Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data

by Cédric Béguin and Beat Hulliger

June 2008

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 91-103
Statistics Canada, Catalogue No. 12-001-X

The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data

Cédric Béguin and Beat Hulliger [1]

Abstract

With complete multivariate data the BACON algorithm (Billor, Hadi and Vellemann 2000) yields a robust estimate of the covariance matrix. The corresponding Mahalanobis distance may be used for multivariate outlier detection. When items are missing the EM algorithm is a convenient way to estimate the covariance matrix at each iteration step of the BACON algorithm. In finite population sampling the EM algorithm must be enhanced to estimate the covariance matrix of the population rather than of the sample. A version of the EM algorithm for survey data following a multivariate normal model, the EEM algorithm (Estimated Expectation Maximization), is proposed. The combination of the two algorithms, the BACON-EEM algorithm, is applied to two datasets and compared with alternative methods.

Key Words: Forward search method; Outlier detection; Multivariate data; Missing value; Sampling; Robustness; EM-algorithm.

1. Introduction

The problem underlying the methods presented in this article is a sample survey on quantitative data like sales of different products where missing values and outliers occur. Often in the editing phase of the survey outliers are detected by inspection of individual questionnaires or by univariate outlier detection methods. However, there are few systematic methods which allow multivariate outlier detection in incomplete survey data.

Outlier detection is an important aspect of statistical data editing. Undetected outliers may have a large and undesirable impact on survey results. Most existing outlier-detection methods are designed for complete univariate or bivariate data. However, real outliers in survey data are often multivariate in nature. The problem of outliers becomes much more difficult in three or more dimensions than in one dimension or two. While an outlier can only be very small or very large in one dimension (at least for unimodal distributions) in higher dimensions the issue of the "direction" of the outlier becomes more and more important. Outliers may be quite close to the bulk of the data or to a model if the distance is measured in a Euclidean metric because this metric only checks the axis directions. However, if a metric appropriate to the correlation structure of the bulk of the data is used the outlier may be far away. Thus in higher dimensions the form of the point cloud of the bulk of the data must be well reflected in the metric used to detect outliers.

Outlier detection needs a model for the bulk of the data to be able to distinguish observations which are not fitted well by the model. Thus outlier detection is inherently tied to models and their robust estimation. In a sampling context the model should be appropriate for the bulk of the population and not only for the bulk of the sample. Therefore, the sample design should be taken into account when detecting outliers in sample survey data. The discussion on the role of sampling weights is taken up again in Sections 1 and 5.

Survey data often contains missing values. Outlier detection with missing items must estimate the model for the bulk of the data taking into account the missingness. This estimation under missing values will be based on the relationship among the observed and missing variables. The relationship must be modeled robustly to protect it from outliers. If an observation would be classified as an outlier based on complete information but the values causing the outlyingness are missing then the outlier will not show up compared with a robust model. Therefore it will be difficult to detect an outlier which is outlying only in its missing values. This is analogue to the conception of missingness at random (MAR) (Little and Rubin 1987): We need information in the observed values to infer that an observation is an outlier. We may call this situation "outlying at random". We can formalize it by stating that the outlier mechanism does not depend on unobserved data, which includes unobserved true values of the outlier in case the outlier is an error. However, for outlier detection this condition is too strict because we may be able to detect outliers in observed values even if the mechanism depends on unobserved values. This is possible because the model must hold for the bulk of the data only and not for the outliers. If the observed values of the outlier deviate enough from the model the outlier will be detected. However, when it comes to imputation of true values for nominated outliers we are in the same situation as for missing values. If the

1. Cédric Béguin, University of Neuchâtel, 2010 Neuchâtel, Switzerland; Beat Hulliger, University of Applied Sciences Northwestern Switzerland, 4600 Olten, Switzerland. E-mail: beat.hulliger@fhnw.ch.

outlier mechanism, conditionally on the observed values, still depends on the true unobserved values of the outlier we cannot estimate a model for the unobserved values. In this article we use imputation only as an ad hoc device for better outlier detection. Nevertheless we assume that, conditionally on the observed data, the non-response mechanism and the outlier-mechanism are independent and that both mechanisms do not depend on unobserved data.

In a workpackage of the EUREDIT project on "The development and evaluation of new methods for editing and imputation" (EUREDIT 2003) the authors developed outlier detection methods which cope with this difficult set-up: multivariate incomplete sample survey data. Two of these methods, Transformed Rank Correlations and the Epidemic Algorithm are presented in (Béguin and Hulliger 2004). The third method, BACON-EEM, is presented here.

In this article we concentrate on outlier detection. The scenario we have in mind is that once an outlier is detected, either it may be checked and treated manually or it may be treated by imputation. Robust estimation would replace both detection and imputation but is less adapted to the practice of official statistics. We do not distinguish between representative and non-representative outliers (Chambers 1986) since both types of outliers have to be detected, though they may have to be treated differently.

For complete data the existing multivariate methods can be classified into two major families. Many methods suppose that the data follow some elliptical distribution and try to estimate robustly the center and the covariance matrix. Then they use a corresponding Mahalanobis distance to detect outliers. The second class of methods does not rely on a distributional assumption but uses some measure of data-depth (see Liu, Parelius and Singh 1999, for a review) to be used as an outlyingness measure. The second family is at first sight more appealing, but, unfortunately, it often fails to yield methods computationally feasible with large datasets.

Many robust estimators of the covariance matrix have been reported in the literature. M-estimators (Huber 1981; Maronna 1976) have the advantage of being relatively simple to compute with a straightforward iteration from a good starting point (Rocke and Woodruff 1993). But their breakdown point - *i.e.*, the smallest fraction of the data whose arbitrary modification can carry an estimator beyond all bounds - is at most $1/(p+1)$ where p is the dimension of the data (Donoho 1982; Maronna 1976; Stahel 1981). This handicap is important when dealing with data from official statistics, which is often high dimensional. Many other affine equivariant robust estimators, *i.e.*, estimators which transform coherently when the data is transformed linearly, were studied by (Donoho 1982) but all have breakdown points of at most $1/(p+1)$. Other approaches ended up with affine equivariant high breakdown point estimators, *e.g.* the Stahel-Donoho (SD) estimator (Stahel 1981; Donoho 1982) or the Minimum Covariance Determinant (MCD) estimators (Rousseeuw 1985; Rousseeuw and Leroy 1987), but had the disadvantage of being computationally expensive. An approach of Gnanadesikan and Kettenring, using a componentwise construction of the covariance matrix, sacrificed affine equivariance but gained simplicity and speed. This approach has been re-actualized in (Maronna and Zamar 2002) and in one of the methods presented in (Béguin and Hulliger 2004), called Transformed Rank Correlations (TRC). TRC calculates an initial matrix of bivariate Spearman Rank correlations. To ensure positive definiteness of the covariance matrix the data is transformed into the space of eigenvectors of the initial matrix. The coordinatewise medians and median absolute deviations in this new space are then backtransformed into the original space to obtain an estimate of the center and a positive definite covariance matrix.

