MARCO RIANI - SERGIO ZANI (*)

An iterative method for the detection of multivariate outliers

1. INTRODUCTION

Let $\mathbf{X}$ be an $n \times p$ data matrix representing n observations on p variates. The deviation of the i-th unit can be evaluated computing its Mahalanobis distance $d_i(n)$ (Dasgupta, 1993) from the centroid of the n observations:

$$d_i(n) = \sqrt{\{x_i - \hat{\mu}(n)\}' \ \hat{\Sigma}^{-1}(n) \ \{x_i - \hat{\mu}(n)\}}\,, \qquad (1)$$

for $(i = 1, \ldots, n)$, where x_i is the vector corresponding to the i-th statistical unit, $\hat{\mu}(n)$ denotes the estimated centroid and $\hat{\Sigma}(n)$ is the sample covariance matrix. The symbol in round brackets denotes the number of units used to compute the centroid and the covariance matrix.

It is known that the presence of outliers can have misleading effects on the results of statistical analyses. In the classical approach, observations with a Mahalanobis distance much greater than the others

(*) Istituto di Statistica, Università di Parma, Italy.

The authors wish to thank A.C. Atkinson for helpful discussions and comments. A preliminary version of this paper was presented at the Conference "Giornate di analisi dei dati multidimensionali" held in Naples in October 1995. Financial support from Ministry of University (MURST) is gratefully acknowledged. Only for bureaucratic reasons we state that S. Zani wrote sections 1 and 2 and M. Riani wrote sections 3, 4, 5 and the Appendix.

are declared as outliers. This method, however, involves serious drawbacks especially when there are more than one outlier (Rousseeuw and van Zomeren, 1990) because:

I. outliers do not necessarily have a high value of $d_i(n)$. As a matter of fact, a cluster of outliers could attract the centroid and consequently increase the value of the variances;

II. it is not always true that observations which present a large Mahalanobis distance are outliers. A cluster of atypical values, indeed, could move the means from the center of the p dimensional cloud where most observations lie, causing a large Mahalanobis distance for these last observations.

These two phenomena are respectively known in the literature as masking and swamping (Barnett and Lewis, 1994).

In order to overcome the drawbacks of the Mahalanobis distance it has been suggested to use robust estimates of the means and of the covariance matrix (Hampel *et al.*, 1986). In this paper we suggest a "forward" procedure (Hadi, 1992; Atkinson, 1994) in which very robust methods based on confidence bivariate contours are used to select an outlier free subset of data. This subset is increased in size using a search which avoids, in the first steps, the inclusion of outliers. During the forward search we monitor particular Mahalanobis distances. The output is presented through plots which are powerful and easy to interpret. The effectiveness of the suggested method in detecting masked multiple outliers and more generally in ordering the data according to their degree of outlyingness is shown by means of data sets widely used in the literature about multivariate outliers.

The program has been written in Gauss, version 3.2 and is available upon request.

2. Steps of the procedure

In this section we give the details of our procedure. In the next sections we will relate our suggestions to the methods currently used in the literature.

The steps of the procedure are the following:

i) *Definition of initial subset*
We build the scatter plot matrix with respect to the $p(p-1)/2$ couples of variables and in each scatter plot we represent a bivari-

ate bo
first is
"relplo
uses th
which
and pr
of the
As a
which
and C
For ea
set" th
contou
a reas
0.90.
order
outlie

For th

DEFI
servations
clean data

We w
initial sub
multivaria

Rema
subset do
atypical
without e
happen, b
the secon

We
sential: i
speak of
the initia
centroid

rious draw-
sseeuw and

(n). As a
entroid and

arge Maha-
ıes, indeed,
ional cloud
วis distance

iterature as

วis distance
ans and of
er we sug-
) in which
rs are used
ıcreased in
ıclusion of
lahalanobis
e powerful
method in
rdering the
means of
ıtliers.
ıd is avail-

n the next
ently used

$(p-1)/2$
ıt a bivari-

ate boxplot. We suggest two variants of bivariate boxplots. The first is based on robust confidence ellipses and is similar to the "relplot" suggested by Golberg and Iglewitz (1992). The second uses the procedure suggested by Riani, Zani and Corbellini (1997) which is based on convex hull peeling and B-spline smoothing and produces outer contours which adapt to the differing spread of the data in the different directions.

As a robust centroid, in this paper we use the bivariate median which minimizes the L_1 norm in $\mathbb{R}^2$ (Small, 1990). Zani, Riani and Corbellini (1998) discuss other choices of a robust centroid. For each couple of variables we define as "bivariate clean data set" the subset which contains the units inside a $(1-\alpha)$ confidence contour. Usually outliers are a small part of the n units. Therefore a reasonable choice, if the sample is large, might be to set $1-\alpha = 0.90$. In small samples this level can conveniently be decreased in order to compute the distances from a subset certainly free from outliers.

For the total of the p variables we propose the following:

DEFINITION. We call the initial subset of multivariate clean observations the one formed by the intersection of the subsets of bivariate clean data in each of the $p(p-1)/2$ pairs of variables.

