{ file praha.tex for *Prague Stochastics* & *Information Theory, Aug. 23–28, 1998* }

FINDING MULTIVARIATE OUTLIERS BY DYNAMIC GRAPHICS

Anna Bartkowiak

Institute of Computer Science, University of Wrocław Przesmyckiego 20, 51–151 Wrocław, Poland
e-mail: aba@ii.uni.wroc.pl

Abstract

AMS91 classification: 62-07, 62-09, 68U05
Key words: multivariate outlier, rotation, graphical methods, grand tour, linked plots, ellipse of concentration

We present shortly the most important methods used for detecting multivariate outliers, among them the recently developed grand tour method based on dynamic graphics.

The advantage of the graphical approach is stressed.

An example – the bushfire data – is analyzed using some sequentially trimmed Mahalanobis distances and the grand tour. The last approach provides much better insight into the shape and content of the analyzed cloud of data points.

1 THE PROBLEM

When performing a statistical analysis one should be very careful not to permit the analysis be influenced by some outliers hidden in the data. The outliers may mask each other and be difficult to detect. There are many examples that a careless data analysis may yield quite wrong solutions. The statisticians have been giving much attention to this problem and there is a large amount of publications on this topic, also several books (see e.g. [12, 8, 2, 3] for further references).

It is said that an outlier is an observation (data vector) not matching the pattern suggested by the majority of the observations. Generally outliers represent data vectors atypical in magnitude or interdependence structure of their components. The outlyingness may be due to a specially large error in measurement, a typing error, or just coming from an atypical object.

The methods considered in the statistical literature have been based mostly on an approach exploiting covariance matrices of the considered data. An index of outlyingness was then constructed indicating which observations (data vectors) might be suspected as 'outliers'. The difficulty was that the covariance matrix, when evaluated from the entire data set with the non-identified outliers included, might yield quite wrong information and indication for the plausible outliers. Therefore much attention was given to find a 'clean' data set, not contaminated by the outliers, and next use that set as a start for building a robust covariance matrix.

With the developments of the nowadays information technology a breakthrough in analysis of observational data has occurred; in particular some special techniques for visualizing multivariate data have been developed. One of them is the *grand tour* proposed firstly by Asimov [1], and next elaborated and implemented a.o. by Tierney [14]. The method was specifically adapted for detecting multivariate outliers by Bartkowiak & Szustalewicz [4].

In the following we firstly recall in section 2 some of the recent mathematical approaches used for finding outliers. These are the methods proposed by Hadi [8], Atkinson [2], and Rocke & Woodruff [12].

In Section 3 we present briefly the dynamic graphic approach as proposed by Bartkowiak & Szustalewicz [4]. The method is illustrated in section 4 by the bushfire data used in [2, 12].

2 SOME MATHEMATICAL MODELS

Let $\mathbf{A}$ be a positive definite symmetric $p \times p$ matrix. The Mahalanobis distance between points $\mathbf{x} = (x_1, \ldots, x_p)$ and $\mathbf{y} = (y_1, \ldots, y_p)$ with respect to $\mathbf{B}$ is defined as

$$d^2_{\mathbf{B}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})\mathbf{B}^{-1}(\mathbf{x} - \mathbf{y})^T. \qquad (1)$$

One says that the matrix $\mathbf{B}^{-1}$ constitutes the metric (distance measure) in R^p for calculating the square (Mahalanobis) distances.

Now suppose, we have a sample of n points from R^p and it happens that a majority of the points come from a well-behaved population with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. These points are called "good points" and constitute a fraction $1 - \epsilon$ of the entire data set ($\epsilon > 0$).

The remaining fraction ϵ comes from a different population with mean $\boldsymbol{\mu}_0 + \boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$.

Then the expected sample mean $(\bar{\mathbf{x}})$ and covariance matrix $(\mathbf{S})$ – evaluated from the mixture – are [12]:

$$
\begin{aligned}
E(\bar{\mathbf{x}}) &= \boldsymbol{\mu}_0 + \epsilon\boldsymbol{\mu}, && (2)\\
E(\mathbf{S}) &= (1-\epsilon)\boldsymbol{\Sigma}_0 + \epsilon\boldsymbol{\Omega} + \epsilon(1-\epsilon)\boldsymbol{\mu}^T\boldsymbol{\mu}. && (3)
\end{aligned}
$$

It is obvious that any inference using $\bar{\mathbf{x}}$ and $\mathbf{S}$ is truly based on the mixture of the good and bad points and as such it may be unable to discover the bad points.