Another idea from (Gnanadesikan and Kettenring 1972) is related to the so-called forward search methods, which are closely related to the method proposed in this paper. These so called forward search methods are based on the concept of "growing a good subset of observations". By "good subset" one means a subset free or almost free of outliers. The idea is to start with a small subset of the data and then to add non-outlying observations until no more non-outliers are available.

The idea of a forward search algorithm was first suggested in (Wilks and Gnanadesikan 1964) and described in detail in (Gnanadesikan and Kettenring 1972). The articles of (Hadi 1992) and (Atkinson 1993) demonstrated the efficiency of such methods. In both articles the "good subset" grows one point at a time using Mahalanobis distances to rank the observations. Then research concentrated on developing faster and more sophisticated methods based on the same idea. The last two and most efficient were developed in (Billor, Hadi and Vellemann 2000) and (Kosinski 1999). These algorithms were compared in (Béguin 2002) and the BACON algorithm (Billor *et al.* 2000) turned out to be the most robust and fastest forward search method with complete multivariate normal data. In particular the breakdown point turned out to be high in practical applications. Also when comparing with other Mahalanobis type methods the performance of BACON on complete data is very good (Béguin and Hulliger 2003).

None of the above methods is designed to deal with incomplete data stemming from surveys, *i.e.*, with missing values and sampling weights. The first article to address the problem of multivariate outlier detection in incomplete data is (Little and Smith 1987). The authors propose Mahalanobis distances to detect outliers, with robust

estimations of center and scatter obtained by the ER algorithm. The ER algorithm replaces the maximum-likelihood estimator in the maximization step of the EM algorithm (Dempster, Laird and Rubin 1977) by a robust one-step *M*-estimator. However, the starting point of the ER algorithm is the classical non-robust mean and covariance and therefore the breakdown point of the ER algorithm is 0. In other words even one single outlier can carry the estimator beyond any limit. To correct the low breakdown point of that algorithm (Cheng and Victoria-Feser 2000) used an MCD algorithm for the maximization step of the EM-algorithm. However, the combination of the iterative procedures of MCD and EM makes the computation for large datasets too slow for practical applications. Moreover the introduction of sampling weights is not straightforward.

The TRC algorithm in (Béguin and Hulliger 2004) uses robust linear regression imputations by the best univariate predictor to cope with missing values. The Spearman rank correlations are expressed as functionals of the empirical distribution function of the sample to obtain estimates for the Spearman rank correlations in the population.

The BACON algorithm is based on the multivariate normal distribution and thus the EM algorithm for multivariate normal data was chosen to impute missing values within the BACON iterations. To take into account the sampling aspect, the estimates of the BACON algorithm have to be replaced by Horvitz-Thompson type estimators and a special version of the EM algorithm is developed where the expectations on the population level are estimated from the sample. Section 2 sets up the notation, recalls quickly the BACON algorithm and presents its adaptation to sampling weights. Section 3 introduces the Estimated-EM (EEM) algorithm and Section 1 discusses the adaptation of the Mahalanobis distance to missing values. Section 4 explains how BACON and EEM are merged in an efficient way to become the BACON-EEM algorithm. Section 5 shows the application of BACON-EEM to two datasets. The results are compared to the competitor methods, Transformed Rank Correlations, developed in (Béguin and Hulliger 2004), the ER-algorithm and a baseline algorithm which uses MCD after non-robust imputation based on the EM-algorithm.

2. The BACON algorithm

The BACON algorithm is presented in (Billor *et al.* 2000). Two versions are described: one for multivariate data in general and one for regression data. Only the first case will be considered here.

The data are stocked in a matrix X of n rows (observations $x_1, ..., x_n$) and p columns (variables $x^1, ..., x^p$). We assume that the bulk of the data is unimodal and roughly elliptical symmetric. The coordinatewise mean (resp. covariance matrix) computed on X is denoted by m_X (resp. C_X). The squared Mahalanobis distance of a point y based on m_X and C_X is $\mathrm{MD}_X^2(y) = (y - m_X)^\top C_X^{-1}(y - m_X)$. If the mean and covariance are calculated only on a subset G of the data then we denote them m_G and C_G with corresponding Mahalanobis distance MD_G.

The first step of the algorithm is the choice of an initial subset G of "good data". Two versions are proposed in the literature. The first version simply selects the cp points with smallest Mahalanobis distances $\mathrm{MD}_X(x_i)$, $i \in \{1, ..., n\}$, with c being an integer chosen by the data analyst. It may be set to $c = 3$ by default. The second version selects the cp points with smallest Euclidean distances from the coordinatewise median, with c as before. The second version is more robust but it looses affine equivariance. Other starting points than the coordinatewise median might be considered like a spatial median. In this article we concentrate on the second version of the basic good subset. In both versions if C_G is singular then the basic subset is increased by adding observations with smallest distances until C_G has full rank. Then an iterative process starts.

Denote by $\chi^2_{p,\beta}$ the $1-\beta$ percentile of the χ^2 distribution with p degrees of freedom and by $|G|$ the number of elements in the set G. The steps of the BACON algorithm are:

1. Compute the squared Mahalanobis distances $\mathrm{MD}_G^2(x_i)$ for $i \in \{1, ..., n\}$;
2. Define a subset G' including all points with $\mathrm{MD}_G^2(x_i) < c_{npr}\, \chi^2_{p,\alpha/n}$, where $c_{npr} = c_{np} + c_{hr}$ is a correction factor with $c_{np} = 1 + (p+1)/(n-p) + 1/(n-h-p)$, $c_{hr} = \max\{0, (h-r)/(h+r)\}$, $h = \lceil (n+p+1)/2 \rceil$ and $r = |G|$.
3. If $G' = G$ then stop, else set G to G' and go to Step 1.

Note that the correction factor c_{npr} is close to 1 for large n. The observations that are not contained in the final G are declared outliers. Alternatively a threshold for the Mahalanobis distance, above which observations are nominated outliers, may be chosen by inspecting the distribution of the Mahalanobis distance.

The computing effort required by the BACON algorithm depends on the configuration of the data. Compared with other algorithms it is small and in particular this effort grows slowly with increasing sample size (see also Section 5). This makes the BACON method particularly well suited for large datasets.

Note that the original selection criterion of Step 2 is designed for a multivariate normal distribution, which

implies that the squared Mahalanobis distances follow asymptotically a χ^2 distribution with p degrees of freedom. Suppose all points follow a multivariate normal distribution and that the Mahalanobis distance is computed using the sample mean and covariance matrix. The test $\mathrm{MD}_X^2(x_i) > \chi^2_{p,\alpha}$ declares about 100α percent of the points as outliers. Instead of α we often use α/n. Using Bonferroni inequalities one can show that under normality the test with level α/n will declare no outlier with probability larger than $1-\alpha$ (*i.e.*, $P(\mathrm{MD}_X^2(x_i) < \chi^2_{p,\alpha/n}, \forall i \in \{1,...,n\}) \geq 1-\alpha$). The test with α/n very rarely detects points that are not outliers but it also reduces its sensitivity to close outliers when n becomes large. One may also want to run the method with both types of the test level and compare the results.

