[oi]}) \begin{bmatrix} \mathbf{S}_{[ooi]} & \mathbf{S}_{[omi]} \\ \mathbf{S}_{[moi]} & \mathbf{S}_{[mmi]} \end{bmatrix} - \nu_3(d_{[oi]}) \Sigma \right\} = \mathbf{0}, \tag{23}$$

where

$$\begin{aligned} \mathbf{S}_{[ooi]} &= (\mathbf{y}_{[oi]} - \mu_{[oi]})(\mathbf{y}_{[oi]} - \mu_{[oi]})^{\mathrm{T}}, \\ \mathbf{S}_{[omi]} &= \mathbf{S}^{\mathrm{T}}_{[moi]} = (\mathbf{y}_{[oi]} - \mu_{[oi]})(\mathbf{y}_{[oi]} - \mu_{[oi]})^{\mathrm{T}} \Sigma^{-1}_{[ooi]} \Sigma_{[omi]}, \\ \mathbf{S}_{[mmi]} &= \Sigma_{[moi]} \Sigma^{-1}_{[ooi]} (\mathbf{y}_{[oi]} - \mu_{[oi]})(\mathbf{y}_{[oi]} - \mu_{[oi]})^{\mathrm{T}} \Sigma^{-1}_{[ooi]} \Sigma_{[omi]} \\ &\quad + \Sigma_{[mmi]} - \Sigma_{[moi]} \Sigma^{-1}_{[ooi]} \Sigma_{[omi]}. \end{aligned}$$

The TBS estimator for missing data can be found by using an iterative procedure like the one proposed by Rocke (1996) for S-estimators, to which we add an expectation step for computing the conditional expectations $\hat{\mathbf{y}}_i$ and $\mathbf{C}_i$ given current values of μ and Σ. In other words, the TBS estimator can be computed using an ER-type algorithm in which, given current values for μ and Σ, the quantities $\hat{\mathbf{y}}_i$, $\mathbf{C}_i$ and $d_i = d(\mathbf{y}_{[oi]})$ are computed in the E-step using respectively (3), (4) and (21). In the R-step, using $\hat{\mathbf{y}}_i$, $\mathbf{C}_i$ and d_i computed in the E-step, the quantities k and the weights w_i^{μ}, w_i^{δ} and w_i^{η} are computed by means respectively of (19) (or (20)), (15), (16) and (17), and finally the values of μ and Σ in (12) are updated considering $\hat{\mathbf{y}}_i$, $\mathbf{C}_i$, w_i^{μ}, w_i^{δ} and w_i^{η} as fixed.

We will call the resulting algorithm the expectation–robust algorithm based on the TBS estimator with missing data (ERTBS). It should be noted that in the case where all the weights are equal to 1, one obtains the ML estimator with missing data. We also propose using the MCD estimator as a starting point. The resulting estimates will be identified by the subscript ERTBS.

The robust estimator proposed by Little and Smith (1987) does not have the same form as in (12). Actually, the weights are not directly applied to the correction matrices $\mathbf{C}_i$, which in our opinion does not make it consistent. Therefore, to be fair in our comparisons, we slightly modify the R-step of the ER algorithm to

$$\mu^{(t+1)} = \frac{\sum_{i=1}^{n} w_i \hat{\mathbf{y}}_i^{(t)}}{\sum_{i=1}^{n} w_i}$$

$$\mathrm{vech}(\Sigma)^{(t+1)} = \frac{\sum_{i=1}^{n} w_i^2 \, \mathrm{vech}((\hat{\mathbf{y}}_i^{(t)} - \mu^{(t+1)})(\hat{\mathbf{y}}_i^{(t)} - \mu^{(t+1)})^{\mathrm{T}} + \mathbf{C}_i^{(t)})}{\sum_{i=1}^{n} w_i^2 - 1},$$

with w_i defined in (8).

4. Examples

In this section, we examine the estimators discussed in the previous sections on simulated and real data and compare the results.