We will denote by m the number of the units belonging to the initial subset. The observations not included in this initial subset of multivariate clean units constitute the set of potential outliers.

Remark. The proposed definition does not guarantee that the initial subset does not contain multivariate outliers. It may happen that an atypical value can always fall near to the threshold of the ellipses without ever lying outside. Undoubtedly this situation is unlikely to happen, but in such cases this unit could be conveniently detected in the second step of the procedure.

We note that the hypothesis of multivariate normality is not essential: if the data do not follow the normal law, we cannot anymore speak of $(1-\alpha)$ confidence levels. Instead the units belonging to the initial subset are those whose Mahalanobis distance from a robust centroid does not exceed a prefixed threshold.

ii) *Iterative inclusion of the units*

(a) We calculate the Mahalanobis distances from the centroid for each of the n units of the sample using as estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$, that is the vector of the means and the covariance matrix calculated using the m units of the initial subset.

(b) Given a subset of dimension $m \geq p$ we move to dimension $m+1$ by selecting the $m+1$ units which show the smallest Mahalanobis distances. Because n distances are calculated and ordered for each move from m to $m+1$, observations can leave the subset used for computing the centroid and the covariance matrix as well as joining it as m increases. This enables us to detect atypical observations that never fall outside of the bivariate confidence contours. However, we note once again that this situation is very unlikely.

Steps (a) and (b) are repeated until $m = n$ that is when we include all the units in the subset.

If just one unit enters the subset at each move from m to $m+1$ then the algorithm provides a new method for ordering multivariate data according to their degree of outlyingness (Barnett, 1976; Zani, Riani and Corbellini, 1997). As we will see in the next session, multiple outliers or more generally observations which lie far from the bulk of the data can be detected by inspection of simple plots of a variety of statistics.

3. Comparison with existing methods

In the three following subsections our method is related to the existing procedures in the literature in terms of: definition of *initial subset*, iterative inclusion of the units and outlier detection.

3.1. Definition of initial subset

Rousseeuw (1985) in order to find a robust centroid proposed to search for the minimum volume ellipsoid which contains at least half of the observations. However in order to select the minimum volume ellipsoid we must compare $\binom{n}{[n/2]}$ subsamples. Therefore in large samples (e.g. when $n > 200$), this method becomes burdensome and computationally infeasible. Rousseeuw and Leroy (1987) in order to overcome these drawbacks suggested an approximate procedure

which is l
the determ
ever, assu
Rousseeuw
volume ell
singular co
does not s

Our pr
observatio
provides c
potential c
use this m
fact, since
of transfor
points lyin
of the data
transforma
be decreas

3.2. Iterati

Atkins
izes the sq
treating di
outside. In
Consequen
of exchang
of units. C
have this r
just one r
the data a
search the
extremely
search the
likely to

3.3. Outl

The
on p deg

ıe centroid for
stimates $\hat{\mu}(m)$
the covariance
al subset.

to dimension
v the smallest
are calculated
, observations
centroid and
m increases.
:hat never fall
However, we
ely.

ən we include

ı m to $m+1$
g multivariate
t, 1976; Zani,
next session,
h lie far from
' simple plots

əlated to the
on of *initial*
ən.

ɔid proposed
ains at least
ıe minimum
Therefore in
burdensome
1987) in or-
te procedure

which is based on resampling techniques and on the calculation of the determinants of the subsamples of size $p+1$. This method, however, assumes the hypothesis that every subsample has a full rank. Rousseeuw and van Zomeren (1990), in the search for the minimum volume ellipsoid, do not include those subsamples which have a nearly singular çovariance matrix. Our method, compared to this approach, does not suffer from these drawbacks and is computationally easier.

Our procedure allows us to start with a "clean data set" of several observations, if the percentage of contamination is not high. This provides computational savings and simplifications in the analysis of potential outliers. Riani and Atkinson (1998) found convenient to use this method for the analysis of multivariate transformations. In fact, since it is the extreme observations which provide the evidence of transformations, the initial subset found as the intersection of all points lying within a robust contour containing a specified proportion of the data provides a good start to the search for many values of the transformation parameter. In addition, the size of the subset can easily be decreased or increased by changing the level of the contour.

3.2. Iterative inclusion of the units

Atkinson (1994), in every step of the forward search, normalizes the squares of Mahalanobis distances using simulation techniques treating differently the units which are inside the subset from those outside. In Atkinson's procedure the initial subset is chosen randomly. Consequently, in this case is necessary to guarantee a certain degree of exchangeability between the initial subset and the remaining group of units. Our initial subset is free from outliers, therefore we do not have this requirement. In our method in most moves from m to $m+1$ just one new unit joins the subset. This provides a natural order of the data according to their degree of outlyingness. With this kind of search the event in which in one step one unit leaves as two join is extremely unusual. If this occours in the last steps of the forward search these two units might belong to a cluster of outliers and are likely to be highly influential.