2.1 Adaptation to sampling weights

For the sampling context we use the following notation. The data stem from a random sample s of the finite population U with N elements. The sample of size n is drawn with the sample design $p(s)$ and the first order inclusion probabilities are denoted $\pi_i = \sum_{s|i\in s} p(s)$. The weights will be the inverse of the inclusion probabilities of the observations $w_i = 1/\pi_i$ such that the Horvitz-Thompson estimator of the population total, $\sum_{i\in U} x_i$, is $\sum_s w_i x_i = \sum_{i=1}^n w_i x_i$. Furthermore it is assumed that $\sum_s w_i \approx N$. The mean m_X and the covariance matrix C_X may be estimated by the Hájek estimators

$$\hat{m}_X = \frac{\sum_s w_i x_i}{\sum_s w_i}$$

and (1)

$$\hat{C}_X = \frac{\sum_s w_i (x_i - \hat{m}_X)(x_i - \hat{m}_X)^\top}{\sum_s w_i}.$$

The sample estimate of the median is defined as in (Béguin and Hulliger 2004): let x_u^k be the smallest value such that $\sum_s w_i 1_{x\leq x_u^k}(x_i^k) \geq 0.5\sum_s w_i$ and x_v^k the smallest value such that $\sum_s w_i 1_{x\leq x_v^k}(x_i^k) > 0.5\sum_s w_i$, then the estimate is given by

$$\widehat{\mathrm{med}}_X = (w_u x_u^k + w_v x_v^k)/(w_u + w_v). \quad (2)$$

To adapt the BACON algorithm to sampling the initial subset is selected using Hájek estimators $\hat{m}_X$ and $\hat{C}_X$ or the median $\widehat{\mathrm{med}}_X$. For the iterative process, denote by s_G the selected "good observations" of the sample. These observations are representatives of a "virtual good subset" G of the whole population with estimated size $\hat{r} = \sum_{s_G} w_i$. The mean and covariance matrix of this subset are estimated by the Hájek estimators

$$\hat{m}_G = \frac{\sum_{s_G} w_i x_i}{\sum_{s_G} w_i}$$

and (3)

$$\hat{C}_G = \frac{\sum_{s_G} w_i (x_i - \hat{m}_G)(x_i - \hat{m}_G)^\top}{\sum_{s_G} w_i}.$$

These estimates are used to compute the estimates of the Mahalanobis distances $\widehat{\mathrm{MD}}_G(x_i), x_i \in s$. Finally the correction factor $c_{Npr} = c_{Np} + c_{hr}$ of the selection criteria is computed using the estimates $\hat{N} = \sum_s w_i$ and $\hat{r} = \sum_{s_G} w_i$. If N is known, its actual value is used.

If there are no missing values in the data the BACON algorithm can be used to estimate the population mean and covariance. The basic assumption for the BACON algorithm is still that the bulk of observations of the population has an elliptical distribution. We may use the BACON algorithm without weighting and compare the result with the weighted version. Different results indicate that the design-variables or a model used for non-response weighting are not well reflected in the model. We advocate the use of weights, in particular in routine applications, to give some protection against miss-specification of the model. In any case the estimand should be the mean and covariance of the bulk of the population.

Note that the Mahalanobis distance does not involve the sampling weights directly. The weight of a possible outlier influences the Mahalanobis distance only through the model, *i.e.*, the mean and the covariance.

3. The EEM algorithm

Nonresponse issues are important in official statistics and many surveys cannot deliver a complete dataset. The problem of unit-nonresponse, *i.e.*, completely missing observations, is usually dealt with by using appropriate weights and is not treated here. Item-nonresponse, *i.e.*, observations with only partially available information, cannot be treated by discarding all incomplete observations because too much information is lost. The approach followed here will retain high efficiency under multivariate normal data. At each BACON iterative step the mean and covariance matrix of the good subset of observations will be computed using a modified version of the EM algorithm for multivariate normal data. The expectations computed in the E-step are replaced by sample estimates. The modified algorithm is therefore named the EEM (Estimated-Expectation/Maximization) algorithm. Note that this adaptation is presented here for multivariate normal data but the results can be generalized to other distributions of the regular exponential family.

This paragraph re-uses the description and notation of the EM algorithm given in (Schafer 2000). All details about EM not given here can be found within the first three chapters and in Section 5.3 of this book. The following abuse of notation is also used here: X will denote simultaneously a p-dimensional random variable and the $N \times p$ matrix containing the realized values of the variable X of the population U. If a census were taken of the whole population to measure the variable X it would result in some observed and missing values $X = X_o \cup X_m$. The EM-algorithm assumes that the missingness mechanism is ignorable (Schafer 2000, section 2.2). Here we assume in addition that the missingness is independent from the sampling. The observations of the data can be modeled as independent, identically distributed (iid) draws from a multivariate normal probability distribution with density $f(x, \theta)$. Using the assumptions and the factorization $P(X \mid \theta) = P(X_o \mid \theta)P(X_m \mid X_o, \theta)$ the complete-data log-likelihood can be written as $l(\theta \mid X) = l(\theta \mid X_o) + \log(P(X_m \mid X_o, \theta)) + c$, where $l(\theta \mid X_o)$ is the observed-data log-likelihood and c is an arbitrary constant. The term $P(X_m \mid X_o, \theta)$ captures the interdependence between X_m and θ on which the EM-algorithm capitalizes. Because $P(X_m \mid X_o, \theta)$ is unknown the average of $l(\theta \mid X)$ over $P(X_m \mid X_o, \theta^{(t)})$ is taken at each E-step, where $\theta^{(t)}$ is a preliminary estimate of the unknown parameter. The next estimate $\theta^{(t+1)}$ is found by maximizing the result of the expectation step (M-step). The sequence of E and M-steps is iterated until convergence. Conditions under which this sequence $\theta^{(t)}$ converges to a stationary point of the observed-data likelihood are provided in (Dempster *et al.* 1977). In well-behaved problems this stationary point is a global maximum.

For a probability distribution of the regular exponential family the complete data log-likelihood may be written as

$$l(\theta \mid X) = \eta(\theta)^{\top} \cdot T(X) + Ng(\theta) + c, \qquad (4)$$

where $\eta(\theta) = (\eta_1(\theta), \eta_2(\theta), ..., \eta_k(\theta))^{\top}$ is the canonical form of the parameter θ and $T(X) = (T_1(X), T_2(X), ..., T_k(X))^{\top}$ is the vector of complete-data sufficient statistics. Moreover, each of the sufficient statistics has an additive form $T_j(X) = \sum_{i=1}^{N} h_j(x_i)$, for some function h_j. Because $l(\theta \mid X)$ is a linear function of the sufficient statistics, the E-step replaces $T_j(X)$ by $E(T_j(X) \mid X_o, \theta^{(t)})$. In other words the E-step fills in the missing portions of the complete-data sufficient statistics. For a multivariate normal distribution $X = (X^1, \ldots, X^p)$ the sufficient statistics are composed of two types of elements: the sums $\sum_{i=1}^{N} x_i^k$ and the sums of products $\sum_{i=1}^{N} x_i^k x_i^l$, $1 \le k, l \le p$. The E-step reduces to computing the conditional expectations of these sums given the observed data X_o and the preliminary parameter $\theta^{(t)}$.