4.1. Simulated data

We turn now to the simulated data presented in Section 2.3. We first compute the MCD estimator, which will be used as starting point for ER or ERTBS. We choose to base the MCD on $h = \lfloor 0.75n \rfloor$ and do 100 forward searches. The resulting estimate is used as starting point for the ER (modified version), and we obtain the following results:

$$\mu_{\text{ERMCD}} = [-0.259 \quad 0.122 \quad 0.113 \quad -0.0355 \quad 0.205],$$

$$\Sigma_{\text{ERMCD}} = \begin{bmatrix} 1.23 & 0.18 & 0.25 & 0.14 & 0.05 \\ & 0.75 & 0.39 & -0.26 & 0.12 \\ & & 1.02 & -0.02 & 0.02 \\ & & & 0.96 & 0.00 \\ & & & & 0.82 \end{bmatrix}.$$

It is clear that the results have been improved by a good starting point for ER. The estimates are of the same order of magnitude as the true values. A normal probability plot of the transformed distances Z_i is given in Figure 2 (ERMCD) and it shows that the outlying observations have been found and therefore their influence upon the estimates downweighted.

For the ERTBS estimator, a choice also needs to be made a priori for the breakdown point and ARP. We tried several combinations of values, which all lead to similar results. The corresponding normal probability plots are given in Figure 2 (the first value in parentheses is for the breakdown point, the second for ARP) where one can see that for all combinations, the outlying observations are detected and therefore their influence upon the estimates downweighted. We also found that the estimates are of the same order of magnitude as the true values and as the estimates provided by ERMCD.

4.2. Working memory data

In Section 2.2, we saw that the ML and a robust estimator gave quite different results on the working memory data. The robust estimates given in (5) and (6) are the ERTBS with breakdown point 50% and ARP 1%, i.e. a very robust estimator. Computing ERMCD using $h = \lfloor 0.75n \rfloor$ gives

$$\mu_{\text{ERMCD}} = [13.4 \quad 12.5 \quad 18.6 \quad 12.5 \quad 12.8 \quad 20.4 \quad 20.3 \quad 31.7],$$

$$\Phi_{ERMCD} = \begin{bmatrix} & \text{BCXS} & \text{BCXD} & \text{ML1TOT} & \text{ML2TOT} & \text{CMITTMSS} & \text{CMITTMSD} & \text{CMITTMCS} & \text{CMITTMCD} \\ \text{BCXS} & 1.0 & 0.84 & 0.43 & 0.51 & -0.36 & -0.49 & -0.49 & -0.46 \\ \text{BCXD} & & 1.0 & 0.29 & 0.42 & -0.44 & -0.48 & -0.51 & -0.48 \\ \text{ML1TOT} & & & 1.0 & 0.85 & -0.09 & -0.16 & -0.27 & -0.28 \\ \text{ML2TOT} & & & & 1.0 & -0.19 & -0.29 & -0.37 & -0.35 \\ \text{CMITTMSS} & & & & & 1.0 & 0.80 & 0.68 & 0.69 \\ \text{CMITTMSD} & & & & & & 1.0 & 0.75 & 0.77 \\ \text{CMITTMCS} & & & & & & & 1.0 & 0.90 \\ \text{CMITTMCD} & & & & & & & & 1.0 \end{bmatrix},$$

which are similar estimates to those of ERTBS. Compared to the classical ML estimator, a high-breakdown point estimator gives a different look at the data.

5. Conclusions

We have considered two alternatives for high-breakdown robust estimation of the mean and covariance of multivariate data when there are missing data. One is a modification of the ER algorithm of Little and Smith (1987) for which we propose to use as our starting point a high-breakdown estimator, namely the MCD estimator for missing data. It is computed by means of a modification of the forward search algorithm. The other is a generalization of the ML estimator for missing data to the class of S-estimators in which we propose the use of the TBS estimator (Rocke, 1996) which is known to have a high-breakdown point. It is also computed by means of an ER-type algorithm and, to make it really robust, we propose using a high-breakdown point estimator such as the MCD estimator as the starting point. For the simulated data set and the real data set, we found that both procedures give similar results, and are robust to relatively large numbers of outliers, which is not the case for ER with a non-robust starting point.

It should again be stressed that a robust estimation of the mean and covariance of multivariate data is important if one wants to conduct statistical analyses such as factor analysis that are not too much influenced by extreme data. Yuan and Bentler (1998) showed that the influence of such data on covariance structure analysis is limited if the covariance matrix is robustly estimated. Jöreskog (1979, p. 109) mentions the problem of robustness when presenting the ML estimator for covariance structure analysis which depends on the sample covariance matrix, saying that 'if the distribution deviates far from the multinormal it is probably wise to 'robustify' the (sample) variances and covariances'. We can only endorse this type of statement.