Rocke & Woodruff [12] discuss in more detail which are the worst cases of outliers do detect. They show that the hardest kind of outliers to identify is that kind that has a covariance matrix with the same shape as the good data. Such outliers are called often *shift outliers*. The same authors give also formulae for $E\{d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_0 + \epsilon\boldsymbol{\mu})\}$, the expected values of square Mahalanobis distance evaluated for the mixture composed from the "good" and "bad" points — in the metric $\boldsymbol{\Sigma}$, with $\boldsymbol{\Sigma}$ denoting the expected covariance matrix of the mixture , when the good and bad points follow **multivariate normal** distributions $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $N(\boldsymbol{\mu}_0 + \epsilon\boldsymbol{\mu}, \lambda\boldsymbol{\Sigma}_0)$ respectively.

Much attention was given to find (establish) a subset containing good points only ("clean" subset) and obtain on their basis some estimates of the parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ useful for constructing square Mahalanobis distances

$$
d^2_{\boldsymbol{\Sigma}_0} = (\mathbf{x} - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\mathbf{x} - \boldsymbol{\mu}_0)^T \qquad (4)
$$

for all data vectors (i.e. the good and bad as well) with the hope that the calculated distances will permit to detect in such way the bad points. Widely in use for that purpose are the methods based on elemental sets, like the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) [13] combined with random search, some iterative reweighting schemes (e.g. method proposed by Campbell, 1980, see Maronna & Yohai [11]) and some ascending methods beginning with a small "clean" subset (preferably near the median) and augmented sequentially by adding to the clean subset other data vectors not contradicting the postulate of a homogeneous subset (see e.g. the papers by Hadi [8, 6, 9, 10], the stalactite plot by Atkinson [2]). There are also some hybrid methods, see e.g. Rocke & Woodruff [12].

Generally these methods proceed as follows:

Substituting in (4) for $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ the values $\bar{\mathbf{x}}_0$ and $\mathbf{S}_0$ obtained from a clean subset, the Mahalanobis distances for all data points are calculated. The calculated distances follow – under the assumption that the clean subset is representative for the good points, which in turn are representative for the population under investigation – approximately the χ^2 distribution with p degrees of freedom (this is in principle true for data having multivariate normal distributions or not differing much from such distributions). A quantile from the distribution χ^2_p, e.g. a quantile of order $\beta = 0.95$ or $\beta = 0.99$, may serve as a guideline, whether the evaluated points are distant from the center $\bar{\mathbf{x}}_0$. This may be shown graphically by drawing a *Q–Q plot* with the χ^2 distribution as a reference function ([7], pp. 228–230). The $Q-Q$ plot with the χ^2 distribution taken as the reference distribution is called also *chi2–plot*. It is constructed in such way, that in the "y" axis we mark the calculated squared distances, and in the "x" axis the respective quantiles from the corresponding χ^2 distribution.

If the referenced distribution suits the data, then the plot exhibits a linear pattern. Points not matching the linear pattern are considered as outliers.

Some illustrations containing chi2–plots are shown in Fig. 1 in section 4.

3 A DYNAMIC GRAPHICS APPROACH

When finding the outliers analytically, we come to the situation: Applying one or more methods we get some data vectors indicated as "bad points" which do not match the pattern of the part of the data described as "good points".

Working analytically we are never completely sure whether we got a clean subset and whether we have found all the atypical points. Our analysis would appear more trustworthy, if we could somehow display graphically the analyzed data set and the identified outliers. For two- or three-dimensional data vectors we may do that by displaying the data points in a scatterdiagram (2D graph) or a spin-plot (3D graph).

What about a truly multidimensional data, i.e. with more than 3 dimensions?

A method which permits to perceive points located in a multivariate space is the *grand tour*.

The concept of the grand tour has appeared firstly in a paper by Asimov [1]. His idea was to obtain a dense set of projections which would become dense in the manifold of all projections, thus permitting to obtain views which contain possibly many (a dense set of) views of the data points included in the analyzed data cloud.

Asimov proposed some algorithms for performing a sequence of projections with the desired property. Next, in subsequent years, the problem was elaborated by other researchers (for references, see [4]), a.o. also by Tierney [14], who implemented in XLisp–Stat a simple procedure, `tour-plot`.