For a single summand i one can show (Schafer 2000, section 5.3) that these expectations depend only on the observed components of the same observation, *i.e.*, on x_i^{obs}. This leads to

$$E\left(\sum_{i=1}^{N} x_i^k \,\middle|\, X_o, \theta^{(t)}\right) = \sum_{i=1}^{N} E(x_i^k \mid X_o, \theta^{(t)}) = \sum_{i=1}^{N} E(x_i^k \mid x_i^{\text{obs}}, \theta^{(t)}), 1 \le k \le p \qquad (5)$$

and the analogue form of the sum of products. Of course $E(x_i^k \mid x_i^{\text{obs}}, \theta^{(t)}) = x_i^k$ if $x_i^k \in x_i^{\text{obs}}$. If x_i^k is missing, then this expectation is the fitted value of a regression of x^k given the parameter $\theta^{(t)}$ on the variables which are observed for observation i. Thus the sufficient statistics are composed of population sums of observed values (T_o) and sums of fitted values (T_m).

In the situation where our data stem from a sample of a finite population we consider the finite population as a realization of a multivariate normal distribution and the sums (5) and sums of products have to be estimated from the sample. The form of (5) allows the use of simple Horvitz-Thompson estimators. The estimate of (5) is

$$T^{k0} = \sum_{s} w_i E(x_i^k \mid x_i^{\text{obs}}, \theta^{(t)}), \; 1 \le k \le p, \qquad (6)$$

and $E(\sum_{i=1}^{N} x_i^k x_{i_s}^l \mid X_o, \theta^{(t)})$ is estimated by

$$T^{kl} = \sum_{s} w_i E(x_i^k x_i^l \mid x_i^{\text{obs}}, \theta^{(t)}), \; 1 \le k, l \le p. \qquad (7)$$

In short: We replace the population sums of T_o and T_m by their Horvitz-Thompson estimators $\hat{T}_o$ and $\hat{T}_m$. We call the calculation of the T^{k0} and T^{kl} the estimated expectation step (EE-step). Plugging these estimators into (4) we obtain an estimator of the average population likelihood function.

For the M-step, the maximization of the estimate of the average population likelihood, the weighted normal equations have to be solved. The solution is found by a simple matrix operation using the sweep operator (Schafer 2000, section 5.3) applied to the symmetric $(p+1) \times (p+1)$ matrix $(T^{kl})_{0 \le k, l \le p}$ of the estimated expectations of the sufficient statistics (with T^{00} set to 1) divided by N, which is estimated by the sum of weights if unknown:

$$\theta^{(t+1)} = \text{SWP}[0]\left(\frac{(T^{kl})_{0 \le k, l \le p}}{\sum_s w_i}\right), \qquad (8)$$

where SWP[0] is the sweep operator on the first line/column of the matrix.

The EEM algorithm iterates the EE and the M-step. Computationally the difference between the EE-step and the E-step of the original EM-algorithm comes down to using

weighted sums instead of un-weighted sums with weights that do not change over the iterations. We therefore expect that the convergence of the EEM-algorithm will remain similar to the EM-algorithm. For the BACON-EEM algorithm we only need a rough approximation to the solution in each BACON-step. Thus we use only a small number of iterations of the EEM algorithm.

3.1 Mahalanobis distance with missing values

The Mahalanobis distance is developed for complete observations and needs to be adapted to missing values. One option is to use the EEM estimate to impute the conditional mean for the missing values given the observed values and then calculate the Mahalanobis distance with imputed values. Under a MAR (Missing At Random) assumption there is a valid model based on the observed part of the data to impute the missing values. In the case of outlier detection we suppose that the imputation model may hold for the bulk of the data only and is estimated in a robust way. But then we may not expect that an outlier value is predicted by the model, except if already the observed part of an observations is outlying. Therefore there is no advantage to use imputation before outlier detection and we prefer to directly adapt the Mahalanobis distance to missingness. Two different versions of the Mahalanobis distance are possible in this situation.

We call the first version *marginal* Mahalanobis distance. It uses the Mahalanobis distance in the space of observed variables and scales it up with a factor p/q, where $q = \sum_k r_{ik}$ is the number of non-missing variables and p is the total number of variables. More precisely, we assume an observation x is partitioned into $x = (x_o^\top, x_m^\top)^\top$ (after possible rearrangement), where x_o denotes the observed part and x_m the unobserved part of the observation. Then the marginal Mahalanobis distance is

$$\mathrm{MD}^2_{\mathrm{marg}} = \frac{p}{q}(x_o - m_o)^\top (S_{oo})^{-1}(x_o - m_o), \qquad (9)$$

where S_{oo} is the part of the covariance matrix corresponding to x_o. This version is also used in (Little and Smith 1987).

The second version of Mahalanobis distance with missing values is obtained by reducing the contribution of the missing values to the Mahalanobis distance to zero. This amounts to replacing all missing values by their mean, *i.e.*, $x_m = m_m$. In other words we would impute a mean without consideration of the covariance matrix and the above arguments against outlier detection with imputed values apply here as well. Nevertheless we tested this second version of Mahalanobis distance. It yields erratic Mahalanobis distances (Béguin 2002) and (Béguin and Hulliger 2003) and we did not use it any further.

4. The BACON-EEM algorithm

Both algorithms, BACON and EEM, are computationally demanding. By merging them in a convenient way we gain performance. The "growing" structure of the BACON algorithm implies redundancies which may be used to avoid extra-computations in the EEM-algorithm at each step. The crucial point at each BACON step is that the estimations of the mean and the covariance matrix from the EEM-algorithm allow the exclusion of outlying points from the good subset and this does not need extremely precise estimates. Thus it is not necessary to iterate EEM to convergence each time the mean and covariance are needed. We use only 5 iterations by default. Furthermore we use the result of the last EEM-iteration of the last BACON-step as a starting value for EEM.

As much information from past iterations as possible should be reused. In fact the sufficient statistic T^G computed on some good subset G have an observed part of the sum T_o^G and a missing part of the sum T_m^G. The expectation computed by the E-step can therefore be written as

$$E(T^G \mid X_o^G, \theta) = T_o^G + E(T_m^G \mid X_o^G, \theta). \qquad (10)$$

As the subsets G are usually growing, $\hat{T}_o^G$ is not recomputed at each step of the BACON loop, but a global variable for $\hat{T}_o^G$ is updated each time G changes (usually only adding points, sometimes removing a few).

At each iteration of the BACON-EEM algorithm, once the EEM algorithm has obtained the estimations of the center and the scatter of the good subset, marginal Mahalanobis distances for all observations are used in step 2 of the BACON algorithm.

Note the crucial point for the robustness of the algorithm: EEM is not robust, but at each BACON-step EEM is run only on points that have the smallest and therefore non-outlying marginal Mahalanobis distance in the preceding step. In other words the observation x will be used by EEM if and only if x_o is sufficiently small for the metric given by $(S_{oo})^{-1}$ at the preceding step. Therefore if the first subset of good points is free or almost free of outliers, the imputation process in EEM will never create outlying values throughout the whole BACON-EEM algorithm. In other words, the non-robust EEM-algorithm is protected by the general forward search approach of the BACON algorithm in the same way as the non-robust mean and covariance of the original BACON algorithm is protected.

