EUREDIT
Deliverable D4/5.2.1/2 Part C

Robust Multivariate Outlier Detection and Imputation with Incomplete Survey Data

28 March 2002

Cédric Béguin and Beat Hulliger

Swiss Federal Statistical Office

Statistical Methods Unit

CH-2010 Neuchâtel

This report describes work for the EUREDIT project of the IST-Programme under the Framework Programme 5 for Research and Technical Development of the European Union. The contribution of the Swiss Federal Statistical Office is financed by the Swiss Federal Office of Education and Science.

New material

This version contains all the texts written by SFSO up to 06.03.2002 and updated at 28.03.2002. The numbering of the different parts/sections has changed, we therefore indicate the new material added to the last version of the report (31.08.01). Other slight changes have been made to the pre-existing parts/sections.

The following parts/sections have been added to the last version (31.08.01):

- Part I, Section 4
- Part V
- Part VI

Contents

Foreword

This report describes the work of the Swiss Federal Statistical Office (SFSO) for EUREDIT project workpackages 4.2 and 5.2 under the Information Society Technology Program (IST) of Framework Program 5 of the European Union. The participation of SFSO to EUREDIT is financed by the Swiss Federal Office of Education and Science.

EUREDIT workpackages 4.2 and 5.2 have been unified into workpackage x.2, now called ”Develop and evaluate new methods for statistical outlier detection and outlier robust multivariate imputation”. The main effort of SFSO for EUREDIT goes into this workpackage and SFSO is the leader of it.

This draft of 31 August 2001 describes the outlier detection methods that SFSO has explored or developed until that date. These methods have been tested with real and artificial data sets and they have been adapted to cope with sampling weights.

Future updates of the report will describe the adaption of the outlier detection methods to missing values and the development of imputation methods based on the outlier detection methods. The evaluation with the data sets and in the formal frame work established in EUREDIT workpackages 2 and 6 will also be added.

We would like to thank Werner Stahel, Ali Hadi and Yves Tillé as well as our partners in workpackage x.2 and in EUREDIT overall for fruitful discussions on multivariate outlier detection. We would like to thank our colleagues from the Statistical Methods Unit of SFSO for their support and understanding.

Summary

EUREDIT will develop, evaluate and disseminate new tools aimed at improving the quality of statistical data through improved data editing and imputation. In EUREDIT the term editing means error localization, i.e. identifying doubtful or erroneous data values. In this report we are looking at a particular type of error, namely outliers. Error localization is usually achieved via the calculation of indices that measure the potential for particular data values to be in error. In our case such an index is a measure of outlyingness. Previously, in many cases these indices have been based on strong assumptions about the nature of the population from which the data values were obtained. For example, with univariate continuous data one can apply an outlier test based on the standard deviation. Such tests typically assume that the data are generated from a low dimensional symmetric distribution (e.g. the bivariate normal). This is at odds with the high dimensional mixed categoric-continuous nature of modern data sets. EUREDIT will evaluate and compare a range of both currently used as well as new methods for outlier detection and robust imputation.

The objectives of the EUREDIT project as a whole are described in six different points.

1. To establish a standard collection of datasets.
2. To develop a methodological evaluation framework.
3. To evaluate current ”in-use” methods for data editing and imputation and to develop and evaluate a selected range of new or recent techniques for data editing and imputation.
4. To compare all methods tested and develop a strategy for users of edit and imputation leading to a ”best practice guide”. This evaluation is made using criteria developed in 2. applied to the results given by the methods selected in 3. acting on the data sets chosen in 1.
5. To disseminate selected methods on a project-wide basis by developing prototype software.
6. To exploit the results of the project by developing planned routes to exploitation.

This report will concentrate on points 3 and 4 and editing is interpreted as outlier detection while imputation is interpreted as robust imputation. In order to avoid excessive ”tuning” of methods to a particular situation (one of the major concerns in EUREDIT) SFSO’s strategy is to clearly separate these two phases. Therefore all methods selected for the project are developed totally independently of the two datasets on which they will be evaluated. The first chapters are concerned with the development phase (point 3). Future updates of this report will describe the evaluation phase (point 4).

After a short introduction recalling the classical knowledge and well known concepts of outlier detection and introducing the notations used in this report, the second part will explain how the different multivariate outlier detection methods chosen for EUREDIT were selected. Five methods are emphasized, one classical method (minimization of scale), two modified existing methods (forward search and projection pursuit) and two new methods (simple and nonparametric). The third part will show a comparison of these methods applied to development data sets (none of the evaluation datasets of EUREDIT). The fourth part will describe how these methods have to be modified to account for sampling weights. The fifth part adds the problem of missing values, but by lack of resources and time only three methods are modified to cope with missing value. Finally the sixth part will introduce an imputation method that takes into account outliers, edit failures and missing values.

Part I

Introduction

A very important aspect of statistical data editing is outlier detection. Besides graphical tools, robust mathematical algorithms can be used to detect outliers. Imputation in the presence of outliers has to control the influence of the outliers on the imputation model and must prevent from imputing (non-representative) outliers. Dealing with outliers is considered an essential part of the edit and imputation process. Most outlier-detection and imputation methods are univariate or bivariate in nature and can handle only continuous data. However, real errors in data are usually multivariate and consist of a mix of categorical and highly skewed continuous variables. Furthermore real data usually have missing values. Often the data stem from sample surveys, therefore the sample design should be taken into account by outlier-detection methods and by imputation methods. The idea here is to concentrate on the outlier-detection methods and then to develop relatively simple imputation methods based on the outlier-detection methods. The aim of the combination of outlier-detection and imputation will be to develop procedures that preserve the distributional structure as far as possible while remaining robust to outliers in the data.

The problem of outliers becomes much more difficult in two or more dimensions than in only one dimension. While an outlier can only be very small or very large in one dimension (at least for unimodal distributions) in higher dimensions the ”direction” of the outlier becomes more and more difficult because there are infinitely many directions. Outliers may be quite close to the bulk of the data or to a model if the distance is measured in a Euclidean metric. However, if a metric appropriate to the distribution of the bulk of the data is used it may immediately show up. Thus in higher dimensions the form of the point cloud of the bulk of the data must be well represented in the metric used to detect outliers.

In what concerns sampling the approach of SFSO is mainly design-based. However, models are inherently necessary for a meaningful discussion of outliers. Even if the model can be as vague as ”outliers are far from a center of the data” the definition of what ”far” and ”center” mean needs a model.

An important aspect of the models used for outlier-detection is the sub-population that it applies to. For larger data sets one usually has to subdivide the data set in order to obtain a meaningful model for the bulk of the data and then to detect outliers. We call such a sub-population a **reference population**. In other words usually our model is a mixture of models for the different reference populations. The definition of the reference populations is a crucial point in outlier-detection and robust imputation. In this version of the report we shall only treat the case where the reference population is fixed beforehand.

For finite population sampling in addition to the problem of accounting for the sample design, and related to the problem of the modelling of the bulk of the data, we face the question of **representative** and **non-representative** outliers (Chambers, 1986). In fact, we may have outliers in the population with respect to a model for an infinite underlying super-population. For the purpose of outlier-detection the distinction between representative and non-representative outliers is not of prime importance because even if an outlier is a correct observation belonging to the finite population, we would like to detect it because we will have to check it, it may be influential and we may want to treat it specially in the estimation procedure. Anyway, in the face of a detected outlier one usually is not sure whether it is representative or not. The nice thing would be to have a measure of the degree of belief we can have that the outlier is a good observation, some sort of a value of representativity. However, usually we do not have such a value on a continuous scale and we have to take a dichotomous decision: representative or not. Thus after checking an outlier to a certain extent one often assumes that an outlier is representative. Nevertheless, when it comes to imputation and estimation, one treats these representative outliers specially. For example in imputation one would not impute representative outliers in the same way as normal observations because they probably are rare in the population.

For our outlier-detection methods we do not distinguish between representative and non-representative outliers at all. We will introduce some flexibility to consider the ”representativity” of an outlier for the imputation phase.

When selecting outlier-detection methods for this study we had four guiding principles in mind:

Good detection capability: Ideally all outliers are detected but no good observations declared outliers.

Sufficient speed: The algorithmic complexity should make the methods feasible also for large data sets. The computing time should be at most moderate.

High versatility: The assumptions on the data (how much missingness, categoric and continuous variables) should be low, adaption to sampling and missing values should be feasible.

Simplicity: The methods should be simple to teach and apply. Few tuning should be necessary, the know-how needed by users should be limited and simple to explain.

For robust imputation methods the first principle is replaced by

High preservation capability: Ideally the imputed data should be as close as possible to the true data.

1 Definitions and notations

This section will set up a list of the definitions and notations that will be used throughout all this report. The reader should be able to refer to it whenever he’ll need it.

General notations All matrices will be denoted by capital letters, e.g. A, while vectors will always be column vectors and denoted by small letters, e.g. a_i. I_p will denote the identity matrix in dimension p and 1_p the vector of 1’s in the same dimension.

Usual distributions The univariate normal distribution with mean μ and variance σ^2 will be denoted by $N(\mu, \sigma^2)$. Similarly the multivariate normal distribution will be denoted by $N(\mu, \Sigma)$ where this time μ is the vector mean and Σ the covariance matrix. The chi square distribution with p degrees of freedom will be denoted by χ^2_p and its $1-\alpha$ percentile by $\chi^2_{p,\alpha}$.

Data The data will be encoded in a $n \times p$ matrix X. The n lines of X denoted by x_i will correspond to the n observations of the dataset and the p columns denoted by x^j to the p variables observed.

Equivariances Let $x_1, ..., x_n$ be a set of observations in $\mathbb{R}^p$, let $b \in \mathbb{R}^p$ be any point in the Euclidean space and let A be any non singular $p \times p$ matrix. Let $y_1, ...y_n$ be the images of the x_i's through the affine transformation

$$\begin{array}{ccc} \mathbb{R}^p & \longrightarrow & \mathbb{R}^p \\ x & \longmapsto & y = Ax + b. \end{array}$$

Let M be some estimator of location and let S be some estimator of scatter. Then M and S are said to be **affine equivariant** if

$$M(y_1, ..., y_n) = A \cdot M(x_1, ...x_n) + b \text{ and } S(y_1, ..., y_n) = A \cdot S(x_1, ...x_n) \cdot A^t.$$

If the property is true when restricted to orthogonal transformation (A orthogonal and $b = 0$) the estimators are said to be **orthogonal equivariant**.

If the property is true when restricted to scale transformation ($A = aI_p$ a non zero scalar times the identity matrix and $b = 0$) the estimators are said to be **scale equivariant**.

If the property is true when restricted to shift transformation ($A = 0$) the estimators are said to be **shift or location equivariant**.

2 Robust editing

Outlier detection requires a ”metric” that somehow measures the ”outlyingness” of a data point. Typically, the metric arises from some model for the data (for example, a center or a fitted equation) and some measure of discrepancy for that model. A classical way of computing a measure of discrepancy and identifying multivariate outliers is to calculate the Mahalanobis distance. Recall that this distance uses estimators M of location and S of scatter of a set of observations and is defined for an observation x by:

$$MD_{M,S}(x) = (x - M)^t S^{-1}(x - M).$$

Unfortunately both estimators of location and scatter are very sensitive to outlying observations. Therefore robust estimators of both location and scatter have to be used to remedy that problem. Several methods have been reported in the literature for a number of different approaches always with their advantages and disadvantages. Smooth estimators such as maximum likelihood and 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 on the other hand 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 almost eliminatory when dealing with official statistics, most of them being high dimensional data. M-estimators were therefore not considered further in this study. Many other affine equivariant estimators were studied by Donoho (Donoho, 1982) but all have breakdown points at most $1/(p+1)$. Other approaches ended up with affine equivariant high breakdown point estimators but had the disadvantage of being computationally expensive. The first of these approaches was related to the projection pursuit principle: the Stahel-Donoho (SD) estimator (Stahel, 1981), (Donoho, 1982). Other approaches followed like the ones based on the minimization of a robust scale like the Minimum Volume Ellipsoid (MVE), the Minimum Covariance Determinant (MCD) estimators (Rousseeuw, 1985), (Rousseeuw and Leroy, 1987) and S estimators (Davies, 1987). The affine equivariance and high-breakdown point properties seem clearly to imply very high or even infinite computer costs, therefore a robust outlier detection must either approximate the solution, like the ”Fast MCD” (FMCD) (Rousseeuw and van Driessen, 1999) or the Modified Stahel-Donoho (MSD) (Patak, 1990) (both methods will be part of this study in section 8 and 7) or sacrifice affine equivariance. Different ideas for the second solution can already be found in (Gnanadesikan and Kettenring, 1972). Two approaches of Gnanadesikan and Kettenring will be further developed in this study. The first one is based on the fact that each component of a covariance matrix can be computed as the covariance between two variables. Gnanadesikan and Kettenring proposed to robustify this component by component computation and then use a final transformation of the obtained matrix to ensure positive definiteness. We used this idea to define new simple robust estimators of location and covariance in section 5. Note that Maronna and Zamar have also worked in the same direction re-actualizing the ideas of Gnanadesikan and Kettenring, see (Maronna and Zamar, 2001). Another idea found in (Gnanadesikan and Kettenring, 1972) gave birth to the so-called forward search methods (Hadi, 1992), (Atkinson, 1993). The two most recent forward search methods (Kosinski, 1999) and (Billor et al., 2000) are studied in section 6, and a slightly modified version of the BACON (Billor et al., 2000) algorithm is selected for the rest of the study.

The methods based on the Mahalanobis distance will be adapted to cope with missing values by an EM-algorithm. For the MCD-method this has been done by Cheng and Victoria-Feser (Cheng and Victoria-Feser, 2000). The adaption to sampling is relatively easy for these methods.

Nonparametric or semi-parametric approaches of outlier detection like data depth (Liu et al., 1999) or multivariate quantiles seem also very attractive and promising, but unfortunately due to the lack of resources these methods were not included in SFSO’s work for EUREDIT. Nevertheless an alternative nonparametric method that seems to be new is introduced in section 9 (Hulliger and Béguin, 2001). The idea is to start an epidemic in the population at some well chosen point and let it grow. The last infected points should be outliers.

Some authors do think that only hybrid methods using elements from the different approaches quoted above will have a chance to extend the practical boundaries of outlier detection capabilities. Trying to combine the different methods was not an option chosen for this study because it runs contrary to the guiding principles above, in particular simplicity. the reader who's willing to measure the effect of a hybrid method is advised to read (Rocke and Woodruff, 1996). Note that Kosinsky has compared the method proposed by Rocke and Woodruff with his algorithm (see Section 6).

3 Robust imputation

The idea is to use the outlier-detection methods for the purpose of ”outlier”-imputation as well. Since we have a division of the data in outliers and good data we will impute good data for the outliers. If we think that some of the outliers might be representative then we might relax the boundary of the good data somewhat compared with the outlier-detection phase. Missing values will have to be imputed for observations which are not considered outliers. We will not use any sophisticated method like logistic regression or neural networks here. Obviously these could be applied once the outliers are imputed.

The methods that end up with a robust estimate of the center and the covariance of the data lead to two simple ways of imputation for outliers. The first method is to take a limit of the good data described by an ellipsoid of equal Mahalanobis distance and to project an outlier to the closest point of the ellipsoid. In other word we censor the outliers or still in other words we winsorize the outliers metrically. The second imputation method would be to impute (may be with probability proportional to the distance) an observation from the good (non-outlying) observations which is close to the outlier. Thus this is a nearest neighbour imputation with a restriction on the donors. The limiting distance for winsorizing or the border of good data for nearest neighbour imputation is a parameter that can be used to adapt for representative outliers.

The missing values of observations which ar not declared outliers can be imputed randomly by a Nearest Neighbour from the good data.

The Epidemic Algorithm can be run backwards starting from a detected outlier until the epidemic infects one or several good and complete observations. Then among these infected good observations we may select one at random for imputation. The same process may be used for non-outlying observations with missing values. Thus the epidemic algorithm run backwards is a nearest-neighbour imputation method with a very particular type of distance.

4 A modular system for data preparation

The treatment of data from raw input to data which is of defined quality is very complex. Usually several phases interact and there are loops which individual data or the whole of the data go through several times. Ideally the system would be completely automated but in practice manual controls and corrections often must occur. Also the integration of true values due to call backs is possible. Every survey has its own specialities and therefore there cannot be a system which covers all of the tasks in the sequence needed. The only way to make the building of such a system easier is to have modules at hand, which do specific subtasks, which are parameterised and which can be built easily into a system. A simple example of such a modular system is shown here. It is merely developed for the purposes of the EUREDIT project. But of course the modules may be used in a more complex system.

4.1 The system

We first describe the system in general terms and then look closer at the modules it contains. Modules we may consider are

E: A control module which flags missing values and applies edit rules that control which of the values of a record might be in error.

C: A correction module which corrects failing items or missing items which fulfill specific conditions in a deterministic way. For example we may fill in a missing total if all subtotals are given by just summing the subtotals. Or we may recalculate the age from the year of birth if there is a contradiction between the given age and the year of birth.

L: An error localisation module which narrows down the set of values which might be in error.

D: An outlier detection module which flags possible outliers or calculates a robustness weight.

I: An imputation module which imputes for missing values, outliers and failing items.

M: A manual correction module which allows correction and imputations by human intervention.

The data that should be treated may be composed of observations on categorical (ordered and unordered) and continuous variables.

Each of these modules should have a defined standard input and output, a defined set of parameters and a defined set of informations for the user to judge its performance. Of course there may be several different possible methods and algorithms for a module. For example imputation may be done with the help of linear models or with a nearest neighbor method. Outlier detection may use non-parametric or parametric methods. The point is that the input and output of each module should be defined in such a way that different methods can be chained as modules to form a system.

A system like NIM from Statistics Canada resolves the tasks of several of the above modules in a more interconnected way. E.g. NIM does a check on whether a possible imputation actually resolves all edit failures at the very moment of the imputation. Thus the E and I module of NIM are intimately connected. The disadvantage is that NIM cannot be combined easily with other modules like a D module or an M module.

The system we use for EUREDIT consists of the following sequence: DEIE or EDIE. A system like EDIE checks only after imputation whether the edit failures actually have been resolved. In other words, after applying the system EDIE we cannot be sure to obtain failure free records! We then may have to add a manual correction module followed by the E module again. This would amount to a EDIEME system. Of course we might also change certain parameters of the E, D, I modules and rerun the EDIE system in the hope to get a result we can live with.

The main effort for this report is concentrated on a set of D modules. The I module is needed to have at least a minimum output to be evaluated with the EUREDIT criteria.

4.2 The modules

4.2.1 Module E

Module E is the module that controls the correctness of data with edit rules.

Input: The $n \times p$ matrix of Data X. The n vector of weights w.

Parameters: A set of rules $C_k, k = 1, \dots, K$.

Output: The $n \times p$ matrix R of response indicators r_{ij}. The $n \times p$ matrix E of indicators e_{ij} of edit passes.

Each rule C_k is a function which maps x_i to 0 or 1. If an observation fails the rule, its result is 1, if it passes its result is 0. Let J_k be the sub-set of variables on which the function C_k depends. We define a p vector $c_{kj}(x_i)$ as follows:

$$c_{kj}(x_i) = \begin{cases} 1 & \text{if } j \in J_k \text{ and } C_k(x_i) = 1 \\ 0 & \text{if } j \in J_k \text{ and } C_k(x_i) = 0, \\ 0 & \text{if } j \notin J_k. \end{cases} \tag{1}$$

In other words the $c_{kj}(x_i) = 1$ if the observation fails rule k and rule k involves variable j. Of course a rule cannot be applied to an observation if $\prod_{J_k} r_{ij} = 0$, i.e. if it depends on a missing observation. We then set $c_{kj}(x_i) = 0$.

The entries of the matrix E are calculated as

$$e_{ij} = 1\{\sum_{k=1}^{K} c_{kj}(x_i) = 0\} = \begin{cases} 1 & \text{if } \sum_{k=1}^{K} c_{kj}(x_i) = 0, \\ 0 & \text{otherwise}. \end{cases}. \tag{2}$$

Another measure which might be useful as output would be

$$\tilde{e}_{ij} = \frac{\sum_{k=1}^{K} c_{kj}(x_i)}{\sum_{k=1}^{K} 1\{j \in J_k\} r_{ij}}. \tag{3}$$

Thus $\tilde{e}_{ij}$ is the proportion of rules that fail and contain item x_{ij} among the rules that actually can fail for this item. Thus $\tilde{e}_{ij}$ might be useful for error localisation or later on in the distances.

4.2.2 Module D

Module D is the module for outlier detection.

Input: The data X, the weights w. The matrix of edit passes E.

Parameters: Tuning constants for the severity of outlier detection. Type of weighting functions. Number of iterations or convergence criterion.

Output: The vector of robustness weights u.

4.2.3 Module I

Module I is the module for imputation.

Input The data X, the sampling weights w, the robustness weights u, the matrix of edit passes E, the matrix of response indicators R.

Parameters Tuning constants for severity of outlier imputation. Tuning constants for conditions on donors.

Output The imputed data $\tilde{X}$.

Part II

Selected methods for multivariate outlier detection

As described in the introduction the first four sections of this chapters furnish outlier detection methods based on robust Mahalanobis distances. Recall that for an estimate M of location and an estimate S of scatter the Mahalanobis distance of an observation x is computed as

$$MD_{M,S}(x) = (x - M)^t S^{-1} (x - M).$$

The first section will introduce new simple robust estimators of location and scatter based on ideas of Gnanadesikan and Kettenring (Gnanadesikan and Kettenring, 1972). The second one will report the selection made between the two most recent forward search method, namely Kosinski algorithm (Kosinski, 1999) and BACON algorithm (Billor et al., 2000). The third one will describe a modified version of the first high breakdown point affine equivariant method related to the projection pursuit principle (Stahel, 1981), (Donoho, 1982). The fourth one will recall one of the most popular and well used high breakdown point affine equivariant method based on the minimization of a robust scale of Mahalanobis distances (Rousseeuw, 1985), (Rousseeuw and Leroy, 1987). Finally the last section will introduce a nonparametric method based on an approach that seems to be new, the epidemic algorithm.

5 A simple method

In order to evaluate sophisticated methods used to detect multivariate outliers we try to find simple estimators of the mean and the covariance matrix. We seek computationally non-expensive estimators that are suitable for detection in large and high dimensional datasets. In other sections we shall study and compare sophisticated methods with high breakdown point but also with heavy computation needs: methods based on the minimization of a robust scale (Minimum Covariance Determinant, MCD), based on projections (Modified Stahel-Donoho, MSD) or based on an epidemic spread through the data (Epidemic Algorithm, EA). Only one of the studied methods seems to be computationally economic: the forward search method (BACON). Here the idea is to define estimators of mean and scatter that do not need any fancy algorithm to be computed and that retain some direct statistical meaning.

A first step in this direction was made by Gnanadesikan and Kettenring (Gnanadesikan and Kettenring, 1972). The authors used the fact that the components of the covariance matrix can be written as:

$$\text{cov}(x, y) = \frac{1}{4} \left(\sigma^2(x + y) - \sigma^2(x - y) \right),$$

where x and y are two univariare random variables. Using a robust estimator of univariate variance σ^* (they used trimmed or Winsorized variance) they replaced the usual variance σ by σ^* in the above formula. Doing so they obtained some ”covariance” or ”correlation” matrix that is not necessarily positive definite. They then used some transformation to ensure positive definiteness and obtain an estimator of the covariance matrix; such transformations are detailed in (Rousseeuw and Molenberghs, 1993).

We develop here quite similar ideas. We use rank statistics as robust estimate of correlation between variables and we do a different transformation to ensure positive definiteness using principal components. Then we propose to add one reweighting M-step to improve performance.

5.1 Approximation of correlation coefficients

Our idea is to use the Spearman rank correlation R to approximate the usual correlation ρ. We use the following proposition; see (van der Waerden, 1971) §70.

Proposition 1 *Let X, Y be two normal variables, let ρ be the correlation coefficient between X and Y, let x and y be two samples of X and Y, let $R(x,y)$ be the Spearman Rank correlation of the two samples. The following estimator is consistent for ρ:*

$$\widetilde{R}(x,y) = 2\sin\left(\frac{\pi}{6}R(x,y)\right)$$

This estimator will be used to construct the correlation matrix coefficient by coefficient.

5.2 Construction of the estimators (SMP and RSMP)

Our construction of simple robust estimators of the mean and the covariance matrix is as follows:

Let X be the $n \times p$ matrix of the data, with n observations and p variables. All vectors will be written in column. Denote by x_i, $i = 1, .., n$, the i^{th} line (observation) of the matrix X and by x^j, $j = 1, .., p$, the j^{th} column (variable). Let $\widetilde{\mu}$ and $\widetilde{\sigma}^2$ be robust estimators of the mean and variance for univariate data.

(i) Construct the $p \times p$ symmetric matrix $\widetilde{S}_1 = \widetilde{\Sigma}\widetilde{R}\widetilde{\Sigma}$ where $\widetilde{\Sigma} = diag(\widetilde{\sigma}(x^j))$ and $\widetilde{R}_{jk} = \widetilde{R}(x^j, x^k)$.

(ii) Let B be the orthogonal matrix such that $\widetilde{S}_1 = B\Lambda B^t$, with Λ diagonal. Define m with $m_j = \widetilde{\mu}((XB)^j)$ and $\Xi = diag(\widetilde{\sigma}^2((XB)^j))$.

(iii) The simple robust estimators (SMP) for the mean and covariance matrix are $\widetilde{m} = Bm$ and $\widetilde{S} = B\Xi B^t$.

In other words this algorithm computes in (i) some robust but not necessarily positive definite estimation of the covariance matrix. The ”principal components” of this matrix are then used in (ii) to robustly estimate univariate location and scatter in these directions. The SMP estimators are eventually constructed from the estimates of location and scatter obtained on these robust estimates of the principal components by a back transformation into the original basis.

Remarks:

a) If besides outlier detection variance problematic is of interest we could possibly add one reweighting step to improve efficiency. Denote by $d_i = (x_i - \widetilde{m})^t \widetilde{S}(x_i - \widetilde{m})$ the Mahalanobis distances and let u be a weight function, the new estimators (RSMP) would then just be weighted mean and covariance:

$$\widetilde{m}_u = \frac{\sum_{i=1}^n u(d_i) x_i}{\sum_{i=1}^n u(d_i)} \qquad \widetilde{S}_u = \frac{\sum_{i=1}^n u(d_i)(x_i - \widetilde{m}_w)(x_i - \widetilde{m}_w)^t}{\sum_{i=1}^n u(d_i)}$$

As a weight function we may use Huber weights $u \ : \ \mathbb{R}^+ \to \mathbb{R}^+,\ d \mapsto u(d) = \begin{cases} d & \text{if } d \le k \\ k & \text{if } d > k \end{cases}$, where k is chosen to give an estimator with reasonable performance, or other redescending weights function.

b) In our simulations we use $\widetilde{\mu} = median$ and $\widetilde{\sigma} = mad$ with the mad scaled by a multiplicative constant to be a consistent estimator of the standard deviation at the Gaussian model. These particular simple (resp. reweighted) estimators will be denoted by m_{SMP} (resp. m_{RSMP}) and S_{SMP} (resp. S_{RSMP}) in the next sections. Other SMP estimators defined for example with trimmed or Winzorised mean and variance would have to be explored.

