Robust Multivariate Outlier Detection Based on Forward Search Methods

EUREDIT Workpackage 4.2

EUREDIT Deliverable D4.2.1 (Sections 1 and 2)
EUREDIT Deliverable D4.2.2 (Section 3)

Adão L. Hentges

Department of Social Statistics
University of Southampton
Highfield, Southampton, SO17 1BJ, U.K.

1. Introduction

The problem of identifying outliers in a data set has been subject of extensive research. For a quite comprehensive study see, for example, Barnett and Lewis (1984). When we deal with an univariate sample any outliers in the data are usually detected regarding their extremeness relative to the basic model F from which all observations came from. Tests of the discordancy of the outliers with respect to the fully specified distribution F are then performed.

An observation x_i can be judged through deviation/spread statistics, where a measure of its distance to the whole data uses some measure of the central tendency in the data and the spread of the sample. One of the extreme values from $x_1, x_2, \ldots, x_n$ could be declared a contaminant (an observation from some other distribution G) if its extremeness disagrees to what it was expected from the basic model F. The usual tests are of the form: declare unit i as an outlier if

$$u(x_i) \;=\; \frac{x_i - \bar{x}}{s} \;> c_\alpha, \tag{1}$$

where $\bar{x}$ and s denote the sample mean and the sample deviation, and c_α some appropriate cutoff with significance level α. The median deviation $s_m = median(|x_i - \bar{x}|$

may also be used as a sample measure of dispersion instead of the sample deviation s.

In the simple univariate case it is quite clear what the definition of extremeness means. A large value for $u(x_i)$ is an indication that the observation x_i may be an outlier or at least a suspected unit in the sample. For data related to linear models in seeking outliers it is generally common to examine the relative size of the residuals. For example, in a simple linear regression like $y_j = \beta_0 + \beta_1 x_j + \epsilon_j$, where the set of n observations $y_1, \ldots, y_n$ of independent random variables Y_j has means depending linearly on values $x_1, \ldots, x_n$ of X, the residuals ϵ_i can be estimated as

$$\hat{\epsilon}_i \;=\; y_i \;-\; \hat{y}_i \;=\; y_i \;-\; (\hat{\beta}_0 + \hat{\beta}_1 x_i), \tag{2}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates usually provided by least squares. The studentized residuals $e_i = \frac{\hat{\epsilon}_i}{s_i}$, where

$$s_i \;=\; s\,\sqrt{1 - \frac{1}{n} - (x_i - \bar{x})^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{3}$$

and $s^2 = \sum \hat{\epsilon}_i^2/(n-2)$ is an unbiased estimate of σ^2, are then used to declare the observation y_i as an outlier when they are sufficiently large. For multivariate linear response models, where $\mathbf{Y} \;=\; \mathbf{X}^T\beta + \epsilon$, regression outliers can be detected through measures like the Mahalanobis distance

$$D_i \;=\; \sqrt{\{(\mathbf{y}_i - \hat{\mathbf{y}}_{i(r)})^T \;\hat{S}_{(r)}^{-1}\; (\mathbf{y}_i - \hat{\mathbf{y}}_{i(r)})\}}, \quad i = 1, \ldots, n, \tag{4}$$

where $\hat{\mathbf{y}}_{i(r)}$ and $\hat{S}_{(r)}$ are obtained from regression of $\mathbf{Y}$ on $\mathbf{X}$ based on the n observations from the sample data.

Discordancy tests for outliers may suffer from the *masking* problem, in which a testing procedure to identify a single outlier may be ineffective when the data set has several suspected values. Usually the most suspected observations form small subgroups and then it is difficult to locate the true outliers of the sample from that particular subset. An outlying subset thus goes undetected because of the presence of another, usually adjacent, subset. A *swamping* problem occurs when "good" observations are incorrectly identified as outliers because of the presence of another, usually remote, subset of observations. Since the outlying units attract the estimates towards them, reasonable observations then become suspicious.

It is possible to make tests consecutively in order to test first the most extreme or most suspected observation to the last one from that subset. In case the most extreme, $x_{(n)}$, is declared outlier when compared to the remaining $n-1$ cases, we then move to test the next one, $x_{(n-1)}$, where $x_{(j)}$ denotes an order statistic. Another approach is to use a block test, where a group of k units are tested all together. If some statistic U_k exceeds some critical value then the k units $x_{(n-k+1)}, \ldots, x_{(n)}$ are declared outliers.