The approach of [4] follows that of Tierney's. The approach is based on the concept of *rotation*, which may be considered both from an algebraical and from a geometrical point of view.

Algebraically rotation is considered as a transformation of the data matrix $\mathbf{X}$ by an orthogonal matrix $\mathbf{A}$. Applying the transformation $\mathbf{A}_{p\times p}$ to the data matrix $\mathbf{X}^{(0)}_{n\times p}$ we obtain a transformed matrix $\mathbf{X}^{(1)}_{n\times p} = \mathbf{X}^{(0)}_{n\times p}\mathbf{A}_{p\times p}$.

Geometrically rotation is considered as a physical act of rotating the space in which the data points are located. The data cloud may be imagined as enclosed in a p–dimensional sphere (ball). Somewhere outside the sphere an observer is located. The observer has at any moment a view of all the points; his sees the projection of all the points onto a 2–dimensional plane perpendicular to the direction of his look. To obtain different views the observer has (a) either to move round the sphere, or (b) to rotate the sphere. In both cases the observer perceives some projections of the data points onto some 2–dimensional planes.

The method elaborated by Bartkowiak & Szustalewicz [4] combines both the algebraical and the geometrical approach. The procedure `out-tour`, developed by these authors, works as follows:

1. A direction $\vec{UV}$ established by choosing randomly some points U and V on the p–dimensional hypersphere is established. Given the points U, V and α, the rotation angle, the rotation matrix $\mathbf{A}$ is constructed.
2. The actual data matrix $\mathbf{X}_{n\times p}$ is transformed yielding $\mathbf{X}^{(tr)}_{n\times p}$:
$$\mathbf{X}^{(tr)} = \mathbf{X}\mathbf{A}.$$
3. A scatterplot of the points $(x^{(tr)}_{i1}, x^{(tr)}_{i2})$, $i = 0, \ldots, (n-1)$ is drawn.
4. A concentration ellipse is superimposed on the scatterplot. The role of the ellipse is to focus our attention on the points located far – in the Mahalanobis metric – from the data center. The size of the ellipse is determined by a value q_β, taken as the βth quantile from the χ^2 distribution. Such an ellipse is expected to cover a β part of the displayed points.
5. A linked count–plot is opened in the screen. The count–plot is constructed as an indexplot, where for each i ($i = 0, \ldots, n-1$) the frequency for the i-th point of falling beyond the borders of the concentration ellipse is shown.

Performing such rotations in sequence we notify more and more points, which – in various rotations – appeared at outlying positions.

After performing several hundreds (or thousands) of such rotations, we have a record of data points which appeared at least one time at outstanding position, thus might be considered as suspected outliers.

On the other hand, points which were not notified as falling beyond the border of the concentration ellipse (i.e. with frequency count equal zero) may be considered as typical ones. They may constitute a "clean", homogeneous subset of data.

4 SOME APPLICATIONS

4.1 The stackloss and the Hawkins-Bradu-Kass data

The grand tour method was tested in [4] by using two data sets, which are generally considered as bench mark data for detecting outliers. These were the stackloss data and the Hawkins-Bradu-Kass data. Taking a 95% confidence ellipse, all outliers were identified correctly and without doubts, so no further confirmative stage for these data sets were needed.

4.2 The bushfire data

The data were originally used by Campbell (CSIRO, Australia). We have taken their values from [11]. The data contain values of satellite measurements on five frequency bands, corresponding to each of 38 pixels. The observations had as aim to locate bushfire scars. The data were analyzed also by Maronna & Yohai [11] and Rocke & Woodruff [12]. It was found that the data set contains outliers which appear in two subgroups, denoted in the following as sets

S5 – containing the points 7, 8, 9, 10, 11,

S7 – containing the points 32, 33, 34, 35, 36, 37, 38.

The outliers belonging to the set *S5* are relatively easy to detect by various robust methods. The set *S7* is difficult to detect, because it is masked by the set *S5*.

Sequentially trimmed Mahalanobis distances

We have firstly evaluated Mahalanobis distances using the ordinary covariance matrix evaluated from the entire data set. The respective chi2-plot (using χ^2 quantiles evaluated on 5 degrees of freedom) is shown in Fig. 1, upper plot (with y-axis labelled as *mm1*).

The outstanding position of points no. 7, 8, 9 is clearly visible. Let us recall that the critical quantiles from the χ^2_5 distribution are: for $\beta = 0.95$: $q_{0.95} = 11.07$ and for $\beta = 0.99$: $q_{0.99} = 12.83$. Clearly the squared distances evaluated for the points 7,8,9 are higher than these values.