3.3. Outlier detection

The asymptotic distribution of the Mahalanobis distances is χ^2 on p degrees of freedom. Hadi (1992) and Atkinson (1994) suggest outlier detection rules based on the χ^2 r.v. However, even if it seems that the results based on Mahalanobis distance remain qualitatively unaltered when the initial distribution of the data belongs to the elliptic family (Mitchell and Krzanowski, 1985), the rules defined above are only asymptotically valid and must therefore be considered as simple approximations. In addition, a "universal" threshold for declearing a unit as outlier does not exist.

Wilks (1963), in order to detect isolated outliers suggested to use the following ratio:

$$|\hat{\Sigma}(n-1)|/|\hat{\Sigma}(n)|\,. \tag{2}$$

If the original data follow a p-variate normal distribution, the ratio in equation (2) is distributed as a Beta r.v. If we suspect that r units are outliers, an approximate test in order to appraise if these r observations must be considered as atypical can be based on the ratio $|\hat{\Sigma}(n-r)|/|\hat{\Sigma}(n)|$. Between the ratio of these two determinants and the Mahalanobis distance there is the following relation (see proof in the Appendix):

$$\begin{aligned}\frac{|\hat{\Sigma}_{(x_1,\dots,x_r)}(n-r)|}{|\hat{\Sigma}(n)|} &= \left(\frac{n-1}{n-r-1}\right)^p \left[1-\frac{n}{(n-1)^2}d^2_{x_1}(n)\right] \times \\ &\times \left[1-\frac{n-1}{(n-2)^2}d^2_{x_2(x_1)}(n-1)\right] \times \cdots \times \\ &\times \left[1-\frac{n-r+1}{(n-r)^2}d^2_{xr(x_1,\dots,x_{r-1})}(n-r+1)\right]\,,\end{aligned} \tag{3}$$

where $d_{x_2(x_1)}(n-1)$, for example, denotes the Mahalanobis distance of observation x_2 with the centroid calculated excluding unit x_1. The greater the Mahalanobis distances of the r units considered as potential outliers, the lower is the ratio considered in (3). The rejection area of the null hypothesis of the absence of outliers, therefore, is in the left tail of the distribution. The distribution of the ratio in equation (3) becomes difficult to estimate because, as it is possible to see from the results in the appendix, the Mahalanobis distance of unit i excluding unit j is equal to a complicated expression. In addition, if the number of outliers is not known a priori, multiple deletions can become computationally cumbersome. These are the reasons why the former ratio has scarsely been used for multiple outlier detection.

ven if it seems
n qualitatively
s to the elliptic
ned above are
ered as simple
r declearing a

ggested to use

(2)

tion, the ratio
uspect that r
ise if these r
d on the ratio
erminants and
(see proof in

$_{\text{l}}(n)\Big] \times$

(3)

obis distance
unit x_1. The
d as potential
ction area of
is in the left
equation (3)
see from the
i excluding
if the num-
can become
y the former
1.

However, given that we start with a subset surely free from outliers, we can monitor the value of expression (3) in each step of the forward procedure. A simple plot which reports on the x-axis the size of the subset and on the y-axis the former ratio must present an upward jump in correspondence of the inclusion of the first outlier. More generally: the inspection of this plot enables us to monitor the effect of the inclusion of each unit on the former ratio.

Other quantities which we found useful to monitor during the forward search are:

1.
$$\mathrm{tr}(\hat{\Sigma}(m)) / \mathrm{tr}(\hat{\Sigma}(n)) \,. \tag{4}$$
where symbol $\mathrm{tr}(\hat{\Sigma}(m))$ denotes the trace of the covariance matrix based on a subset of m observations. An upward jump in this plot points out that we have included an observation wich causes a strong increase in the variance of the variables.

2.
$$d_{[m+1]}(m) \,, \tag{5}$$
where symbol $d_{[m+1]}$ denotes the $(m+1)$-th ordered Mahalanobis distance. Usually this distance refers to the smallest among those of the group of potential outliers. (This might not happen only when more than one unit joins the subset at the same step). The plot which reports on the y-axis this distance and on the x-axis the number of units forming the subset should increase monotonously and present a peak in correspondence to the step prior to the inclusion of the first outlier. A subsequent decrease in the curve is generally due to the masking effect.

3.
$$d_{[m]}(m) \,. \tag{6}$$
Usually this distance refers to the largest Mahalanobis distance among those of the subset. The plot of this quantity must present a peak in correspondence of the inclusion of the first outlier.

4.
$$d_{[m+1]}(m) - d_{[m]}(m) \,. \tag{7}$$
This quantity must present a peak in correspondence to the step prior to the inclusion of the first outlier.