Finally, to estimate means and covariance matrices for multivariate data with missing values in practice, we have put the EM and ERTBS routines into an S-PLUS library that is available at http://www.hec.unige.ch/professeurs/VICTORIAFESER_Maria-Pia/pages_web/Recherche/Spluslib.htm. A 'readme' file is also provided that explains how to install the library and how to use the different functions. The data analysed in this paper are also available at the same site.

Acknowledgements

The research for this paper took place while the first author was a Ph.D. student at the London School of Economics. The second author is partially supported by the Fond National Suisse pour la Recherche Scientifique. We would like to thank A. C. Atkinson, M. A. Knott, P. Rousseeuw,

E. Ronchetti and two anonymous referees for their comments and suggestions, as well as M. Rocke for providing S-functions for the TBS, and S. Copt for putting the routines into an S-PLUS library.

References

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

Atkinson, A. C., & Cheng, T.-C. (2000). *The forward search for the minimum covariance determinant estimator*. Paper presented at the 22nd Biennial Conference of the Society for Multivariate Analysis in the Behavioural Sciences, London, 17–19 July, UK.

Baddeley, A., Della Sala, S., Gray, C., Papagno, C., & Spinnler, H. (1997). Testing central executive functioning with a pencil-and-paper test. In P. Rabbitt (Ed.), *Methodology of frontal and executive function* (pp. 61–80). Hove: Psychology Press.

Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B*, *37*, 129–146.

Butler, R. W., Davies, P. L., & Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *Annals of Statistics*, *21*, 1385–1400.

Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics*, *29*, 231–237.

Davies, P. L. (1987). Asymptotic behaviour of S-estimators of multivariate location parameters and dispersion matrices. *Annals of Statistics*, *15*, 1269–1292.

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

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, *62*, 531–545.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Department of Statistics, Harvard University.

Hampel, F. R. (1973). Robust estimation: A condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *27*, 87–104.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

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. (1977). Robust covariances. In S. S. Gupta & D. S. Moore (Eds.), *Statistical decision theory and related topics*, Volume 2 (pp. 1753–1758). New York: Academic Press.

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

Jöreskog, K. G. (1979). Structural equation models in the social sciences: Specification, estimation and testing. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models* (pp. 105–127). Cambridge, MA: Abt Books.

Kendall, M. G., & Stuart, A. (1969). *The advanced theory of statistics, Vol. 1*. New York: Hafner Press.

Kent, J. T., & Tyler, D. E. (1996). Constrained M-estimation for multivariate location and scatter. *Annals of Statistics*, *24*, 1346–1370.

Kray, J., Frensch, P., & Lindenberger, U. (1996). *Age differences in cognitive control ability*. Paper presented at the 14th biennial meeting of the ISSBD, Quebec City, Canada.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics*, *37*, 23–38.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

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

Lopuhaä, H. P. (1991). τ-estimators for location and scatter. *Canadian Journal of Statistics*, *19*, 307–321.

Lopuhaä, H. P., & Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Annals of Statistics*, *19*, 229–248.

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

Ribaupierre, A., & Ludwig, C. (2000). Attention divisee et vieillissement cognitif: Differences d'age dans 5 epreuves duelles de memoire de travail. In D. Brouillet & A. Syssau (Eds.), *Le vieillissement cognitif normal: Vers un modèle explicatif du vieillissement* (pp. 27–51). Brussels: De Boeck Universite.

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

Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, *91*, 1047–1061.

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

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, & W. Wertz (Eds.), *Mathematical statistics and applications, Vol B* (pp. 283–297). Dordrecht: Reidel.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

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

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

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*, 403–428.

Stahel, W. A. (1981). Breakdown of covariance estimators. Technical Report 31, Fachgruppe für Statistik, ETH, Zurich.

Tamura, R., & Boos, D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, *81*, 223–229.

Tyler, D. E. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika*, *70*, 411–420.

Tyler, D. E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics. *Annals of Statistics*, *22*, 1024–1044.

Wechsler, D. (1987). *Manual for the Wechsler Memory Scale-Revised*. San Antonio, TX: Psychological Corporation.

Wilcox, R. R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology*, *51*, 1–39.

Woodruff, D. L., & Rocke, D. M. (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., & Bentler, P. M. (1998). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, *51*, 63–88.

Received 25 March 1999; revised version received 29 October 2001