5.3 Properties of the estimators

Lemma 5.1 *Suppose that $\widetilde{\mu}$ and $\widetilde{\sigma}^2$ are shift and scale equivariant then the SMP estimators are shift and scale equivariant.*

Proof 1. Shift equivariance
Denote by $y_i = x_i + b$ the shifted observations with $b = (b^1, ..., b^p) \in \mathbb{R}^p$, i.e $Y = X + 1_n b^t$ where 1_n is the n-vector with all components equal to 1. By definition we have that $\widetilde{R}(Y) = \widetilde{R}(X)$. As $\widetilde{\sigma}^2$ is shift equivariant we also have that $\widetilde{\Sigma}(Y) = \widetilde{\Sigma}(X)$. Therefore $\widetilde{S}_1(Y) = \widetilde{S}_1(X)$ implying $B(Y) = B(X)$. Finally using the assumptions on $\widetilde{\mu}$ and $\widetilde{\sigma}$ we

have

$$\begin{aligned}
m_j(Y) &= \widetilde{\mu}((YB(Y))^j) = \widetilde{\mu}(((X + 1_n b^t)B(X))^j) \\
&= \widetilde{\mu}((XB(X) + 1_n b^t B(X))^j) = \widetilde{\mu}(XB(X))^j) + (B^t(X)b)_j \\
&= m_j(X) + (B^t(X)b)_j \\
\Longrightarrow & \\
m(Y) &= m(X) + B^t(X)b \\
\Longrightarrow & \\
\widetilde{m}(Y) &= B(Y)m(Y) = B(X)(m(X) + B^t(X)b) = \widetilde{m}(X) + b
\end{aligned}$$

and

$$\begin{aligned}
\Xi(Y) &= diag(\widetilde{\sigma}^2((YB(Y))^j)) = diag(\widetilde{\sigma}^2(((X + 1_n b^t)B(X))^j)) \\
&= diag(\widetilde{\sigma}^2((XB(X) + 1_n b^t B(X))^j)) = diag(\widetilde{\sigma}^2((XB(X))^j)) = \Xi(X) \\
\Longrightarrow & \\
\widetilde{S}(Y) &= B(Y)\Xi(Y)B^t(Y) = B(X)\Xi(X)B^t(X) = \widetilde{S}(X)
\end{aligned}$$

2. Scale equivariance
Denote by $y_i = ax_i$ the scaled observations with $a \in \mathbb{R}\backslash\{0\}$, i.e $Y = aX$. By definition we have that $\widetilde{R}(Y) = \widetilde{R}(X)$. As $\widetilde{\sigma}^2$ is scale equivariant we also have that $\widetilde{\Sigma}(Y) = a\widetilde{\Sigma}(X)$. Therefore $\widetilde{S}_1(Y) = a^2\widetilde{S}_1(X)$ implying $B(Y) = B(X)$. Finally using the assumptions on $\widetilde{\mu}$ and $\widetilde{\sigma}$ we have

$$\begin{aligned}
m_j(Y) &= \widetilde{\mu}((YB(Y))^j) = \widetilde{\mu}((aXB(X))^j) \\
&= a\widetilde{\mu}(XB(X))^j) = am_j(X) \\
\Longrightarrow & \\
m(Y) &= am(X) \\
\Longrightarrow & \\
\widetilde{m}(Y) &= B(Y)m(Y) = B(X)(am(X)) = a\widetilde{m}(X)
\end{aligned}$$

and

$$\begin{aligned}
\Xi(Y) &= diag(\widetilde{\sigma}^2((YB(Y))^j)) = diag(\widetilde{\sigma}^2(aXB(X))^j)) \\
&= diag(a^2\widetilde{\sigma}^2((XB(X))^j)) = a^2\Xi(X) \\
\Longrightarrow & \\
\widetilde{S}(Y) &= B(Y)\Xi(Y)B^t(Y) = B(X)a^2\Xi(X)B^t(X) = a^2\widetilde{S}(X)
\end{aligned}$$

■

Remark However as the rank statistics do change when the data are rotated, the SMP estimators are neither orthogonal nor affine equivariant.

The construction was made to make the estimators consistent at the multivariate normal model:

Lemma 5.2 *If $\widetilde{\mu}$ and $\widetilde{\sigma}$ are consistent estimators for resp. the location and the scale at the univariate normal model $N(\mu, \sigma^2)$ then the SMP estimators are consistent for the location and the shape at the multivariate normal model $N(\mu, \Sigma^2)$.*

Proof By proposition 1 and the fact that $\widetilde{\sigma}$ is consistent, we have that $\widetilde{S}_1$ is a consistent estimator for the covariance matrix under multivariate normal distribution. By continuity of the eigenvectors of a matrix, the estimated principal components will be consistent for true real principal components. Therefore B will be a consistent estimator of the matrix that orthogonally diagonalizes the covariance matrix. The assumption that $\widetilde{\mu}$ and $\widetilde{\sigma}$ are consistent concludes the proof. ■

6 A forward search method

In this section we deal with methods based on the concept of ”growing a good subset of observations”. By ”good subset” we mean a subset free of outliers. The idea is to start with a small subset of the data and then add non-outlying observations until no more non-outliers are available.

The first criterion to check the outlyingness of one single point in multivariate data can be tracked back to the article of Wilks in 1963 (Wilks, 1963). The author used the so called one-outlier scatter ratio as a measure of outlyingness. This ratio is defined as a ratio of determinants of sample covariance matrices in the following way. Let $x_1, ..., x_n$ be a set of points in $\mathbb{R}^p$, denote by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t$ the usual sample mean and covariance matrix. Let's add $y \in \mathbb{R}^p$ to the set of points and denote by $\bar{x}_y$ and S_y the new sample mean and covariance matrix. The one-outlier scatter ratio of Wilks is defined as

$$R_y = \frac{|S_y|}{|S|}$$

where $|\cdot|$ is the determinant function. Wilks studied this criterion and extended it to two or three added points but did not include any iterating process in his article. The idea of a forward search algorithm was suggested by Wilks and Gnanadesikan in 1964 (Wilks and Gnanadesikan, 1964). We report here the description made in (Gnanadesikan and Kettenring, 1972).

The first step in the procedure is to rank the multiresponse observations $x_1, ..., x_n$ in term of their Euclidean distance $\| x_i - x^* \|$ from some robust estimator of location x^*. A subset G_0 of the observations whose ranks are the smallest $100(1-\beta_0)\%$ is then chosen and used to compute a sum-of-product matrix

$$A_{G_0} = \sum_{i \in G_0}(x_i - x^*)(x_i - x^*)^t.$$

The size of G_0 is chosen big enough in order to ensure that A_{G_0} is not singular. Then all n observations are ranked in terms of the values of the quadratic form $(x_i - x^*)A_{G_0}^{-1}(x_i - x^*)^t$. A new subset G_1 of the observations whose ranks are the smallest $100(1-\beta_1)\%$ is chosen.

The steps are then repeated with new β_i and G_i. The process is iterated until a ”stable” estimator of the covariance matrix is obtained :

$$S^*_{G_i} = \frac{k}{n(1-\beta_i)} \sum_{i \in G_i} (x_i - x^*)(x_i - x^*)^t,$$

where k is some constant chosen to make the estimator unbiased.

Probably due to the lack of computer resources these ideas were not developed any further by Wilks and Gnanadesikan. Let us remark here that to grow the good subset we need some ranking of all the observations based on the good ones. It would seem possible here to use either the Wilks one outlier scatter ratio or the Mahalanobis distances type criterion. These two rankings are actually equivalent. This result is very well known to all specialists but we felt that it was worthwhile to write it once in details.

Lemma 6.1 *Let $G = \{x_1, ..., x_n\} \subset \mathrm{I\!R}^p$ and $B = \{y_1, ..., y_m\} \subset \mathrm{I\!R}^p$ be two sets of observations, let $R_{y_i} = \frac{|S_{G,y_i}|}{|S_G|}$, $y_i \in B$, be the one outlier scatter ratios of the elements of B based on G, let $d_i^2 = MD^2_{\bar{x}_G, S_G}(y_i)$, $y_i \in B$, be the Mahalanobis distances of the elements of B based on G, then*

$$R_{y_i} = \left(\frac{n-1}{n}\right)^p \left(1 + \frac{n}{n^2-1} d_i^2\right)$$

in particular the rankings of the observations in B associated to R_{y_i} and d_i are the same.

Proof To simplify the notations, let us denote $\bar{x} = \bar{x}_G = \frac{1}{n}\sum_{i=1}^n x_i$ and

$$S \; = S_G = \tfrac{1}{n-1}\textstyle\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^t.$$

Similarly for $y \in B$ denote $\bar{x}_y = \frac{1}{n+1}\left(\sum_{i=1}^n x_i + y\right)$ and

$$S_y \; = \tfrac{1}{n}\left(\textstyle\sum_{i=1}^n (x_i - \bar{x}_y)(x_i - \bar{x}_y)^t + (y - \bar{x}_y)(y - \bar{x}_y)^t\right).$$

We have the trivial relations $\bar{x}_y = \frac{n}{n+1}\bar{x} + \frac{1}{n+1}y = \bar{x} + \frac{1}{n+1}(y - \bar{x})$ and with $\varepsilon = \frac{1}{n+1}(y - \bar{x})$

$$\begin{aligned} nS_y &= \textstyle\sum_{i=1}^n (x_i - \bar{x} - \varepsilon)(x_i - \bar{x} - \varepsilon)^t + (y - \bar{x} - \varepsilon)(y - \bar{x} - \varepsilon)^t \\ &= (n-1)S - \varepsilon \textstyle\sum_{i=1}^n (x_i - \bar{x})^t - \sum_{i=1}^n (x_i - \bar{x})\varepsilon^t + n\varepsilon\varepsilon^t \\ &\qquad +(y - \bar{x})(y - \bar{x})^t - \varepsilon(y - \bar{x})^t - (y - \bar{x})\varepsilon^t + \varepsilon\varepsilon^t \\ &= (n-1)S - 0 - 0 + n\varepsilon\varepsilon^t \\ &\qquad +(n+1)^2\varepsilon\varepsilon^t - (n+1)\varepsilon\varepsilon^t - (n+1)\varepsilon\varepsilon^t + \varepsilon\varepsilon^t \\ &= (n-1)S + n(n+1)\varepsilon\varepsilon^t \end{aligned}$$

i.e.

$$S_y \; = \tfrac{n-1}{n}S + \tfrac{1}{n+1}(y - \bar{x})(y - \bar{x})^t$$

A classical result of determinants computation states that for any $n \times n$ square matrix A and any n vector b we have

$$|A + bb^t| = |A|(1 + b^t A^{-1} b).$$

Applying this result to the last equality gives

$$\begin{aligned} |S_y| &= \left|\tfrac{n-1}{n} S\right| \left(1 + \tfrac{1}{n+1}(y - \bar{x})^t \tfrac{n}{n-1} S^{-1}(y - \bar{x})\right) \\ &= \left(\tfrac{n-1}{n}\right)^p |S| \left(1 + \tfrac{n}{n^2-1}(y - \bar{x})^t S^{-1}(y - \bar{x})\right). \end{aligned}$$

And finally

$$\begin{aligned} R_{y_i} = \frac{|S_{y_i}|}{|S|} &= \left(\tfrac{n-1}{n}\right)^p \left(1 + \tfrac{n}{n^2-1}(y_i - \bar{x})^t S^{-1}(y_i - \bar{x})\right) \\ &= \left(\tfrac{n-1}{n}\right)^p \left(1 + \tfrac{n}{n^2-1} d_i^2\right). \qquad \blacksquare \end{aligned}$$

After the articles of Wilks and Gnanadesikan almost 30 years will pass before the interest for a forward search algorithm grew up again. Articles by Hadi (Hadi, 1992) and Atkinson (Atkinson, 1993) will start to demonstrate the efficiency of such methods. In both articles the growth of the ”good subset” is one point at a time using Mahalanobis distances to rank the observations. Several articles will follow developing faster and more sophisticated methods based on the same idea. The last two and most efficient were developed by Billor, Hadi and Velleman (Billor et al., 2000) and Kosinski (Kosinski, 1999). Both will be presented in the next two subsections. The third subsection will present a comparison that was made to select the most efficient one for our purpose.

6.1 BACON algorithm

The BACON algorithm is presented in (Billor et al., 2000). Two versions are included: one for multivariate data in general and one for regression data. Our interest here will be the first case. The BACON acronym (Blocked Adaptative Computationally-efficient Outlier Nominators) was chosen after the last name of Sir Francis Bacon who wrote in 1620:

> *”Whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways.”*

The idea of the algorithm is similar to the ones presented above. We shall present the detailed algorithm and some properties underlined by Billor et al.

The algorithm The first step will be the choice of an initial basic subset of ”good data”. Two versions are proposed. Let us first describe these two initializations and then state the steps of the algorithm.

The data are stocked in a matrix X of n rows (observations) and p columns (variables). The assumption on the data is that they should be unimodal and roughly elliptical symmetric.

Version 1 (V1) (*Initial subset selection based on Mahalanobis distances*)
For $i = 1, ..., n$ compute the Mahalanobis distances

$$d_i(\bar{x}, S) = \sqrt{(x_i - \bar{x})^t S^{-1} (x_i - \bar{x})}, \quad i = 1, ..., n$$

where $\bar{x}$ and S are the mean and covariance matrix of the n observations. Identify the $m = cp$ observations with the smallest values of d_i. Nominate these as a potential basic subset. c is an integer chosen by the data analyst and set by default to 3.

Version 2 (V2) (*Initial subset selection based on distances from the medians*)
For $i = 1, ..., n$ compute $\| x_i - med \|$, where med is a vector containing the coordinatewise median, x_i is the ith row of X and $\| \cdot \|$ is the Euclidean norm. Identify the $m = cp$ observations with the smallest values of $\| x_i - med \|$. Nominate these as a potential basic subset.

In both versions if S_G (the covariance matrix of the selected data) is singular then increase the basic subset by adding observations with smallest distances until S_G has full rank.

6.1.1 Steps of the algorithm

Step 1 Select an initial basic subset G of size m using either V1 or V2.

Step 2 Compute the discrepancies

$$d_i(\bar{x}_G, S_G) = \sqrt{(x_i - \bar{x}_G)^t S_G^{-1} (x_i - \bar{x}_G)}, \quad i = 1, ..., n$$

where $\bar{x}_G$ and S_G are the mean and covariance matrix of the observations in G.

Step 3 Set a new subset G to all points with discrepancy less than $c_{npr}\chi_{p,\alpha/n}$, where $\chi^2_{p,\beta}$ is the $1 - \beta$ percentile of the chi square distribution with p degrees of freedom, $c_{npr} = c_{np} + c_{hr}$ is a correction factor with

$$c_{hr} = max\{0, (h - r)/(h + r)\}, \; h = \lceil (n + p + 1)/2 \rceil, \; r = |G|$$

$$c_{np} = 1 + \frac{p + 1}{n - p} + \frac{1}{n - h - p} = 1 + \frac{p + 1}{n - p} + \frac{2}{n - 1 - 3p}.$$

Step 4 *The stopping rule:* Iterate Steps 2 and 3 until the size of the basic subset no longer changes.

Step 5 Nominate the observations excluded by the final G as outliers.

6.1.2 Properties of the algorithm

We report here properties of the methods presented in (Billor et al., 2000).

This outlier detection method is computationally efficient. The version with starting point V1 is affine equivariant but less robust. Nevertheless simulations show that it has an empirical breakdown point near 20%. It has a lower computational cost than the other version. The second one with starting point V2 is more robust but only nearly affine equivariant. In simulation trials it offered a breakdown point in excess of 40%.

The small computing effort required by the BACON algorithm, and in particular the fact that this effort grows slowly with increasing sample size, makes this method particularly well-suited for large datasets.

6.1.3 Remark and modification of the step 3 selection criteria

The selection criteria of step 3 is designed for a multivariate normal distribution. In fact under normality it is well known that the Mahalanobis distances follow asymptotically a χ^2 distribution with p degrees of freedom. Suppose all points are derived from a multivariate normal distribution and that the Mahalanobis distance is computed using the all sample mean and covariance matrix, therefore testing the number of points with $MD(x_i) > \chi^2_{p,\alpha}$ should end up with about 100α percents of points detected. The test defined in step 3 is designed in a different way, testing the number of points with $MD(x_i) > \chi^2_{p,\alpha/n}$. Using Bonferroni inequalities we can show that under normality this test will not detect any point with probability $1-\alpha$ (i.e. $P(MD(x_i) < \chi^2_{p,\alpha/n}, \forall i \in \{1,...,n\}) = 1-\alpha$). Now if this test defined this way detects very rarely points that are not outliers it also reduces its sensitivity to close outliers when n becomes large. As we shall have to deal with very large datasets and we are worrying about contamination close to the ”good data” we shall prefer a test using $\chi^2_{p,\alpha}$ instead of $\chi^2_{p,\alpha/n}$. This solution decreases the number of non detected outliers but accepts that under normality about 100α percents of good points are detected as outliers. As BACON algorithm is computationally cheap the analyst should always have the possibility to run the method with both tests and compare the results.

6.2 Kosinski algorithm

In 1999 Kosinski tried to push further the ideas of Hadi and Atkinson to create a method that could cope with high contamination (Kosinski, 1999). We shall present the algorithm in detail after having given several new notations and definitions required for it's understanding. Finally we shall report some conclusion drawn by Kosinski.

Definitions and notations As usual n observations $x_1, ..., x_n \in \mathbb{R}^p$ are considered. For any $E \subseteq D = \{1, ..., n\}$ the number of element in E will be denoted by $|E|$. A partition-based Mahalanobis distance of elements of D is given by a partition (G, B) of D and the distances

$$MD_i(G,B) = (x_i - \bar{x}_G)^t (c^2_{|G|p} S_G)^{-1} (x_i - \bar{x}_G).$$

where the constant $c_{|G|p}$ is defined as in BACON and was originally suggested by Hadi in (Hadi, 1994). An α-partition of D is a partition of D such that

1. $|G| \geq h = \lfloor (n+p+1)/2 \rfloor$;
2. $MD_i(G,B) \geq \chi^2_{p,\alpha}$ for $i \in B$;
3. $\underset{i \in G}{max}\ MD_i(G,B) < \underset{i \in B}{min}\ MD_i(G,B)$;
4. if $|G| > h$ then $MD_i(G,B) < \chi^2_{p,\alpha}$ for all $i \in G$.

The level γ of an α-partition is defined as $\gamma = \underset{i \in B}{max}\ P_i(G,B)$ where

$$P_i(G,B) = Prob\{\chi^2_p \geq MD_i(G,B)\}.$$

Remark here that by property 2 the level γ of an α-partition has to satisfy $\gamma < \alpha$. This fact will be used in the algorithm. The part G is named for the ”good data points” and the part B for the ”bad data points”.

The algorithm is rather sophisticated. Before giving all the technical steps that might not help greatly the understanding of the method we shall try to clarify the progress of the method.

6.2.1 Progress of the algorithm

The algorithm will try to find the α-partition with all the good points in G and all the bad points in B.

1. Start The ideal algorithm would start with all the so called elemental partitions ($|G| = p+1$) and would try to construct the sought α-partition from each of them. But this solution would be computationally too expensive, therefore only a random subset of all these elemental partitions will be used. The number of these starting elemental partition, denoted by $J(n,p,0.99,g)$, will ensure with a 0.99 probability that at least one of the chosen elemental partition has its ”good part” G free of outliers (g denotes the number of good points in the full dataset).

2. Forward search (outer cycle) The algorithm then applies to each of the selected elemental partition the classical forward search algorithm (Hadi, 1992) adding observations one by one until it reaches an α-partition. $J(n,p,0.99,g)$ α-partitions are obtained. At that point the algorithm may have obtained the sought α-partition as well as non-valid α-partition (obtained if the initial partition already contained outliers). A treatment of the resulting partitions is therefore needed.

3. Treatment of the α-partitions Three different cases can occur:

a) All obtained α-partitions are trivial ($B = \emptyset$). In that case the algorithm declares no outlier at the α level.

b) Only one non-trivial partition (G, B) is obtained. In that case the algorithm declares the points in B as outliers at the α level.

c) Different distinct non-trivial α-partitions are obtained. Here is the point where the algorithm differs from other existing ones. Kosinski argues that simply choosing one of the partition using for example a criteria like minimizing a volume (like MCD or MVE) may occasionally fail to detect the correct outliers in particular under high contamination. Therefore he eliminates first the more extreme outliers: the algorithm computes all the levels of these α-partitions and select the minimum value γ (recall that $\gamma < \alpha$). The algorithms then applies again the classical forward search methods to the obtained α-partitions but this time to obtain γ-partitions and it goes back to the beginning of 3 (inner cycle).

4. Treatment of detected outliers If no inner cycle have been used all the outliers are detected at the α level and the algorithm proceeds to the final check. If one or more inner cycles have been used then all the outliers are detected at a γ level with $\gamma < \alpha$ therefore the algorithm removes them from the data and starts all over again at point 1 but with a smaller dataset.

5. Final check If several outer cycle have been used (i.e. the α-partition has been found on a smaller dataset after removing the more extreme outliers) then the algorithm applies one more time a forward search to this partition to be sure to obtain an α-partition of the whole dataset (in simulations this check has never changed anything).

Comments By taking several starting partitions Kosinski tries to solve the main problem of the classical forward search method, namely the choice of a small subset of good points. His treatment of the possible distinct found partitions is not based on a criteria like MVE or MCD but first removes the more outlying points and then reapplies the algorithm. We shall see later that the simple forward search methods are rather fast algorithms therefore clearly the speed of Kosinski's method will depend on the number $J(n, p, 0.99, g)$ of starting partitions. As an example, using Kosinski's formula, we computed the number of starting partitions with $n = 10'000$ observations, $g = 9'000$ good points and $p = 100$ variables. We got : $J(10000, 100, 0.99, 9000) = 203'840$. This number shows that we have to be aware that with large dataset we might have to take a probability smaller than 0.99 : for example $J(10000, 100, 0.95, 9000) = 132'601$.

We are now able to describe the algorithm with all the technical details.

6.2.2 The algorithm

Even if the author does not state any assumption required by the algorithm it is clear that as the classical ideas of a forward search methods are used we should assume that the data is unimodal and roughly elliptical symmetric.

Consider type I error $\alpha = 0.01$ and assume that there are at most $N - h$ outliers. Start with outer cycle number $m = 0$ and $D(0) = \{1, 2, ..., N\}$.

Step 1 Increment m by one and set the inner cycle number to $w = 0$. Randomly form $J(|D(m-1)|, p, 0.99, h)$ distinct elemental (i.e. $|G| = p+1$) partitions of data $D(m-1)$. To each elemental partition apply the classical forward search algorithm adding one observation at a time and stop when you get an α-partition of $D(m-1)$. Let $K(m,w)$ be the number of resulting distinct α-partitions. If $K(m,w) = 0$ then define $D(m) = D(m-1)$ and go to step 5, otherwise move to step 2.

Step 2 If $K(m,w) = 1$, denote the single available partition of $D(m-1)$ by $(G(m), B(m))$ and go to step 4, otherwise move to step 3.

Step 3 Denote the levels of the $K(m,w)$ available partitions by $\gamma_k(m,w)$ with $k = 1, ..., K(m,w)$. Chose the partition corresponding to the most significant level $\gamma(m,w) = \underset{k}{min}\ \gamma_k(m,w)$. Apply the forward search procedure to all available partitions with the new $\alpha = \gamma(m,w)$. Increment w by one. Denote by $K(m,w)$ the number of resulting distinct α-partitions of $D(m-1)$ and return to step 2.

Step 4 Form the reduced data $D(m) = G(m)$. If $w \geq 1$, i.e. step 3 was used, then return to step 1 as long as $|G| > h$, otherwise ($w = 0$ or $|G(m)| = h$) move to last step.

Step 5 If $D(m) = \{1, .., N\}$ declare no outlier. If observations were removed only during the first outer cycle, declare $B(m)$ as outliers. If observations were removed in more than one outer cycle, then apply one last time the forward search with α to the partition $(D(m), D - D(m))$ of D and declare as outliers the ”bad part” of the resulting partition.

6.2.3 Properties

Kosinski does not state many properties of its algorithm. It seems to have empirically a very high breakdown point but may be computationally intensive for large datasets due to the large number of elemental partitions. Simulations were run to compare the algorithm to an hybrid method given by Rocke and Woodruff (Rocke and Woodruff, 1996). Kosinski's methods performed better than the Rocke and Woodruff's one. These tests are used in the next section to select which method between Kosinski and BACON will be chosen for the rest of the study.

6.3 Comparison between BACON and Kosinski

Kosinski's method and Bacon have been compared individually to the original forward search methods (Hadi and Atkinson) and have performed better. As we wished to study only one forward search method in the following, we ran some tests to select the most efficient one. We used the tests ran by Kosinski himself in his 1999 article. That saved us the time to implement the Kosinski algorithm. Let us start by describing these simulations.

6.3.1 Description of the tests

Recall that these tests are designed and described in (Kosinski, 1999). For each test $T = 100$ datasets are generated with g "good data" points and b outliers, i.e $N = g + b$ and the contamination fraction $f = b/N$. The performance is evaluated on three criteria:

$$p_1 = \frac{1}{T}\sum_{i=1}^{T} \mathbb{1}(m_i = 0),\ p_2 = \frac{1}{T}\sum_{i=1}^{T} \frac{m_i}{b},\ \text{and}\ p_3 = \frac{1}{T}\sum_{i=1}^{T} \frac{s_i}{g},$$

where m_i is the number of undetected outliers, s_i the number of swamped "good observations" and $\mathbb{1}(m_i = 0) = 1$ if and only if $m_i = 0$. In other words, p_1 is the proportion of simulation runs which resulted in identification of all the outliers, p_2 is the average proportion of undetected outliers, and p_3 is the average proportion of swamped "good observations". A perfect method would get $p_1 = 1$, $p_2 = 0$ and p_3 close to its nominal significance level α. Remark here that $p_2 \leq 1 - p_1$ and that the equality occurs only when in every run where not all the outliers were detected actually none was detected.

Initial tests were run to check if the value of p_3 is close to the nominal significance level when no outlier is present. Tests were therefore run with $g = 100$ and $b = 0$. The significance level was set to $\alpha = 0.01$ and tests were run in dimensions from $p = 2$ to 10. Table 1 shows the results.

Two similar series of tests were then run, one in dimension $p = 2$ (see Table 2) and one in dimension $p = 5$ (see Table 3). The number of "good observations" was fixed at $g = 100$, the contamination fraction varies from $f = 0.10$ to 0.45 by steps of 0.05. The "good points" were generated from a multivariate normal distribution $N_p(0, \sigma_1^2 I_p)$, and the outliers from $N_p(d \cdot 1_p, \sigma_2^2 I_p)$, where 1_p is the p-vector of 1's and I_p the identity matrix. The tests were run with $\sigma_1^2 = 40$, $\sigma_2^2 = 1$ and $d = 20$, 25 or 30. The significance level was set to $\alpha = 0.01$.

6.3.2 Results of the tests

The following tables display the results obtained by Kosinski for his algorithm (KOS) and reported in his paper (Kosinski, 1999) and the ones we obtained for BACON with non-robust start (V1) and robust start (V2).