For multivariate sample problems it is much more difficult to locate outliers since for univariate sample the definition of extremeness came from some form of ordering the data. Now a multivariate observation can be an outlier without need to be an extreme in any of its components. A common principle is specified in terms that the most extreme observation $\mathbf{x}_i$ yields the largest incremental increase in the maximized likelihood under the model F for the remaining data, when it was omitted from the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$. If that increase was very large it then leads to the declaration of $\mathbf{x}_i$ as outlier.

Following this idea, a sensible criterion suggested by Wilks for declaration of an outlier is to choose the observation whose omission leads to the least value for

$$R_{(j)} \;=\; \frac{\left|\mathbf{A}^{(j)}\right|}{|\mathbf{A}|}\;, \tag{5}$$

where $\mathbf{A}$ is the matrix of sum of squares and cross products of the observations about the component sample means, and $\mathbf{A}^{(j)}$ a similar matrix omitting unit j; see, Barnett and Lewis(1994).

We can also look for a set of k outliers through $R_{(k)} \;=\; |\mathbf{A}^{(k)}|/|\mathbf{A}|,\ k \geq 1$ when seeking outliers in a multivariate data through the use of such "leave-one-out" methods. In this case we evaluate the effect the deletion of one unit or block of unit vectors may cause on the remaining matrix. However, for a large data set it may not be feasible to look for all sets of reasonable sizes of potential outlier observations. The computational complexity for this kind of search may limit the use of the method to only small data set problems or to a case where the number of suspected observations in a large sample is quite small.

2. Proposed methods for multivariate outlier detection

From the brief overview of some approaches for multivariate outlier detection the

complexity for applying the Wilks' criterion (5) is clear, specially for large datasets. We then suggest three alternative approaches to deal with the general problem of identifying multivariate outliers.

The first approach for the detection of outliers is to use forward methods, which start from a small subset of the data and observations are added to the subset until finally all the sample is included. The starting subset of size $m, (m < n)$ is chosen to be *clean*(free from outliers) and this kind of algorithm monitors the effect that each new observation causes in the estimates of the parameters. The aim is to avoid the masking effect of multiple outliers possibly present in the sample that can be a disadvantage and provide a poor performance for backward methods. Hadi and Simonoff(1993) and also Riani and Atkinson(2000) provide algorithms to perform this "include in" searches, where the outlier–free data set is found by starting from small subsets and moving to larger subsets containing only observations that have small residuals and thus are unlikely to be outliers.

Generally the identification has to be carried out relative to some assumed model for the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ in the sample. The standard linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ is a natural option but we also want to structure the outlier identification process so that it makes minimal assumptions about the nature of the model relating $\mathbf{Y}$ to $\mathbf{X}$.

A non–parametric option is to work with regression trees; see, for example, Breiman, Friedman, Olshen and Stone (1984), and Tsai and Chambers (2000). In this second approach a vector of categorical variables $\mathbf{X}$ is used to build a regression tree that describes the distribution of the response variable in terms of the categories of the explanatory variables. The tree is built in a way that terminal nodes, defined by classes of $\mathbf{X}$, tend to group together homogeneous values for the response variable. Methodology and experimental results for this approach are reported in Xinqiang and Chambers (2002).

Finally, the third proposed approach uses the M–quantile idea (Breckling and Chambers, 1998; Lübke, Kokic and Breckling, 2001), where the modelling of the data is associated with extreme points in a sample. This algorithm can be designed in a way to detect the local behaviour of the data either in the centre of the sample or in the tails. Ideally, this method for identifying outlying points should be able to relate each sample unit to a certain probability and direction, which has an orientation within the whole data set. From the sets of probabilities associated to the n sample units, the outliers could be then located through some degree of outlyingness. Methodology for

this approach is described in Kokic (2002).

3. The forward search method

When masked multiple outliers are present in the data generally it is difficult to locate the true outlying units. Classical identification methods do not always find those observations since they are generally based on the sample mean and covariance matrix, which are estimates already affected by the outliers. A relatively small cluster of outliers may attract the estimate of the location and then would inflate the estimation of the variability in its direction. Outliers would not then have a large value for the usual Mahalanobis distance. In the case of regression data the ordinary least squares approach may mask the outliers in a similar way.