Summing up, the steps of the BACON-EEM algorithm are the following:

1. Calculate the weighted coordinate-wise median $\widehat{\mathrm{med}}(x)$ ignoring missing values in each variable separately. Determine the Euclidean distance from the median of

each observation omitting missing values but standardizing for the number of present values: $a_i = \|x_i - \widetilde{\text{med}}(x)\|\sqrt{p/q}$. Select the $m = cp$ observations with least a_i to constitute the initial subset G.

2. Compute a center $\hat{m}_G$ and scatter $\hat{C}_G$ using the EEM-algorithm and update the estimate of the sufficient statistic of the observed part $\hat{T}_o^G$.
3. Compute the squared marginal Mahalanobis distances $\text{MD}_G^2(x_i)$ for $i = 1, ..., n$. The new set G' contains the observations with $\text{MD}_G^2(x_i) < c_{\hat{N}p\hat{r}}\ \chi^2_{p,\alpha}$.
4. If $G' = G$ then stop, else set G to G' and go to step 2.

If instead of outlier detection the mean and covariance estimates of BACON-EEM are the main objectives the EEM-algorithm may be iterated further without changing G. In step 3 one may alternatively use α/n instead of α (see Section 2).

5. Applications

In this section we compare the BACON-EEM algorithm (BEM) with Transformed Rank Correlations (TRC) from (Béguin and Hulliger 2004) and the ER-algorithm from (Little and Smith 1987). As a further benchmark we use an imputation under the multivariate normal model with estimates of the mean and covariance by the EM algorithm. In other words we create a non-robust imputation. Then robust estimates of the multivariate location and the covariance matrix are obtained by the Minimum Covariance Determinant estimator computed on the imputed data and finally outliers are detected using the corresponding Mahalanobis distances. The benchmark method is called GIMCD for "Gauss Imputation followed by MCD detection". The algorithms are implemented in `R` (R Development Core Team 2006) with the help of the `R`-packages `norm` (Novo and Schafer 2002) and `MASS` (Venables and Ripley 2002).

5.1 Bushfire data

The reaction of the BACON-EEM to the introduction of missing values is illustrated with a real dataset of 38 observations and 5 variables. It was used by (Maronna and Zamar 2002) to locate bushfire scars. This well known example is also studied in (Maronna and Yohai 1995) and (Maronna and Zamar 2002). It allows a two dimensional plot (in variable 2 and 3) that reveals most of the outliers (see Figure 1). The data contains an outlying cluster of observations 33 to 38 a second outlier cluster of observations 7 to 11 and a few more isolated outliers, namely observations 12, 13, 31 and 32. We have added observation 31 to the list of potential outliers because it is indicated as a borderline case by MCD, BACON and also other methods studied in (Maronna and Zamar 2002). Missing values are created with a MCAR (Missing Completely At Random) mechanism. Two datasets are created with respectively 20 and 40% of missing items. The dataset with 40% of missing values have observations with up to 4 out of 5 missing values and therefore are a challenge for any method. As the size n of the dataset is small, BACON-EEM is run with the $\chi^2_{p,\alpha/n}$ test. The results are given in Table 1. Observations 7 to 13 and 31 to 38 are individually shown as detected or not, while for the other 23 good points the number of observations declared as outliers is indicated. The limit above which a Mahalanobis distance indicates an outlier, was determined for each run by inspection of the quantile plot of the Mahalanobis distance.

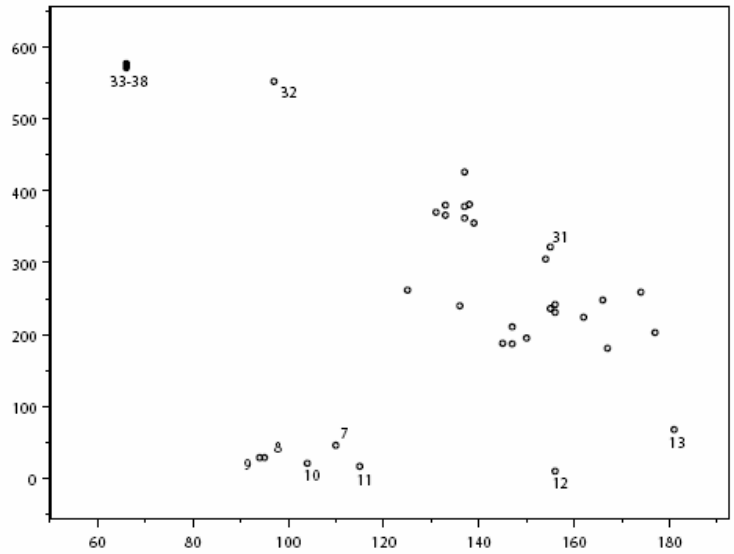

Figure 1 Bushfire Data

With complete observations none of the methods declared any good observations as outlier. The Mahalanobis distance with non-robust mean and covariance detects 3 outliers but misses the others. MCD and BACON-EEM end up with the same subset of data as good points and therefore give the same result, *i.e.*, exactly the same Mahalanobis distance for all observations. Both do not declare observations 12 and 13 as outliers but all the others. ER does detect the group of outliers 7,...,12 but none of the other outliers. TRC detects the group 32,...,38 and two more outliers.

With 20% missing values BEM and TRC declare one good data point as outlier. TRC detects 14 of the 15 potential outliers. ER misses all outliers except observation 7. GIMCD detects the same 13 outliers as without missing values. Note however, that there is some variability in the results for GIMCD due to the random imputation. BEM detects the same outliers as with complete data except observation 11.

Table 1 Outliers detected for 3 missingness rates

(1 - *q*)%	Method	7-11	12,13,21	32-38	n. good
0	MD	11100	000	0000000	0
0	MCD	11111	001	1111111	0
0	ER	11111	100	0000000	0
0	BEM	11111	001	1111111	0
0	TRC	01100	000	1111111	0
20	GIMCD	11111	001	1111111	0
20	ER	10000	000	0000000	0
20	BEM	11111	000	1111111	1
20	TRC	01111	111	1111111	1
40	GIMCD(1)	11100	000	1111111	0
40	GIMCD(2)	11100	000	0100000	5
40	ER	10000	010	0000000	2
40	BEM	11111	000	1111111	1
40	TRC	11111	010	1111111	1

MD: Classical Mahalanobis distance, MCD: Minimum Covariance Determinant, GIMCD: Non-robust imputation under Gaussian model followed by MCD (GIMCD(1) and GIMCD(2) are two realisations of the GIMCD-algorithm), ER: Expectation-Maximization with one M-step at maximization, BEM: BACON-EEM, TRC: Transformed Rank Correlations. The first column indicates the proportion of missing values, the last column gives the number of other points (non-outliers) declared outliers, the intermediate columns are detection indicators for the observations in the first row.

With 40% missing values ER nominates observation 7, 13 and two good observations as outliers. Since the imputation is random the result of GIMCD has some variability. Two realizations, GIMCD(1) and GIMCD(2) are reported in Table 1. In a good case GIMCD detects 10 of the 15 outliers and does not declare any good observation as outlier. In a bad case GIMCD detects only 4 outliers but declares 5 good observations as outliers. BEM detects 12 of the outliers and declares one good observation as outlier. TRC detects 13 outliers and declares one good observation as outlier.