Table 1: Values of p_3 in initial tests, significance level set to $\alpha = 0.01$

Method	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$	$p = 9$	$p = 10$
KOS	0.012	0.009	0.010	0.009	0.007	0.009	0.009	0.008	0.007
V1	0.012	0.011	0.011	0.009	0.008	0.007	0.007	0.006	0.006
V2	0.011	0.011	0.010	0.008	0.009	0.008	0.008	0.007	0.006

Table 2: Tests in dimension $p = 2$, significance level set to $\alpha = 0.01$

Values of f	p_1			p_2			p_3		
	KOS	$V1$	$V2$	KOS	$V1$	$V2$	KOS	$V1$	$V2$
Distance $d = 30$									
0.45	1.000	0.970	1.000	0.000	0.030	0.000	0.013	0.013	0.011
0.40	1.000	0.990	1.000	0.000	0.010	0.000	0.011	0.011	0.013
0.35	1.000	0.990	1.000	0.000	0.010	0.000	0.011	0.012	0.014
0.30	1.000	1.000	1.000	0.000	0.000	0.000	0.012	0.014	0.012
0.25	1.000	1.000	1.000	0.000	0.000	0.000	0.012	0.012	0.014
0.20	1.000	1.000	1.000	0.000	0.000	0.000	0.012	0.015	0.014
0.15	1.000	1.000	1.000	0.000	0.000	0.000	0.010	0.015	0.013
0.10	1.000	1.000	1.000	0.000	0.000	0.000	0.013	0.012	0.011
Distance $d = 25$									
0.45	1.000	0.860	1.000	0.000	0.140	0.000	0.012	0.226	0.015
0.40	1.000	0.930	1.000	0.000	0.070	0.000	0.010	0.015	0.015
0.35	1.000	0.890	1.000	0.000	0.110	0.000	0.010	0.023	0.014
0.30	1.000	0.970	1.000	0.000	0.030	0.000	0.011	0.015	0.014
0.25	1.000	1.000	1.000	0.000	0.000	0.000	0.012	0.012	0.013
0.20	1.000	1.000	1.000	0.000	0.000	0.000	0.012	0.013	0.012
0.15	1.000	1.000	1.000	0.000	0.000	0.000	0.011	0.012	0.013
0.10	1.000	1.000	1.000	0.000	0.000	0.000	0.009	0.012	0.014
Distance $d = 20$									
0.45	0.530	0.700	0.810	0.470	0.300	0.190	0.428	0.054	0.022
0.40	0.970	0.620	0.990	0.030	0.380	0.010	0.036	0.053	0.014
0.35	0.990	0.730	1.000	0.010	0.270	0.000	0.019	0.018	0.013
0.30	1.000	0.890	1.000	0.000	0.110	0.000	0.010	0.013	0.013
0.25	1.000	0.920	1.000	0.000	0.080	0.000	0.013	0.013	0.011
0.20	1.000	1.000	1.000	0.000	0.000	0.000	0.009	0.014	0.011
0.15	1.000	0.980	1.000	0.000	0.020	0.000	0.010	0.013	0.013
0.10	1.000	1.000	1.000	0.000	0.000	0.000	0.013	0.020	0.011

6.3.3 Conclusions of the tests

The initial tests confirm that p_3 is very close to the nominal significance level for all methods.

Table 3: Tests in dimension $p = 5$, significance level set to $\alpha = 0.01$

Values of		p_1			p_2			p_3	
f	KOS	$V1$	$V2$	KOS	$V1$	$V2$	KOS	$V1$	$V2$
Distance $d = 30$									
0.45	1.000	0.000	1.000	0.000	0.996	0.000	0.008	1.000	0.010
0.40	1.000	0.000	1.000	0.000	0.998	0.000	0.010	0.806	0.008
0.35	1.000	0.000	1.000	0.000	1.000	0.000	0.008	0.806	0.012
0.30	1.000	0.200	1.000	0.000	0.800	0.000	0.010	0.102	0.012
0.25	1.000	0.990	1.000	0.000	0.010	0.000	0.008	0.013	0.011
0.20	1.000	1.000	1.000	0.000	0.000	0.000	0.009	0.011	0.010
0.15	1.000	1.000	1.000	0.000	0.000	0.000	0.010	0.009	0.012
0.10	1.000	1.000	1.000	0.000	0.000	0.000	0.011	0.011	0.009
Distance $d = 25$									
0.45	1.000	0.000	1.000	0.000	0.996	0.000	0.008	1.000	0.009
0.40	1.000	0.000	1.000	0.000	0.999	0.000	0.009	0.999	0.010
0.35	1.000	0.000	1.000	0.000	1.000	0.000	0.009	0.883	0.010
0.30	1.000	0.110	1.000	0.000	0.890	0.000	0.010	0.110	0.011
0.25	1.000	0.950	1.000	0.000	0.050	0.000	0.008	0.013	0.012
0.20	1.000	1.000	1.000	0.000	0.000	0.000	0.009	0.010	0.012
0.15	1.000	1.000	1.000	0.000	0.000	0.000	0.007	0.010	0.012
0.10	1.000	1.000	1.000	0.000	0.000	0.000	0.009	0.008	0.010
Distance $d = 20$									
0.45	0.690	0.000	0.980	0.310	0.996	0.020	0.279	1.000	0.012
0.40	1.000	0.000	1.000	0.000	0.999	0.000	0.008	0.999	0.010
0.35	1.000	0.000	1.000	0.000	0.999	0.000	0.008	0.872	0.011
0.30	1.000	0.010	1.000	0.000	0.990	0.000	0.009	0.154	0.010
0.25	1.000	0.880	1.000	0.000	0.120	0.000	0.008	0.014	0.010
0.20	1.000	0.920	1.000	0.000	0.080	0.000	0.010	0.012	0.013
0.15	1.000	0.950	1.000	0.000	0.050	0.000	0.009	0.012	0.010
0.10	1.000	1.000	1.000	0.000	0.000	0.000	0.008	0.012	0.010

The main tests showed clearly that BACON with a non-robust start (V1) is not as efficient as Kosinski's method (KOS). Looking for example at the tests run with $p = 2$ and $d = 25$ we see that V1 is no longer perfect when the contamination proportion is higher than 25% while KOS remains perfect. With the same distance in dimension $p = 5$ V1 breaks down even with 25% of contamination. This breakdown comes from the fact that the overall mean is attracted more and more by the contamination cloud when the latter grows. It is even so attracted by it in some cases that V1 will end by considering the outliers as ”good data” and the remainder as ”outliers”: you can see this particularity for example in the test with $p = 5$, $d = 30$ and $f = 0.45$ where $p_3 = 1$, which means that all good points are always considered as outliers.

On the contrary the main tests showed that BACON with a robust start (V2) is even more efficient than Kosinski's algorithm. V2 is almost perfect in all cases. It only omits a few outliers when $d = 20$ and the contamination is very high $f = 40$ or 45. But in any cases it is as efficient as KOS. Moreover, even if we did not implement KOS we can see that V2 has to be quicker: for example Kosinski presented the results on the Bushfire dataset (Maronna and Yohai, 1995) and showed that it took several outer and inner cycle to find the outliers; BACON took 4 iterations (in 0.12s in S-Plus on a 600MHz PC with 128Mb RAM) to get the same outliers (see next section).

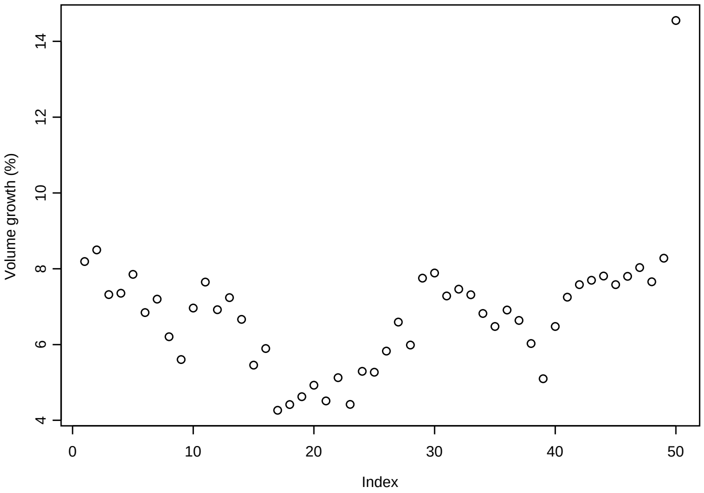
6.3.4 Summary

Simulations with the same test bed as Kosinski used (but of course with different realisations) showed that BACON algorithm with a robust start is superior to the Kosinski algorithm. For the rest of this study, BACON with a robust start was selected as our forward search method. In all tests ran by Kosinski to show the superiority of his algorithm over the hybrid method designed by Rocke and Woodruff in (Rocke and Woodruff, 1996) BACON with a robust start performed always as well and even better when the contamination is high and relatively close to the good data. BACON is a very fast algorithm and is very efficient when the good data comes from some unimodal multivariate normal distribution (in that case it's the best algorithm we have tested). BACON with a robust start has a very high empirical breakdown point and is computationally very efficient but is not affine equivariant (see the introduction for some comments on that fact).

6.4 A new graphical tool based on forward search to analyze outliers

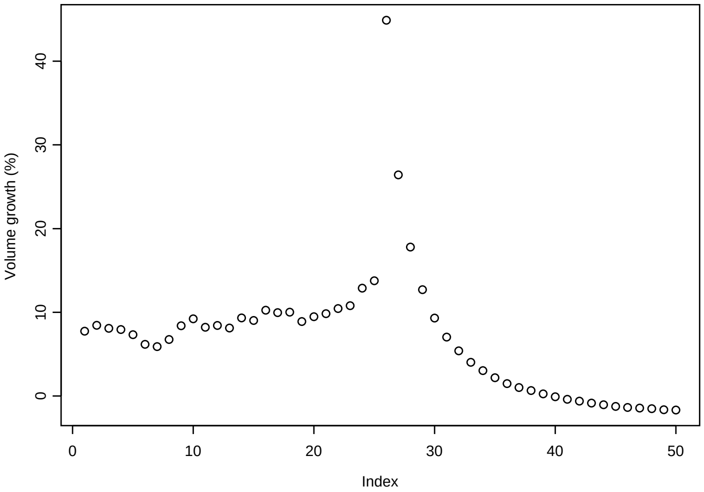
The methods exposed above give us tools to detect outliers by splitting the data in two parts: ”good” and ”bad” points. Robust estimates of the mean and covariance matrix are obtained by taking the sample mean and covariance matrix of the subset of the ”good points”. These estimates allow us to calculate the Mahalanobis distances and identify outliers but do not give any more information on these outliers. We are proposing here to use a plot of the oldest criterion on outlyingness to get a more precise overall picture of the outliers situation. Atkinson used also graphical techniques in his article but only as a detection tool: he kept for all points the history of the Mahalanobis distance d_i. As he used a forward algorithm growing one observation at each step he had to stock $n \times (n - k)$ distances where k is the number of observations used for the first estimate of the covariance matrix. What we propose here is to memorize at each step only the Wilks's one outlier scatter ratio of the added observation. This will give us an idea on the growth of the ellipsoid volume when the observation is added. To visualize this ”Volume History” (VH) we plot the percentage of growth at each step for the second added half of the data. To illustrate this we used the VH of one example of the above tests: case $p = 5$, $d = 20$, $f = 0.25$. We plotted first the VH of such a set without outliers and then the one generated for the test. We clearly see on the outlier-free VH (see Chart 1) that only one point seems suspicious with a volume growth of about 14% which is higher than the other ones. On the other hand the second VH history (see Chart 2) shows a typical pattern of concentrated contamination: we see an ”hyperbole-shaped” curve indicating the presence of clear outliers close to each other. The first detected outlier has a volume growth of

Chart 1: Volume history for a multivariate normal distribution with 100 points in dimension 5



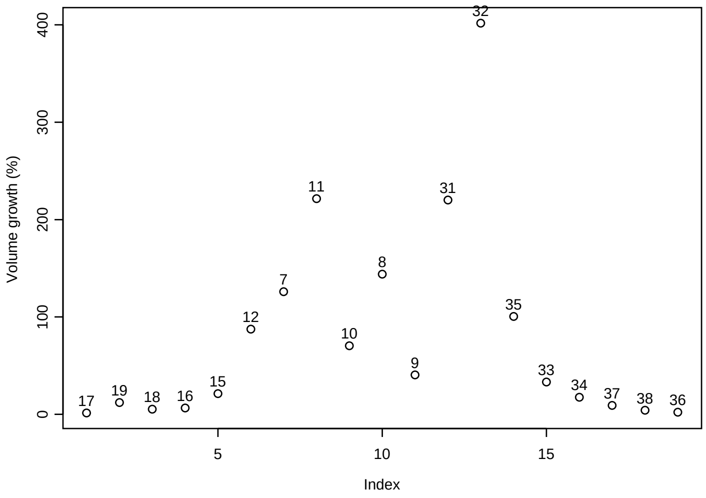
more than 40%. This pattern is exactly the same with a real point mass contamination. VH gives us a general picture of proximity of the outliers to each other. Let us look at our

Chart 2: Volume history for a multivariate normal distribution with 75 points in dimension 5 contaminated by 25 points as in Kosinski's test with $d = 20$



favorite example of the Bushfire data (see the second chapter) to see the utility of VH (see Chart 3). The first outlier seems isolated (12) with a big growth rate (87%) followed by

Chart 3: Volume history for the Bushfire data



observation 7 also isolated (126%). Then outlier 11 (221%) might be close to outlier 10 (70%). Similarly 8 (144%) might mask 9 (40%) while outlier 31 seems isolated (220%). Finally observation 32 (402%) seems to lead by far a concentrated contamination with observations 33 to 38.

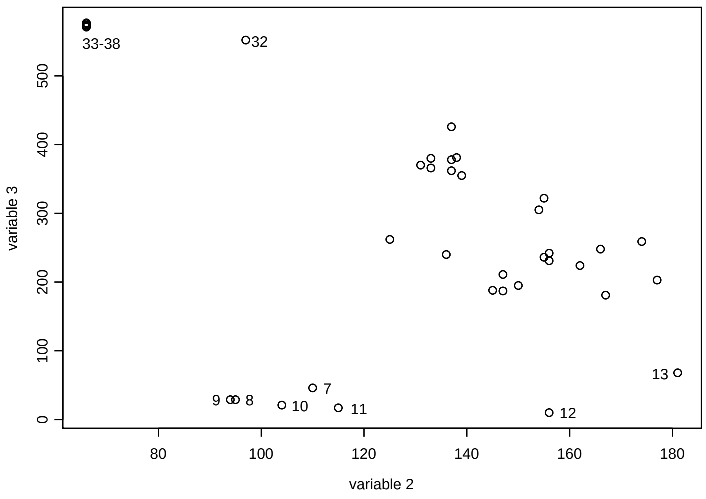
The Bushfire dataset has 38 observations in dimension 5 and allows a two dimensional plot (in variable 2 and 3) that reveals almost all the outliers (see Chart 4). On this scatter plot we see that the VH diagnostic is pretty accurate. Observation 7 is actually far from observation 10 and 11 on other variables than 2 and 3, 31 is outlying also on other variables, and 32 is not very close to $33 - 38$ but indicates the direction of the contamination.

The disadvantage of the Volume History is of course the speed of the algorithm. In fact using the relation of Lemma 6.1 the computation of the Wilks's one outlier scatter ratios correspond to the computation of the smallest Mahalanobis distances, therefore the speed of such an algorithm is the same as the first versions of Hadi and Atkinson of forward search methods. But with moderate size the VH could give interesting information on the outliers.

7 A projection pursuit method

Among the methods for computing a robust estimate of the covariance matrix for a unimodal elliptical distribution some of them are using a simple geometrical idea: ”If a point is a multivariate outlier, then there must be some one-dimensional projection of the data for which the point is a univariate outlier”. These methods fall under projection pursuit techniques. Two different approaches are here possible. The first approach computes

Chart 4: Bushfire dataset



directly estimates of the eigenvectors and eigenvalues of the covariance matrix using a robust measure of univariate scatter. This method of robust principal component analysis has been mentioned by Huber (Huber, 1985), developed by Li and Chen (Li and Chen, 1985) and studied further by Croux and Ruiz-Gazen (Croux and Ruiz-Gazen, 2000). The other approach use the geometric idea to find the ”interesting directions for outlyingness”, to identify outliers and then to compute an estimate of the covariance matrix using this information. This second approach gave birth to the first affine equivariant multivariate estimators of location and scatter robust enough to tolerate up to 50% of outliers in the sample before they break down. They were discovered independently by Stahel (Stahel, 1981) and Donoho (Donoho, 1982).

In this work only the second approach is followed. It was selected because it has already been used in official statistics by a national statistical office (Statistics Canada) in (Franklin et al., 2000). Moreover at the beginning of that study we were not aware of the existence of the new algorithm given by Croux and Ruiz-Gazen and therefore didn't compare its performances to the Stahel-Donoho method. We implemented a modified version of the original Stahel-Donoho estimator, starting from a version given by Patak (Patak, 1990) and reported in (Franklin et al., 2000).

7.1 Modified Stahel-Donoho (MSD) estimators

We start by recalling the construction of the original Stahel-Donoho (SD) estimators, and some properties obtained by Maronna and Yohai in (Maronna and Yohai, 1995).

7.1.1 Original SD estimators and some properties

The SD estimators are defined as weighted mean and covariance matrix, where each has a weight that is a function of an outlyingness measure, with points having large outlyingness receiving small weights.

As usual let X be the $n \times p$ data matrix with n observations $(x_1, ..., x_n)$ and p variables. Let μ and σ^2 be affine equivariant univariate estimator of location and scatter, the outlyingness measure r_i of each observation x_i is given by

$$r_i = \underset{\|a\|=1}{sup} \frac{|a^t x_i - \mu(a^t X^t)|}{\sigma(a^t X^t)}.$$

Each r_i measures the maximum standardized one-dimensional deviation from the estimated location μ for all directions in $\mathbb{R}^p$. Then the weights are computed as

$$u_i = u(r_i) \text{ where } u \,:\, \mathbb{R}^+ \to \mathbb{R}^+ \text{ is a weight function.}$$

The SD estimators are then defined as

$$m_{SD} = \frac{\sum_{i=1}^n u_i x_i}{\sum_{i=1}^n u_i} \text{ and } S_{SD} = \frac{\sum_{i=1}^n u_i (x_i - m_{SD})(x_i - m_{SD})^t}{\sum_{i=1}^n u_i}.$$

By definition and by the assumptions on μ and σ^2 the estimators are affine equivariant. Actually if μ and σ^2 are the usual mean and variance and if u is the identity then the SD estimators are the usual sample mean and covariance matrix. Stahel (Stahel, 1981) showed that the SD estimators have an asymptotic breakdown point of $1/2$ at continuous multivariate model if μ and σ have the same property and Donoho (Donoho, 1982) derived the finite-sample breakdown point in the case in which $\mu = median$ and $\sigma = mad$. In (Maronna and Yohai, 1995) Maronna and Yohai studied the finite sample breakdown point of the latter estimator but with the outlyingness measure r_i taken only on a random subset of size N of all $a \in \mathbb{R}^p$ with $\| a \| = 1$. They computed the size N needed for the breakdown of this approximate estimator to be as good as the usual one with a probability of 0.999. They showed that N grows exponentially with p implying unavoidable computing difficulties for large p. For example, for $p = 4, 6, 8,$ and 10 one needs $N = 210, 1'050, 5'000,$ and $26'260$. Their study also determined what was the best weight function to use according to their quality measures (biases and efficiencies) and the following "Huber-like" weight was selected:

$$u \,:\, \mathbb{R}^+ \to \mathbb{R}^+,\; r \mapsto u(r) = \begin{cases} 1 & \text{if } r \le c \\ \left(\frac{c}{r}\right)^2 & \text{if } r > c \end{cases} \quad \text{with } c = \sqrt{\chi^2_{p,0.95}}$$

7.1.2 Modified SD estimators

We start by giving the modified Stahel-Donoho estimators proposed by Patak (Patak, 1990) as reported and used in (Franklin et al., 2000). This construction is as follows:

1. The data are centered using the L_1-estimate of the location vector. The L_1-estimate of the location vector is defined as the solution of the minimization problem: $\min_T \sum_{i=1}^{n} \| x_i - T \|_2$. It is often named the spatial median.
2. The initial weights are all set to one: $u_i = 1$, $i = 1, ..., n$.
3. For $k = 1$ to m (m usually set to 10) do
    a) Randomly generate a unit vector $v_1 \in \mathbb{R}^p$ using a uniform distribution on the unit sphere in $\mathbb{R}^p$.
    b) Calculate $v_2, ..., v_p$ in such a way that the v_i's form an orthonormal basis of $\mathbb{R}^p$
    c) For $i = 1, ..., n$ and $j = 1, ..., p$ compute
    $$r_{ij} = \frac{|v_j^t x_i - \text{med}(v_j^t X^t)|}{\text{mad}(v_j^t X^t)} \text{ and then } \widetilde{r}_{ij} = \begin{cases} r_{ij} & \text{if } 0 \leq r_{ij} < 2.5 \\ 2.5 & \text{if } 2.5 \leq r_{ij} < 4 \\ 0 & \text{if } 4 \leq r_{ij} \end{cases} .$$
    Finally compute
    $$u_i^k = \prod_{j=1}^{p} \frac{\widetilde{r}_{ij}}{r_{ij}}.$$
    d) If $u_i^k < u_i$ then set $u_i = u_i^k$.
4. Compute the weighted estimates of location and scatter using the weights u_i.
5. Reset all weights to one: $u_i = 1$, $i = 1, ..., n$.
6. Redo the loop in 3. but this time by replacing the random orthogonal basis (points a) and b)) by the computation of the principal components of the current weighted covariance matrix. Stop when the weights do not change significantly (in practice one iteration has been found to be sufficient).

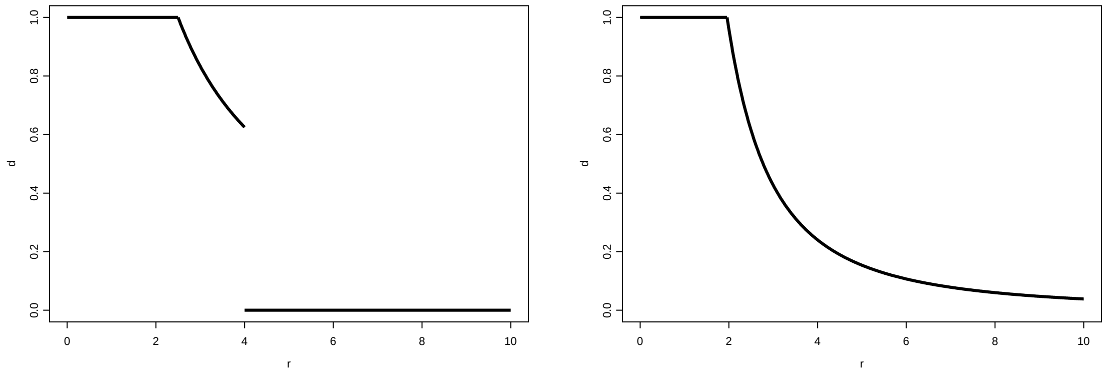
Our version of the modified Stahel-Donoho will differ in several points from the Patak's algorithm:

(i) As the computation of the weights use some centering on the one-dimension projections, the weights are location invariant. Therefore the weighted estimates of location and scatter are location equivariant and the initial centering is useless. We removed it from our algorithm.

(ii) Following Maronna and Yohai we decided to use ”Huber-like” weight function instead of the non-continuous weight function proposed by Patak (see Chart 5 for the picture in dimension 1), i.e. we change the computation of the $\widetilde{r}_{ij}$ into:

$$\widetilde{r}_{ij} = \begin{cases} r_{ij} & \text{if } 0 \leq r_{ij} < c \\ \frac{c^2}{r_{ij}} & \text{if } c \leq r_{ij} \end{cases} \quad \text{with } c = \sqrt{\chi^2_{p,0.95}}.$$

Chart 5: Weights used by Patak and "Huber-like" weights in dimension 1



(iii) Following Maronna and Yohai m is set by default to $m = \lfloor \exp(2.1328 + 0.8023 * p)/p \rfloor$. Of course in high dimension the user might have to choose a much smaller m.

(iv) We did not reset the weights to one in 5. The reason here is that according to our experience outliers that are not on the principal components directions might be masked if we do reset the weights.

8 A minimization of scale method

After Stahel and Donoho, Rousseuw (Rousseuw, 1984), (Rousseeuw, 1985) introduced a second affine equivariant estimator with maximal breakdown point, by putting "$T(X) =$center of the minimal volume ellipsoid covering (at least) h points of X", where h can be taken equal to $\lfloor n/2 \rfloor + 1$. This estimator is called the minimum volume ellipsoid estimator (MVE). The corresponding covariance estimator is given by the ellipsoid itself, multiplied by a suitable factor to obtain consistency at multivariate normal data. Rousseuw noticed however that for $p = 1$ the MVE reduces to the shortest half, so $T(X)$ becomes the one-dimensional least median of squares which converges like $n^{1/3}$, see Theorem 3 in Section 4 of Chapter 4 in (Rousseeuw and Leroy, 1987). Assuming that MVE will not have a better rate Rousseuw then proposed to generalize the least trimmed squares which converges like $n^{1/2}$, see Theorem 4 in Section 4 of Chapter 4 in (Rousseeuw and Leroy, 1987), and ended up with the minimum covariance determinant estimator (MCD) defined this time by minimizing the determinant of the covariance matrix computed from the h points. This estimator will be included in this study or actually a reweighted form of it that is standard implemented in S-Plus.

8.1 Reweighted minimum covariance determinant estimators (RMCD)

As usual let X be a sample of n observations $(x_1, ..., x_n)$ with p variables. The MCD estimators are determined by selecting the subset $\{x_{i_1}, ..., x_{i_h}\}$ of size h which minimizes the determinant of the covariance matrix computed from that subset. The location and scatter estimators are then defined as

$$m_{MCD} = \frac{1}{h}\sum_{j=1}^{h} x_{i_j} \text{ and } S_{MCD} = c_p \frac{1}{h}\sum_{j=1}^{h}(x_{i_j} - m_{MCD})(x_{i_j} - m_{MCD})^t$$

with c_p the consistency factor at multivariate normal. Now h can be chosen by the user determining the breakdown point of the estimator: if $h = n(1-\beta)$ the estimator has a breakdown point of β. Typically β is set to 0.5 or 0.25. As it is usually not feasible to find the exact minimum several algorithms have been proposed to approximate the solution. The best one was proposed by Rousseuw and van Driessen (Rousseeuw and van Driessen, 1999), it is called the FAST-MCD algorithm. The major drawback of the MCD estimators remains its low efficiency at the normal distribution (Croux and Haesbroeck, 1999). To overcome this problem a reweighting step can be added to the MCD estimators. Weights are computed using a cut-off value on the Mahalanobis distances:

$$u_i = \begin{cases} 1 & \text{if } (x_i - m_{MCD})^t S_{MCD}^{-1}(x_i - m_{MCD}) \leq \chi^2_{p,\alpha} \\ 0 & \text{otherwise} \end{cases}$$

Then the reweighted minimum covariance determinant estimators (RMCD) are defined by

$$m_{RMCD} = \frac{\sum_{i=1}^n u_i x_i}{\sum_{i=1}^n u_i} \text{ and } S_{RMCD} = d_p \frac{\sum_{i=1}^n u_i (x_i - m_{RMCD})(x_i - m_{RMCD})^t}{\sum_{i=1}^n u_i}$$

with d_p the consistency factor at multivariate normal. The RMCD estimators inherit the breakdown point of the MCD estimators. The RMCD estimators are standard implemented in S-Plus as the ”cov.mcd” function with $\alpha = 0.025$.

8.2 FAST-MCD algorithm

We report here the FAST-MCD algorithm as described in (Rousseeuw and van Driessen, 1999). We shall need this description in the next sections when we'll adapt the algorithm to sampling weights and missing values. In this algorithm a C-step is like a BACON-step but with the number of point in the subset fixed: if you have a subset of k observations, compute the Mahalanobis distances of all the points in the set using the mean and covariance matrix based only on the subset and select a new subset of size k corresponding to the k smallest obtained Mahalanobis distances.

1. By default set $h = (n+p+1)/2$ or let the user choose, report the breakdown point of $(n-h+1)/n$.

2. If $h = n$ return the usual mean and covariance matrix and stop.

3. If $p = 1$ compute the exact MCD using the algorithm given in (Rousseeuw and Leroy, 1987), pages 171-172, then stop.

4. If $n < 600$ then

    - repeat 500 times:
        - construct an initial subset of size h starting from $p + 1$ randomly chosen points then adding randomly one point at a time until the covariance matrix of this subset is non-singular and finally selecting the h smallest Mahalanobis distances based on these randomly chosen points,
        - carry out two C-steps,
    - among these 500 subsets select the 10 with lowest determinant of the covariance matrix,
    - apply C-steps until convergence to all these subsets,
    - among these 10 subsets select the one with lowest determinant of the covariance matrix,
    - report the mean m and covariance matrix S based on that subset and go to point 7.