Robust methods are then necessary to overcome this problem since robust distances suit better to expose the true outliers in the data. Some approaches are designed in that way. Rousseuw (1984) uses standardized least median of squares (LMS) residuals

$$\frac{r_i}{\hat{\sigma}} \;=\; \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{k\,\sqrt{\mathsf{median}(r_1^2,\ldots,r_n^2)}} \tag{6}$$

where k is a positive constant. The estimator $\hat{\beta}$ is defined by

$$\mathsf{min}_{(\beta)} \;\; \mathsf{median}(r_i^2(\hat{\beta}),\; i=1,\ldots,n) \tag{7}$$

where $r_i(\hat{\beta}) = y_i - \mathbf{x}_i^T\hat{\beta}$ is the residual for the i–th observation. Large standardized residuals (6) may indicate the regression outliers. Unbiased estimates of the regression line are still provided by the LMS method for large n, even if almost half of the data are outliers or come from some other model. The LMS has then an asymptotic breackdown point of 50%.

In multivariate data, for a dataset $\mathbf{Y} \;=\; (\mathbf{y}_1,\ldots,\mathbf{y}_n)$ of n points in p dimensions, the sample mean and the sample covariance matrix may be not adequate as estimators for the center and scatter of $\mathbf{Y}$. Rousseuw and Zomeren (1990) define robust distances to identify multivariate outliers. The measure RD_i is obtained by inserting the minimum volume ellipsoid (MVE) estimates for the sample mean and covariance matrix on the classical Mahalanobis distance,

$$RD_i \;=\; \sqrt{(\mathbf{y}_i - \bar{\mathbf{Y}}_{(mve)})\; S^{-1}_{(mve)}\; (\mathbf{y}_i - \bar{\mathbf{Y}}_{(mve)})}. \tag{8}$$

The minimum volume ellipsoid estimator is defined as the pair (A, **B**) such that the determinant of **B** is minimized subject to

$$\#\{i;\;\; (\mathbf{y}_i - A)\; \mathbf{B}^{-1}\; (\mathbf{y}_i - A)^T \;\leq a^2\} \;\geq h \tag{9}$$

where h is the integer part of $(n + p + 1)/2$, A is a p–vector and **B** is a positive semidefinite p–by–p matrix. The number a^2 is a fixed constant, usually chosen as $\chi^2_{(p;0.5)}$ when it is expected the majority of the data come for a normal distribution. Small samples will require a factor $c^2_{(n,p)}$, which depends on n and p. The MVE has also a breakdown point of nearly 50%, which means that the location estimate A will remain bounded and the eigenvalues of **B** will stay away from zero and infinity when a little less than half of the data are replaced by arbitrary values. Even if those arbitrary values contains outliers, robust estimates would still be provided by the MVE method.

Although the RD_i is a robust measure to detect the outlying observations, it is computationally expensive to be computed. Even for modest sample sizes it may not be feasible to find the MVE since we need to select the ellipsoid with the minimum volume from all the $n!/(h!(n-h!))$ possible combinations from the n observations. Approximate algorithms for the MVE may be used to overcome the computational cost using resampling methods. By drawing subsamples of $p+1$ different observations $\{i_1, \ldots, i_{p+1}\}$, indexed by J, the mean and covariance matrix are then

$$A_j \;=\; \frac{1}{p+1}\; \sum_j\; \mathbf{y} \qquad \mathsf{and} \qquad \mathbf{B}_j \;=\; \frac{1}{p}\; \sum_j\; (\mathbf{y}_i - A_j)^T (\mathbf{y} - A_j). \tag{10}$$

We then compute

$$m_j^2 \;=\; \{(\mathbf{y}_i - A_j)\; \mathbf{B}_j^{-1}\; (\mathbf{y}_j - A_j)^T)\; \}_h \tag{11}$$

as the corresponding ellipsoid should contain exactly h points. The squared volume of the j–th resulting ellipsoid is proportional to $m_j^{2p}\; \mathsf{det}(\underline{\mathrm{B}}_j)$, of which the smallest value is recorded. Finally, the best j selected subset provides

$$\hat{A} \;=\; A_j \qquad \mathsf{and} \qquad \hat{\mathbf{B}} \;=\; (\chi^2_{(p,0.5)})^{-1}\; c^2_{(n,p)}\; m_j^2\; \mathbf{B}_j \tag{12}$$

as approximation for the MVE estimators. A weighted mean,

$$A_1 \;=\; (\sum_{i=1}^{n} w_i)^{-1} \;\; \sum_{i=1}^{n} w_i \mathbf{y}_i, \tag{13}$$

and a weighted covariance matrix,

$$\mathbf{B}_1 \;\;=\;\; (\sum_{i=1}^{n} w_i - 1)^{-1} \;\; \sum_{i=1}^{n} (\mathbf{y}_i - A_1)^T \, (\mathbf{y}_i - A_1), \tag{14}$$

where the weights $w_i \;=\; w(RD)$ depend on the robust distances (8), are computed later in a reweighting step.