5.2 MU281 data

The MU284 data set from (Särndal, Swensson and Wretman 1992) contains data about Swedish municipalities. We use the variables *population in 1975* and *population in 1985* (pop75 and pop85), *revenue from municipal taxes 1985* (RMT85), *number of municipal employees 1984* (ME84) and *real estate value 1984* (REV84). The largest three cities according to pop75 are discarded because they are huge outliers and would be treated separately in practice. The remaining municipalities are supposed to be a stratified sample of a larger population. Strata are defined according to $0 < \text{pop75} < 20, 20 \le \text{pop75} < 100, 100 \le \text{pop75}$. Table 2 shows the assumed population sizes and the corresponding weights. This sample design reflects a typical stratification for establishment surveys with a take-all stratum of the largest establishments, where in the end 8 of 10 establishments answer the survey.

Table 2 MU281 population and sample sizes

	stratum		
	1	2	3
pop75	0-19	20-99	100+
N	1,600	250	10
n	171	102	8
w	9.36	2.45	1.25

The three variables RMT85, ME84 and REV84 are divided by pop85 to obtain figures per capita. The per capita variables are denoted by lower case names (rmt85, me84 and rev84). Figure 2 shows the distribution of these 3 variables plus the auxiliary variable pop75. The per capita figures are roughly elliptically distributed. There is a linear relationship between rmt85 and me84 and a slightly non-linear relationship of these variables with pop75. There is no apparent relationship between rev84 and rmt85 and between rev84 and me84 but there is clearly more variability in rev84 for low pop75. The distributions of variable pop75 and rev84 are skew. There is a large outlier in rmt85 and me84 and at least two in rev84.

We include pop75 in all our calculations. In practice one would include the auxiliary variable which defines the sample design in a model. Note that pop75 has no missing values.

The qq-plot of Mahalanobis distances based on MCD shows only the two clear outliers in rev84. The large outlier jointly in rmt85 and me84 has 25th largest Mahalanobis distance. We call these largest 25 observations the unweighted basic outliers. In the original MU284 dataset these unweighted basic outliers have LABEL 3, 4, 29, 31, 46, 47, 56, 79, 83, 117, 126, 131, 140, 158, 199, 211, 222, 246, 248, 252, 254, 260, 262, 272, and 273. With classical non-robust Mahalanobis distances only 12 of the basic outliers are nominated outliers, *i.e.*, are among the 25 observations with

largest (classical) Mahalanobis distance. Robust methods are necessary to detect the outliers in the MU281 data. To allow a comparison between the methods we fix the number of observations which are to be considered outliers to 25 for the moment. Thus we consider for each method the 25 observations with largest Mahalanobis distance as the outliers. Note that this may not be the threshold one would choose after inspection of the qq-plot of the Mahalanobis distances.

Table 3 shows the number of basic outliers detected by the methods run on the complete dataset. MCD detects its own 25 outliers, of course. The ER, BEM and TRC algorithm detect 25 or 24 of these outliers if no weights are applied but only 11or 15 if weighted. A suffix "w" behind the acronym of the method indicates that the sampling weights were used. Since the small municipalities have more weight the estimates are attracted towards them and other outliers will come up among the 25 largest Mahalanobis distances for ERw, BEMw and TRCw (see also Table 4). Closer inspection shows that many of the unweighted outliers are located in the tail of rev84 while most of the weighted outliers are located in the tail of pop75. The weighted methods coincide on 20 observations as outliers. We will call them weighted basic outliers. The weighted basic outliers have original MU284-LABEL 16, 28, 36, 45, 46, 55, 97, 113, 115, 121, 155, 185, 196, 208, 233, 241, 245, 265, 267, and 270. Only 10 of these observations are also among the unweighted basic outliers.

Table 4 shows the number of weighted and unweighted basic outliers in the strata. There are 12 unweighted but only 2 weighted basic outliers in stratum 1. Thus the weights have a clear influence on outlier detection. The influence is on the model primarily which is attracted towards the small observations with larger weights. Of course this can be seen as a sort of masking of outliers but in the context of modeling a better explanation is that the model is not completely adequate over all the strata and the weighted model fits the population better than the unweighted.

The second row of Table 3 gives the computation time for the algorithms. The ER algorithm is much slower than its competitors. This may be due to an inefficient implementation, however. The fastest algorithm is BEM, followed by TRC and, at some distance MCD. TRC may become slow, however, when the missingness rate is high.

Table 3 Complete MU281, number of detected unweighted basic outliers

Method	MCD	ER	ER_W	BEM	BEM_W	TRC	TRC_W
Number detected	25	25	11	24	15	24	15
Computation time	0.81	3.17	2.52	0.07	0.04	0.14	0.14

Suffix w indicates that the algorithm is run with sampling weights.

Table 4 Number of basic outliers per stratum

stratum	1	2	3	Total
unweighted	12	5	8	25
weighted	2	10	8	20

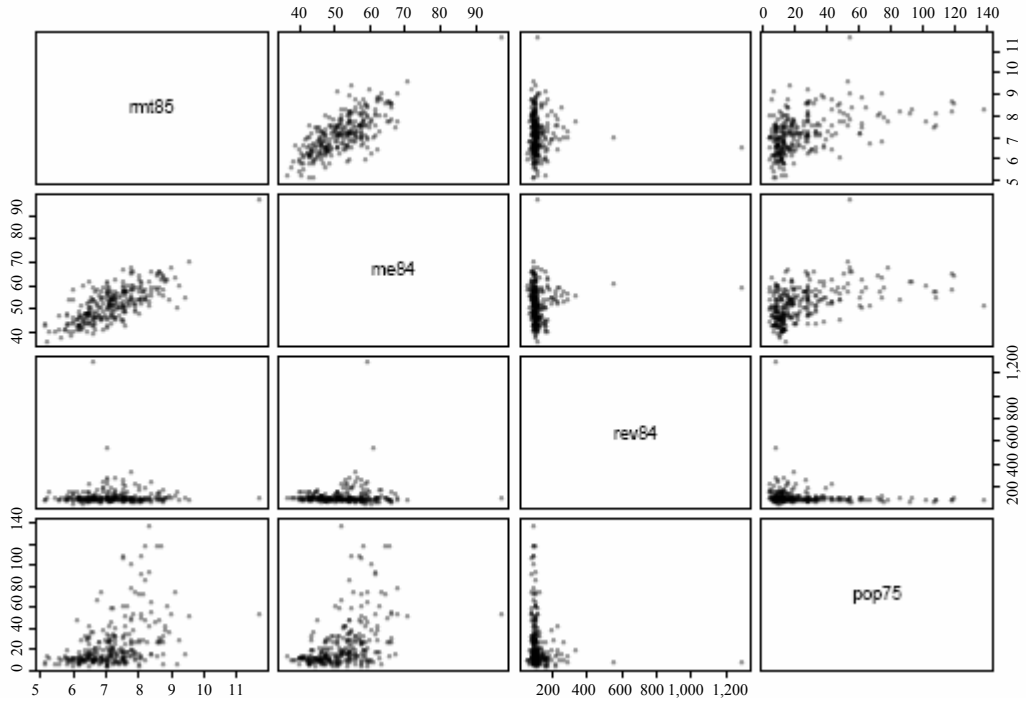

Figure 2 MU281 per capita figures and pop75

5.2.1 Missing values

We now introduce missingness into the variables rmt85, me84 and rev84 according to a mild MAR mechanism. Missingness patterns are assigned to the observations according to the modulo of their label in the original MU284 data. The MAR mechanism is reflected in higher missingness rates for stratum 1 and 2 (see Appendix). The response patterns and missingness rates per stratum are shown in Table 5. For example rev84 is the only missing value in 15 observations of stratum 1 and 2 of stratum 2. Overall 187 observations remain complete and the proportion of observations with missing values (missingness rate) is 33%.