After trimming off (i.e. removing) the 3 points we have calculated a new mean and a new covariance matrix and used them in turn for calculations of Mahalanobis distances in the revised metric. The new distances are shown in Fig. 1 (plot with the y-axis labelled as *mm2*).

One can see now, that for the points no. 8 and 9 the Mahalanobis distances – when evaluated in the new metric – have grown considerably, up to values about 50.0 or 60.0 shown in the y-axis. Apart from them, three other points (no. 10, 11, 7) have emerged at very outstanding position and with values certainly above the critical quantiles.

After removing all the 5 points (i.e. no. 7–11) we obtain a much more differentiated chi2–plot (*mm3*) with Mahalanobis distances approaching the value 400.0 for the extreme points. However, the outliers from the set $S7$ were not detected.

The chi2–plot for Mahalanobis distances evaluated from a covariance matrix, obtained from a reduced data set with observations no. 7–11 and 32–38 removed, is shown in Fig. 1. bottom plot. The outlyingness of points belonging to the subsets $S5$ and $S7$ is clearly visible.

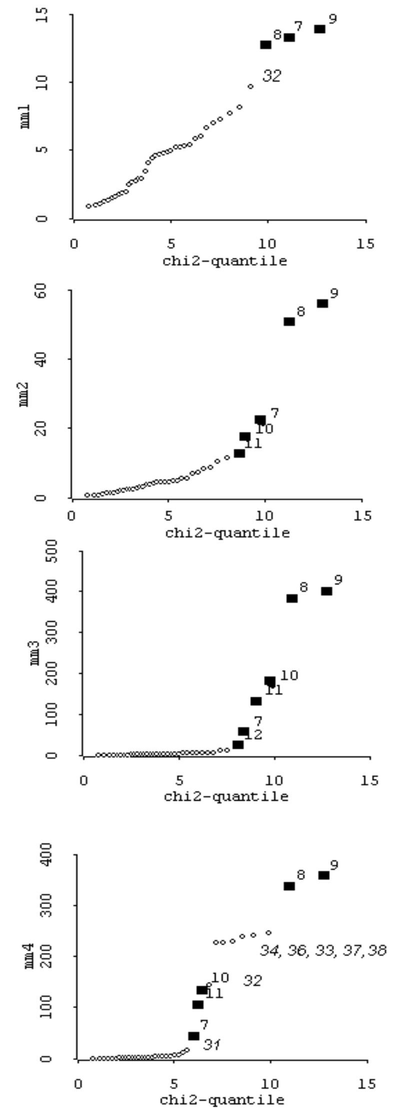


Figure 1. Chi2 plots for Mahalanobis distances constructed with covariance matrices calculated from trimmed data. From top to bottom:
set mm1 – entire data set, no trimming,
set mm2 – points 7,8,9 removed,
set mm3 – points 7–11 removed,
set mm4 – points 7–11 and 32–38 removed.

Using the grand tour

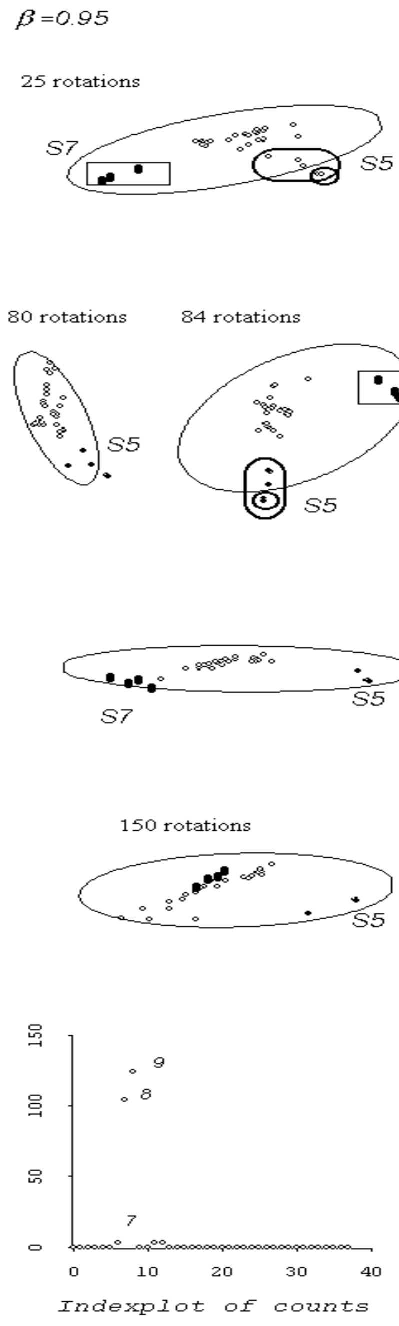
When using the dynamic graphic approach, the structure of the bushfire data may be revealed quite easy . Running the grand tour algorithm we see already after 25 rotations the structure of the data. It is shown in Fig. 2, upper plot. A more clear differentiation of the sets $S5$, $S6$ and the remainder was obtained after 84 rotations – see Fig. 2. Another instances, when one of the subsets is overshadowed by the good points, are shown in other plots in Fig. 2. One may see clearly, that the data set contains two smaller subsets, differing from the main part of data. In some projections these subsets are visible, in other not. In some of the concentration ellipses in Fig. 2 we have framed the subset *S7* with a rectangle, the subset *S5* with an oval, and the points 8, 9 by a circle.