The implementation of the MVE method via resampling may be also very expensive since it demands a lot of different samples to reach good estimates.

Starting with the full sample and removing sequentially all the suspected units untill no more outliers are present in the data may be an appealing method. However, the swamping problem may also be present in the sample, affecting discordancy test for blocks of two or more suspected units. A carefull investigation of all possible combinations of suspected units at some particular stage, through tests like (5), in order to get a clean data and move to the next trimming step, may also become not feasible. Although the underlying idea is simple, the combinatorial explosion of the number of cases of potential outliers to be considered at once is a severe drawback of such backward approach.

Many methods for the detection of multiple outliers therefore use very robust methods to split the data into a clean part and the potential outliers. One option is the forward search method, which seems to overcome the problems faced by for one single–step search or a backward outlier detection; see, for example, Hadi (1992), Hadi and Simonoff (1993), Atkinson (1994), Hadi (1994), and Riani and Atkinson (2000) . The basic idea is to start with a relatively clean data set of size m, defined from a robust method, and include observations until only the outlying units remain out. As Atkinson (1994) points, the forward algorithm rapidly leads to the detection of multiple outliers. The exact calculation of robust parameter estimates, like the MVE method attempts to do, does not seem to be necessary for outlier detection.

Variants of this idea include the BACON (Blocked Adaptive Computationally efficient

Outlier Nominators algorithm; Billor, Hadi and Veleman, 2000), starting with one initial subset and iterating until the data is separated in two parts, the outlier–free subset and the outlying units. The Kosinski algorithm (Kosinski, 1999; and De Boer and Feltkamp, 2000) follows the same principle, where several small subsets are selected as starting points of a two levels iterating algorithm, ending with the partition of the dataset in two parts, the good points and the outliers. Hulliger (2000) presents an algorithm starting from a randomly chosen point. The epidemic then spreads through the data and eventually all points are infected, the outliers usually are infected late in this process due to their outlying isolation. A general comparison of the available methods is not the focus of this work at this stage and then these last mentioned algorithms will not be explored.

3.1. The initial clean subset

Let

$$C_{(m)} \;=\; \{(\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \ldots, m\ (m < n)\} \tag{15}$$

be the initial clean data, supposedly outlier–free, and $(\mathbf{y}, \mathbf{x})$ the sample values of the multivariate response variable $\mathbf{Y}_p$ and the covariate vector $\mathbf{X}_q$. Usually the size m is chosen as the integer part of $h = (n + k - 1)/2$, where k is the number of parameters in the model.

This starting subset of data may be defined in different ways. Hadi and Simonoff (1993) suggests two procedures: the first, by fitting a regression model to the full data and then ordering the n observations by an appropriate diagnostic measure. The first $k+1$ units form the initial basic subset. A regression model is then fitted to this basic subset and all n observations are again arranged in ascending order according to

$$d_i \;=\; \begin{cases} \dfrac{y_i \;-\; \mathbf{x}_i^T\ \hat{\beta}_{(m)}}{\sqrt{1 \;-\; \mathbf{x}_i^T (\mathbf{X}_{(m)}^T \mathbf{X}_{(m)})^{-1}\ \mathbf{x}_i}}\ , & \text{if}\ \ i \in B \\ \dfrac{y_i \;-\; \mathbf{x}_i^T\ \hat{\beta}_{(m)}}{\sqrt{1 \;+\; \mathbf{x}_i^T (\mathbf{X}_{(m)})^{-1}\ \mathbf{x}_i}}\ , & \text{if}\ \ i \notin B. \end{cases} \tag{16}$$

where B is the basic subset with the $k + 1$ starting units. The new subset will then include the first $(k + 1) + 1$ and will grow up this way until the basic subset contains h observations, becoming the initial clean data $C_{(m)}$.

The second method involves a backward selection by constructing a single linkage clustering tree and ordering the clusters from most to least extreme by the order of joining. The principle is that the later a cluster joins, the more extreme it is. At each joining, the cases in the smaller cluster are identified as the more outlying. When the number of the most extreme identified cases reaches $n - h$, the h remaining cases constitute the initial clean data $C_{(m)}$.