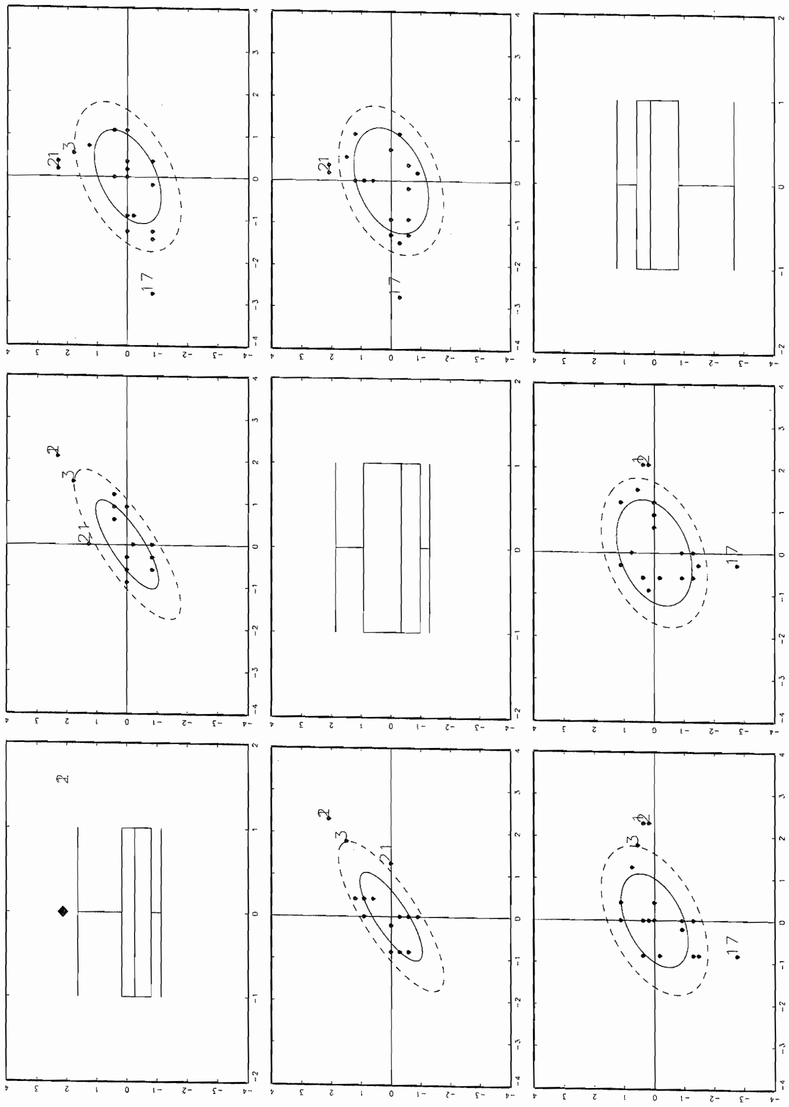


Fig. 1. 50% and 75% confidence ellipses of Mahalanobis distances from a robust centroid on standardized stack-loss data.

5.

An ou
the va
distan
with t
of an
report
becaus
distan

In the
equations
ing", "trac
ing", "gap
monitoring

In ord
in each st
associates
(or more g
the data) s
the words

Finall
based on a
ranked in
present a
outliers.

4. Exam

We ha
literature c
ing effects
al. simula
cerns Haw
on the for
The analy

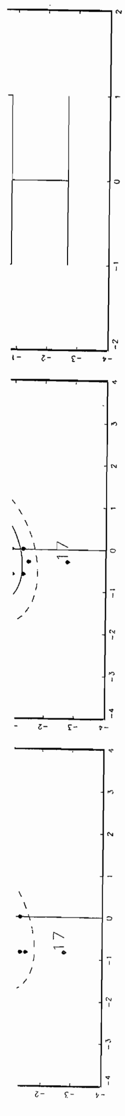


Fig. 1. 50% and 75% confidence ellipses of Mahalanobis distances from a robust centroid on standardized stack-loss data.

5.

$$\text{(a)}\quad \frac{\sum_{j=1}^{m} d_{[j]}(m)}{m}; \qquad \text{(b)}\quad \frac{\sum_{j=1}^{m-1} d_{[j]}(m)}{m-1}. \qquad (8)$$

An outlier is likely to produce an increase in the variance of the variables once it is included into the subset. In Mahalanobis distance the deviation of the unit from the centroid is weighted with the inverse of the covariance matrix, therefore the inclusion of an outlier is likely to cause a sharp decrease in the quantities reported in equation (8). Formula (8b) must be preferred to (8a) because it does not consider in the computation of the mean the distance associated to the unit just included (potential outlier).

In the next section we will refer to the seven plots associated with equations (3)-(8) respectively with the words: "determinant monitoring", "trace monitoring", "minimum monitoring", "maximum monitoring", "gap monitoring", "average monitoring" and "trimmed average monitoring".

In order to have a complete picture of the Mahalanobis distances in each step of the forward seach we can also produce a plot which associates a curve to every unit. The curves referred to the outliers (or more generally to the observations which lie far from the bulk of the data) stand apart from the others. We will refer to this plot with the words "distance monitoring".

Finally, another way to monitor Mahalanobis distances can be based on a plot which reports the values of the Mahalanobis distances ranked in non decreasing order ("scree monitoring"). This plot must present a sharp increase — an "elbow" — in correspondence to the outliers.