The illustrations in Fig. 2 were obtained when using concentration ellipses expected to cover 95% of all points.

In the bottom of Fig. 2 we show the indexplot of frequency counts – captured after *ca.* 150 rotations. One may see, how frequently each of the 38 points was notified as falling beyond the borders of the concentration ellipse.

Looking at the count plot one can see, that the points 8 and 9 were quite frequently notified at outstanding positions. There were single instances, when also the points 7, 12 and 13 were notified. None of the outliers from the subset *S7* was identified as outlier. To detect these outliers we should draw smaller ellipses, or – alternatively, use a robustified covariance matrix for construction of the concentration ellipse.

In Fig. 3 we show views obtained when using ellipses of concentration with expected coverage of $\beta = 0.90$ and $\beta = 0.85$.



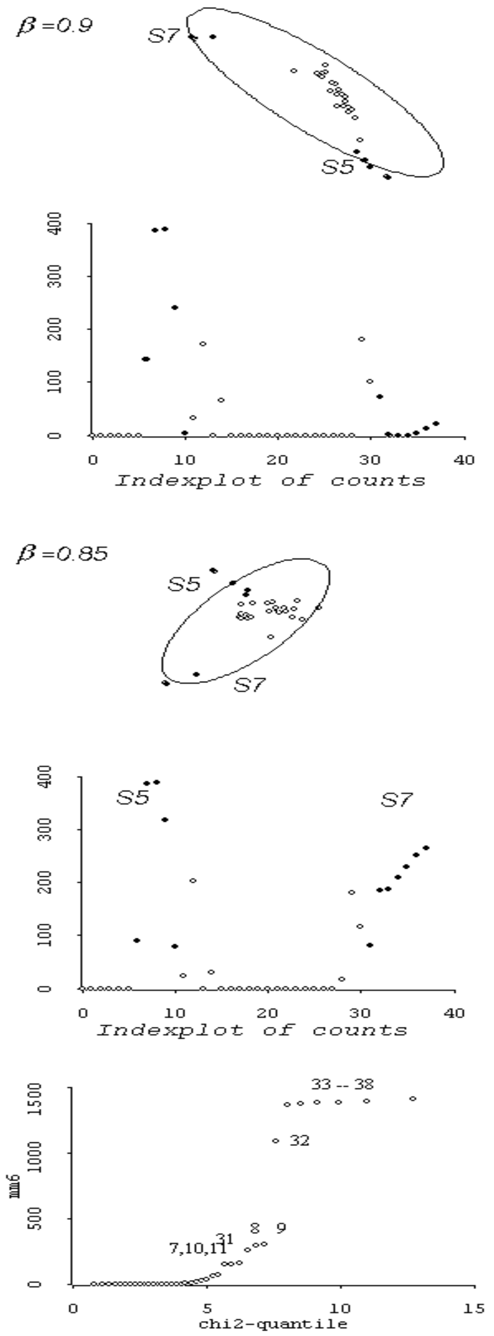


Figure 2. Some ellipses of concentration – with expected coverage $\beta = 0.95$ – displayed when running the grand tour for the bushfire data.
Set S5 (oval) contains the points 7–11. Set S7 (rectange) contains the points 32–38.
The lower indexplot displays, for a given point, how frequently this point was notified outside the concentration ellipse.