Riani and Atkinson (2000) start the forward search in the univariate case with the selection of a subset of size $q + 1$ units, equal to the number of parameters in the model, where $\mathbf{X}$ is the $n \times q$ matrix of explanatory variables. For moderate sample size n, the choice of the best clean subset of size $q+1$ can be performed by exhaustive enumeration of all $\binom{n}{q+1}$ distinct subsets. The initial clean data is defined by the subset which provides the minimum median for the least squares residuals from regression. A larger number of samples is required for the definition of this starting subset as the best initial clean data if $\binom{n}{q+1}$ is too large. This criterion gives a least median of squares (LMS) approach for regression models with independent errors but may be very expensive in computational terms.

For multivariate data from a p–dimensional population, Hadi (1992) defines the basic subset by selecting the first $p+1$ observations from the n units arranged in an ascending order according to a robust distance, using

$$D_i(L_R, S_R) \;=\; \sqrt{\{(\mathbf{y}_i - L_R)^T \; S_R^{-1} \; (\mathbf{y}_i - L_R)\}}, \quad i = 1, \ldots, n, \tag{17}$$

where L_R and S_R are robust location and covariance matrix estimators from the fit in the full sample. Riani and Atkinson (2000) define a larger initial subset than $m = p+1$. After transforming the data they perform a robust analysis of the matrix of bivariate scatterplots and take as the initial subset those observations that are not outlying on any scatterplot. The selected units to compose $C_{(m)}$ are found as the intersection of all points lying within a robust contour containing a specified proportion of the data.

3.2. The main algorithm

Suppose a clean subset $C_{(m)}$ is already available. The forward search then moves from m observations to $m + 1$ by choosing the $m + 1$ observations with the smallest residuals from the fit on data of $C_{(m)}$. Standardized residuals from the estimate $\hat{\beta}$ for the linear regression model $E(\mathbf{Y}) = \mathbf{X}\beta$ are computed by Hadi and Simonoff (1993) in the univariate case ($p = 1$) as

$$d_i \;=\; \begin{cases} \dfrac{y_i \;-\; \mathbf{x}_i^T\;\hat{\beta}_{(m)}}{\hat{\sigma}_{(m)}\;\sqrt{1\;-\;\mathbf{x}_i^T(\mathbf{X}_{(m)}^T\mathbf{X}_{(m)})^{-1}\;\mathbf{x}_i}}\;, & \text{if}\;\; i \in C_{(m)} \\ \dfrac{y_i \;-\; \mathbf{x}_i^T\;\hat{\beta}_{(m)}}{\hat{\sigma}_{(m)}\;\sqrt{1\;+\;\mathbf{x}_i^T(\mathbf{X}_{(m)}^T\mathbf{X}_{(m)})^{-1}\;\mathbf{x}_i}}\;, & \text{if}\; i \notin C_{(m)}, \end{cases} \tag{18}$$

where $\hat{\beta}_{(m)}$ are the estimated *least squares* regression coefficients computed from fitting the linear model to $C_{(m)}$ and

$$\hat{\sigma}^2_{(m)} \;=\; \frac{\sum_{i=1}^{m}\;(y_i - \mathbf{x}_i^T\;\hat{\beta}_{(m)})^2}{m-q} \tag{19}$$

the corresponding residual mean square. When $i \in C_{(m)}$, d_i is then the internally studentized residual and when $i \notin C_{(m)}$, d_i is the scaled prediction error based on the subset $C_{(m)}$.

Atkinson (1994) uses almost similar residuals, except that for $i \in C_{(m)}$ d_i is defined as the least squares residuals but in their comparison no evidence was found in favor of one type. For $p > 1$, Riani and Atkinson (2000) uses the squared Mahalanobis distances

$$d_i^2 \;=\; \{(\mathbf{y}_i - \hat{\mathbf{y}}_{i(m)})^T\;\hat{S}^{-1}_{(m)}\;(\mathbf{y}_i - \hat{\mathbf{y}}_{i(m)})\}, \quad i = 1, \ldots, n, \} \tag{20}$$

to order observations for the forward search, where $\hat{\mathbf{y}}_{i(m)}$ and $\hat{S}_{(m)}$ are obtained from regression of $\mathbf{Y}$ on $\mathbf{X}$ based on the m observations from the basic clean data $C_{(m)}$. Similarly, Hadi (1994) performs the forward search based on the Mahalanobis distances calculated from the residuals from the regression.