5. If $600 \leq n < 1500$ then

    - construct as many disjoint random subsets as possible with all these subsets being of size $n_{sub} \geq 300$ (or $n_{sub} + 1$), denote by k the number of these subsets (i.e. $2 \leq k \leq 4$),
    - inside each subset repeat $500/k$ times:
        - construct an initial subset of size $h_{sub} = n_{sub}h/n$ as in point 4,
        - carry out two C-steps, using n_{sub} and h_{sub},
        - keep the 10 subsets with lowest determinant of covariance matrix,
    - from these $10k$ subsets of size h_{sub} form $10k$ subsets of size h using the smallest Mahalanobis distances,
    - apply two C-steps to all these subsets,
    - among these $10k$ subsets select the 10 with lowest determinant of the covariance matrix
    - apply C-steps until convergence to all these subsets,
    - among these 10 subsets select the one with lowest determinant of the covariance matrix,
    - report the mean m and covariance matrix S based on that subset and go to point 7.

6. If $n \geq 1500$ select a random subset of size $n_1 = 1499$, then apply point 5 to that subset with n_1 and $h_1 = 1499h/n$ except that when the last 10 subsets of size h_1 are selected (fifth step) their sizes are extended to h using Mahalanobis distances and the last steps are applied to the all dataset.

7. In order to obtain consistency under multivariate normal distribution set

$$m_{MCD} = m \qquad \text{and} \qquad S_{MCD} = \frac{\text{med}_i(MD_{m,S}(x_i))}{\chi^2_{p,0.5}} S$$

8. To improve efficiency under normal distribution set finally

$$m_{RMCD} = \frac{\sum_{i=1}^n u_i x_i}{\sum_{i=1}^n u_i} \text{ and } S_{RMCD} = \frac{\sum_{i=1}^n u_i (x_i - m_{RMCD})(x_i - m_{RMCD})^t}{\sum_{i=1}^n u_i}$$

   with

$$u_i = \begin{cases} 1 & \text{if } MD_{m_{MCD},S_{MCD}}(x_i) \leq \chi^2_{p,0.025} \\ 0 & \text{otherwise} \end{cases}$$

9 A nonparametric method

9.1 Introduction and motivation

As noticed in the introduction our first intention was to include diverse nonparametric or semi-parametric approaches of outlier detection like data depth (Liu et al., 1999) in this study but we had to renounce by lack of resources. Nevertheless we are proposing a new non-parametric method for the detection of multivariate outliers, the Epidemic Algorithm (Hulliger and Béguin, 2001).

The idea of the Epidemic Algorithm (EA) is the following: We want to detect outliers in a population of n points in p-dimensional space. We start a simulated epidemic from a well chosen point. The epidemic will spread through the population and eventually all points will be infected. In this process the outliers should either not be infected or be infected late due to their isolation. We use the infection time to judge on the outlyingness of a point. In other words the epidemic defines a random mapping from the population into the time axes which should give high values for outliers.

9.2 Distances, center and infection probability

The probability of transmission of the epidemic depends on the distance between observations and decreases with it. The transmissions are independent. The time is discrete. An infected point can transmit the epidemic as long as the epidemic lasts.

Denote the population with U. The points are described by the vector valued variable $x_i \in \mathbb{R}^p$, $(i = 1, ..., n)$. The distance between points i and j is the Euclidean distance:

$$d_{ij} = d(x_i, x_j) = \| x_i - x_j \|_2 = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2} = ((x_i - x_j)^t (x_i - x_j))^{1/2}.$$

The matrix of these distances is D. To avoid unbalanced effects of the different variables, their variances shall be standardized before calculating the distances, e.g. by

$$\widetilde{x}_{ik} = \frac{x_{ik} - \text{med}(x_{ik})}{\text{mad}(x_{ik})}.$$

Alternatively one may weight the contribution of each variable to the distance by the inverse of a robust measure of scale:

$$d_{ij} = d(x_i, x_j) = \left(\sum_{k=1}^{p} q_k (x_{ik} - x_{jk})^2 \right)^{1/2},$$

where e.g. $q_k = (\text{mad}(x_{ik}))^{-2}$.

The starting point of the epidemic shall be the ”sample spatial median” c, namely the sample point that has the characterizing minimal property of the usual spatial median:

$$c = \{x_i : \text{ where } i \text{ is such that } \sum_{j \in U} d_{ij} = \min_{k \in U} \left(\sum_{j \in U} d_{kj} \right) \} = \arg \min_{i \in U} \sum_{j \in U} d(x_i, x_j).$$

Note that the sample spatial median is not necessarily close to the real spatial median. E.g. for a uniform distribution on a circle the spatial median will be near the center and the sample spatial median will be on the circle. However the sample spatial median will be in the bulk of the data. Moreover as all the distances d_{ij} will be needed anyway for the Epidemic Algorithm, the computation of ssm is cheap.

Given a point i that is infected, the probability that a non-infected point j is infected by i at any time t is

$$P[j|i] = h(d_{ij}) = P[i|j],$$

where the function h is monotone decreasing for growing d and $0 \leq h(d_{ij}) \leq 1$. We write $h_{ij} = h(d_{ij})$ for brevity. There are many possible choices for the transmission function h. Three examples are:

a) The step function

$$h(d) = \begin{cases} 1 & \text{if } d \leq d_0 \\ 0 & \text{if not} \end{cases}$$

corresponding to a total infection in the ball with radius d_0 and no possible infection outside this ball. This yields a deterministic epidemic or rather a minimum journey with day-trips between points at maximal distance d_0.

b) A simple linear transmission function

$$h(d) = \begin{cases} (1-\beta d) & \text{if } d \leq \frac{1}{\beta} \\ 0 & \text{if not} \end{cases}$$

This function becomes exactly 0 at $d_{ij} = 1/\beta$ and thus no transmission is possible beyond this distance. The parameter β may be chosen in the following way. Calculate the maximum distance to a nearest neighbor $d_0 = \max_i \{ \min_j \{d_{ij}\}\}$. Then $\beta = (1-1/n)\min\{d_0, 2\sqrt{p}\}$. Thus β is chosen such that the transmission probability is $1/n$ at d_0 or at $2\sqrt{p}$ if d_0 is inflated by one or several single outliers.

c) The inverse power function:

$$h(d) = 1/(\beta d + 1)^p.$$

We propose to choose β such that $h(d_0) = 1/(\beta d_0 + 1)^p = 1/n$, i.e. $\beta = (n^{1/p} - 1)/d_0$.

d) The logistic function:

$$h_{ij} = \frac{\exp(\alpha + \beta d_{ij})}{(1 + \exp(\alpha + \beta d_{ij}))}$$

with $\alpha > 0$ and $\beta < 0$. The transmission probability is close to 1 for $d_{ij} = 0$ and $= 0.5$ at $d_{ij} = -\alpha/\beta$. The slope at this latter distance is $\beta/4$. We propose to choose the parameters α and β in such a way that the transmission probability is 0.5 at the median of the interpoint distances and $1/n$ at the maximal distance d_0.

In the following examples, the transmission function a) is used. The choice of the transmission function and its parameters is crucial for the detection capability of the algorithm and for its speed.

If a subset $I \subset U$ of points is infected at a certain time then the total infection probability that an uninfected point j is infected at the next step is

$$P[j|I] = 1 - \prod_{i \in I}(1 - P[j|i]) = 1 - \prod_{i \in I}(1 - h_{ij}).$$

Thus we do not have to simulate each infection from point to point but only from the set of infected points to the each non-infected point.

9.3 The steps of the Epidemic Algorithm

Denote by I_t the subset of all the points infected up to time t: $I_t = \{i : 0 < t_i \leq t\}$. Denote the index of the sample spatial median c with $i(c)$.

1. Set the infection time of all points to zero: $t_j := 0, \forall j \in U$.
2. Set the time to one : $t := 1$. Choose the sample spatial median c as the starting point, i.e. set its infection time to one: $t_{i(c)} := 1$ and thus $I_1 = \{i(c)\}$.
3. Increase the infection time by one: $t := t + 1$.
4. Calculate the total infection probability $P[j|I_{t-1}]$ for all non-infected points $j \notin I_{t-1}$:
   $$P[j|I_{t-1}]\,, \forall j \notin I_{t-1}.$$
5. Realise independent Bernoulli trials with success probability $P[j|I_{t-1}]$ for the points $j \notin I_{t-1}$. A success means that the point is infected at time t and its infection time t_j is set to t: $t_j := t$.
6. If $|I_t| = n$ or $t - \max\{t_i : i \in I_t\} > l$ then set $t_{\max} = t$ and stop. Otherwise go to step 3.

The algorithm stops if all points are infected or if no infection occurs during a period of length l. The non-infected points will keep infection time $t_j = 0$. The integer number l is chosen by the statistician. In the next Section it is set to 10. Alternatively the choice of l may be guided by an upper bound on the probability of no infection in l trials: $(1-h(d_0))^l$. In the following we sometimes abbreviate Epidemic Algorithm to EA.

9.4 Computational complexity

In the beginning we have to calculate the $n(n-1)/2$ distances, each involving $p+1$ operations. We cannot speed up this part because we need all distances.

However, we can avoid the recalculation of the products involved in the total infection probability because the sets I_t are nested. For this we have to introduce a vector of products $H_{j,t} = \prod_{i \in I_t}(1 - P[i|j])$ for each time point and we have to change the Epidemic Algorithm slightly:

In step 1) set $H_{j,0} = 1\ \forall j \in U$.

In step 4) do the following for each $j \notin I_{t-1}$: Set $H_{j,t-1} := H_{j,t-2} \prod_{i \in I_{t-1} \setminus I_{t-2}}(1 - h_{ij})$ and calculate the total infection probability $Pr[j|I_{t-1}] = 1 - H_{j,t-1}$.

The point is of course that for computer implementation one needs to keep in memory only one vector H which is updated.

At each stage t there are $k_t = |I_{t-1}|$ infected points and $(n - k_t)$ non infected points. For each non infected point the total infection probability must be calculated. This involves a product with $(k_t - k_{t-1}) + 1$ factors. Thus for the whole epidemic for each observation at most $n + t_{\max}$ multiplications are needed and at most $t_{\max}$ experiments are needed. Therefore the order of complexity of the epidemic is n^2. Together with the initial distance calculation the epidemic is of complexity $n^2(p+1)$. In other words the order of complexity of the Epidemic algorithm is quadratic in n but only linear in the number of dimensions p!. The dimension of the space only affects the initial calculation of the distances. Nevertheless for large populations the computation may be very slow.

9.5 Behavior of the Epidemic algorithm with normally distributed data

To analyze the behavior of the algorithm in the absence of outliers several datasets were simulated with a multivariate normal distribution in $\mathbb{R}^p$, with mean at the origin and covariance matrix equal to I_p (identity matrix). The following table gives the total number of infected points at each infection time for 10 different datasets with n ranging from 100 to 2000 and p from 2 to 100 (see Table 4).

Table 4: Infection times for multivariate normal distribution

Data	n	100	100	500	500	1000	1000	1000	2000	2000	2000
sets	p	2	10	10	20	10	20	50	20	50	100
Infection time (t)	1	1	1	1	1	1	1	1	1	1	1
	2	13	15	53	81	78	79	75	199	96	136
	3	52	61	369	435	715	665	516	1758	1027	1335
	4	78	89	477	489	948	943	900	1981	1815	1887
	5	89	95	490	495	980	965	950	1990	1909	1963
	6	95	97	494	497	989	976	970	1996	1938	1975
	7	97	97	494	498	992	987	980	1998	1952	1982
	8	99	97	496	499	992	991	985		1962	1984
	9		98	497		994	992	989		1972	1987
	10			497		994	992	990		1976	1987
	11			498		995	992	991		1977	1989
	12					996	992	992		1982	1990
	13					996	993	993		1984	1990
	14					996	996	992		1985	1990
	15					997		993		1988	1990
	16							993		1990	1991
	17							993		1990	1991
	18							996		1990	1992
	19							997		1991	1993
	20							997		1991	1995
Largest inf. time		8	9	11	8	15	14	25	7	47	34
Non-infected		1	2	2	1	3	4	2	2	3	2
Comp. time		0.7	0.8	3.4	3.4	9.2	10.4	15.0	388.5	776.1	252.3

This table shows that under normal distribution the median infection time is always 3 and that after time 7 more than 95% of the population has been infected in all cases for any values of n and p (the worst case occurred when $n = 100$ where only 97% is detected at $t = 7$). We therefore use $t = 7$ as critical time under normal distribution. The number of non-infected points does not seem to depend on n or p; in all simulations it has never exceeded 5. In contrast the length of the epidemic does vary very much, even if half of the population has been infected after time 3 in all cases! It seems that for a fixed n the largest infection time increases with p. The three computing times for $n = 2000$ are not too relevant because a large part of them is due to memory swapping.

9.6 Remarks

- The distance matrix $D = (d_{ij})$ contains all the necessary information on the population. Thus if two point clouds have the same distance matrix the Epidemic Algorithm should detect the same outliers apart from random variation. This is in order. However, in a situation where the good observations follow a model like a multiple regression the Epidemic Algorithm may be worse than an algorithm which builds on this model (see the Stackloss data example in the next section).

- We may integrate ordinal categorical variables in the distance by introducing some scale. For nominal categorical variables we may set the distance to 0 if the categories coincide and to 1 if not. Other possibilities exist for example with the nomenclature of economic activities. There you may count the nodes you have to pass in the classification tree for moving from one category to the next.

- An observation which is outlying in only one or two dimensions but an inlier in all other dimensions may have an overall Mahalanobis distance which does not show it as an outlier. This sort of outliers could be detected better with distances like L_∞ or L_1 instead of the Euclidean distance.

- The infection process is a Markov process but it is not time homogeneous because the infection probability changes over time. In fact for the infection probability of a point at a certain time the whole history of the epidemic is important. And this history depends on the spatial configuration of the points as it is reflected by the distance matrix. The infection probability of a point j when it is the only remaining non-infected point, i.e. $P[j|U\backslash j] = 1 - \prod_{i \neq j}(1 - h_{ij})$, gives no direct hint to its infection time because the infection time of j depends on which of the points in $U\backslash j$ become infected at what time.

- Theoretically one could calculate the expected infection time $E[t_j]$ by considering all possible epidemics which lead to the infection of point j. However, since the number of possible epidemics is exponential in n this is not feasible in practice.

The Epidemic algorithm is computationally feasible. It is somewhat slower than the most efficient algorithms. However its computing time does not grow exponentially with the number of dimensions. It does not need any assumption on the data except that the good data is not divided into well separated clusters. No transformation is necessary to apply EA. It is based on the intuitive notion of an outlier as an isolated point or group of points. The starting point of a sample spatial median seems to be very fruitful.

The EA has connections to clustering algorithms and to nearest neighbor methods. However, by exploiting the dynamics of the epidemic, it takes into account local and global properties at the same time.

The choice of the transmission function is crucial for the efficiency of the algorithm. Our simple and first choice will have to become more sophisticated to be able to cope with all types of masking problems.

Part III

Application to real and synthetic datasets

All the above selected methods were developed and tested on several datasets that are not the ones chosen in EUREDIT for the evaluation phase. Most of them have been found in the literature and were known to be somehow challenging for multivariate outlier detection. Some of them were created to test particular configurations (compact contamination, non-elliptical data). In most articles where a new method is proposed, the authors usually present one particular dataset on which their method behaved relatively well. Our goal here is to gather several of these datasets and compare the results of all the above methods on all of them. The results are presented below, with cases of real and synthetic datasets as well as symmetric and non-symmetric datasets. Conclusions are drawn in the last subsection.

The results obtained by the methods using a robust Mahalanobis distance (SMP, BACON, MSD and RMCD) will be illustrated by Q-Q plots of transforms of Mahalanobis distances (MD_i) using the following approximation for normal data :

$$D_i = F^{-1}(0.5, p, n-p)\frac{MD_i}{\text{median}(MD_i)} \approx f_i = F^{-1}(\frac{i}{n+1}, p, n-p)$$

where $F^{-1}(\alpha, k, l)$ is the α-quantile of the F distribution with k and l degrees of freedom. For the epidemic algorithm the infection times are plotted versus the indices of the observations. Points which are not infected are plotted with an infection time of $t_i = \lceil 1.2 \cdot t_{\max} \rceil$ instead of $t_i = 0$ to show their outlyingness.

It is difficult to compare detection capabilities of different methods for real data sets because no ”gold method” tells us which are the ”true” outliers. What we do is to compare the sets of points which are declared good and outlying by the different methods and eventually we will come up with a consensus measure to quantify the degree of coincidence a particular method has with all the other competing methods.

All algorithms have been implemented in S-Plus 2000, on a PC with a 600 MHz Intel Pentium Processor and 128 Mb RAM. The S-language is not efficient for EA and MSD as any use of loops should be avoided in S. Therefore one should not consider the comparison of computing times as totally relevant. Moreover memory problems were sometimes encountered in particular with EA when dealing with the $n \times n$ distance matrix: the 128 Mb RAM were not enough as soon as $n = 2000$ and the processor used virtual memory on the hard disk making the computing time explode.

Let’s emphasize finally that parameters could vary according to the data in most of the methods to get better results. As we are trying to develop some automatic editing procedure we decided to fix once for all the parameters of the method throughout the tests. Of course this decision is open to criticism but its justification is the fact that EUREDIT tries to develop methods that users could use without any specific statistical knowledge. Only in one of the last examples we emphasized how important the parameters' tuning can be.

Let's recall the parameters used in the following:

SMP No parameter, version with median and mad as described in the preceding section.

BACON The version with a robust start, a starting subset of size $3p$ and a signification level of 0.01 (see the preceding section).

MSD Huber's weights are used. The number of projections is just reduced for high dimension to avoid very long computations.

RMCD Standard implementation in SPlus with a 50% breakdown point, reweighting with a cut-off point with $\alpha = 0.025$.

EA With a simple linear transmission function and a maximum transmission distance automatically computed as described in the preceding section.

10 The Bushfire data

The first real dataset has 38 observations in dimension 5. It was used by Campbell in 1989 (Campbell, 1989) to locate bushfire scars. It contains satellite measurements on five different frequency bands corresponding to each of 38 pixels. It has the advantage of having been well studied (Maronna and Yohai, 1995) and of allowing a two dimensional plot (in variable 2 and 3) that reveals almost all the outliers (see Chart 4). The data contains an outlying cluster of observations 32 to 38 and a few other outlying values 32 and 7 to 11, eventually also 12 and 13.

A classical multivariate analysis using the sample mean and covariance estimator would not detect anything. Chart 6 shows that the results obtained from the three comparative methods are quite similar. Table 5 gives the observations with the largest MD_i in decreasing order for the three methods. All of them detect the above mentioned outliers. MSD does not consider the $32-38$ group as more outlying than the other outliers and MCD detects also 31 as an outlier. The EA applied to the Bushfire data did not infect any points after time $t = 6$ (see Chart 7). Only non-infected observations will therefore be declared as outliers, namely points 7 to 11 and 32 to 38. Clearly in that case all methods are equivalent. Finally, due to the small size of the dataset all computing times are moderate : SMP 0.11s, BACON 0.08s, MSD 6.7s (500 projections), MCD 0.22s and Epidemic 0.40s.

Table 5: Highest Mahalanobis distances for the Bushfire data

SMP	38	37	36	35	34	33	9	8	32	7	10	11	
BACON	38	35	37	33	34	36	32	9	8	10	11	7	
MSD	9	8	7	32	38	10	37	35	36	34	33	11	
MCD	33	35	34	38	37	36	32	9	8	31	10	11	7

Chart 6: D_i for SMP, BACON, MSD, RMCD for the Bushfire dataset

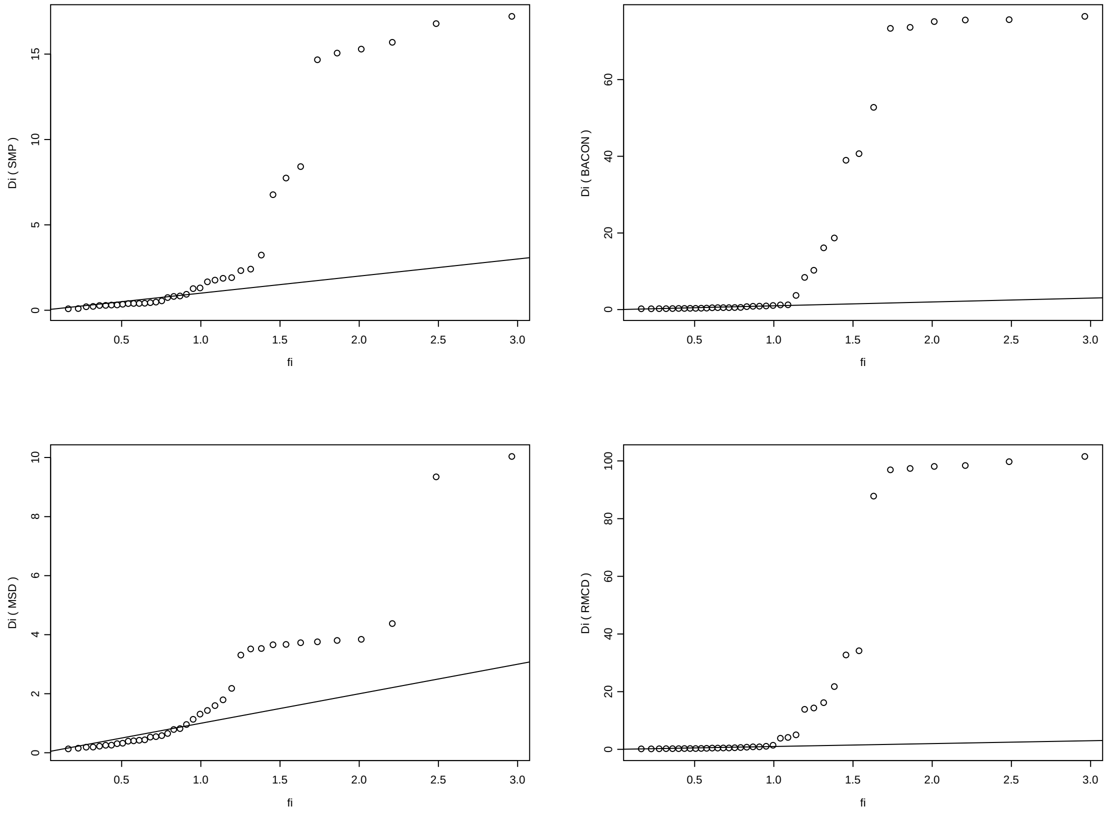
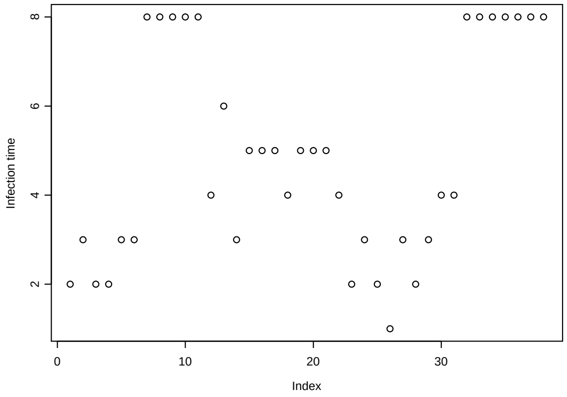


Chart 7: EA infection time on the Bushfire dataset



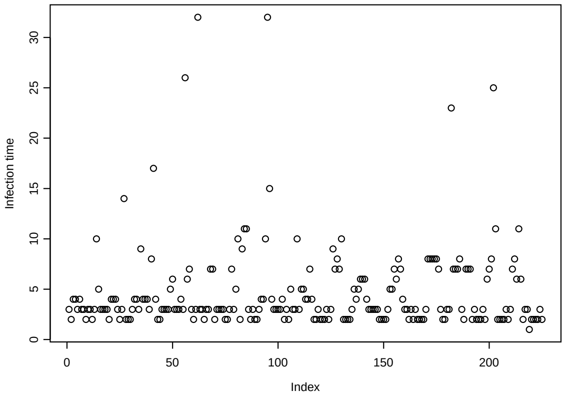
11 The Ionosphere data

The second real dataset was taken from the UCI Machine Learning Database Repository (Bay, 1999) and was suggested to us by Ricardo Maronna (Maronna and Zamar, 2001).

This dataset was part of a study of the Ionosphere carried out by the Space Physics Group of the Applied Physics Laboratory of the Johns Hopkins University (Sigillito et al., 1989). Radar data were collected by a system in Goose Bay, Labrador. The targets were free electrons in the ionosphere. ”Good” radar returns were those showing evidence of some type of structure in the ionosphere. These good radar measurements form the dataset which is studied here: there are 225 observations in dimension 32 (two variables with no variance were eliminated).

The EA was run first and gave results shown on chart 8. Two observations were not infected (62 and 95) and 10 others were infected after time $t = 10$. To compare these results with the other methods, the Q-Q plots are given in Chart 9. Note here that according to Maronna and Yohai MSD should have used about $1.19 * 10^{12}$ different directions which is computationally unfeasible, therefore we restricted ourselves to 5000. These plots show that about $60\%(= 135$ observations) of the data behave like normally distributed. The picture for SMP differs from the other ones as SMP is the only estimator not based (by selection or downweighting) on only this supposed normal part. Note that after time $t = 3$ the EA had infected 134 observations! Clearly something is happening for the remaining data. Choosing a value where to cut for outlyingness would require more knowledge of the data.

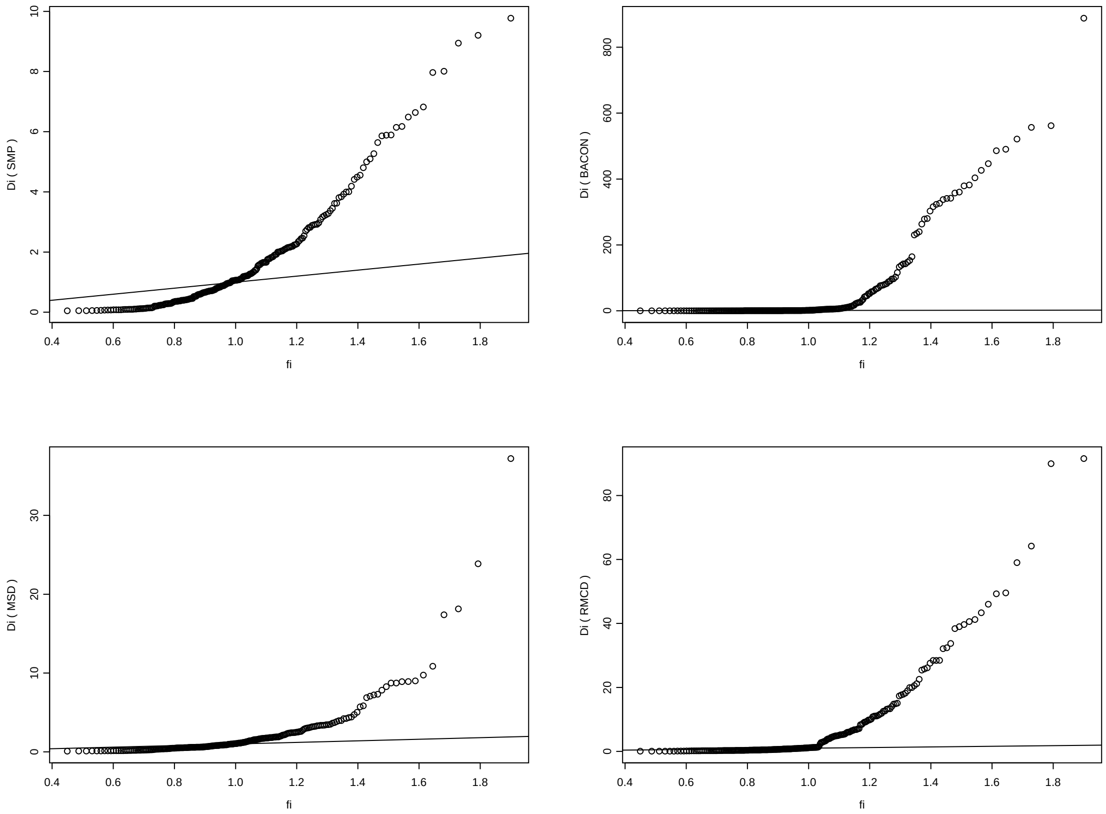
Chart 8: EA infection time on the Ionosphere dataset



To compare all the results we give two tables with the number of common points in the ”central part” of each method and in the ”extreme part” (see Table 6). The central part of a method consists of the 134 observations which are least outlying (lowest MD_i or infection time ≤ 3) while the extreme part consists of the 12 most outlying observations (highest MD_i or infection time > 10 or non-infected).