4. Examples

We have applied our method to some data sets well known in the literature concerning outliers, in which there are masking and swamping effects. They are known by the following names: Hawkins *et al.* simulated data, Body and brain weight, Stack loss data. As concerns Hawkins *et al.* simulated data, our procedure like all those based on the forward search enables to immediately detect the atypical cases. The analysis of the body and brain weight data through the technique of the bivariate boxplot can be found in Zani, Riani and Corbellini (1998). Here we simply concentrate on the stack loss data.

Stack loss data set (Chatterjee and Hadi, 1987, p. 228; Atkinson, 1985, p. 129), contains the values of 3 explanatory variables and one dependent variable obtained from 21 days of operation of a plant for the oxidation of ammonia to nitric acid. Similarly to Hadi (1992) we concentrate just on the 3 explanatory variables. This data set is peculiar, because Mahalanobis distances calculated on all the observations would not show any outlier. Hadi (1992) showed that observations 1, 2, 3 and 21 must be considered as multivariate outliers.

In this paper we use the simple version of bivariate boxplot based on robust confidence ellipses because we want to show the power of our method even in its simplest version. Figure 1 shows 50% and 75% confidence ellipses for each pair of variables and the boxplots on the main diagonal. The 75% threshold leads us to initially exclude from the initial subset units 1, 2, 3, 17, 21. If we want to start with an initial subset of smaller dimension we simply must decrease the outer threshold. For example, an initial subset of dimension 8 corresponds to a threshold of 55%. After finding this initial subset, we iteratively increase it by one unit as described in section 2 up to reach the end of the sample. Figure 2 shows statistics (3), (5), (6) and (7). Panel (*a*) (determinant monitoring) shows a change in slope passing from $m = 17$ to $m = 18$. Panel (*b*) and (*d*) (minimum and gap monitoring) point out that the maximum in each of the two curves is reached when $m = 17$. Finally, panel (*c*) (maximum monitoring) shows an upward jump passing from $m = 17$ to $m = 18$. Figure 3 shows statistics (8a) and (8b). Panel (*a*) (average monitoring) and even more clearly panel (*b*) (trimmed average monitoring) show a downward jump passing from $m = 17$ to $m = 18$. The joint examination of all the plots seems to point out clearly that the units included for $m > 17$ cause a strong modification in the statistcs. Table 1 reports the units included in the last 12 steps of the forward search. In fact, the units included in the last 4 steps of the forward search refer to the 4 outliers.

TABLE 1: UNITS INCLUDED IN THE 12 STEPS OF THE FORWARD SEARCH.

Steps	9-10	10-11	11-12	12-13	13-14	14-15
Unit included	9	4	8	7	10	18
Steps	15-16	16-17	17-18	18-19	19-20	20-21
Unit included	19	17	21	3	1	2

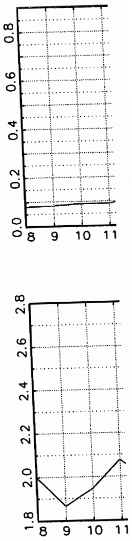


Fig. 2. De
(c)

The p
figure 4 (d
tances ref
search. Th

I. *the un*
examp
units

II. *what*
tially
ture s
mono
unit i
plot"
ing f
is co
influ
of fi
chan
most

and Corbellini
data.
228; Atkinson,
iables and one
a plant for the
1992) we con-
set is peculiar,
rvations would
ations 1, 2, 3

boxplot based
the power of
ows 50% and
l the boxplots
itially exclude
want to start
must decrease
dimension 8
initial subset,
ction 2 up to
; (3), (5), (6)
ange in slope
minimum and
he two curves
n monitoring)
18. Figure 3
ing) and even
v a downward
ination of all
d for $m > 17$
orts the units
act, the units
the 4 outliers.

14-15 18
20-21 2

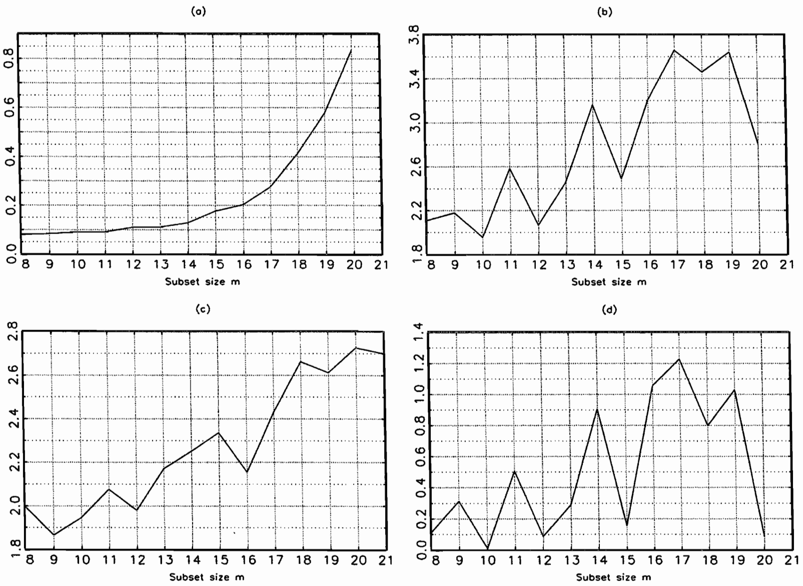


Fig. 2. Determinant monitoring" (a), "minimum monitoring" (b), "maximum monitoring" (c), "gap monitoring" (d).