The observations are rearranged in ascending order according to similar measure based on $L_{(m)}$ and $S_{(m)}$, the mean and covariance matrix of the basic subset. At the next step the clean data increases its size to $m+1$ using the n distances obtained from $C_{(m)}$. Observations can leave the subset for fitting as well as joining it as m increases as n distances are evaluated and ordered to define each move from m to $m+1$. However, from our computational experience with different data sets (Hentges, 2001–b) it seems that for a few s steps ahead the clean data $C_{(m+s)}$ keeps all the former $(m+s-1)$ units, just including a new one, and then becomes very stable in its composition.

Instead of growing the size of the clean data $C_{(m)}$ by just $u = 1$ unit at each step the algorithm remains basically the same if the clean data is increased by more than one

observation. For large sample sizes in the first moves the value of u could be large in the beginning and smaller (or equal to 1) by the end, when changes in the distance measures are more likely to appear.

When some stopping criterion is met at some particular stage s^*, then all the units not included on $C_{(s^*)}$ are declared outliers and the algorithm stops. If the search augments $C_{(m)}$ up to the full sample size n without a stopping requirement being met then the data is declared outlier free.

3.3. Outlier identification and stopping rules

Under the assumption that the random errors ϵ_i in the general linear model

$$\mathbf{Y} \;=\; \mathbf{X}^T\beta \;+\; \epsilon$$

are $iid\ N(0,\sigma^2)$ for the univariate response variable case ($p = 1$), then the residuals d_i^2 in (18) would follow a t distribution. However, as they involve the estimate $\hat{\beta}$ they are dependent. Using $\hat{\sigma}_{(m)}$ (19) to scale the residuals, assuming normality and if $\hat{\beta}_{(m)}$ and $\hat{\sigma}_{(m)}$ were independent, then d_i would have a Student's t distribution with $s - p$ degrees of freedom for each subset of size s for $i \notin C_{(m)}$. Although $\hat{\beta}_{(m)}$ and $\hat{\sigma}_{(m)}$ are dependent, Hadi and Simonoff (1993) use the t distribution as a benchmark from witch to determine cutoff values. The residuals d_i in (18) are then compared to $t_{(\alpha/(2(s+1)),s-p)}$ in the main algorithm in order to point out the outlying units.

A stopping criterion is used for Hadi and Simonoff (1993) in the forward search. For $p = 1$, $d_{(s+1)}$ is defined as the $(s+1)$–th order statistic of the n absolute residuals $|d_i|$, where s is the size of the current subset $C_{(m)}$. If

$$d_{(s+1)} \geq t_{(\alpha/2(s+1),s-q)} \tag{21}$$

then all observations satisfying $|d_i| > t_{(\alpha/2(s+1),s-q)}$ are then declared outliers and the forward search finishes.

For $p > 1$, Hadi (1992) suggested two possible stopping rules when the basic data set is increased. The first criterion is to stop when $min\{D_i(L_b, S_b);\ i \notin C_{(b)}\} \geq c_\alpha$ where the critical value c_α can be chosen such that $Pr[min\{D_i(L_b, S_b); i \notin C_{(b)}\} \geq c_\alpha|\mathbf{Y}$ contains no outliers$] \;=\; 1 - \alpha$. The problem is that c_α depends on the distribution of $D_i(L_b, S_b)$, which is difficult to obtain. The other rule stops when the basic subset

$C_{(b)}$ is augmented and contains h observations. However, both rules need to work with resampling for the MVE since the covariance matrix S_b used to evaluate $D_i(L_b, S_b)$ in (17) has a correction factor depending on m_j.

With a modification on his former algorithm, Hadi (1994) orders the n evaluated squared Mahalanobis measures $D_i^2(L_b, S_b)$, where now S_b does not depend on m_j, and defines $D^2_{(s+1)}$ as the $(s+1)$–th order statistic of the D_i^2. In a regression model, residuals from fitting of $\mathbf{Y}$ on $\mathbf{X}$ are used to evaluate D_i^2. The multivariate search stops if

$$D^2_{(s+1)} \geq \chi^2_{(p,\alpha/n)}, \tag{22}$$

and then and all observations with $D_i^2 \geq \chi^2_{(p,\alpha/n)}$ are identified as outliers. If the basic data set increases to $C_{(m)} = C_{(n)}$, without the stopping criterion being met, then the data set is declared outlier free.