Amazingly SMP is the most consensual estimator for the central part sharing always more than 103 points (77%) with any other estimator. The four other methods seem to pair off: MSD and RMCD share 125 points (93%) of their central parts while BACON and EA share 119 points (89%). But the two pairs of methods seem to diverge somehow: for example RMCD and EA only share 80 points (59%) of their central parts. A possible

Chart 9: D_i for SMP, BACON, MSD and RMCD for Ionosphere



explanation to that phenomena could be the ideas behind the methods: both MSD and RMCD are based on geometrical ideas while both BACON and EA are based on growing the good part of the data.

For the extreme part there is no consensus, but if we look closer at the Q-Q plots or the infection times, SMP has five clear outliers (27, 62, 85, 95 and 202), BACON has only one (27), MSD has four (27, 62, 95 and 96), MCD has also four (18, 27, 95 and 96) and EA has eight (27, 41, 56, 62, 95, 96, 182, 202). If all methods detected observation 27, BACON missed everything else. The other four methods detected also 95, while two other observations where only missed by one method: 96 missed by SMP (but ranked only one observation behind) and 62 missed by RMCD (but ranked only two observations behind). Observation 202 was detected by both SMP and EA. Finally RMCD added 18, SMP added 85 and EA added 41, 56 and 182. If we except BACON that probably fails because of the total lack of normality of the data we see that only four observations appear in all twelve most outlying points for all methods: 27, 62, 95 and 96 (all detected as more outlying by MSD).

To give another way to see these results we introduce a new measure called a consensus measure. For a fixed number k, denote by $X(k)$ the set of k first outliers declared by the method X, $X \in \{$SMP,BACON,MSD,RMCD,EA$\}$ and by $all(k)$ the union with

Table 6: Comparison of central and extreme parts for the Ionosphere data

Central part (134 points)						Extreme part (12 points)					
	SMP	BAC	MSD	RMCD	EA		SMP	BAC	MSD	RMCD	EA
SMP	134	118	111	103	108	SMP	12	2	7	7	7
BAC	118	134	98	90	119	BAC	2	12	2	2	2
MSD	111	98	134	125	87	MSD	7	2	12	9	7
RMCD	103	90	125	134	80	RMCD	7	2	9	12	6
EA	108	119	87	80	134	EA	7	2	7	6	12

repetition (i.e. $\{a;b\} \cup \{a;c\} = \{a;a;b;c\}$) of the $X(k)$'s. Our consensus measure is defined as:

$$cm(X,k) = \frac{1}{k} \sum_{x \in X(k)} \frac{\text{\#occurrences of } x \text{ in } all(k) - 1}{\text{\#methods} - 1}$$

In other words $cm(X,k)$ measure the average frequency that a given outlier in $X(k)$ is detected by another method. Note that if you have the above table, $cm(X,k)$ is just the average of the quotients of the non-diagonal elements of the line for X divided by k. When all methods detects the same first k outliers then $cm(X,k) = 1$ for all X and when for a given method X none of the $X(k)$ is detected by another method then $cm(X,k) = 0$. Table 7 gives the values of the $cm(X,12)$ and confirm that for the Ionosphere data BACON is very isolated and that MSD is the most consensual.

Table 7: Consensus measures for the Ionosphere data

X	SMP	BACON	MSD	RMCD	EA
cm(X,12)	0.48	0.17	0.52	0.5	0.46

The computing times diverge. SMP took 0.6s, BACON 0.41s, MSD 342s, MCD 22s and EA took 2.1s. Note that even if our implementation of MSD is not optimized we can see that when the dimension of the data grows, the computing time of MCD and MSD grows too. This was expected as well as the fact that the computing time of EA is not much affected by the growth of dimension (remember that the dimension appears in the algorithm only in the distance computation). SMP and BACON remain by far the fastest but in such a case with a large part of non-normal data BACON seems to fail to detect the outliers.

12 The Low Resolution Spectrometer (LRS) data

The third real dataset is also taken from the UCI Machine Learning Database Repository. These data were gathered in the Infra-Red Astronomy Satellite (IRAS) project, that was the first attempt to map the sky at infra-red wavelengths. It consists of 531 high quality spectra measured on 93 different frequencies.

We encountered two problems when running the different algorithms. As the number of points (531) is not important relatively to the dimension (93) of the data, BACON totally failed to work out : all the considered subsets did have a singular covariance matrix and therefore the algorithm was unable to compute Mahalanobis distances. Moreover, the S-Plus function cov.mcd does not allow more than 50 variables but as the LRS dataset has already been analyzed using RMCD by Maronna and Zamar (Maronna and Zamar, 2001) we are just referring to these results for RMCD. MSD was run with 2000 different directions. We do not show the Q-Q plots of the D_i's or the infection times as they are similar to the preceding ones except that this time the normally behaving part of the data seems bigger. For example only 8 observations were infected after time 7 and only 3 not infected with EA. As the other methods also had 11 or 12 clear outliers, we give the comparative table of the extreme part in Table 8. The results are here very similar.

Table 8: Comparison of the extreme parts for the LRS data

Extreme part (11 points)				
	SMP	MSD	RMCD	EA
SMP	11	10	10	10
MSD	10	11	9	9
RMCD	10	9	11	9
EA	10	9	9	11

SMP is the most consensual method and eight observations are simultaneously detected by all methods. The differences here are rather the measures of outlyingness given by the methods. Table 9 lists the 11 observations in decreasing order of their measure of outlyingness.

Table 9: Most outlying observations for the LRS data

SMP	210	90	112	173	307	281	451	193	2	67	382
RMCD	210	173	112	90	307	2	281	193	451	67	370
MSD	307	382	210	281	280	90	173	112	2	67	451
EA	210	307	281	451	398	90	382	67	112	173	193

The consensus measures are here very high (see Table 10).

Table 10: Consensus measures for the LRS data

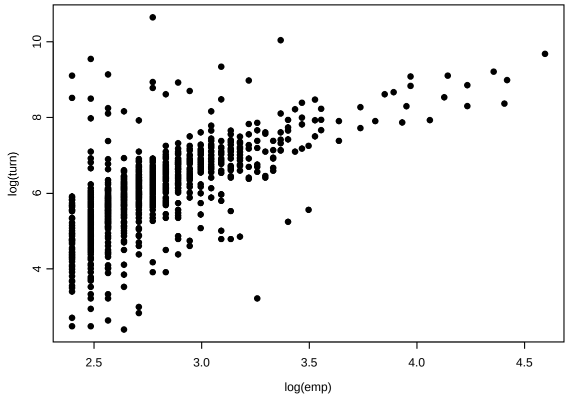
X	SMP	MSD	RMCD	EA
cm(X,11)	0.91	0.85	0.85	0.85

The computing times diverge. SMP took 2.1s, MSD 398s, MCD 616s and EA took 5.7s. With that dimension the computing time of MCD and MSD start to get very big while EA is not much affected by the growth of dimension. SMP keeps performing fast and well.

13 The Restaurants data

As business surveys are often encountered in official statistics we felt that it was necessary to include in these preliminary tests a dataset of such a kind. The problematic point of such data is that they always need some transformation, usually some log transformation, prior to any analysis and that they often do not have some nice elliptical or symmetric distribution. The following dataset is a subsample of restaurants of the 1995 Swiss census of the enterprises. The largest restaurants were removed for confidentiality reasons. As we wished to present graphically the results only two variables were retained: emp will denote the number of employees and $turn$ the turnover of the restaurants. As usual a log transformation is performed first. A scatter plot of the 1271 observations is given in Chart 10.

Chart 10: Scatter plot of the restaurants data after a log transformation



Such a picture is common in business surveys. No symmetry appears in the dataset and therefore the methods needing that assumption will clearly have trouble to cope with that characteristic. Looking at the plot we could consider as potential outliers the restaurants with a high number of employees or for the other ones with high or low turnover.

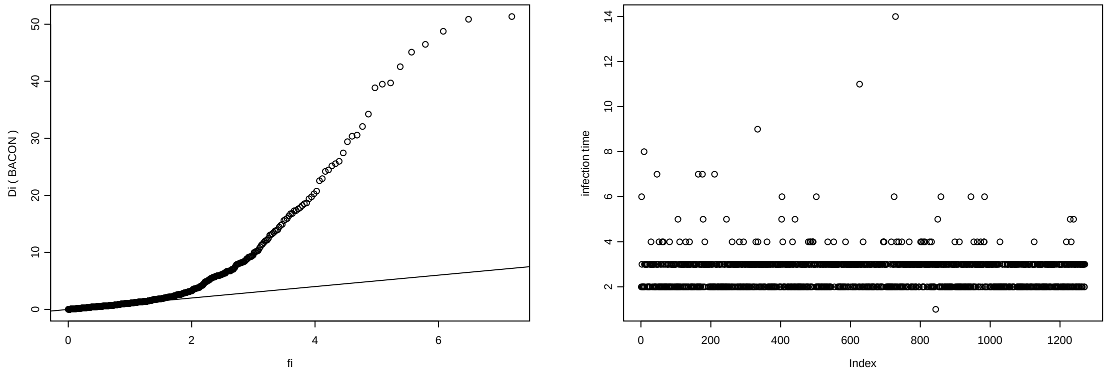
In that case the results obtained by all the methods using a Mahalanobis distance are so close that there is no point to try to compare them. To illustrate that fact we gave some consensus measures for these methods in Table 11.

Table 11: Consensus measures (without EA) for the restaurants

X	SMP	BACON	MSD	RMCD
cm(X,10)	1	1	1	1
cm(X,50)	0.97	0.97	0.97	0.97
cm(X,100)	0.94	0.96	0.94	0.96
cm(X,150)	0.98	0.98	0.96	0.98

Therefore we restrict our comparison between one of them (BACON) and EA. We gave first the Q-Q plot of the D_i for BACON and the infection history for EA (see Chart 11).

Chart 11: D_i for BACON and infection times for the restaurants dataset



Looking at these charts we could consider that 22 observations seem to be really outlying for BACON while EA found 23 observations with infection time greater than 4. We plotted the data with these outliers for BACON and EA (see Chart 12). As EA infected only 75 observations after time 3, we also plotted the 75 most outlying points for both methods (see Chart 13).

On these pictures we clearly see the difference between EA and the other methods. BACON bases its measure of outlyingness using what should be the symmetric (elliptical) part of the good data. Therefore here we clearly see that BACON does not detect as well as EA the observations located in the direction of the main axis of the ellipsoid (high emp and high turn) because these observations seem to fit the normal model sought by

Chart 12: Outliers for BACON (22) and EA (23) for the restaurants dataset

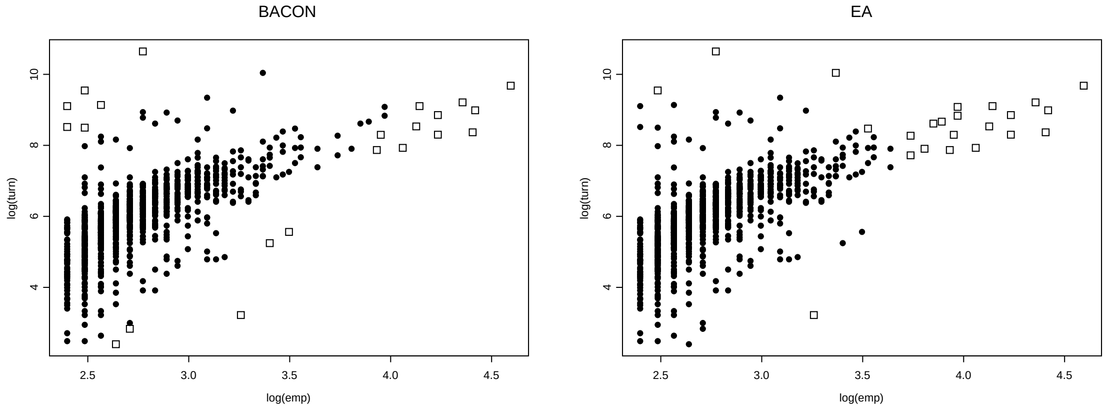
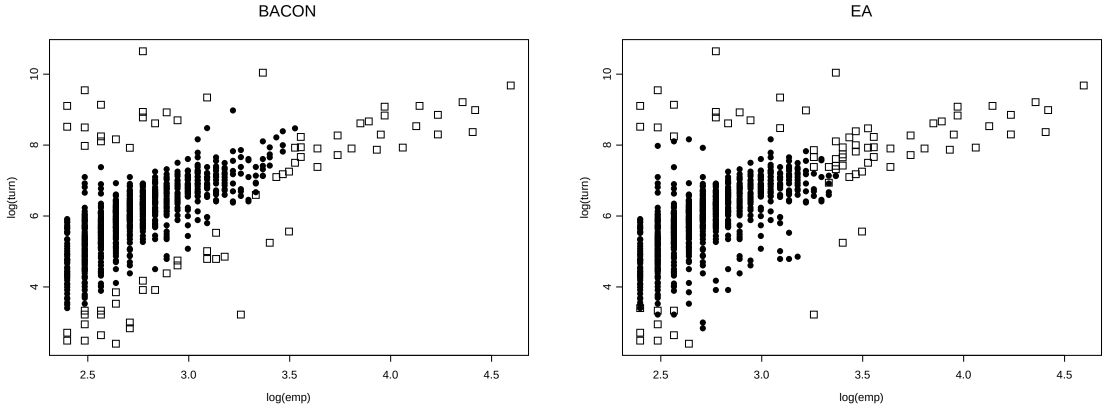


Chart 13: Outliers (75) for BACON and EA for the restaurants dataset



BACON. On the contrary EA doesn't look for a model and therefore found very well the observations that we considered as outliers when we first looked at the scatter plot.

The computing times here show clearly that EA is more affected by the number of observations than other methods. SMP took 0.6s, BACON 0.5s, MSD 2.1s, MCD 0.7s and EA took 11s.

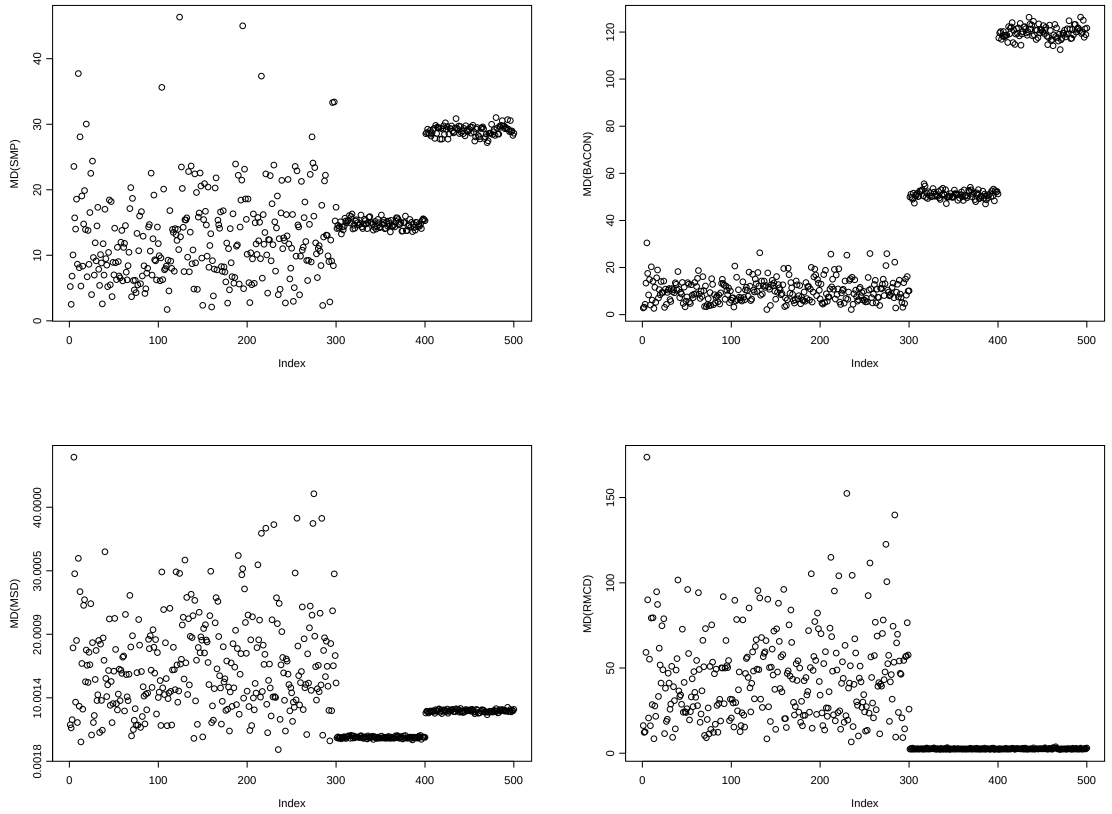
14 Dataset with high concentrated contamination

In (Rocke and Woodruff, 1996) Rocke and Woodruff made two observations: 1) it is very hard to detect outliers in data with a contamination fraction of 35% or higher; 2) compactly spaced outliers are harder to find. To test the quality of the different methods we combined here the two difficulties: we generated a dataset with 500 observations in

$\mathbb{R}^{10}$ with observations 1 to 300 that followed a multivariate normal distribution centered at the origin with a covariance matrix set to $10 * I_{10}$ and two contaminations formed by two other clouds centered at two randomly chosen points in $\mathbb{R}^{10}$, one at distance 70 (observations 301 to 400) and the other at distance 100 (observations 401 to 500), both with multivariate normal distribution with covariance matrix of I_{10}.

Here, as we know the indices of the outliers, the results of all methods are just plotted with the Mahalanobis distance or the infection time versus the index (see Chart 14). We restricted MSD to 5000 projections.

Chart 14: MD_i or infection time for SMP, BACON, MSD, RMCD, EA for the dataset with concentrated contamination

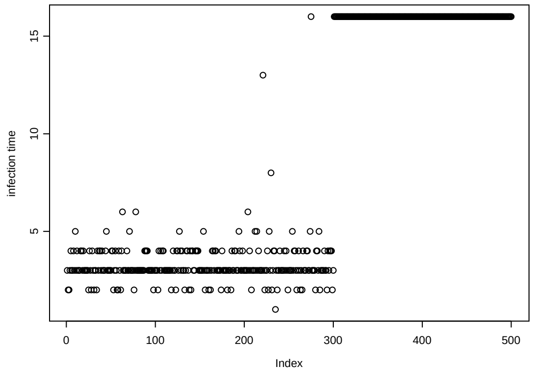


The results are very different:

SMP The more distant outlier cloud and some other good points were detected with high Mahalanobis distances, but the closest cloud was not.

BACON The detection is perfect even adding the distinction between the two clouds. This is no surprise since BACON is designed to be perfect in such cases.

MSD Nothing is detected except good points. Of course here by changing the weighting function the results could be totally different.



MCD The 200 outliers got the smallest Mahalanobis distances and had no chance of being detected. The Q-Q plot looks very strange but can only tell that there is a problem...

EA The 200 outliers have not been infected and they are therefore perfectly detected. Three other points are infected after time 6 and are therefore suspicious. The algorithm did not make any difference between the two clouds.

The computing times were the following: SMP took 0.18s, BACON 0.14s, MSD 140s, MCD took 5.1s and EA 11s.

15 Other datasets

We have tried the methods on several other datasets found in the literature and considered as challenging for multivariate outlier detection. The methods tested here worked perfectly well in most of the cases. Only with few data relative to the dimension some methods failed to identify the outliers. We do not report in all details the tests, only the references for the data, the computing times and the encountered problems are given.

The Hertzsprung-Russell data This dataset is given in (Rousseeuw and Leroy, 1987), Table 3, Chapter 2. A scatter plot can be found on page 261. The dataset has 47 points in dimension 2. All the methods found perfectly the 6 clear outliers with computing times: SMP in 0.08s, BACON in 0.09s, MSD in 0.63s, RMCD in 0.15s and EA in 0.32s.

The Hawkins-Bradu-Kass data This dataset is given in (Rousseeuw and Leroy, 1987), Table 9, Chapter 3. The dataset has 75 points in dimension 3 (we did not use the response variable). All the methods found perfectly the 14 outliers with computing times: SMP in 0.14s, BACON in 0.17s, MSD in 1.36s, RMCD in 0.23s and EA in 0.45s.

The Modified Wood Specific Gravity data This dataset is given in (Rousseeuw and Leroy, 1987), Table 8, Chapter 6. The dataset has 20 points in dimension 5 (we did not use the response variable). BACON (with a smaller starting subset, i.e. $c = 2$), MSD and RMCD found perfectly the 4 artificial outliers with computing times: BACON in 0.5s, MSD in 7s and RMCD in 0.18s. EA did not infect the four outliers but also four other good points. BACON with the default starting subset of size $3p$ and SMP did not detect anything.

Remarks The last example shows that with very small datasets SMP, BACON and EA might encounter some problems while the two methods based on geometric ideas are performing relatively well.

In the cases of regression data robust multiple regression methods should rather be used and in several cases the multivariate methods we present here will totally fail to detect the regression outliers. Nevertheless we can always try to find the outliers in the explanatory variables using the multivariate methods just simply by deleting the response variable as we did in the two preceding examples. Sometimes the multivariate methods will also end up with the right outliers in a regression context as in the well known next example.

The Stackloss data This dataset is well-known and can be found in several articles but also in (Rousseeuw and Leroy, 1987), Table 1, Chapter 3. The dataset has 21 points in dimension 4. Most analysts agree that observations 1, 3, 4 and 21 are outliers and some of them add observation 2. SMP, BACON, MSD and RMCD found the five outliers with 2 as the least outlying. EA was run several times, always found 1, 2 and 3 as outliers but sometimes missed 4 or 21 and sometimes added 17. The computing times were: SMP in 0.09s, BACON in 0.06s, MSD in 2.8s, RMCD in 0.15s and EA in 0.4s.

The Philips data This dataset is an illuminating example to show how important the parameters' tuning can be. This dataset has been used by Rousseeuw and van Driessen (Rousseeuw and van Driessen, 1999) to test their FAST-MCD. The analysis using RMCD shows 78 clear outliers (observations 297, 298, 433 and some concentrated contamination from 491 to 565) and some other suspicious points in 3.1s. SMP detects clearly 297 and 298 as well as some other suspicious points (among them 433) but totally fails to detect the concentrated contamination of observations 491 to 565. BACON with our default parameters detects 638 outliers indicating that we have to take a much smaller significative level than 0.01. Actually with this level set to 0.0001 BACON detects exactly the same outliers as MCD in 0.66s (In fact with its original parameters BACON would have obtained these results). MSD got the same results as SMP, missing the concentrated contamination. By changing the weighting function we could of course improve the detection of closed contamination of MSD but then of course we would also increase drastically the number of good observations declared as outliers. EA with the default settings detects only three points as clear outliers (175, 297 and 298). The maximum transmission distance is then $d_0 = 3.05$. By setting $d_0 = 2.4$ all the concentrated contamination also appears clearly as outlying, nevertheless EA missed 433.

16 Conclusions

Let's try to summarize the diverse advantages and disadvantages pointed out up to that point.

SMP The behavior of SMP has been quite a good surprise. By construction it is very fast in all cases and seems to get very satisfying results comparing to the other methods. It has problems to deal with very small datasets and with close concentrated contamination. No parameters are needed. It does need the assumption of symmetric data.

BACON BACON is very fast in all cases. It is the best method when the good part of the data is normally distributed. It starts to behave strangely when an important part of the data is not normal (Ionosphere data). It's major problem is that it cannot work when the number of observations n is not large relatively to the dimension p. It has problems to deal with very small datasets. Some knowledge of the algorithm is needed for a good parameters' tuning. It does need the assumption of symmetric data.

MSD MSD is a relatively slow algorithm. It's computing time explodes with the dimension p and therefore approximation using less projections has to be taken. The choice of the weighting function decides the sensitivity to close outliers. It does need the assumption of symmetric convex data.

RMCD RMCD is a relatively slow algorithm. It's computing time explodes with the dimension p. It has major problems to deal with concentrated contamination. It does need the assumption of symmetric data.

EA EA is a relatively slow algorithm. It's computing time does not grow with p but with n. It has problem to deal with very small datasets. The choice of the maximum transmission distance is crucial. It compares very well with the other methods. It is the only tested methods that work well with non-symmetric data.

Part IV

Adaptation to sampling weights

All the methods developed in EUREDIT will have to be able to analyze data from sample surveys. In consequence they should all take sampling weights into account. This section is dedicated to the adaptation of the methods selected above to the sampling weights. We shall keep the same notations throughout the section. The population will be denoted by U and will have N units numbered by $1, 2, ..., N$. We shall assume that the sample s is drawn from U according to a sample design $p(s)$. The size of s will be n and its units labeled by $1, 2, ..., n$. This is a slight abuse of notation as the sample indices should rather be written as $i_1, ..., i_n$ with $i_j \in \{1, ..., N\}$. The first and second order inclusion probabilities will be denoted by π_i and π_{ij}. We shall assume that the weights w_i given with the data are just the sampling weights $w_i = 1/\pi_i$. If a quantity is measured on the sample with values $x_1, ..., x_n$ the classical Horvitz-Thompson estimator of the total $X = \sum_{i=1}^{N} x_i$ is then given by $\widehat{X}_{HT} = \sum_{i=1}^{n} w_i x_i$. If there is a more complex procedure behind the weights, e.g. calibration, we simply assume that $\sum_{i=1}^{n} w_i x_i$ yields a good estimate of the total $\sum_{i=1}^{N} x_i$.

17 SMP

The adaptation of SMP to sampling weights will require some more sophisticated estimation methods. We shall only give a general outline of the construction of the diverse estimators. For more details on the estimators and the estimation of their variances the reader should refer to (Deville, 1999).

17.1 Substitution estimators

We shall work in a measure space on $\mathbb{R}^p$ denoted by $\mathcal{M}$ containing at least all point measures denoted by δ_x with $x \in \mathbb{R}^p$ (in the following we shall only deal with discrete measures). A functional T on $\mathcal{M}$ associates to every measure $M \in \mathcal{M}$ a real number $T(M)$. We shall work only with homogeneous functionals, i.e. those for which there exists some $\alpha = \alpha(T) \in \mathbb{R}^+$ such that $T(tM) = t^\alpha T(M)$. A set of real values $\{x_1, ..., x_N\}$ taken on the population U defines a measure $M_U = \sum_{i=1}^{N} \delta_{x_i} \in \mathcal{M}$. Similarly the values $\{x_1, ..., x_n\}$ taken on the sample s with given sampling weights w_i defines a measure $M_s = \sum_{i=1}^{n} w_i \delta_{x_i} \in \mathcal{M}$.

Definition With the above notations the *substitution estimator* of some functional value $T(M_U)$ is $T(M_s)$.