The plots of figures 2 and 3 must be examined together with figure 4 (distance monitoring plot) which reports the Mahalanobis distances referred to the all the 21 units in each step of the forward search. This figure enables us to detect:

I. *the units whose Mahalanobis distance is far from the others.* For example, up to $m = 16$ it is evident that the curves referred to units 2,1,3,21 and 17 stand apart from the others;

II. *what are the units whose Mahalanobis distance decreases substantially once they are included into the subset.* For example this picture shows that Mahalanobis distance of unit 17 seems to decrease monotonously from $m = 11$ to $m = 17$. Table 1 shows that this unit is included when $m = 17$. In fact the "distance monitoring plot" shows a big decrease in the curve referred to unit 17 passing from $m = 16$ to $m = 17$. This plot also shows that unit 17 is consistent with the others when $m = 17$ and is not a highly influential observation because, as we have seen from the plots of figures 2 and 3, its inclusion does not cause an appreciable change in the statistics. However, unit 17 must be considered the most remote after excluding the four outliers. More generally:

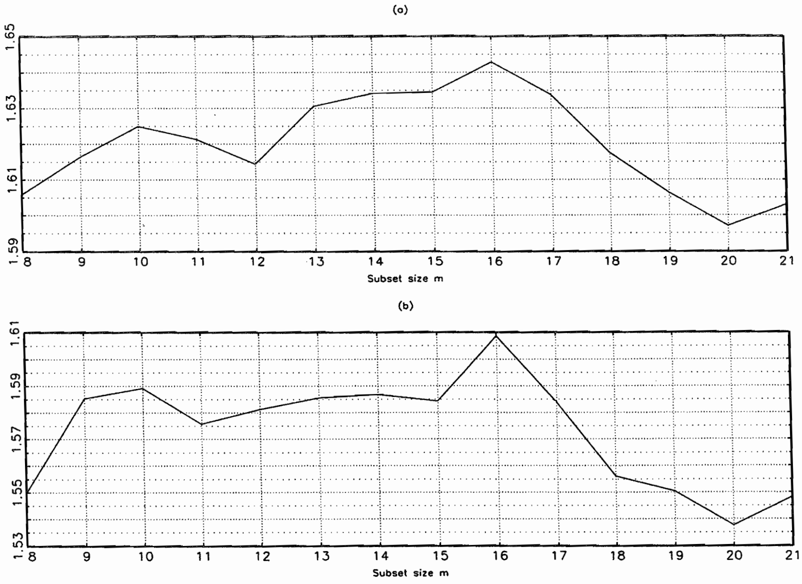


Fig. 3. Average monitoring" (a), "trimmed average monitoring" (b).

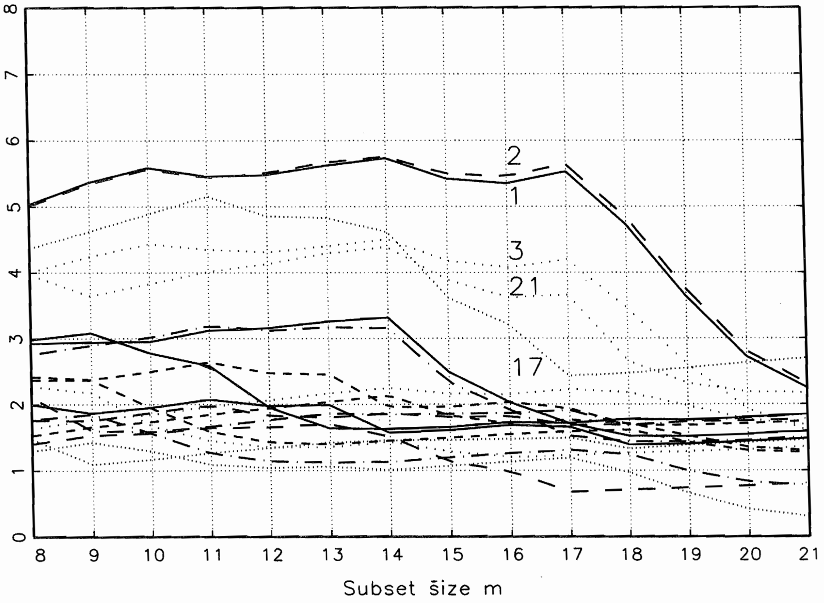


Fig. 4. "Distance monitoring"

the exa
units a
their de

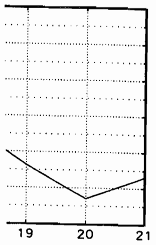
III. *the pre.*
biggest
in almc
distanc
only th

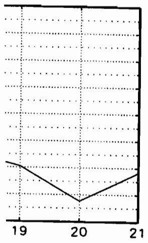
IV. *eventu*
that the
are ver
This hi
shows
ellipses
is alwa
include
Riani a
to link
origina
physica