Figure 3. Some ellipses of concentration – with expected coverage $\beta = 0.90$ (top) and $\beta = 0.85$ (middle) – and count plots displayed when running the grand tour for the bushfire data.
Bottom: Chi2-plot of Mahalanobis distances obtained when the metric was evaluated from a clean subset taken as points with 0 frequencies exhibited by the middle count plot

When using the ellipse determined by $\beta = 0.85$ all the outliers from the subsets $S5$ and $S7$ were diagnosed as suspected outliers (i.e. they appeared at least once outside the concentration ellipse).

Taking the points with zero counts as the clean set we have evaluated on their basis the covariance matrix **S** and taken its inverse as the metric for a new evaluation of the Mahalanobis distances. The respective chi2-plot is shown in in the bottom plot of Fig. 3. One may recognize clearly the atypicalness of the points belonging to the subsets $S5$ and $S7$.

5 FINAL REMARKS

The method of the grand tour, as proposed and implemented in the procedure `out-tour` [4], permits to gain some insight into the shape of a multivariate data cloud.

It was the goal of many researchers to obtain a clean set of data vectors. Such a clean set could then be used to construct a kind of robust covariance matrix. The grand tour procedure permits to discover the hidden interrelation structure for data having elliptical or nearly elliptical distributions. For many situations it permits to obtain a clean subset. For other situations the original grand tour, as described by [4, 5], should be combined with constructing concentration ellipses based on robust covariance matrices and overlaid on the two-dimensional projection planes. In any case the problem of constructing multivariate (robust) covariance matrices of size $p \times p$ is reduced to the problem of constructing multivariate (robust) covariance matrices of size 2×2, which is a greatly reduced task.

What concerns the suspected outliers, another confirmative stage of the analysis is needed – to confirm their outlyingness and throw more light on the fact that they appeared at outlying positions. A work on this topic is in progress.

Acknowledgement

The work on the grand tour algorithm was partially sponsored from the (Polish) KBN grant no. 8T11C 04011 elaborated in the years 1997–1998.

References

[1] Asimov D.: The grand tour. A tool for viewing multidimensional data. *SIAM J. Sci. Stat. Comput.*, 6, 1985, 128–143.

[2] Atkinson A.C.(1994): Fast very robust methods for the detection od multiple outliers. JASA 89, 1329-1339.

[3] Bartkowiak A.: Some basics for detecting multivariate outliers in regressional context. *Biocybernetics and Biomedical Engineering*, 17, 1997.

[4] Bartkowiak A., Szustalewicz A.: Detecting outliers by the grand tour. *Machine Graphics & Vision*, 6, 1997, 487–505.

[5] Bartkowiak A., Szustalewicz A.: Watching steps of a grand tour implementation. Machine Graphics & Vision, 1998, submitted, Manuscript 1–19.

[6] Chatterjee S., Hadi A.S., Mächler M.: A reweighted least squares method for robust regression estimation and outlier selection. 1998. Manuscript 1–17.

[7] Gnanadesikan R. : *Methods for Statistical Data Analysis of Multivariate Observations.* Second edition. Wiley 1997.

[8] Hadi A.S. : Identifying multiple outliers in multivariate data. *J. R. Statist. Soc., B, 1992, 54, no. 3, 761–771*

[9] Hadi A.S., Velleman P.F.: Computationally efficient adaptive methods for the identification of outliers and homogeneous groups in large data sets. Paper read at: *Joint Statistical Meetings, Anaheim/Orange County, 1997, August 10–14, Session 246.*

[10] Hadi A.S., Simonoff J.: A more robust outlier identifier for regression data. *Bull. of the Int. Stat. Inst.*, 1997, Vol. LVII, 51st Session Istanbul, Contr. papers, Book 2, 281–282.

[11] Maronna R.A., Yohai V.J.: The behaviour of the Stahel-Donoho robust multivariate estimator. *JASA* V.90, 1995, 330–341.

[12] Rocke D.M., Woodruff L.: Identification of outliers in multivariate data. *JASA* V.91, 1996, 1047–1061.

[13] Rousseeuv P.J., Leroy A.M.: *Robust Regression and outlier detection.* 1997, Wiley N.Y .

[14] Tierney L.: *Lisp–Stat, an Object–Oriented Environment for Statistical Computing and Dynamic Graphics.* Wiley, New York 1990.