In the case of a total this definition is nothing else than the classical expansion estimator (π-estimator or Horvitz-Thompson estimator): the functional is defined by $T(M) =$

$\int x dM(x)$. The value of the functional on the population distribution is the sought total $T(M_U) = \sum_{i=1}^{N} x_i = X$ and therefore the substitution estimator is $\widehat{X} = T(M_s) = \sum_{i=1}^{n} w_i x_i$. Several estimators cannot be directly defined as a functional value but are actually solution of an implicite functional equation (maximum likelihood estimators for example). Generally the estimating equation can be written as $T(M_U, \lambda) = 0$ where this time the functional has a real parameter λ. This equation is supposed to have a unique solution for M_U fixed. In that case the substitution estimator of λ is the solution of the equation $T(M_s, \widehat{\lambda}) = 0$.

Even if we shall not estimate the variance of our estimators in this report, let us note here that a tool developed in the field of robustness becomes a very powerful tool in estimation theory for variance computation. Actually the influence function of a functional defined here as $IT(M, x) = \lim_{t \to 0} \frac{1}{t}(T(M + t\delta_x) - T(M))$ can define a linearized version of the substitution estimate and therefore can be used to compute the variance of the estimate using classical formulas. The variance of all the estimators we shall use here can be computed this way, see (Deville, 1999).

Substitution estimators will be used here to adapt SMP to sampling weights. In fact as SMP uses the Spearman Rank correlation we do need an estimation of the ranks in the population to be able to compute the estimator. An easy way to estimate the ranks is to express them as functionals and use substitution estimators. Similarly the median and the mad will be expressed as solution of implicit functional equation and the substitution estimators are nothing else than the classical weighted median and mad.

As usual denote by $\{x_1, ..., x_N\}$ and $\{y_1, ..., y_N\}$ the values of two quantities measured on the population and by $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_n\}$ the values in the sample ($x_i, y_j \in \mathbb{R}$). Define the two following functionals

$$R_i(M) = \int \mathbb{1}_{x \le x_i}(x) dM(x) - \frac{1}{2} \int \delta_{x_i}(x) dM(x) + \frac{1}{2}$$

and

$$Q_j(M) = \int \mathbb{1}_{x \le y_j}(x) dM(x) - \frac{1}{2} \int \delta_{y_j}(x) dM(x) + \frac{1}{2}.$$

The two functionals evaluated on the two population measures given by the x_i's and the y_i's are nothing else than the ranks in the population:

$$R_i(M_U^x) = \sum_{k=1}^{N} \mathbb{1}_{x \le x_i}(x_k) - \frac{1}{2} \sum_{k=1}^{N} \delta_{x_i}(x_k) + \frac{1}{2} = R_i$$

and

$$Q_j(M_U^y) = \sum_{k=1}^{N} \mathbb{1}_{y \le y_j}(y_k) - \frac{1}{2} \sum_{k=1}^{N} \delta_{y_j}(y_k) + \frac{1}{2} = Q_j$$

where R_i (resp. Q_j) is the rank of x_i (resp. y_j) in the whole population values. Note that in the literature the formula for the ranks is often simply given as $\sum_{k=1}^{N} \mathbb{1}_{x \le x_i}(x_k)$. The formula we proposed here is slightly more complicated but has two advantages. Firstly the formula is exact when some values are tied giving in that case the mean rank of these values (when no equality appears it reduces to the usual formula) and secondly when we shall look at the estimation this formula gives a better estimation in particular with very large weights. The substitution estimators are

$$\widehat{R_i} = R_i(M_s^x) = \sum_{k=1}^{n} \mathbb{1}_{x\leq x_i}(x_k)w_k - \frac{1}{2}\sum_{k=1}^{n}\delta_{x_i}(x_k)w_k + \frac{1}{2} = \sum_{\substack{1\leq k\leq n\\ x_k<x_i}} w_k + \frac{1}{2}\sum_{\substack{1\leq k\leq n\\ x_k=x_i}} w_k + \frac{1}{2}$$

and

$$\widehat{Q_j} = Q_j(M_s^y) = \sum_{k=1}^{n} \mathbb{1}_{y\leq y_j}(y_k)w_k - \frac{1}{2}\sum_{k=1}^{n}\delta_{y_j}(y_k)w_k + \frac{1}{2} = \sum_{\substack{1\leq k\leq n\\ y_k<y_j}} w_k + \frac{1}{2}\sum_{\substack{1\leq k\leq n\\ x_k=x_j}} w_k + \frac{1}{2}.$$

Using these estimated ranks we are now in a position to calculate the Spearman Rank correlation. Recall that by definition

$$R(x,y) = \frac{\sum_{i=1}^{N}(R_i - \overline{R}_i)(Q_i - \overline{Q}_i)}{\sqrt{\sum_{i=1}^{N}(R_i - \overline{R}_i)^2 \sum_{i=1}^{N}(Q_i - \overline{Q}_i)^2}}.$$

Using the relations $\sum_{i=1}^{N} R_i = \sum_{i=1}^{N} i = N(N+1)/2$ and $\sum_{i=1}^{N} R_i^2 = \sum_{i=1}^{N} i^2 = N(N+1)(2N+1)/6$ it reduces to

$$R(x,y) = \frac{12}{N(N^2-1)}\sum_{i=1}^{N} R_iQ_i - 3\frac{N+1}{N-1} \cong \frac{12}{N^3}\sum_{i=1}^{N} R_iQ_i - 3.$$

When N is large the last approximation is accurate. Setting $N(M) = \int dM$ we define the functional

$$R(M) = \frac{12}{N^3(M)}\int R_i(M)Q_i(M)dM - 3$$

which satisfy $R(M_U) = R(x,y)$ and we obtain the estimator

$$\widehat{R}(x,y) = R(M_s) = \frac{12}{N^3(M_s)}\int R_i(M_s)Q_i(M_s)dM_s - 3 = \frac{12}{(\sum_{i=1}^{n} w_i)^3}\sum_{i=1}^{n} w_i\widehat{R_i}\widehat{Q_i} - 3.$$

Note that by using the functional form of R we actually have double integrals involved in this formula. But this is straightforward once the functional form is used. Inserting the above formula for $\widehat{R_i}$ and $\widehat{Q_i}$ we have finally

$$\widehat{R}(x,y) = \frac{12}{(\sum_{i=1}^{n} w_i)^3}\sum_{i=1}^{n} w_i \left(\sum_{\substack{1\leq k\leq n\\ x_k<x_i}} w_k + \frac{1}{2}\sum_{\substack{1\leq k\leq n\\ x_k=x_i}} w_k + \frac{1}{2} \right) \cdot \left(\sum_{\substack{1\leq k\leq n\\ y_k<y_j}} w_k + \frac{1}{2}\sum_{\substack{1\leq k\leq n\\ x_k=x_j}} w_k + \frac{1}{2} \right) - 3.$$

Note that we have used the fact that the mean and variance of the ranks are known on the population to simplify the correlation formula. We might obtain a more efficient estimator if we estimate these quantities using the sample values. Finally let us underline the fact that we have no guarantee that the value of our estimator is between -1 and 1 but in such a case we would clearly set the value to -1 or 1.

To end the adaptation of SMP to sampling weights we still have to estimate the median and mad for univariate data. In the above context this is done very easily. Define the functional $T(M,\lambda) = \frac{1}{M(\mathbb{R})} \int \mathbb{1}_{x \leq \lambda}(x) dM(x)$. Then the median of the population data is the solution of the functional equation $T(M_U, \lambda) = 0.5$ and therefore its estimator is the solution of $T(M_s, \widehat{\lambda}) = 0.5$, i.e. $\left(\sum_{\substack{1 \leq k \leq n \\ x_k \leq \lambda}} w_k \right) = 0.5 \cdot \sum_{1 \leq k \leq n} w_k$. Now in general this equation does not have a solution. Different approximations can be used, the one we choose is defined as follows. Let x_j be the smallest value such that

$$\left(\sum_{\substack{1 \leq k \leq n \\ x_k \leq x_j}} w_k \right) \geq 0.5 \cdot \sum_{1 \leq k \leq n} w_k,$$

and let x_l be the smallest value such that

$$\left(\sum_{\substack{1 \leq k \leq n \\ x_k \leq x_l}} w_k \right) > 0.5 \cdot \sum_{1 \leq k \leq n} w_k,$$

then the weighted median is defined as

$$\widehat{med}(x) = weighted.med(x,w) = \begin{cases} x_j & \text{if } x_j = x_l \\ \frac{w_j x_j + w_l x_l}{w_j + w_l} & \text{if } x_j < x_l \end{cases} .$$

As the mad is defined using medians only, its estimation follows in the same way.

18 BACON

The adaptation of the BACON algorithm is almost straightforward. The initial subset is selected the same way except that the usual median is replaced by its estimate defined in the preceding section 17, namely the weighted median. For the main iterations of the algorithm the mean and covariance matrix of the population are estimated each time by m_G and S_G and the observations are ranked using this estimation. We only have to follow the same scheme except that we shall work in the sample. Suppose that we randomly chose k element of the sample s. We can estimate the mean and the covariance matrix of the population with the Hájek estimator using the fact that the probability that the observation x_i appears in this subset G of the sample s is simply given by $\widetilde{\pi}_i = k\pi_i/n =$

$k/(w_i n)$. The estimates are therefore

$$m_G = \frac{\sum_{i\in G} \widetilde{\pi}_i^{-1} x_i}{\sum_{i\in G} \widetilde{\pi}_i^{-1}} = \frac{\sum_{i\in G} w_i x_i}{\sum_{i\in G} w_i}$$

and

$$S_G = \frac{\sum_{i\in s_G} \widetilde{\pi}_i^{-1}(x_i - m_G)(x_i - m_G)^t}{\sum_{i\in G} \widetilde{\pi}_i^{-1}} = \frac{\sum_{i\in G} w_i(x_i - m_G)(x_i - m_G)^t}{\sum_{i\in G} w_i}.$$

Finally we have to determine the correction factors used to do the selection of Step 3 of the algorithm. The first factor c_{hr} measures the correction if the size r of the part on which we made the estimation is smaller than "half" of the observations $h = \lceil (N+p+1)/2 \rceil$. As r (resp. h) can be estimated using the Horvitz-Thompson estimator $\widehat{r} = \sum_{i\in G} w_i$ (resp. $\widehat{h} = (\sum_{i\in s} w_i + p + 1)/2$) we may estimate this correction and use

$$\widehat{c}_{hr} = \max\left\{0, \frac{\sum_{i\in s} w_i + p + 1 - 2\sum_{i\in G} w_i}{\sum_{i\in s} w_i + p + 1 + 2\sum_{i\in G} w_i}\right\}$$

instead of c_{hr}. In the same way we use the estimate

$$\widehat{c}_{Np} = 1 + \frac{p+1}{\sum_{i\in s} w_i - p} + \frac{2}{\sum_{i\in s} w_i - 1 - 3p}$$

instead of c_{Np} to take into consideration the size of p proportionally to the size of the population.

19 MSD

In (Franklin et al., 2000) a comparaison of the effects of multivariate outlier detection using MSD with and without considering sampling weights is made. The approach chosen by Franklin et al. will not be followed here: to avoid burdensome reprogramming they decided simply to multiply each observation by its sampling weights and then to apply the algorithm. It didn't seem to us that we could find a theoretical justification to that scheme.

We propose to make the following adaptations to the algorithm given in 7.1. The projections are unchanged but the computation of the weights for a given one-dimensional projection need the value of the median and the mad for the whole population. We replace here these two values by their estimate obtained using the estimators defined in the preceding section, namely the weighted median and the weighted mad. With this correction Points 1 to 5 of the algorithm remain the same. Finally Point 6 and the final estimation are obtained using the usual estimates of the mean and covariance matrix of the population computed with robustness weights. We shall do the same just by replacing the usual estimators by the Hájek estimators, i.e. using the following estimates:

$$m_{MSD} = \frac{\sum_{i=1}^n u_i w_i x_i}{\sum_{i=1}^n u_i w_i} \text{ and } S_{MSD} = \frac{\sum_{i=1}^n u_i w_i (x_i - m_{MSD})(x_i - m_{MSD})^t}{\sum_{i=1}^n u_i w_i}$$

20 RMCD

The adaptation of the FAST-MCD algorithm described in section 8 is also straightforward. As in BACON the C-steps use computations of the mean m_K and covariance matrix S_K of a subset $K \subset U$ to rank all the observations according to the Mahalanobis distance. In each step the idea is that the mean and covariance matrix of the population are estimated by m_K and S_K and the observations are ranked using this estimation. We only have to follow the same scheme except that we shall use the sample. Suppose that we randomly chose k element of the sample s. We can estimate the mean and the covariance matrix of the population with the Hájek estimator using the fact that the probability that the observation x_i appears in this subset s_k of the sample s is simply given by $\widetilde{\pi}_i = k\pi_i/n = k/(w_i n)$. The estimates are therefore

$$m_k = \frac{\sum_{i \in s_k} \widetilde{\pi}_i^{-1} x_i}{\sum_{i \in s_k} \widetilde{\pi}_i^{-1}} = \frac{\sum_{i \in s_k} w_i x_i}{\sum_{i \in s_k} w_i}$$

and

$$S_k = \frac{\sum_{i \in s_k} \widetilde{\pi}_i^{-1} (x_i - m_k)(x_i - m_k)^t}{\sum_{i \in s_k} \widetilde{\pi}_i^{-1}} = \frac{\sum_{i \in s_k} w_i (x_i - m_k)(x_i - m_k)^t}{\sum_{i \in s_k} w_i}.$$

In the case $h = n$ (point 2 of the algorithm) the Hájek estimates m_h and S_h are returned. In the case $p = 1$, the same arguments give a clear adaptation of the algorithm given in (Rousseeuw and Leroy, 1987) replacing the $n - h + 1$ means by their Hájek estimates and the sum of squares by the Hájek estimate of the corresponding variances. With these corrections the structure of Points 1 to 6 of the algorithm remains unchanged. Note here that $h = (n + p + 1)/2$ is computed using the sample size n and therefore the breakdown point is expressed according to the proportion of outliers in the sample and not in the population. Once the subset s_h is chosen the Hájek estimates m_h and S_h are used and points 7 and 8 become:

7. In order to obtain consistency under multivariate normal distribution set

$$m_{MCD} = m_h \qquad \text{and} \qquad S_{MCD} = \frac{weighted.med_i(MD_{m_h,S_h}(x_i))}{\chi^2_{p,0.5}} S_h$$

   where the $weighted.med$ denotes the weighted median defined in 17.

8. To improve efficiency under normal distribution set finally

$$m_{RMCD} = \frac{\sum_{i=1}^n u_i w_i x_i}{\sum_{i=1}^n u_i w_i} \text{ and } S_{RMCD} = \frac{\sum_{i=1}^n u_i w_i (x_i - m_{RMCD})(x_i - m_{RMCD})^t}{\sum_{i=1}^n u_i w_i}$$

   with

$$u_i = \begin{cases} 1 & \text{if } MD_{m_{MCD},S_{MCD}}(x_i) \leq \chi^2_{p,0.975} \\ 0 & \text{otherwise} \end{cases}$$

21 EA

As usual we assume that a sample s of size n is drawn from the population U of size N according to the sample design $p(s)$. The first and second order inclusion probabilities are denoted π_i and π_{ij}. We assume that the sampling and the epidemic are independent.

The initial standardization of the data, designed to avoid unbalanced effect of the different variables, should be done using the median and mad computed on the population data. We therefore estimate these quantities using the sample data with the weighted.med (defined in 17) and the weighted.mad (defined as the mad replacing the median by the weighted.median), i.e

$$\tilde{x}_{jk} = \frac{x_{jk} - weighted.med_{i\in s}(x_{ik}, w_i)}{weighted.mad_{i\in s}(x_{ik}, w_i)}.$$

To determine the starting point of the epidemic according to the algorithm we should use the population spatial median $c = \arg\ \min_i\ \sum_{j\in U} d(x_i, x_j)$. As the sum over the population is not known we use its Horvitz-Thompson estimate and therefore our starting point will be

$$c = \arg\ \min_i\ \sum_{j\in s} w_j d(x_i, x_j).$$

Denote by $I_{U,t}$ the set of infected points in the population U at time t. The set $I_{U,t}$ is a domain. Its intersection with the sample s is $I_{s,t} = s \cap I_{U,t}$ the set of infected points in the sample. What we actually observe is $I_{s,t}$. In order to infer on the infection in the population, we have to estimate the infection probabilities $P[j|I_{U,t}] = 1 - \prod_{i\in I_{U,t}}(1 - h_{ij})\ , \forall j \in s \setminus I_{s,t}$. Thus we have to estimate the product $\prod_{i\in I_{U,t}}(1 - h_{ij})$ from the sample and from knowing $I_{s,t}$. Taking the log of this estimand we can see that we have to estimate the exponential of $\sum_{I_{U,t}} \log(1 - h_{ij})$. This sum can be estimated by the Horvitz-Thompson estimator

$$\sum_{I_{s,t}} \frac{1}{\pi_{ij}} \log(1 - h_{ij}).$$

Exponentiation of this unbiased and consistent estimator leads to a consistent estimator of the product. Thus the estimator of the transmission probability becomes

$$\hat{P}[j|I_{U,t}] = 1 - \prod_{i\in I_{s,t}} (1 - h_{ij})^{1/\pi_{ij}}.$$

In theory the problem is solved: We use these transmission probabilities for the epidemic in the sample. Since the transmission probabilities estimate the transmission probabilities of the population infection, the infection times will estimate the corresponding infection times in the population.

In practice we seldom have the second order inclusion probabilities π_{ij} at hand. Often we just have for each point a sampling weight w_i, which is approximately the inverse $1/\pi_i$ of the inclusion probabilities. We propose to use the approximations $1/\pi_{ij} \approx 1/(\pi_i\pi_j) \approx w_i w_j$. The first approximation holds exactly for simple random sampling with replacement and for Poisson Sampling. It holds approximately for large samples, where the dependence of inclusion between elements usually is small. This leads to the following estimator of the population infection probability

$$\hat{P}[j|I_{U,t}] = 1 - \prod_{i \in I_{s,t}} (1 - h_{ij})^{w_i w_j}.$$

A more heuristic approach assumes that at the same place as the sampled point i there are w_i points of the population which are infected and transmit the infection at the same time as the sampled point. In other words we assume an immediate transmission if the distance is zero. Thus one would have w_i points which are already infected and instead of one candidate at x_j to be infected there are w_j of them. The transmission probability becomes $1 - \prod_{i \in I_{S,t}} (1 - h_{ij})^{w_i w_j}$, exactly as above.

We may standardize the weights to sum to n to obtain an infection probability which compares better with an epidemic in the sample alone. This may also help in the choice of the maximal infection distance d_0.

Another heuristic approach compares the density in the population with the density in the sample. The density is decreased by a factor which corresponds to the sampling fraction for simple random sampling. In the same way the average distance decreases by the sampling fraction. Thus an approach for accounting for sampling would be to transform the interpoint distance d_{ij} to $d'_{ij} = 2d_{ij}/(w_i + w_j)$. This would correct the distance by the average sampling rate at points i and j.

Part V

Adaptation to missing values

In this part we approach one of the critical problem encountered with real data: the missing values. In survey data we can distinguish two kinds of non-response that lead to missing values in a dataset. In fact not all units in the sample respond to all the study variables; some co-operate with the survey, but fail to supply answers to some question - we talk about *item non-response* - and others do not co-operate at all - *unit non-response*.

Different sampling techniques exist to deal with unit non-response. The methods developed here will not cope with that kind of non-response, it will always be assumed that the unit non-response has been taken into account by sampling techniques and that the sampling weights have been corrected according to unit non-response. All units that have all items missing will therefore be removed from the dataset.

Most of the edit methods that deal with item non-response do need strong assumptions on the missingness mechanism. That will also be the case here even if we still have to study further two methods to see if the hypothesis could be weakened. The first section will fix the notations and definitions for the missingness mechanism, while the next three sections will present the proposed solutions for three methods, each of them retaining the philosophy of the initial method: SMP will be adapted using simple imputations based on bivariate statistics, BACON will use a method designed to estimate a covariance matrix for incomplete multivariate normal data (BACON is best designed for this framework) and EA will simply compute distances using the available coordinates and correcting them with a proportionality factor to calibrate for the fraction of missing information. We did renounce to go further with the other two methods developed in the preceding parts. The projections in MSD couldn't be applied without some previous imputation of the missing data and we were not willing to merge together the edit and imputation phases at that point. Regarding MCD, an algorithm was developed in (Cheng and Victoria-Feser, 2000) using MCD at each step of the ER algoritm (Little and Smith, 1987) which combines the EM algorithm and and M-estimator. However this algorithm was not designed for survey data and we were lacking ressources to make the adaptation to sampling weights. Finally a very short exploration of that algorithm seems to show that it could have some difficulties to treat large size datasets.

22 Missingness mechanisms

The notions and notations for this section are largely taken from (Schafer, 2000). To make the following text readable we shall use the following abuse of notation: X will denote simultaneously a p-dimensional random variable (we shall always refer to the "variable X") and the $N \times p$ matrix containing the realized values of the variable X of the population U. The variable X follows a p-dimensional probability model with parameters θ. If a census was taken of the whole population to measure the variable X

it would result in some observed and missing values $X = X_o \cup X_m$. We shall model this behavior by a zero-one response variable R with the same abuse of notation: R also denotes the $N \times p$ matrix containing the values of the variable R on the population U,

$$\text{i.e. } r_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is given,} \\ 0 & \text{if } x_{ij} \text{ is missing.} \end{cases}$$

The parameters of the missingness mechanism will be denoted by ξ. We would not in general expect the distribution of the variable R to be unrelated to the variable X, so we posit a probability model $P(R|X, \xi)$. We shall always assume that the parameters θ of the data model and the parameters ξ are distinct. In most methods the assumption is that the missing data are ”missing at random” (MAR) or ”missing completely at random” (MCAR). The reader should be aware that the definition of MAR may vary depending on the context and the author, while the definition of MCAR is standard. We shall use the definition given in (Rubin, 1976) and used in (Schafer, 2000).

Definition 2 *The missing data are MAR if the distribution of R does not depend on X_m, i.e.*

$$P(R|X_o, X_m, \xi) = P(R|X_o, \xi).$$

If both MAR and the distinctness of the parameters θ and ξ hold, then the missing-data mechanism is said to be ignorable.

Definition 3 *The missing data are MCAR if the distribution of R does not depend on X, i.e.*

$$P(R|X, \xi) = P(R|\xi).$$

MCAR is a particular case of MAR, occuring for example when the missing data are a simple random sample of all data.

As the methods will use the survey data and not the population data we shall need an assumption on the relation between the missingness and the sampling mechanisms. If we denote by S the sampling variable, we shall always assume that S and R are independent variables : in other words we suppose that the missingness patterns do not depend on the sample: one unit x_i of the population would have the same observed and missing items regardless of the sample. If s is the sample obtained as a realization of S we shall simply use X_o^s (resp. X_m^s) to denote the observed (resp.missing) values of the survey data.

23 SMP

Almost every step of the SMP method is perturbed by missing values. We shall assume here that the data are MCAR; a more careful study should be carried out to see if this hypothesis can be weakened. Two different kinds of problems are encountered: the computation of univariate or bivariate statistics ($\tilde{\sigma}$ and R) and the projection of the observations onto the new basis B. The first issue is solved just by restriction to the observed cases. The second issue could be avoided by using another way of transforming the matrix $\tilde{S}_1$ into a definite positive matrix. But we prefer our transformation, which has some statistical interpretation, to a purely algebraic transformation. We propose a solution that keeps the ”robust bivariate” spirit of SMP. A missing item in an observation is imputed by a robust regression using another observed variable selected by the robust bivariate rank correlations. This imputation is then used to obtain the coordinates of the data in the new basis and the end of the algorithm remains unchanged. The final measure of outlyingness is the Mahalanobis distances computed on the observations without the imputed values. All the details are given by going through the algorithm step by step. The notations of Section 5.2 remain unchanged.

(i) The univariate statistics $\tilde{\sigma}$ of x^l is computed on the i’s such that $r_{il} = 1$. For our choice of $\tilde{\sigma}$ we therefore have to define how the estimation of the median is computed.

Let x_{il} be the smallest value of x^l such that $r_{il} = 1$ and

$$\sum_{\substack{1\leq k\leq n\\ r_{kl}=1\\ x_{kl}\leq x_{il}}} w_k \geq\ 0.5 \cdot \sum_{\substack{1\leq k\leq n\\ r_{kl}=1}} w_k.$$

and let x_{jl} be the smallest value of x^l such that $r_{jl} = 1$ and

$$\sum_{\substack{1\leq k\leq n\\ r_{kl}=1\\ x_{kl}\leq x_{jl}}} w_k >\ 0.5 \cdot \sum_{\substack{1\leq k\leq n\\ r_{kl}=1}} w_k,$$

the estimation of the median is given by

$$\widehat{med}(x^l, w) = \begin{cases} x_{il} & \text{if } x_{il} = x_{jl} \\ \frac{w_i x_{il} + w_j x_{jl}}{w_i + w_j} & \text{if } x_{il} < x_{jl} \end{cases}.$$

As the mad is defined using medians only, its estimation follows in the same way. For the Spearman Rank correlation we restrict all the computations to the common observed values of two variables. Using the formula developed in 17.1 we obtain

$$\widehat{R}(x^l, x^m) = \frac{12}{(\sum_{\substack{1\leq i\leq n\\ r_{il}r_{im}=1}} w_i)^3} \sum_{\substack{1\leq i\leq n\\ r_{il}r_{im}=1}} w_i \left(\sum_{\substack{1\leq k\leq n\\ r_{kl}=1\\ x_{kl}<x_{il}}} w_k + \frac{1}{2} \sum_{\substack{1\leq k\leq n\\ r_{kl}=1\\ x_{kl}=x_{il}}} w_k + \frac{1}{2} \right)$$
$$\cdot \left(\sum_{\substack{1\leq k\leq n\\ r_{km}=1\\ x_{km}<x_{im}}} w_k + \frac{1}{2} \sum_{\substack{1\leq k\leq n\\ r_{km}=1\\ x_{km}=x_{im}}} w_k + \frac{1}{2} \right) - 3,$$

if $\{i : r_{il}r_{im} = 1\} \neq \varnothing$. If there is no common observed variable between x^l and x^m then a warning is sent to the user and the correlation rank is set to zero

$$\widehat{R}(x^l, x^m) = 0 \text{ if } \{i : r_{il}r_{im} = 1\} = \varnothing.$$

The sizes of the set on which the correlations are computed are kept in the variable

$$c_{lm} = \sum_{i=1}^{n} r_{il}r_{im}.$$

(ii) The second step contains the projection problem. The computation of the new basis B is straightforward but the matrix product XB corresponding to the change of basis is impossible as soon as one item is missing. We use imputation by fitting a value using a robust regression. We set the following ”quality” condition for a variable x^k to be a regressor for a variable x^j:

$$c_{jk} = \sum_{i=1}^{n} r_{ij}r_{ik} > \gamma n \quad \text{for some parameter} \quad 0 < \gamma < 1.$$

For each variable x^j the algorithm will impute a value for a missing value x_{ij} ($r_{ij} = 0$) with a robust fit using the variable which has the highest $\tilde{R}(x^k, x^j)$ among the variables x^k satisfying the ”quality” condition and $r_{ik} = 1$. The following pseudo-code describes this imputation process.

```
- for all variables $x^j$ having missing values ($\sum_{i=1}^{n} r_{ij} < n$) do
    - select the $m$ ($0 \leq m \leq p-1$) variables $x^k$ such that $c_{jk} > \gamma n$;
    - if $m = 0$ next;
    - rank these variables according to  $\tilde{R}(x^k, x^j)$:
            $\tilde{R}(x^{k_1}, x^j) \geq \tilde{R}(x^{k_2}, x^j) \geq ... \geq \tilde{R}(x^{k_m}, x^j)$;
    - $reg = 1$;
    - while ($\sum_{i=1}^{n} r_{ij} < n$) and $reg \leq$  $m$ do
        - if $\sum_{i=1}^{n}(1 - r_{ij})r_{ik_{reg}} > 0$ fit a robust regression of $x^j$ on
            $x^{k_{reg}}$ and impute all $x_{ij}$ where $(1 - r_{ij})r_{ik_{reg}} = 1$ with the
            robust fit plus a randomly chosen residual error;
        -  $reg = reg + 1$ ;
        - next;
    - next;
- if some missing values are left ask the user to relax his quality
  condition or to exit.
```

Once all missing values have been imputed all the computations of the step can be performed.