5. CONCLU

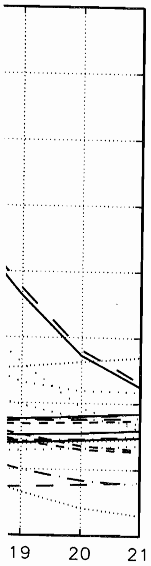
In this
lier detectio
which orde
ness. Our
and does
cesful in f
same time
computatio

Result
tics monit
are both
interpret.
inal data.
atypical v
procedure





' (b).



the examination of the table which reports the order in which the units are included enables to rank the observations according to their degree of outlyingness;

III. *the presence of masking effect.* For example, when $m = 21$ the biggest Mahalanobis distance is referred to unit 17. Unit 3, which in almost all the steps of the forward search shows a Mahalanobis distance which stands apart from the others, when $m = 21$ has only the 11-th Mahalanobis distance;

IV. *eventual clusters of outliers.* For example it is immediate to see that the curves referred to units 1 and 2 (the last 2 units included) are very close to each other in all the steps of the forward search. This highlights that these two atypical units are similar. Figure 1 shows that these two observations always lie outside the 75% ellipses and are very close to each other. In other words: there is always a strong link between the order in which the units are included and the characteristic of the original data. As noted by Riani and Atkinson (1998), the forward search algorithm enables to link the effect of each observation back to features of the original data. In this way every perturbation can be traced to a physical source.

5. Conclusions

In this paper we have suggested a simple iterative method for outlier detection. Our technique is based on a forward search algorithm which orders the observations according to their degree of outlyingness. Our method of choice of the initial subset is easy to implement and does not involve computational problems. Moreover, it is succesful in finding an initial subset which is not too small and at the same time is free from atypical observations. This provides noticeable computational savings when the sample size is large.

Results are displayed through simple graphs of a variety of statistics monitored along the forward algorithm. These graphical displays are both powerful in revealing the structure of the data and easy to interpret. Influential observations can be related to patterns in the original data. The application to a cornerstone data set in the literature of atypical values (the stack loss data) has shown the advantages of our procedure with respect to traditional backward methods.

Appendix: relation between Mahalanobis distance and the determinant of covariance matrix

In this appendix we prove equation (3). The first step consists in obtaining the relation between $\hat{\Sigma}_{(k)}(n-1)$ (the covariance matrix excluding the k-th row of the data matrix), and $\hat{\Sigma}(n)$. After simple algebra we obtain that:

$$\hat{\Sigma}_{(k)}(n-1) = \hat{\Sigma}(n)\frac{n-1}{n-2} - \frac{n}{(n-1)(n-2)}\{x_k - \hat{\mu}(n)\}\{x_k - \hat{\mu}(n)\}' \quad (9)$$

where

$$\hat{\Sigma}_{(k)}(n-1) = \frac{1}{n-2}\sum_{i(\neq k)=1}^{n}(x_i - \hat{\mu}_{(k)}(n-1))(x_i - \hat{\mu}_{(k)}(n-1))' ,$$

$$\hat{\mu}_{(k)}(n-1) = (\hat{\mu}_{1(k)}, \ldots , \hat{\mu}_{p(k)})' ,$$

$$\hat{\mu}_{j(k)} = \frac{1}{n-1}\sum_{i(\neq k)=1}^{n} x_{ij}, \qquad j = 1, 2, \ldots , p .$$

Calculating the determinant of both sides of equation (9) and recalling the properties of this operator (e.g. Mardia *et al.*,1979; p. 457) shows that:

$$|\hat{\Sigma}_{(k)}(n-1)| = |\hat{\Sigma}(n)|\left(\frac{n-1}{n-2}\right)^{p}\left[1 - \frac{n}{(n-1)^2}d_k^2(n)\right] . \quad (10)$$

Going back iteratively excluding each time one additional unit we end up with equation (3). The right hand side of equation (3) can be modified by calculating the relation between the square of the Mahalanobis distance of the j-th unit with the centroid calculated excluding the k-th unit, using quantities based on the whole sample.

In order to find this last relation we have to apply the inverse to both sides of equation (9). Using a standard inversion lemma (e.g. Chatterjee and Hadi, 1987; p. 21) we obtain:

$$\hat{\Sigma}_{(k)}(n-1)^{-1} = \hat{\Sigma}(n)^{-1}\frac{n-2}{n-1} + \frac{\frac{(n-2)n}{(n-1)}\hat{\Sigma}(n)^{-1}(x_k - \hat{\mu}(n))(x_k - \hat{\mu}(n))'\hat{\Sigma}(n)^{-1}}{(n-1)^2 - n d_k^2(n)} . \quad (11)$$