Atkinson (1994) and Riani and Atkinson (2000) perform forward searches but without a stopping rule. The emphasis there is analyzing plots of the residuals obtained from a full search, starting from the clean data and increasing up to the full sample size.
At each particular stage $m, m+1, m+2, \ldots, n$ each observation $\mathbf{y}_i$ is tested if it is an outlier according to the Mahalanobis distances from (20). The cuttof value used is the maximum expected value from a sample of n chi–squared random variables on p degrees of freedom, approximated by

$$E(max\ \chi_p^2) \ = \ \chi_p^2\ \{(n-0.5)/n\}. \tag{23}$$

When the full search ahead is performed, the units which have been identified as outliers in most of the steps can have a close examination. Empirically we define a set of outliers by taking the observations which are not on the current clean data when the relative "jump" on the residual variance on the fit on $C_{(m)}$ is maximum. Let

$$\tau_j \ = \ \frac{det(S_{(j)}) \ - det(S_{(j-1)})}{det(S_{(j-1)})}, \quad j = 2, \ldots, n, \tag{24}$$

where $S_{(m)} \ = \ (m-q)^{-1} \ \sum_{i=1}^{m} \ (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \ (\mathbf{y}_i - \hat{\mathbf{y}}_i)$ is the estimated residual covariance matrix based on the clean data with current size m and $det(S)$ its determinant. Since the search is based in Mahalanobis distances in ascending order, the most important

outlier to join the clean data at some stage should cause a breakdown for $S_{(m)}$. At some step j where τ_j is maximum we declare the unit joining the clean data and all those not included yet as outliers. The distribution of τ_j is not available and depends on the sequence $S_{(m)}, S_{(m+1)}, \ldots, S_{(n)}$ with dependent components, since generally units included in the clean data at step m should be present at step $m+1$ too.

Also, considering the number of times each sample unit was declared outlier in the whole search, we apply a binomial test to define a set of outlying units. For example, suppose π_i is the true probability that sample unit i is an outlier in the population. Let $\delta_i = \sum_{k=1}^{n-m} I_{ik}$ be the number of times the unit was identified as outlier based on the $n-m$ steps performed, where I_{ik} is equal 1 when residual d_i is outlying on the k–th step of the search and 0 otherwise.

Assume now that $\delta_i \sim B(n-m, \pi_i)$, at least approximately since the I_{ik} are not independent. Defining $\hat{p}_i = \delta_i/(n-m)$ we then declare unit i as a true outlier (by specifying $\pi_i = 1$) if

$$\frac{\sqrt{(n-m)-1}\,(1-\hat{p}_i)}{\sqrt{\hat{p}_i\,(1-\hat{p}_i)}} \;<\; c_{(\alpha)}, \tag{25}$$

where $c_{(\alpha)}$ is the cutoff given by the asymptotical normal $N(0,1)$ distribution.

These two empirical procedures will be performed just to have some comparison between the outliers defined by the precise stopping rule from Hadi, in a way to use results from the full search and check the if the outlying sets agree.

3.4. Graphical examination for outlier detection

Having performed $n-m$ steps in the full forward search algorithm it is possible to analyze the behaviour of the sequence of the n residuals. Units with a clear outlying pattern could be detected through the analysis of those residuals. Atkinson (1994) and Atkinson and Riani (2000) use *stalactite* plots, displaying standardized residuals for all n sample units throughout the $n-m$ steps of the forward search. For most of the search the largest residuals expose the outliers and on the last steps some dramatic changes happen when outliers are present on the data, since those units are finally included in the augmented clean data and then creates a breakdown of the quite monotonic behaviour of the residuals.

By performing a number of simulations, in each one defining at random a clean starting subset $C_{(m)}$, Atkinson (1994) finds the average value $\bar{\sigma}_{(m)}$, the mean of the LMS variance estimate $\tilde{\sigma}^2_{(m)} = median(e_i^2)$. The stability of the search can be seen from confidence intervals for the average $\bar{\sigma}_{(m)}$ for different starting sizes m. The smooth increase of $\hat{\sigma}_{(m)}$ (19) is typical of what is expected when the data agree with the model and are correctly ordered by the forward search. Although those graphical procedures do not provide a formal test for outlier detection they are a powerful aid to indicate which units are potential outliers and its influence on the breakdown of the clean residual variance $\hat{\sigma}_{(m)}$.

3.5. Transformations on the data

Outliers in the raw original data may not be outliers in another transformed scale and vice versa. If the data are analyzed using the wrong transformation the possible outliers present into it may not be detected or even enter the search well before the end.