(iii) Unchanged.

Remarks:

1) All regressions are fitted with the initial data, no imputed values are included in these computations.

2) In our implementation we use an M-estimator for regression which bounds the influence of residuals and of the explanatory variable x^k.

The detection is performed using Mahalanobis distance only on the initial data.

23.1 Mahalanobis distances

Once estimations of the mean $\tilde{m}$ and covariance matrix $\tilde{S}$ are available the estimation of the full Mahalanobis distance of an observation $x_i = (x_i)_o \cup (x_i)_m$ is based on the partial Mahalanobis distance computed on the components o_i of $(x_i)_o$ and inflated by a factor inversely proportional to the proportion of observed values $\frac{|o_i|}{p}$, i.e.

$$MD^2_{M,S}(x_i) = \frac{p}{|o_i|}((x_i)_o - M_{o_i})^t (S^{-1})_{o_i}((x_i)_o - M_{o_i}),$$

where M_{o_i} is the orthogonal projection of M to the subspace defined by o_i and $(S^{-1})_{o_i}$ is the restriction to that subspace of the quadratic form given by S^{-1}. Using the response variable R, this can be rewritten as

$$MD^2_{M,S}(x_i) = \frac{p}{\sum_{j=1}^{p} r_{ij}}(x_i - M)^t diag(r_i) S^{-1} diag(r_i)(x_i - M),$$

where $diag(r_i)$ is the diagonal matrix with diagonal r_i, the i^{th} line of R corresponding to x_i.

24 BACON

The ”growing a good subset of observations” principle is not disrupted by item nonresponse as long as the measure that is used to grow the subset at each step is available. In BACON this measure is given by Mahalanobis distances based on the Hàjek estimators of the mean and covariance matrix computed on the subset. The missing values will interfere with the three computations: the estimation of the mean, the estimation of the covariance matrix and the computation of the Mahalanobis distances using the other two. One problem - the Mahalanobis distances - is easily solved while the other two - the mean and the covariance matrix - are more delicate to deal with. The solution to the first problem has been presented in the preceeding section 23.1. For the other two problems we had to select estimators of the mean and the covariance matrix computable with missing values. We choose a method that is known to work well for multivariate normal data when applied to the whole population: the EM algorithm. In the second subsection we shall describe how we adapted the EM-algorithm to survey data to obtain EM estimators of the variance and covariance matrix. The reason of the choice of this algorithm was to maintain the efficiency of the BACON algorithm when applied to multivariate normal data. The last subsection will describe how the BACON and EM algorithms were merged together to create the ”BACON-EM for survey data” algorithm.

24.1 EM estimators for survey data under multivariate normal model

In this subsection we shall adapt the EM algorithm to the context of survey data. We shall begin by stating general points on the algorithm. This summary will present briefly the theory underlying the algorithm and some results for regular exponential families. All details can be found in (Schafer, 2000).

24.1.1 Generalities on EM

Model assumptions In order to justify the different steps of the algorithm, some assumptions on an underlying model of the population data are needed. We shall consider population datasets whose observations can be modeled as independant, identically distributed (iid) draws from some multivariate probability distribution $f(x, \theta)$. The probability function of the complete data may therefore be written as

$$P(X|\theta) = \prod_{i=1}^{N} f(x_i, \theta),$$

where N is the size of the population. This is called the complete-data model. Recall that X denotes simultaneously the random variable and the matrix of the values and that the same holds for R the response variable. In the following we shall assume that the missingness mechanism is ignorable, i.e. MAR and distincness of the parameters θ of X and ξ of R.

The EM algorithm The ignorability assumption allows us to factor the distribution of what we really observe $P(R, X_o|\theta, \xi)$ into two pieces:

$$\begin{aligned} P(R, X_o|\theta, \xi) &= \int P(R, X|\theta, \xi) dX_m \\ &= \int P(R|X, \xi) P(X|\theta) dX_m \\ &= P(R|X_o, \xi) \cdot \int P(X|\theta) dX_m \\ &= P(R|X_o, \xi) \cdot P(X_o|\theta). \end{aligned}$$

This factorization shows that likelihood-based inferences about θ can be performed without regard to the missing-data mechanism. The factor pertaining to θ will be called the *observed data likelihood*: $L(\theta|X_o) \propto P(X_o|\theta)$.

The distribution of the complete data X can always be factored as

$$P(X|\theta) = P(X_o|\theta) P(X_m|X_o, \theta).$$

Viewing each term as a function of θ and taking the log, we obtain

$$l(\theta|X) = l(\theta|X_o) + log(P(X_m|X_o, \theta)) + c,$$

where $l(\theta|X) = log(P(X|\theta))$ is the complete-data loglikelihood, $l(\theta|X_o) = log(L(\theta|X_o))$ is the observed-data likelihood and c is an arbitrary constant. The term $P(X_m|X_o, \theta)$ is crucial and plays a central role in EM. It captures the interdependence between X_m and θ on which EM capitalizes. As this predictive distribution $P(X_m|X_o, \theta)$ cannot be calculated each expectation step (E-step) will take an average over $P(X_m|X_o, \theta^{(t)})$, where $\theta^{(t)}$ is a preliminary estimate of the unknown parameter, i.e. if we set

$$Q(\theta|\theta^{(t)}) = \int l(\theta|X)P(X_m|X_o, \theta^{(t)})dX_m$$

and

$$H(\theta|\theta^{(t)}) = \int log(P(X_m|X_o, \theta))P(X_m|X_o, \theta^{(t)})dX_m$$

we then have

$$Q(\theta|\theta^{(t)}) = l(\theta|X_o) + H(\theta|\theta^{(t)}) + c.$$

The maximization step (M-step) will find the maximum $\theta^{(t+1)}$ of $Q(\theta|\theta^{(t)})$. A central result (Dempster et al., 1977) shows that $\theta^{(t+1)}$ is a better estimate than $\theta^{(t)}$ in the sense that $l(\theta^{(t+1)}|X_o) \geq l(\theta^{(t)}|X_o)$. The EM algorithm is then described as follows.

The EM algorithm *Choose a starting value θ^0 of the parameter to be estimated, then iterate the following steps until convergence up to some desired precision:*

E-step $Q(\theta|\theta^{(t)})$ *is calculated by averaging the complete-data loglikelihood* $l(\theta|X)$ *over* $P(X_m|X_o, \theta^{(t)})$*;*

M-step $\theta^{(t+1)}$ *is found by maximizing* $Q(\theta|\theta^{(t)})$*.*

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.

EM for regular exponential families EM uses the interdependence between missing data X_m and the unknown parameters θ. The E-step uses the value of $\theta^{(t)}$ to fill in somehow the missing data and the M-step uses these values to re-estimate the parameters and obtain $\theta^{(t+1)}$. If in most cases the M-step is straightforward (no computational difference from finding the MLE in the complete-data case), the E-step can be a real burden. This is not the case when the complete-data probability model falls in a regular exponential family. For these families the complete data loglikelihood may be written as

$$l(\theta|X) = \eta(\theta)^t T(X) + Ng(\theta) + c,$$

where $\eta(\theta) = (\eta_1(\theta), \eta_2(\theta), ..., \eta_k(\theta))^t$ is the canonical form of the parameter θ and $T(X) = (T_1(X), T_2(X), ..., T_k(X))^t$ 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 $l(\theta|X)$ is a linear function of the sufficient statistics, the

E-step replaces $T_j(X)$ by $E(T_j(X)|X_o, \theta^{(t)})$. In other words the E-step fills in the missing portions of the complete-data sufficient statistics. In the case of multivariate normal data, these expectations will be available in closed form and thus the E-step will also be straightforward.

With these results we are now able to adapt the algorithm to survey data.

24.1.2 EM for survey data

Assumptions on the study population In order to adapt the EM algorithm to the context of survey data, we need assumptions on the study population U. We shall assume that we have an underlying multivariate normal superpopulation model for the variable of interest, i.e.

$$X \sim N(\theta) = N(\mu, \Sigma).$$

Again X (resp. R) will denote simultaneously the random variable (resp. the response variable) of the superpopulation and the $N \times p$ matrix containing the values of the variable on the population U. If we denote by S the sampling variable, we shall assume that S and R are independent variables. If s is the sample obtained as a realization of S we shall simply use X_o^s (resp. X_m^s) to denote the observed (resp.missing) values of the survey data.

Our strategy is then termed as a *full information maximum likelihood approach (Chambers, 2001)* by opposition to a *maximum sample likelihood* approach where the EM algorithm would be run just by using the information contained in X_o^s. Our idea is very simple : every time the EM algorithm run on the whole population would need a quantity T computed from X_o we shall estimate it by $\widehat{T}$ using X_o^s.

The complete data case To establish the notational conventions of this section we shall begin by looking at the complete data case for which we won't need the EM algorithm to estimate θ. Recall that X (resp. X^s) denotes the population (resp. sample) data. An element of the matrix X (resp. X^s) will be denoted by x_{ij} with $i = 1, \ldots, N$ and $j = 1, ..., p$ (resp. x_{ij}^s with $i = 1, ..., n$ and $j = 1, ..., p$). All vectors will be expressed as column vectors, for example the ith row of X is

$$x_i = (x_{i1}, ..., x_{ip})^t.$$

We assume that x_1,...,x_N are independent realizations of the random variable X, i.e.

$$x_1, ..., x_N \sim \text{ iid } N(\theta) = N(\mu, \Sigma).$$

Discarding a proportionality constant the likelihood function is

$$L(\theta|X) \propto |\Sigma|^{-\frac{N}{2}} exp \left\{ -\frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^t \Sigma^{-1} (x_i - \mu) \right\}.$$

Expanding the exponent and taking the logarithm we can write the loglikelihood function as

$$l(\theta|X) = -\frac{N}{2} \log |\Sigma| - \frac{N}{2} \mu^t \Sigma^{-1} \mu + \mu^t \Sigma^{-1} T_1 - \frac{1}{2} tr(\Sigma^{-1} T_2)$$

where

$$T_1 = X^t 1_N = \left(\textstyle\sum_{i=1}^N x_{i1}, \ldots, \sum_{i=1}^N x_{ip}\right)^t = \sum_{i=1}^N (x_{i1}, \ldots, x_{ip})$$

and

$$\begin{aligned} T_2 = X^t X &= \begin{pmatrix} \sum_{i=1}^N x_{i1}^2 & \sum_{i=1}^N x_{i1}x_{i2} & \cdots & \sum_{i=1}^N x_{i1}x_{ip} \\ \sum_{i=1}^N x_{i2}x_{i1} & \sum_{i=1}^N x_{i2}^2 & \cdots & \sum_{i=1}^N x_{i2}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^N x_{ip}x_{i1} & \sum_{i=1}^N x_{ip}x_{i2} & \cdots & \sum_{i=1}^N x_{ip}^2 \end{pmatrix} \\ &= \sum_{i=1}^N \begin{pmatrix} x_{i1}^2 & x_{i1}x_{i2} & \cdots & x_{i1}x_{ip} \\ x_{i2}x_{i1} & x_{i2}^2 & \cdots & x_{i2}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{ip}x_{i1} & x_{ip}x_{i2} & \cdots & x_{ip}^2 \end{pmatrix} \end{aligned}$$

are the sufficient statistics. As these statistics will be needed to find the MLE for θ, we have to estimate them from survey data X^s if the population data X is not available. The Horvitz-Thompson estimates of both quantities are simply given by (recall that ω_i are the sampling weights)

$$\widehat{T_1} = \sum_{i=1}^n \omega_i (x_{i1}^s, \ldots, x_{ip}^s)$$

and

$$\widehat{T_2} = \sum_{i=1}^n \omega_i \begin{pmatrix} (x_{i1}^s)^2 & x_{i1}^s x_{i2}^s & \cdots & x_{i1}^s x_{ip}^s \\ x_{i2}^s x_{i1}^s & (x_{i2}^s)^2 & \cdots & x_{i2}^s x_{ip}^s \\ \vdots & \vdots & \ddots & \vdots \\ x_{ip}^s x_{i1}^s & x_{ip}^s x_{i2}^s & \cdots & (x_{ip}^s)^2 \end{pmatrix}$$

In the complete data case we have seen that because the multivariate normal is a regular exponential family and the loglikelihood function is linear in the elements of T_1 and T_2 we can find the MLE by equating the realized values of T_1 and T_2 to their expectations $E(T_1) = N\mu$ and $E(T_2) = N(\Sigma + \mu\mu^t)$. This leads to the well known MLE estimator of $\theta = (\mu, \sigma)$:

$$MLE(\mu) = \frac{1}{N} T_1$$

and

$$MLE(\Sigma) = \frac{1}{N} T_2 - MLE(\mu) MLE(\mu)^t$$

If N is known (i.e. $\Sigma_{i=1}^n \omega_i = N$) we estimate these quantities by the classical Horvitz-Thompson estimates

$$\widehat{MLE(\mu)} = \frac{1}{N} \widehat{T_1}$$

and

$$\widehat{MLE(\Sigma)} = \frac{1}{N}\widehat{T}_2 - \widehat{MLE(\mu)}\widehat{MLE(\mu)}^t.$$

If N is not known the Hájek estimator is used estimating N by $\Sigma_{i=1}^n \omega_i$.

The incomplete data case - The EM algorithm We shall proceed in the same way to adapt the EM algorithm to the survey data. We shall analyze the EM algorithm for X and at each step where it is needed we shall use estimates based on X^s. The presentation of the EM algorithm given in (Schafer, 2000) is used here. We shall first give some matrix tools that will simplify the description of the algorithm.

The sweep operator If a multivariate normal random z vector distributed as $N(\mu, \Sigma)$ is partitioned in two parts $z^t = (z_1^t, z_2^t)$ then the z_i's are distributed as $N(\mu_i, \Sigma_{ii})$ with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

It is well knows that the conditional distribution of $z_2|z_1$ is normal with moments $\mu_{2\cdot 1} = \alpha_{2\cdot 1} + B_{2\cdot 1}z_1$ and covariance matrix $\Sigma_{2\cdot 1}$ where

$$\begin{aligned} \alpha_{2\cdot 1} &= \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 \\ B_{2\cdot 1} &= \Sigma_{21}\Sigma_{11}^{-1} \\ \Sigma_{2\cdot 1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned} \tag{4}$$

Now specifying the distribution of z (parametrized by μ, Σ) is the same as specifying the distribution of z_1 (parametrized by μ_1, Σ_1) and the conditional distribution of $z_2|z_1$ (parametrized by $\mu_{2\cdot 1} = \alpha_{2\cdot 1} + B_{2\cdot 1}z_1, \Sigma_{2\cdot 1}$). The transformation from the first parameters to the second ones is therefore one-to-one with inverse given by

$$\begin{aligned} \mu_2 &= \alpha_{2\cdot 1} + B_{2\cdot 1}\mu_1 \\ \Sigma_{12} &= \Sigma_{11}B_{2\cdot 1}^t \\ \Sigma_{22} &= \Sigma_{2\cdot 1} + B_{2\cdot 1}\Sigma_{11}B_{2\cdot 1}^t \end{aligned} \tag{5}$$

Both transformations will play a crucial role in the realization of the EM algorithm and the essential tool to implement it in an easy way is the sweep operator. This device was first introduced by (Beaton, 1964) and is commonly used in linear model computations and stepwise regression.

Definition 4 *Let G be a $p \times p$ symmetric matrix with elements g_{ij}, the sweep operator $SWP[k]$ (for $1 \leq k \leq p$) replaces G by another $p \times p$ symmetric $H = SWP[k]G$ matrix with elements given by*

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} &= h_{kj} = g_{jk}/g_{kk} \text{ for } j \neq k \\ h_{jl} &= h_{lj} = g_{jl} - g_{jk}g_{kl}/g_{kk} \text{ for } j \neq k \text{ and } l \neq k \end{aligned}$$

After the application of the operator $SWP[k]$, the matrix is said to have been swept on position k.

It is convenient to define a reverse-sweep operator that returns a swept matrix to its original form.

Definition 5 *Let H be a $p \times p$ symmetric matrix with elements h_{ij}, the reverse-sweep operator $RSW[k]$ (for $1 \leq k \leq p$) replaces H by another $p \times p$ symmetric $G = RSW[k]H$ matrix with elements given by*

$$\begin{aligned} g_{kk} &= -1/h_{kk} \\ g_{jk} &= g_{kj} = -h_{jk}/h_{kk} \text{ for } j \neq k \\ g_{jl} &= g_{lj} = h_{jl} - h_{jk}h_{kl}/h_{kk} \text{ for } j \neq k \text{ and } l \neq k \end{aligned}$$

By definition we have therefore

$$RSW[k]SWP[k]G = G = SWP[k]RSW[k]G.$$

Both operators are commutative, i.e.

$$\begin{aligned} SWP[k_1]SWP[k_2] &= SWP[k_2]SWP[k_1], \\ RSW[k_1]RSW[k_2] &= RSW[k_2]RSW[k_1]. \end{aligned}$$

Thus we can extend the notations to

$$\begin{aligned} SWP[k_1]SWP[k_2]\cdots SWP[k_l] &= SWP[k_1, k_2, \ldots, k_l], \\ RSW[k_1]RSW[k_2]\cdots RSW[k_l] &= RSW[k_1, k_2, \ldots, k_l]. \end{aligned}$$

Among several properties of these operators let us quote the following. If G is partitioned as

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$$

with G_{11} a $p_1 \times p_1$ matrix then the swept matrix on the first p_1 position is given by

$$SWP[1, 2, \ldots, p_1]G = \begin{pmatrix} -G_{11}^{-1} & G_{11}^{-1}G_{12} \\ G_{21}G_{11}^{-1} & G_{22} - G_{21}G_{11}^{-1}G_{12} \end{pmatrix}.$$

In particular we have $SWP[1, \ldots, p]G = -G^{-1}$. Moreover the determinant is obtained through the process of sweeping on all positions by $|G| = \prod_{k=1}^{p} \gamma_k$ with $\gamma_k = (SWP[1, \ldots, k-1]G)_{kk}$.

Both transformations 4 and 5 can be expressed very easily in a matrix form using the sweep and reverse sweep operators. With the above notations let us write the parameter θ as a $(p+1) \times (p+1)$ matrix

$$\theta = \begin{pmatrix} -1 & \mu^t \\ \mu & \Sigma \end{pmatrix} = \begin{pmatrix} -1 & \mu_1^t & \mu_2^t \\ \mu_1 & \Sigma_{11} & \Sigma_{12} \\ \mu_2 & \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

The reason for placing -1 in the upper-left corner is given at the end of the section. To keep unchanged the indices of Σ we shall number the lines and column of this matrix from 0 to p. Using the above properties we sweep θ on positions $1, \ldots, p_1$ and we obtain the following matrix

$$\begin{aligned} SWP[1,\ldots,p_1]\theta &= \begin{pmatrix} -1-\mu_1^t\Sigma_{11}^{-1}\mu_1 & \mu_1^t\Sigma_{11}^{-1} & \mu_2^t-\mu_1^t\Sigma_{11}^{-1}\Sigma_{12} \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} & \Sigma_{11}^{-1}\Sigma_{12} \\ \mu_2-\Sigma_{21}\Sigma_{11}^{-1}\mu_1 & \Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix} \\ &= \begin{pmatrix} -1-\mu_1^t\Sigma_{11}^{-1}\mu_1 & \mu_1^t\Sigma_{11}^{-1} & \alpha_{2\cdot 1}^t \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} & B_{2\cdot 1}^t \\ \alpha_{2\cdot 1} & B_{2\cdot 1} & \Sigma_{2\cdot 1} \end{pmatrix}. \end{aligned}$$

Now as we also have

$$RSW[1,\ldots,p_1]\begin{pmatrix} -1-\mu_1^t\Sigma_{11}^{-1}\mu_1 & \mu_1^t\Sigma_{11}^{-1} \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} \end{pmatrix} = \begin{pmatrix} -1 & \mu_1^t \\ \mu_1 & \Sigma_{11} \end{pmatrix},$$

we see that just by sweeping θ on position $1, 2, \ldots, p_1$ and then by reverse sweeping the upper-left $(p_1+1)\times(p_1+1)$ submatrix on the same position we obtain the matrix

$$\phi = \begin{pmatrix} -1 & \mu_1^t & \alpha_{2\cdot 1}^t \\ \mu_1 & \Sigma_{11} & B_{2\cdot 1}^t \\ \alpha_{2\cdot 1} & B_{2\cdot 1} & \Sigma_{2\cdot 1} \end{pmatrix}.$$

We have then realized the transformation 4 from θ to ϕ with the sweep and reverse-sweep operators.

The reason for placing the -1 results from the following relation

$$RSW[0]\theta = RSW[0]\begin{pmatrix} -1 & \mu^t \\ \mu & \Sigma \end{pmatrix} = \begin{pmatrix} 1 & \mu^t \\ \mu & \Sigma+\mu\mu^t \end{pmatrix}.$$

The last matrix contains the natural representation of the MLE, i.e using the notation developed in the complete data case we have

$$\begin{pmatrix} 1 & MLE(\mu^t) \\ MLE(\mu) & MLE(\Sigma)+MLE(\mu)MLE(\mu)^t \end{pmatrix} = \frac{1}{N}\begin{pmatrix} N & T_1^t \\ T_1 & T_2 \end{pmatrix} = \frac{T}{N}$$

with $T = \begin{pmatrix} N & T_1^t \\ T_1 & T_2 \end{pmatrix}$ being the matrix form of the sufficient statistics. In the case of multivariate normal data we have thus showed that the MLE can be computed from the sufficient statistics using the sweep operator

$$MLE(\theta) = SWP[0]\left(\frac{T}{N}\right).$$

The EM-algorithm for survey data Recall that X is the $N \times p$ matrix of the population data and X^s the $n \times p$ matrix of the survey data. We shall number by $a = 1, \ldots, A$ the missingness patterns appearing among the rows of X. A pattern for a row x_i of X can be represented as a p-vector of 0's and 1's with 0 values corresponding to missing items and 1 values to observed items. For example if $x_i = (23, NA, 2, 7, NA, NA, 12, 8)^t$ its missingness pattern is described as $mis(x_i) = (1, 0, 1, 1, 0, 0, 1, 1)^t$. The number of different possible missingness patterns A is bounded by $2^p - 1$ (the trivial pattern with all values set to 0's will never be used because the completely missing rows of X contribute to nothing to the observed-data likelihood and should be removed from the data). The $A \times p$ matrix M will be the matrix having as rows the missingness patterns m_a with $a = 1, \ldots, A$. Let m_a be one of these missingness patterns we shall need the following notations

$$\begin{aligned} I(a) &= \{i : mis(x_i) = m_a\} = \{\text{row labels of } X \text{ having } m_a \text{ as missingness pattern}\} \\ O(a) &= \{j : m_{aj} = 1\} = \{\text{column labels of pattern } a \text{ with observed items}\} \\ M(a) &= \{j : m_{aj} = 0\} = \{\text{column labels of pattern } a \text{ with missing items}\} \end{aligned}$$

For the pattern a given above as an example we would have $O(a) = \{1, 3, 4, 7, 8\}$ and $M(a) = \{2, 5, 6\}$.

The E-step With a model of the regular exponential family we have seen that the E-step just replaces the sufficient statistics by their expectation over $P(X_m | X_o, \theta)$ for an assumed value of θ. As theses statistics are linear combinations of x_{ij} and $x_{ij} x_{ik}$ the crucial point is to find their expectations.

As the rows x_i are independent for a given θ we have

$$P(X_m | X_o, \theta) = \prod_{i=1}^{N} P(x_{i(mis)} | x_{i(obs)}, \theta)$$

where $x_{i(obs)}$ (resp. $x_{i(mis)}$ denote the observed (resp. missing) subvector of x_i. Now in the case where $P(x_i | \theta)$ is a multivariate normal distribution we have seen that the moments of $P(x_{i(mis)} | x_{i(obs)}, \theta)$ can be obtained using the sweep operator. More precisely for a given pattern s if $i \in I(s)$, $j, k \in M(s)$ and if we set

$$C = SWP[O(s)]\theta$$

with θ the parameters matrix seen above we then have

$$E(x_{ij} | X_o, \theta) = E(x_{ij} | x_{i(obs)}, \theta) = c_{0j} + \sum_{k \in O(s)} c_{kj} x_{ik}$$

and

$$Cov(x_{ij}, x_{ik} | X_o, \theta) = Cov(x_{ij}, x_{ik} | x_{i(obs)}, \theta) = c_{jk}.$$

If $j \in O(s)$, x_{ij} is fixed and we have trivially that

$$E(x_{ij} | X_o, \theta) = E(x_{ij} | x_{i(obs)}, \theta) = x_{ij}$$

and

$$Cov(x_{ij}, x_{ik}|X_o, \theta) = Cov(x_{ij}, x_{ik}|x_{i(obs)}, \theta) = 0.$$

Using $E(xy) = E(x)E(y) + Cov(x, y)$ we obtain the final general expressions for $i \in I(s)$

$$E(x_{ij}|X_o, \theta) = E(x_{ij}|x_{i(obs)}, \theta) = \begin{cases} x_{ij} & \text{for } j \in O(s) \\ x_{ij}^* & \text{for } j \in M(s) \end{cases}$$

and

$$\begin{aligned} E(x_{ij}x_{ik}|X_o, \theta) &= E(x_{ij}x_{ik}|x_{i(obs)}, \theta) \\ &= \begin{cases} x_{ij}x_{ik} & \text{for } j, k \in O(s) \\ x_{ij}^*x_{ik} & \text{for } j \in M(s), k \in O(s) \\ c_{jk} + x_{ij}^*x_{ik}^* & \text{for } j, k \in M(s) \end{cases} \end{aligned}$$

where

$$x_{ij}^* = c_{0j} + \sum_{k \in O(s)} c_{kj}x_{ik}$$

Remark: We emphasize here the fact that in both equations the independence of the observations x_i implies the first equality and in consequence the fact that these moments can be calculated from one x_i without any knowledge of the other ones. This means that these relations are the same for the x_{ij}^s's:

$$E(x_{ij}^s|X_o, \theta) = E(x_{ij}^s|x_{i(obs)}^s, \theta) = \begin{cases} x_{ij}^s & \text{for } j \in O(s) \\ {x_{ij}^s}^* & \text{for } j \in M(s) \end{cases}$$

and

$$\begin{aligned} E(x_{ij}^s x_{ik}^s|X_o, \theta) &= E(x_{ij}^s x_{ik}^s|x_{i(obs)}^s, \theta) \\ &= \begin{cases} x_{ij}^s x_{ik}^s & \text{for } j, k \in O(s) \\ {x_{ij}^s}^* x_{ik}^s & \text{for } j \in M(s), k \in O(s) \\ c_{jk} + {x_{ij}^s}^* {x_{ik}^s}^* & \text{for } j, k \in M(s) \end{cases} \end{aligned}$$

where

$${x_{ij}^s}^* = c_{0j} + \sum_{k \in O(s)} c_{kj}x_{ik}^s$$

We are now in a position to write the E-step in a matrix form (to shorten the expression we shall write $E(\cdots|X_o,\theta) = \overline{\cdots|}$); for the population data:

$$E(T|X_o,\theta) = E\left(\begin{pmatrix} N & T_1^t \\ T_1 & T_2 \end{pmatrix} |X_o,\theta\right)$$

$$= \sum_{i=1}^{N} \begin{pmatrix} 1 & \overline{x_{i1}|} & \overline{x_{i2}|} & \cdots & \overline{x_{1p}|} \\ \overline{x_{i1}|} & \overline{x_{i1}^2|} & \overline{x_{i1}x_{i2}|} & \cdots & \overline{x_{i1}x_{ip}|} \\ \overline{x_{i2}|} & \overline{x_{i2}x_{i1}|} & \overline{x_{i2}^2|} & \cdots| & \overline{x_{i2}x_{ip}|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \overline{x_{ip}|} & \overline{x_{ip}x_{i1}|} & \overline{x_{ip}x_{i2}|} & \cdots & \overline{x_{ip}^2|} \end{pmatrix}$$

By the remark above we know that all coefficients $\overline{\cdots|}$ can be computed the same way for the population and the survey data therefore we can use the Horvitz-Thompson estimator to write the "estimated E-step" for the survey data:

$$\widehat{E}(T|X_o,\theta) = \sum_{i=1}^{n} \omega_i \begin{pmatrix} 1 & \overline{x_{i1}^s|} & \overline{x_{i2}^s|} & \cdots & \overline{x_{1p}^s|} \\ \overline{x_{i1}^s|} & \overline{(x_{i1}^s)^2|} & \overline{x_{i1}^s x_{i2}^s|} & \cdots & \overline{x_{i1}^s x_{ip}^s|} \\ \overline{x_{i2}^s|} & \overline{x_{i2}^s x_{i1}^s|} & \overline{(x_{i2}^s)^2|} & \cdots| & \overline{x_{i2}^s x_{ip}^s|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \overline{x_{ip}^s|} & \overline{x_{ip}^s x_{i1}^s|} & \overline{x_{ip}^s x_{i2}^s|} & \cdots & \overline{(x_{ip}^s)^2|} \end{pmatrix}$$

The M-step The M-step is relatively trivial in the multivariate normal case. We have shown that for a given sufficient statistics matrix T the MLE is simply obtained by $MLE(\theta) = SWP[0]N^{-1}T$. A M-step is therefore nothing else than

$$\theta(k+1) = SWP[0]N^{-1}E(T|X_o,\theta^k) \text{ for the population data}$$

and

$$\theta(k+1) = \begin{cases} SWP[0]N^{-1}\widehat{E}(T|X_o,\theta^k) & \text{if } N \text{ is known} \\ SWP[0]\left(\sum_{i=1}^n \omega_i\right)^{-1}\widehat{E}(T|X_o,\theta^k) & \text{if } N \text{ is unknown} \end{cases}$$

for the survey data.