This result
p. 213). Pr
it by $(x_j -$

$d^2_{j(k)}(n-1)$
$= (n-2)\{d$

where $d_{j(k}$
calculated
can interpr

From equa
and $d_k^2(n)$
As a corol
Mahalanob
we find th

Atkinson, A
phica

Atkinson, A
JAS

Atkinson,
Mul

Atkinson,
vari
Env

Barnett, V
318

Barnett,
Yo

This result is comparable with that reported by Krzanowski (1988; p. 213). Premultiplying equation (11) by $(x_j - \hat{\mu})'$ and postmultiplying it by $(x_j - \hat{\mu})$, after tedious but simple algebra we find that:

$$d^2_{j(k)}(n-1) = \\ = \frac{(n-2)\{d^2_k(n) + d^2_{jk}(n)[2(n-1) + nd^2_{jk}(n)] + d^2_j(n)[(n-1)^2 - nd^2_k(n)]\}}{(n-1)^3 - n(n-1)d^2_k(n)}, \qquad (12)$$

where $d_{j(k)}(n-1)$ is Mahalanobis distance of unit j, with centroid calculated excluding unit k, and $d^2_{jk}(n)$ is a bilinear form which we can interpret as an interaction term between the 2 units:

$$d^2_{jk}(n) = (x_j - \hat{\mu}(n))'\hat{\Sigma}(n)^{-1}(x_k - \hat{\mu}(n)).$$

From equation (12) we can see that $d^2_{j(k)}(n-1)$ depends both on $d^2_j(n)$ and $d^2_k(n)$ (as we could expect), but also on the bilinear form $d^2_{jk}(n)$. As a corollary of equation (12), putting $j = k$, that is calculating the Mahalanobis distance of unit k after excluding it from the centroid, we find the same result obtained by Atkinson and Mulira (1993):

$$d^2_{k(k)}(n-1) = \frac{n^2(n-2)d^2_k(n)}{(n-1)^3 - n(n-1)d^2_k(n)}. \qquad (13)$$

REFERENCES

Atkinson, A.C. (1985) *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Claredon Press, Oxford.

Atkinson, A.C. (1994) Fast Very Robust Methods for the Detection of Multiple Outliers, *JASA*, 89, 1329-1339.

Atkinson, A. C. and Mulira, H.-M. (1993) The Stalactite Plot for the Detection of Multivariate Outliers, *Statistics and Computing*, 3, 27-35.

Atkinson, A. C. and Riani, M. (1997) Bivariate Boxplots, Multiple Outliers, Multivariate Transformations and Discriminant Analysis: the 1997 Hunter Lecture, *Environmetrics*, 8, 583-602.

Barnett, V. (1976) The Ordering of Multivariate Data (With Discussion), *JRSS*, A, 139, 318-339.

Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*, 3rd edn., Wiley, New York.

Chatterjee, S. and Hadi, A.S. (1987) *Sensitivity Analysis in Linear Regression*, Wiley, New York.

Dasgupta, S. (1993) The Evolution of the D^2-Statistic of Mahalanobis, *Sankhya: The Indian Journal of Statistics*, special volume dedicated to the memory of P.C. MAHALANOBIS, pp. 442-459.

Golberg, K.M. and Iglewicz, B. (1992) Bivariate Extensions of the Boxplot, *Technometrics*, vol. 34, n. 3, pp. 307-320.

Hadi, A.S. (1992) Identifying Multiple Outliers in Multivariate Data, *JRSS*, B, vol. 54, pp. 761-771.

Hampel, F.R., Ronchetti, E.M. Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

Krzanowski, W.J. (1988) *Principles of Multivariate Analysis, A User's Perspective*, Claredon Press, Oxford.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*, Academic Press, Londra.

Mitchell, A.F.S. and Krzanowski, W.J. (1985) The Mahalanobis Distance and Elliptic Distributions, *Biometrika*, 72, pp. 464-467.

Riani, M. and Atkinson , A.C. (1998) A unified approach to multivariate transformations and multiple outliers, submitted.

Riani, M., Zani, S. and Corbellini, A. (1997) Robust Bivariate Boxplots and Visualization of Multivariate Data, *in:* I. Balderjahn *et al.* (eds.) *Classification, Data Analysis and Data Highways*, Springer Verlag, Berlin, pp. 93-100.

Rousseeuw, P.J. (1985) Multivariate Estimation with High Breakdown Point, *in Mathematical Statistics and Applications* (eds. W. Grossmann, G. Pflug, I. Vincze and W. Wertz), vol. B, Dordrecht, Reidel, pp. 283-297.

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

Rousseeuw, P.J. and van Zomeren, B.C. (1990) Unmasking Multivariate Outliers and Leverage Points, *JASA*, vol. 85, pp. 633-651.

Small, C.G. (1990) A Survey of Multidimensional Medians, *Intern. Stat. Rev.*, 58, n. 3, pp. 263-277.

Wilks, S.S. (1963) *Mathematical Statistics*, Wiley, New York.

Zani, S., Riani, M. and Corbellini, A. (1997) New Methods for Ordering Multivariate Data, *Proceedings of the VIII International Symposium on Applied Statistic Models and Data Analysis*, Invited and Specialized Sessions Papers, Napoli, pp. 347-354.

Zani, S., Riani, M. and Corbellini, A. (1998) Robust Bivariate Boxplots and Multiple Outlier Detection, *Computational Statistics and Data Analysis* (in press).