Table 5 Frequency of response patterns per stratum for rmt85, me84, rev84

Response indicator			stratum		
rmt85	me84	rev84	1	2	3
0	1	1	11	4	0
1	0	1	13	2	0
1	1	0	15	2	0
1	0	0	13	2	1
0	1	0	14	2	0
0	0	1	11	4	0
1	1	1	94	86	7
missingness rate			0.450	0.157	0.125

Among the 35 weighted or unweighted basic outliers there are 17 observations with missing values. Table 6 shows how many of the basic outliers have been detected after the introduction of missingness. The 20 weighted basic outliers are detected well by the weighted algorithms ERw, BEMw and TRCw. GIMCD detects 4 of the weighted basic outliers and 14 of the unweighted basic outliers. Thus the missingness affects the capability of the MCD algorithm which was actually used to define the unweighted basic outliers. One word of caution: Several runs of random Gaussian imputation have been made and there is some variability in the results of GIMCD. However, also with a favorable imputation outcome GIMCD did not beat ER, BEM or TRC in detecting the unweighted basic outliers. The weighted versions ERw, BEMw and TRCw detect the weighted basic outliers well. The number of complete observations among the outliers nominated by the different methods is indicated in the last row of Table 6. All methods nominate as outliers also observations with missing values. Since the missingness rate is larger in the stratum of small observations and the weighted versions of the methods nominate less outliers in this stratum, the number of complete outliers is usually larger for the weighted algorithms. Overall the introduction of missingness has not altered the capabilities of ER, BEM and TRC by much, while GIMCD is moderately affected.

Table 6 MU281 data set with missing values, number of detected basic outliers

Method	GIMD	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Weighted	14	4	10	20	9	19	12	17
Unweighted	12	14	23	11	22	15	22	17
Complete	16	8	17	20	16	18	19	18

GIMD: Non-robust imputation under a Gaussian model followed by classical Mahalanobis distance.

5.2.2 Additional outliers

In addition to the outliers in the original data we now introduce new outliers. The observations which should become additional outliers are determined by the modulo of the original LABEL. If (LABEL mod $8 = 1$ and pop75 ≥ 10) or (LABEL mod $16 = 1$ and pop75 < 10) then the observation is an additional outlier. Thus the rate of outlyingness is larger for large municipalities. But the outlyingness is not influenced by the values of the other variables. We may say that the outlyingness is at random. Note that we could have taken a random sample instead of the above systematic sample. We preferred the systematic sample to simplify the replication of the results and to avoid additional randomness.

Two of the weighted and one of the unweighted basic outliers happen to be also additional outliers. We continue to treat them as basic outliers. Taking this into account there are 32 additional outliers in the sample. Together with the 25 unweighted or the 20 weighted basic outliers defined above there are 57 or 52 outliers to detect (20.3% or 18.5% outliers). From now on the threshold for the Mahalanobis distances is set at the 57^{th} largest distance to simplify the comparison of the methods.

The values of the additional outliers are created as follows: $\text{rmt} = 0.2 * \text{rmt85} + 8$, $\text{me} = 0.1 * \text{me84} + 50$, $\text{rev} = 0.4 * \text{rev84} + 300$. Note that we omit the suffix indicating the year for the contaminated variables. The dependence on the old values is negligible. It is only used to avoid an explicit model for the error around the point $(\text{rmt}, \text{me}, \text{rev}) = (8, 50, 300)$. This is the type of contamination that is difficult to detect for robust covariance estimators (Rocke and Woodruff 1996): concentrated and close to the point cloud of good observations.

Figure 3 shows the three variables with contamination and the location of the additional outliers.

Table 7 shows the number of detected outliers. GIMCD detects 31 of the 32 additional outliers, while BEM, BEMw, TRC and TRCw detect many of them but not all. ER and ERw detect less of the additional outliers. The weighted basic outliers are all detected by ER and ERw, BEMw and TRCw. The unweighted versions of BEM and TRC detect less of the weighted basic outliers and GIMCD detects only 4 of the weighted basic outliers. BEM and TRC whether weighted or not detect the unweighted basic outliers best.

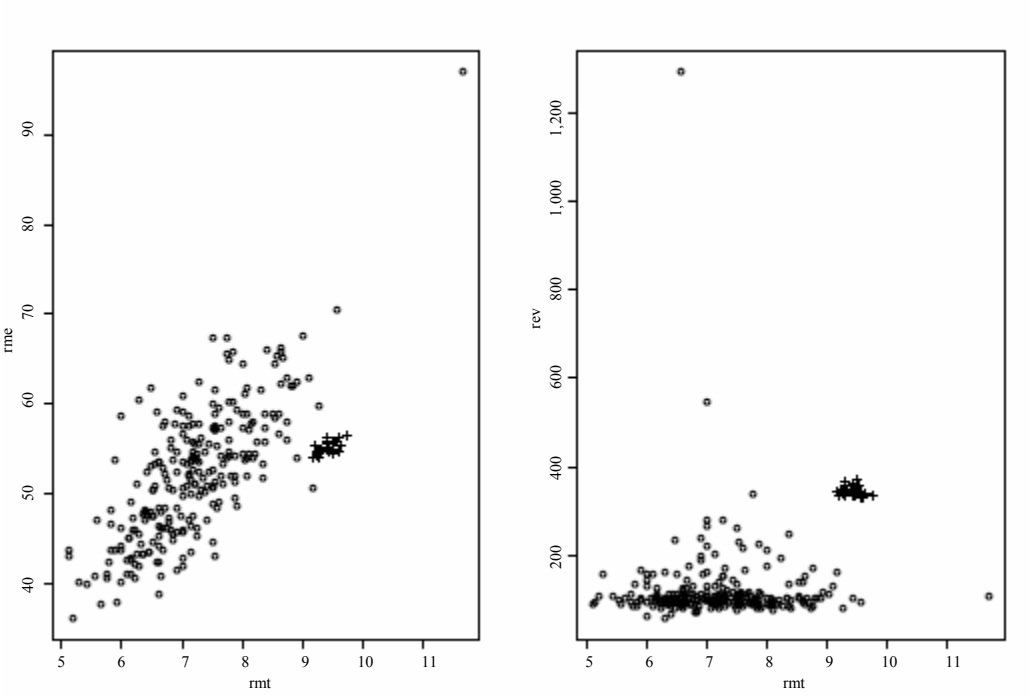

Figure 3 Variables me and rev *vs.* rmt. Additional outliers are marked with a cross

The last row in Table 7 shows that its non-robust imputation leads GIMCD to nominate more outliers with missing values than the other methods which robustify their imputations already before the detection phase.