24.2 The "BACON-EM for survey data" algorithm

Merging both algorithms is relatively straightforward if computation time is not an issue. Each time estimations of the mean and the covariance matrix are needed, the EM algorithm described above is run up to some pre-fixed convergence criteria. Such an approach is clearly too naiv when evaluating the computation time. Firstly the "growing" structure of the BACON algorithm would not be used to avoid extra-computations of EM at each step , secondly a restrictive convergence criteria of EM could slow down much the algorithm only to make improvements of the estimation at each step when they are probably not needed (the crucial point at each step is that the estimations of the mean and the covariance matrix allow the algorithm to exclude outlying points from the good subset and this does not need these estimations to be extremely close to the real values).

The ”BACON-EM for survey data” algorithm is desribed at the end of this subsection. Our approach towards the two issues quoted above is the following. According to our experience of the BACON algorithm we decided tu re-use as much information as we could from one step to the next one. In fact estimation of the sufficient statistics T^G computed on some good subset G (it is actually simply the restriction of the $T_j = \sum_{i=1}^{N} h(x_i j)$ to the elements in G, i.e. $T_j^G = \sum_{i \in G} h(x_i j)$) usually has a part T_o^G with points having no missing values estimated by $\widehat{T}_o^G$ that can be computed straightforward and a problematic part T_m^G with points having missing values estimated by $\widehat{T}_m^G$ that can not be computed. The expectation computed by the E-step can therefore be written as

$$\widehat{E}(T^G | X_o^G, \theta) = \widehat{T}_o^G + \widehat{E}(T_m^G | X_o^G, \theta).$$

As the subsets G are growing, we do not compute $\widehat{T}_o^G$at each step of the BACON loop, but we keep a global variable for $\widehat{T}_o^G$ that is simply updated each time G changes (adding points, and sometimes removing a few to the statistic). Concerning the convergence criteria selection, we choose to fix the number of iteration of EM at each step of the Bacon loop, by default this number is set to 5 but the user is allowed to change it. At the end of the Bacon algorithm EM is run once more but this time with more iterations (by default 10) and this is also the case for the initial subset selection if the user chooses Version 1 of BACON.

24.2.1 The algorithm

```
- Default constants
    $\alpha = 0.95$
    $c = 3$
    $it.em.1 = 10$
    $it.em.2 = 5$
- Starting point
    Version 1
      - Compute $\widehat{M}$ and $\widehat{S}$ using EM with $it.em.1$ iterations on $X$;
      - Compute the $n$ (Mahalanobis) distances $MD_{\widehat{M},\widehat{S}}(x_i)$ (see 23.1);
    Version 2
      - Compute the coordinatewise median $med$ ignoring in each variable
        the missing values;
      - Compute the $n$ distances $||x_i - med||$ based on the observed components
        of $x_i$ and corrected by a factor as in 23.1;
- Select the $m = cp$ smallest distances to form the first good subset $G$;
- Compute $\widehat{M}_G$ and $\widehat{S}_G$ using EM with $it.em.2$ iterations on $G$, and stock $\widehat{T}_o^G$;
- If $\widehat{S}_G$ is singular, exit and ask the user to increase $c$;
- Main loop
    - Compute the $n$ (Mahalanobis) distances $MD_{\widehat{M}_G,\widehat{S}_G}(x_i)$ (see 23.1);
    - Set a new subset $NG$ to all points with Mahalanobis distances smaller
      than $(\widehat{c}_{npr}\chi_{p,\alpha})^2$;
    - If $NG = G$ then exit the loop;
    - Upgrade $\widehat{T}_o^G$ to $\widehat{T}_o^{NG}$;
    - Reset $G = NG$;
    - Compute $\widehat{M}_G$ and $\widehat{S}_G$ using EM (with $\widehat{T}_o^G$ already computed)
      with $it.em.2$ iterations on $G$;
    - If $\widehat{S}_G$ is singular, exit and ask the user to increase $\alpha$;
    - Restart the loop;
- If a better estimation is seeeked $it.em.1$ more iterations of EM on $G$ are
  run with starting parameters $\widehat{M}_G$ and $\widehat{S}_G$;
- Nominate the observations excluded by the final $G$ as outliers.
```

25 EA

Once all the distances (i.e. the infection probabilities) are available, the EA algorithm works regardless of the underlying data values. Therefore only the distances computation has to be adapted to the absence of some values. We shall assume here that the data are MCAR; a more careful study should be carried on to see if this hypothesis can be weakened or not. The adaptation here is done similarly as in 23.1 simply by computing the distance between two points on the common observed variables and inflating it by a factor inversely proportional to the proportion of observed values, if no observed variables are in common the distance is set to infinity. The standardization of each variable is done using only the observed values. Recall that R is the response variable, i.e. $r_{ik} = 1$ if variable k is observed for observation i and $r_{ik} = 0$ if not, then the distance between observations x_i and x_j is given by

$$\tilde{d}_{ij} = \begin{cases} \left(\frac{p}{\sum_{k=1}^{p} r_{ik} r_{jk}} \sum_{k=1}^{p} q_k r_{ik} r_{jk} (x_{ik} - x_{jk})^2 \right)^{1/2}, & \text{if } \sum_{k=1}^{p} r_{ik} r_{jk} \neq 0 \\ \infty, & \text{if not} \end{cases}$$

where

$$q_k = (mad_{k, r_{ik}=1} x_{ik})^{-2}.$$

When $\tilde{d}_{ij}$ is set to infinity the infection probability is forced to be zero forbidding a possible infection between both points. This is actually what we want as we don't have any information on the distance between the two points. Why should we standardise with $\sum_k r_{ik} r_{jk}$? The point is that if an observation is an outlier in some dimensions but has missing values in many other dimensions, then it could be masked without the standardisation. Why should we divide by $\sum_k r_{ik} r_{jk}$ and not by $\sum_k q_k r_{ik} r_{jk}$? The second solution would imply that the distance is a weighted mean of the contributions of the dimensions, the weight being q_k. In the extreme case of an observation with only one observed variable the distance with this observation would correspond to an unstandardized distance in the observed dimension. This is undesirable because then outliers in dimensions with a small dispersion may remain undetected.

Part VI

Robust nearest neighbor imputation

26 Introduction

In this section we describe an algorithm which can impute values for detected outliers and for missing values. Furthermore edit rules and sampling weights should be taken into account. The algorithm should be a module in a system of modules which contains also an edit stage controlling edit rules, an outlier detection stage, and a preliminary stage of imputation which imputes deterministically if possible (e.g. in the case of balance edits). The module should be nearly automatic. Thus we do not want to use any modelling of missing values. This is a serious drawback in many instances. The only device we want to use are distances and therefore the imputation is based on nearest neighbor methods. The Fellegi-Holt principle of minimum change is embedded in the nearest neighbor distance. We use the Mahalanobis distance and assume therefore that the bulk of the data is approximately elliptical. The second method we planned to implement was a backward epidemic algorithm. However, due to lack of ressources, this was not possible.

27 Input

The input to the imputation module is the data, a vector of flags on whether the observation is an outlier, a matrix of the same dimension as the observation which indicates edit failures, and a vector of sampling weights. More formally the inputs are:

1. A $n \times p$ matrix X of observations. In the first place we assume the variables continuous but in principle also categorical variables could be treated. Together with X we get or may calculate a $n \times p$ matrix R of indicators of response with

$$r_{ij} = \begin{cases} 1 & x_{ij} \text{ is given,} \\ 0 & x_{ij} \text{ is missing.} \end{cases}$$

2. A $n \times 1$ vector w of sampling weights.

3. A $n \times 1$ vector u of outlier flags which have been set in a previous outlier detection phase. Instead of the outlier flags u may contain a measure of outlyingness like robustness weights. For the moment we assume that

$$u_i = \begin{cases} 0 & \text{observation } i \text{ is declared an outlier,} \\ 1 & \text{otherwise.} \end{cases}$$

4. A $n \times p$ matrix E of flags (see Section 4.2.1). We assume that any error localisation has been done beforehand. Thus the flags mean

$$e_{ij} = \begin{cases} 0 & x_{ij} \text{ fails one or more edits and is deemed in error}, \\ 1 & x_{ij} \text{ passes all edits}. \end{cases}$$

   If a value is missing, i.e. if $r_{ij} = 0$ then $e_{ij} = 1$. In fact edit rules which involve a missing value usually cannot be applied to an observation.

In the EUREDIT data sets Y_2 which contain only missing values but no errors we have $e_{ij} = 1$ for all i and j but for the data sets Y_3 we have some $e_{ij} = 0$.

The objectives of the imputation module are:

1. Impute x_{ij} if $r_{ij} = 0$, i.e. impute missing values.
2. Impute x_i if $u_i = 0$, i.e. impute outlying observations.

Optionally we may set $r_{ij} = 0$ if $e_{ij} = 0$ beforehand, i.e. we may want to impute a new value whenever a given value failed any edit. A problem with this option is that if no efficient error localisation has been done beforehand it may be very inefficient because too many values are imputed.

28 The imputation module POEM

The idea is to use a weighted Mahalanobis distance which is adjusted for missing values and for edit failures. We call the algorithm POEM for weighted imPutation for Outliers, Edit failures and Missing values.

28.1 Center and standardization

First we calculate the mean of good observations for each variable j:

$$\mu_j = \frac{\sum_i u_i w_i r_{ij} \alpha^{(1-e_{ij})} x_{ij}}{\sum_i u_i w_i r_{ij} \alpha^{(1-e_{ij})}}. \tag{6}$$

Here α is a reduction factor between 0 and 1. Thus if a value failed edits then its weight in the mean is reduced by a factor $\alpha^{(1-e_{ij})} = \alpha$ while there is no weight reduction for $e_{ij} = 1$. Of course this factor is useless if we have set $r_{ij} = 0$ if $e_{ij} = 0$ beforehand. Reasonable values for α are 0, i.e. we treat failures as missings, or 1, i.e. we ignore the matrix E. A factor $\alpha = 0.5$ might represent our relative confidence in the failing items. Missing values are left out by the sums due to $r_{ij} = 0$ and outliers are left out or downweighted due to u_i. Thus we get a robust mean which takes into account as much reliable values as possible.

We will come across the factor $r_{ij}\alpha^{(1-e_{ij})}$ several times and we call it α_{ij} to shorten the notation. Thus the mean of good observations becomes

$$\mu_j = \frac{\sum_i u_i w_i \alpha_{ij} x_{ij}}{\sum_i u_i w_i \alpha_{ij}}. \tag{7}$$

Instead of taking a different weight for each variable we may join the reduction factors of an observation to

$$\tilde{\alpha}_i = \prod_j \alpha^{(1-e_{ij})}. \tag{8}$$

Then we get another estimator of the mean of good observations:

$$\tilde{\mu}_j = \frac{\sum_i u_i w_i r_{ij} \tilde{\alpha}_i x_{ij}}{\sum_i u_i w_i r_{ij} \tilde{\alpha}_i}. \tag{9}$$

In what follows we stick to the first definition of a mean (6).

The different dimensions (variables) should have the same order of magnitude in the distance. This is particularly important because of possible missing values. We calculate the variance of the good observations for each variable:

$$\sigma_j^2 = \frac{\sum_i u_i w_i \alpha_{ij} (x_{ij} - \mu_j)^2}{\sum_i u_i w_i \alpha_{ij}}. \tag{10}$$

Then we standardize the observations:

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}. \tag{11}$$

From now on we work with the standardized observations only.

28.2 Covariance Matrix

The second step is to estimate a variance-covariance matrix of the good observations. To avoid computational problems we set $\tilde{x}_{ij} = 0$ if $r_{ij} = 0$, i.e. we replace missing values by 0 (the mean of the standardized observations).

The terms of the variance-covariance matrix of good observations is calculated as

$$\begin{aligned} (D)_{jk} &= \frac{\sum_i u_i w_i r_{ij} r_{ik} \alpha^{(1-e_{ij})} \alpha^{(1-e_{ik})} \tilde{x}_{ij} \tilde{x}_{ik}}{\sum_i u_i w_i r_{ij} r_{ik} \alpha^{(1-e_{ij})} \alpha^{(1-e_{ik})}} \\ &= \frac{\sum_i u_i w_i \alpha_{ij} \alpha_{ik} \tilde{x}_{ij} \tilde{x}_{ik}}{\sum_i u_i w_i \alpha_{ij} \alpha_{ik}} \end{aligned} \tag{12}$$

Note that this is a slightly different formula as the one used for outlier detection since now we take into account missing values much more simply and we add a downweighting for edit failures.

The covariance matrix for standardized observations $D(\tilde{X})$ is the correlation matrix of the unstandardized observations.

This matrix D may lack positive-definiteness in particular if many values are missing, outlying or failing. In that case we cannot proceed further without analysis of the situation.

The standardization in the numerator of D_{jk} could be even more sophisticated, taking into account the effective degrees of freedom.

28.3 Redefinition of outliers

The observations that have been declared outliers by u_i may be representative. We would like to have a way of relaxing the outlier conditions in order to avoid imputation for representative outliers (or simply for too many outliers). This is necessary for very skew data where rejecting outliers may lead to a large bias.

We calculate the Mahalanobis distance of each observation.

$$d^2 = p^2 \frac{\sum_{j,k} \alpha_{ij} \alpha_{ik} \tilde{x}_{ij} D^{-1}_{jk} \tilde{x}_{ik}}{\sum_{j,k} \alpha_{ij} \alpha_{ik}}. \tag{13}$$

Note that we have included the downweighting for failing items.

Now we may define a second outlier indicator or robustness weight

$$\tilde{u}_i = \begin{cases} 1 & d \leq c, \\ 0 & \text{otherwise}, \end{cases} \tag{14}$$

where c is a tuning constant to be chosen. It is clear that we may choose to use a smooth downweighting of outliers with $u_i = c/d$ for $d > c$ like for an Huber M-estimator.

The total robustness weight is $\sum_i w_i \tilde{u}_i$. The total robustness weight is less than the population size N if the weights w_i are calibrated accordingly. Usually we want that $\sum_i w_i \tilde{u}_i \geq \sum_i w_i u_i$ because of the relaxation of outlyingness. Looking at the total robustness weight may help in choosing c.

28.4 Conditions for donors

Now let i be an observation which has to get imputed values and h a possible donor. We impose the following conditions on the donor:

1. The donor should not be an outlier, i.e. $u_h = 1$. Note that we use the original u_h because we would not want to impute representative outliers.

2. The link between i and h must be sufficiently strong, i.e.

$$\sum_j r_{ij} r_{hj} \alpha^{(1-e_{ij})} \alpha^{(1-e_{hj})} = \sum_j \alpha_{ij} \alpha_{hj} \tag{15}$$

   should be sufficiently large.

3. Donors for outliers must be complete with no failing items and donors for observations with missing or failing items must have enough items, i.e. if $u_i = 1$ the condition is

$$\sum_j (1 - r_{ij})(1 - e_{ij}) r_{hj} e_{hj} = \sum_j (1 - r_{ij})(1 - e_{ij}) \tag{16}$$

   and if $u_i < 1$ the condition is

$$\sum_j r_{hj} e_{hj} = p. \tag{17}$$

We combine the first two criteria into one:

$$\sum_j u_h \alpha_{ij} \alpha_{hj} \geq \beta p, \tag{18}$$

where $0 < \beta \leq 1$ is a parameter to determine the severity of the donor condition. An alternativ would be to use only complete non-outlying observations as donors.

The set of donors H_i may be empty. Then we have to refrain from imputation or relax the donor condition.

28.5 Nearest neighbor

The (squared) distance between an imputand, i.e. the observation to impute, and a donor is

$$d(\tilde{x}_i, \tilde{x}_h)^2 = p^2 \frac{\sum_{j,k} \alpha_{ij} \alpha_{hj} \alpha_{ik} \alpha_{hk} (\tilde{x}_{ij} - \tilde{x}_{hj}) D^{-1}_{jk} (\tilde{x}_{ik} - \tilde{x}_{hk})}{\sum_{j,k} \alpha_{ij} \alpha_{hj} \alpha_{ik} \alpha_{hk}}. \tag{19}$$

Note that it was important to standardize the data beforehand because if different variables are missing for different donors we account for the number of missing variables but not for the variability of the different variables. We calculate the distance $d(\tilde{x}_i, \tilde{x}_h)$ for all h in H_i. Then we choose the donor with minimal distance, i.e.

$$h(i) = \arg\min_{h \in H_i} d(\tilde{x}_i, \tilde{x}_h). \tag{20}$$

Then $h(i)$ is the nearest neighbor of i. Instead of this deterministic version we may determine a small number of nearest neighbors and choose randomly, with probability proportional to the distance, one of them as donor for i.

28.6 Imputation

For non-outliers ($u_i = 1$) impute $x_{ij} = x_{h(i)j}$ for all j with $r_{ij}e_{ij} = 0$, i.e. for all variables with missing or failing items. We may, of course, impute only for missing values. For outliers ($\tilde{u}_i < 1$) impute $x_{ij} = x_{h(i)j}$ for all j. Note that we impute only for the outliers according to the possibly relaxed definition $\tilde{u}_i$. If we accept only complete cases and non-failing observations as donors then we may impute all values always. This results in a loss of information which goes contrary to the Fellegi-Holt principle. However it is the simplest way to ensure that the data does not fail any edits after imputation.

29 Controlling the imputation

We have seen that for the imputation we have to choose several tuning constants:

1. The tuning constant c for the redefinition of outlyingness.
2. The tuning constant α for the downweighting of failing items in the distance.
3. The tuning constant β for the condition on the link to a donor.
4. If we choose random nearest neighbor imputation we have to choose the constant of admissible neighbors.

After imputation we cannot be sure that the imputed data passes the edits. We will have to run the edits again, which results in new values e'_{ij} for the failure indicators and check whether we have been more or less successful. We may also compare the original e_{ij} with the new e'_{ij}. In principle there might be still some missing values left in the imputed data $\tilde{X}$ if no donor could be found. We therefore will have to compute r'_{ij} with an E module to check for missingness.

We need information on

1. The number of remaining missing values per variable $\sum_i r'_{ij}$.
2. The number of good values per variable $\sum_i w_i \alpha_{ij}$.
3. Mean μ_j and variance σ_j^2.
4. The covariance matrix D.
5. The number of outliers $\sum_i 1\{u_i < 1\}$ and $\sum_i 1\{\tilde{u}_i < 1\}$.
6. The total robustness weights $\sum_i w_i u_i$ and $\sum_i w_i \tilde{u}_i$.
7. The number of empty donor sets $\sum_i 1\{|H_i| = 0\}$
8. The maximal number of times a donor is used.

This information is needed to judge the performance of the imputation. To obtain some of the informations we need to run a E module on the output.

References

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

Bay, S. D. (1999). The UCI KDD archive [http://kdd.ics.uci.edu].

Beaton, A. (1964). The use of special matrix operations in statistical calculus. Research Bulletin RB-64-51, Educational Testing Service, Princeton, NJ.

Béguin, C. (2001). Outlier detection in multivariate data. Master's thesis, Université de Neuchâtel. Preprint.

Béguin, C. and Hulliger, B. (2000). Develop and evaluate new methods for statistical outlier detection and outlier robust multivariate imputation. Workplan for EUREDIT workpackage x.2, EUREDIT.

Billor, N., Hadi, A. S., and Velleman, P. F. (2000). BACON: Blocked Adaptative Computationally-efficient Outlier Nominators. To be published in CS DA.

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, 75(4):761–771.

Breckling, J., Kokic, P., and O., L. (2000). A new definition of multivariate M-quantiles based on a generalisation of the univariate estimating equations. Working paper 1, Insiders Financial Technology, Mainz.

Breckling, J., Kokic, P., and O., L. (2001). A semi-parametric approach to multivariate expectiles for outlier detection. Working paper 2, Insiders Financial Technology, Mainz.

Brown, B. M. (1983). Statistical uses of the spatial median. *Journal of the Royal Statistical Society, Series B, Methodological*, 45:25–30.

Brown, B. M. (1988). Spatial median. In *Encyclopedia of Statistical Sciences (9 vols. plus Supplement)*, volume 8, pages 574–575. Wiley (New York).

Brown, B. M. and Hettmansperger, T. P. (1989). An affine invariant bivariant version of the sign test. *Journal of the Royal Statistical Society, Series B, Methodological*, 51:117–125.

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.

Chambers, R. (2001). Likelihood-based methods with complex survey data. To appear in the Proceedings of the Conference on Analysis of Complex Survey Data, Southampton, August 1999.

Cheng, T.-C. and Victoria-Feser, M.-P. (2000). Robust correlation with missing data.

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71:161–190.

Croux, C. and Ruiz-Gazen, A. (2000). High breakdown estimators for principal components: the projection-pursuit approach revisited. Preprint.

Davies, P. (1987). Asymptotic behavior of s-estimates of multivariate parameters and dispersion matrices. *The Annals of Statistics*, 15:1269–1292.

De Waal, T. (2000). New developments in automatic edit and imputation at Statistics Netherlands. In *Proceedings of the Conference of European Statisticians, Cardiff, United-Kingdom, 18-20 October 2000*.

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.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes: linéarisation et techniques des résidus. *Techniques d'enquête, Statistique Canada*, 25(2):?–?

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

Fellegi, I. and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71(353):17–35.

Franklin, S., Thomas, S., and Brodeur, M. (2000). Robust multivariate outlier detection using Mahalanobis' distance and a modified Stahel-Donoho estimator. Technical report, Statistics Canada.

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

Gower, J. C. (1974). [algorithm As 78] The mediancentre (corr: 75v24 p390). *Applied Statistics*, 23:466–470.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *J. R. Statist. Soc. B*, 54(3):761–771.

Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *J. R. Statist. Soc. B*, 56(2):393–396.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88(424):1264–1271.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons.

Huber, P. (1985). Projection pursuit. *Ann. Statist.*, 13:435–525.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21(1):79–87. Statistics Canada.

Hulliger, B. (1999). Simple and robust estimators for sampling. In *Proceedings of the Section on Survey Research Methods*, pages 54–63. American Statistical Association.

Hulliger, B. (2000). ICES II, Invited session on outliers : Discussion.

Hulliger, B. and Béguin, C. (2001). Detection of multivariate outliers by a simulated epidemic. In *Proceedings of the ETK/NTTS 2001 Conference*, pages 667–676. Eurostat.

Hulliger, B. and Kassab, M. (1998). Evaluation of estimation methods for the survey on environment protection expenditures of swiss communes. Methodology report, Swiss Federal Statistical Office.

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

Lee, H. (1995). Outliers in business surveys. In *Business Survey Methods*, pages 503–526. John Wiley and Sons.

Li, G. and Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte-Carlo. *J. Amer. Statist. Ass.*, 80:759–766.

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

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414.

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. (2001). Robust multivariate estimates for high dimensional data sets. Preprint.

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

Maronna, R. A., Stahel, W. A., and Yohai, V. J. (1992). Bias-robust estimators of multivariate scatter based on projections. *Journal of Multivariate Analysis*, 42:141–161.

Maronna, R. A. and Yohai, V. J. (1991). The breakdown point of simultaneous general m estimates of regression and scale. *Journal of the American Statistical Association*, 86:699–703.

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.

Maronna, R. A., Yohai, V. J., and Zamar, R. J. (1993). Bias-robust regression estimation: A partial survey. In *New Directions in Statistical Data Analysis and Robustness*, pages 157–176. Birkhäuser (Basel; Cambridge, MA).

Mesa, D., Tsai, P., and Chambers, R. L. (2000). Using tree-based models for missing data imputation : an evaluation using UK census data.

Milasevic, P. and Ducharme, G. R. (1987). Uniqueness of the spatial median. *The Annals of Statistics*, 15:1332–1333.

Munier, S. (1999). Multiple outlier detection in logistic regression. *Student*, 3(2):117–126.

Oja, H. and Niinimaa, A. (1985). Asymptotic properties of the generalized median in the case of multivariate normality. *Journal of the Royal Statistical Society, Series B, Methodological*, 47:372–377.

Patak, Z. (1990). Robust principal component analysis via projection pursuit. Master's thesis, University of British Columbia, Canada.

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, pages 283–297. Elsevier.

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

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

Rousseeuw, P. J. and Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Commun. Statist.-Theory Meth.*, 22(4):965–984.

Rousseuw, P. (1984). Least median of squares regression. *J. Am. Stat. Assoc.*, 79:871–880.

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

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

Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10:262–266.

Stahel, W. (1981). *Robuste Schätzungen: infinitesimale optimalität und Schätzungen von Kovarianzmatrizen*. Ph.d. thesis, Swiss Federal Institute of Technology.

van der Waerden, B. (1971). *Mathematische Statistik*, volume 87 of *Die Grund. der math. Wiss. in Einzeldarstellungen*. Springer-Verlag.

Wilks, S. S. (1963). Multivariate statistical outliers. *Sankhyã*, 25 A:405–426.

Wilks, S. S. and Gnanadesikan, R. (1964). Graphical methods for internal comparisons in multiresponse experiments. *Ann. Math. Statist.*, 35:623–631.

Yohai, V. J. and Maronna, R. A. (1976). Location estimators based on linear combinations of modified order statistics. *Communications in Statistics, Part A – Theory and Methods*, 5:481–486.