Generally the forward search for outliers in regression models are based on the classical linear model $Y = \mathbf{X}^T\beta + \epsilon$. For transformation on just the response variable Y, Box and Cox (1964) analyzed the normalized power transformation

$$z(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda \; \tilde{y}^{\lambda-1} , & \lambda \neq 0 \\ \tilde{y} \; log(y) , & \lambda = 0 \end{cases}$$

where the geometric mean of the n observations is written as $\tilde{y} = exp(\sum log \; y_i/n)$. The model fitted in this transformation is multiple regression with response $z(\lambda)$,

$$z(\lambda) = \mathbf{X}^T\beta + \epsilon. \tag{26}$$

When $\lambda = 1$ there is no transformation. Another values usually used for λ and supported by empirical reasoning on the analysis of real data sets are: $\lambda = 1/2$, the square root transformation; $\lambda = 0$, the log transformation and $\lambda = -1$, the reciprocal. In this analysis the aim is to find an estimate of λ which provides errors in (26) at least approximately normal distributed with constant variance and for which a simple linear model reasonably fits the data.

Atkinson and Riani (2000) present approximate score test statistics for testing the transformation parameter λ by

$$z(\lambda) \doteq z(\lambda_0) + (\lambda - \lambda_0)w(\lambda_0), \tag{27}$$

where λ_0 is the hypothesized value for λ. The approximate score statistic for testing the transformation, $T_p(\lambda_0)$, is the t statistic regression on $w(\lambda_0)$ in (27). Riani and Atkinson (2000) monitor the score statistic for transformation as the number of observations of the clean data used to fit the model is increased. The *fan* plot displays the influence of each individual observation and the evidence for a transformation.

Although some transformation on the data may be appropriate to expose the true outliers the main point is that outliers identified in that new scale for some λ value, are not necessarily outlying observations in the raw scale ($\lambda = 1$). After the outlier identification is performed, usually inference moves to estimation. If the estimates are evaluated on the transformed data, using robust methods that decrease the impact of those suspected values, it is not straightforward to convert it to the original scale. The possible bias involved in this back transformed estimates may bring an additional complication to the analysis and must be carefully corrected. It may be necessary to transform the data through sophisticated ways but perhaps a simple transformation can be adequate in case it provides a good linear model to be used within the forward search approach, when the main concern is outlier detection.

References

Atkinson, A.C. (1994). "Fast very robust methods for the detection of multiple outliers". *JASA*, **89**, 428, 1329–1339.

Atkinson, A.C. and Riani, M. (2000). *Robust diagnostic regression analysis*, Springer–Verlag, New York.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data*, Wiley.

Billor, N., Hadi, A.S. and Velleman, P.F. (2000). "BACON: Blocked Adaptive Computationally efficient Outlier Nominators", to appear in *Computational Statistics and Data Analysis*.

Box, G.E.P. and Cox, D.R. (1964). "An analysis of transformations", *JRSS*, Ser. B, 26, 211–246.

Chaudhuri, P., Doksum, K. and Samarov, A. (1997). "On average derivative quantile regression", *The annals of statistics*, 25, 2, 715–744.

Daniels, H.E. (1944). "The relation between measures of correlation in the universe of sample permutations", *Biometrika*, 33, 129–135.

De Boer, P. and Feltkamp, V. (2000). "Robust multivariate outlier detection". *Technical report*, Statistics Netherlands, Voorburg.

Hadi, A.S. (1994). "A modification of a method for the detection of outliers in multivariate samples", *J.R. Statist. Soc. B*, **56**, No. 2, 393–396.

Hadi, A.S. and Simonoff, J.S. (1993). "Procedures for the identification of multiple outliers in linear models", *JASA*, **88**, 1264–1272.

Hentges, A.L. (2001)–b. "Robust multivariate outlier detection via the forward search applied to the ABI data set", *Technical report*, EUREDIT.

Hulliger, B. (2000). "An epidemic algorithm for multivariate outlier detection", *Technical report*, Swiss Federal Statistical Office.

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

Ren, R. (2001). "Empirical studies on some outlier robust estimators", *Technical report*, EUREDIT.

Riani, M. and Atkinson, A.C. (2000). "Robust diagnostic data analysis: transformations in regression", *Technometrics*, **42**, 4, 384–398.

Rousseeuw, P.J. (1984). "Least median of squares regression", *JASA*, 79, 871–880.

Rousseeuw, P.J. and van Zomeren, B.C. (1990). "Unmasking multivariate outliers and leverage points", *JASA*, **85**, 411, 633–651.