Table 7 MU281 with missingness and moderate additional contamination

Method	n. out	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Additional	32	31	19	6	27	27	28	27
Weighted basic	20	4	20	20	11	20	12	20
Unweighted basic	25	13	15	12	23	19	23	18
Complete		24	34	43	38	40	40	40

n. out: number of outliers, n. complete: number of complete observations among detected outliers.

In order to check the breakdown of the methods when a high number of additional outliers contaminates the data we set outliers if (LABEL mod 2 = 0 and pop75 $\geq$ 20) or (LABEL mod 3 = 1 and pop75 < 20). Excluding observations which are already basic outliers there are 98 additional outliers. Thus together with the 25 unweighted basic outliers we obtain 43.8% outliers. The threshold for the methods is therefore set at the 123rd largest Mahalanobis distance. Table 8 shows that, due to this large threshold, all weighted basic outliers are detected by all methods. The methods GIMCD, ER and ERw cannot cope with the high rate of outliers. BEM detects most of the outliers with BEMw and TRC only slightly behind. TRCw detects somewhat less of the unweighted basic outliers and of the additional outliers.

Table 8 MU281 with missingness and high additional contamination

Method	n. out	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Additional	98	20	19	37	85	85	85	80
Basic weighted	20	20	20	20	20	20	20	20
Basic unweighted	25	21	19	17	23	18	18	13

6. Conclusions

The EM-algorithm for multivariate normal data can be adapted to a sampling context. The BACON algorithm protects the non-robust EEM-algorithm from outliers when the latter is applied within an iteration of the BACON algorithm. The ER-algorithm uses robustification within the EM-algorithm. The applications showed that this may not yield enough robustification. A possible reason, however, may also be the non-robust starting point of the M-step in the ER-algorithm.

GIMCD, a non-robust EM-algorithm followed by an imputation and detection with MCD covariance worked remarkably well for moderate missingness and contamination. Its variability with high missingness rate is a disadvantage. More stable solutions which also can take into account the sampling design should be explored. The

BACON-EEM algorithm showed very good detection capabilities in particular when the missingness rate and the contamination rate are high.

In spite of its simplicity the TRC algorithm is a good method in many circumstances. Its main problem seems to be the *ad-hoc* imputation with only one covariable, which can be a problem with high missingness rates.

In order to find a good model for the population it is important to use the sampling weights. Nevertheless, it is advisable to use also a non-weighted version and to check the differences. It is possible that outliers are masked by large sampling weights because they may then dominate the model estimate.

Acknowledgement

The EUREDIT research project was part of the Information Society Technology Program (IST) of Framework Program 5 of the European Union. The Swiss participation in EUREDIT was supported by the Swiss Federal Office for Education and Science. A large part of this research was carried out while both authors worked at the Swiss Federal Statistical Office. The authors wish to thank the referees and editors for their valuable remarks.

Appendix

Missingness in MU281

The default response pattern is 111, indicating that the three variables rmt85, me84 and rev84 are all present. A 0 in the string indicates a missing value for the corresponding variable. First, for all strata the response pattern is changed accordingto the following scheme with parameters $(a, b, c) = (1, 2, 3)$:

$$\text{response pattern} = \begin{cases} 011, & \text{if LABEL mod } 20 = a; \\ 101, & \text{if LABEL mod } 20 = b; \\ 110, & \text{if LABEL mod } 20 = c; \\ 100, & \text{if LABEL mod } 30 = a; \\ 010, & \text{if LABEL mod } 30 = b; \\ 001, & \text{if LABEL mod } 30 = c\,. \end{cases}$$

Additionally, the above scheme with parameters $(a, b, c) = (5, 6, 7)$ is applied for stratum 1 again.

References

Atkinson, A. (1993). Stalactite plots and robust estimation for the detection of multivariate outliers. In *Data Analysis and Robustness*. (Eds., S. Morgenthaler, E. Ronchetti and W. Stahel), Birkäuser.

Béguin, C. (2002). Outlier detection in multivariate data. Master's thesis, Université de Neuchâtel.

Béguin, C., and Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data. Deliverable D4/5.2.1/2 Part C, EUREDIT.

Béguin, C., and Hulliger, B. (2004). Multivariate oulier detection in incomplete survey data: The epidemic algorithm and transformed rank correlations. *Journal of the Royal Statistical Society*, A 167(Part 2.), 275-294.

Billor, N., Hadi, A.S. and Vellemann, P.F. (2000). BACON: Blocked Adaptative Computationally-efficient Outlier Nominators. *Computational Statistics and Data Analysis*, 34(3), 279-298.

Campbell, N. (1989). Bushfire mapping using noaa avhrr data. Technical report, CSIRO.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396), 1063-1069.

Cheng, T.-C., and Victoria-Feser, M.-P. (2000). Robust correlation estimation with missing data. Technical Report 2000.05, Université de Genève.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (c/r: P22-37). *Journal of the Royal Statistical Society*, Series B, *Methodological*, 39, 1-22.

Donoho, D. (1982). *Breakdown Properties of Multivariate Location Estimators*. Ph.d. qualifying paper, Department of Statistics, Harvard University.

EUREDIT (2003). *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project*, Volume 1 and 2. EUREDIT consortium. http://www.cs.york.ac.uk/euredit/results/results.html.

Gnanadesikan, R., and Kettenring, J.R. (1972, March). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28, 81-124.

Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society*, B, 54(3), 761-771.

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

Kosinski, A.S. (1999). A procedure for the detection of multivariate outliers. *Computational Statistics & Data Analysis*, 29, 145-161.

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

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

Liu, R.Y., Parelius, J.M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27(3), 783-858.

Maronna, R., and Zamar, R. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4), 307-317.

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

Maronna, R.A., and Yohai, V.J. (1995). The behaviour of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429), 330-341.

Novo, A.A., and Schafer, J.L. (2002). *norm*: *Analysis of multivariate normal datasets with missing values*. R package version 1.0-9.

R Development Core Team (2006). *R*: *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rocke, D., and Woodruff, D. (1993). Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica*, 47, 27-42.

Rocke, D., and Woodruff, D. (1996). Identification of outlier in multivariate data. *Journal of the American Statistical Association*, 91(435), 1047-1061.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, Volume B, 283-297. Elsevier.

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

Schafer, J. (2000). *Analysis of Incomplete Multivariate Data,* Volume 72 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

Stahel, W. (1981). *Robuste Schätzungen: infinitesimale optimalitätund Schätzungen von Kovarianzmatrizen*. Ph.D. Thesis, Swiss Federal Institute of Technology.

Venables,W.N., and Ripley, B.D. (2002). *Modern Applied Statistics with S* (Fourth Ed.). New York: Springer. ISBN 0-387-95457-0.

Wilks, S.S., and Gnanadesikan, R. (1964). Graphical methods for internal comparisons in multiresponse experiments. *Annals of Mathematical Statistics*, 35, 623-631.