file ank2.tex September 02, 2000
to appear in *Biometrical Letters*, Vol. 38 (2000), No. 1

A Semi-Stochastic Grand Tour for Identifying Outliers and Finding a Clean Subset

Anna Bartkowiak

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

Running title: Semi–Stochastic Grand Tour for Identifying Outliers

Summary

The grand tour method with the adds proposed by Bartkowiak & Szustalewicz (1997) has proved to be a very efficient method in detecting outliers.

The present paper proposes further modifications of the grand tour algorithm by constructing robust concentration ellipses. It is also emphasized that the same method can be used for obtaining a "clean" data set. Such a subset may be the start point for robust multivariate procedures.

The method is simple, can be easily implemented on parallel computers, and as such may be used in data mining for large data sets.

The considerations are illustrated with two benchmarks and one real medical data set.

Key Words and Phrases: multivariate outlier, graphical methods, grand tour, linked plots, ellipse of concentration

1 Introduction

An outlier is an observation (data vector) not belonging to the pattern suggested by the majority of the observations. The outlyingness may be due to a gross error in measurement, a typing error, or just an atypical object.

Outliers may be very influential for models constructed on the basis of the analysed data; they may distort the estimated relationships among the considered variables (changing, for example, the magnitude and signs of the calculated correlation coefficients), they may yield quite wrong parameters characterizing the established functions.

Therefore before starting the proper data analysis one should check always the data for outliers.

The problem: how to find and identify outliers hidden in the data was and is still extensively considered and discussed in the statistical literature.

In the following in Section 2 a very short survey of the main recognized methods for outlier detection is presented. The methods outlined in that section are based mainly on some mathematical and statistical considerations.

In Section 3 the grand tour method (Asimov 1985, Bartkowiak & Szustalewicz, 1997) is shortly presented. The grand tour is based mainly on a graphical procedure. Bartkowiak & Szustalewicz (1997, 1998) have added to the classical concept a stochastic element: the concentration ellipse constructed using a confidence parameter. Such a supplemented grand tour is called a semi–stochastic grand tour. So far the semi–stochastic grand tour was considered primarily as a tool to identify suspected outliers. However, the same method may serve also to other purpose: after changing eventually some parameters of the algorithm a 'clean' subset of data may be obtained. This is important for data analysis, because a clean data set is the start point for many robust procedures. The need for a clean subset was stated a.o. by Hadi (1992), Billor and al. (2000), Riani and Atkinson (2000).

Next, Section 4 illustrates the practice of finding outliers and clean subsets: we consider here two typical benchmarks: the Bradu-Hawkins-Kass and the modified wood gravity data (source: Rousseeuw & Leroy, 1987). Additionally a real medical data set, the LIEB125 data, is considered.

In Section 5 final remarks and conclusions are given.

2 A short survey of statistical methods for outlier detection

2.1 Presenting some widely known methods for outlier detection

The problem of detection or identifying outliers has been notified since long; may be so long as data analysis is performed. Many methods have been proposed for this topic and have existed for many years; also many of them have been reported to fail for some cases, especially in detecting clusters of outliers, which may mask each other.

Most of the proposed methods start with calculation of the covariance matrix (ordinary, censored or robustified). The hidden atypical observations may inflate or swamp the covariance matrix, and obtaining a clean subset containing only typical points may be not so easy. A survey of early methods may be found in Gnanadesikan and Kettenring (1972), then in the book by Barnett and Lewis (3rd edition, 1994). More recent results and references to other relevant results may be found in the papers by Rocke and Woodruff (1996), Billor, Hadi and Velleman (2000) and Riani and Atkinson (2000).

Methods, which have gained in the last decade a considerable popularity, are:

1. Mahalanobis distances calculated from robust covariance matrices obtained from weighted data vectors (Huberized or Hampelized wights), see e.g. Campbell (1980).

2. MVD or MCD, i.e. minimum volume determinant or minimum variance determinant methods, proposed by Rousseeuuw and coauthors (see, e.g., Rousseeuuw and Leroy, 1987) with several subsequent modifications, a.g. Rousseeuuw & van Zomeren (1990), Rousseeuuw & and van Driessen (1999).

   A feasible algorithm for a MCD estimator in multivariate data was proposed by Hawkins (1994).

3. Stalactite plots, proposed by Atkinson, see e.g. Atkinson and Mulira (1993), Atkinson (1994).

4. BACON, i.e. blocked adaptive computationally efficient outlier nominators, proposed by Hadi and coauthors (see Hadi and Velleman, 1997, Billor, Hadi and Velleman, 2000).

5. Hybrid algorithm incorporating affine–equivariant methods and random restarts, see Woodruff & Rocke (1993), Rocke and Woodruff (1996).

What concerns the first two enumerated methods – experience has shown that they have failed in many situations and are not thrustworthy. In particular, the MVD and MCD methods, as based on combinatorial evaluations have proved to be computationally expensive and for larger number of variables – unfeasible; also they were often not able to detect *all* outliers, especially when these have occurred in masking clusters; see Billor and. al. (2000) for further references where criticism of the methods may be found.

The three remaining methods are more thrustworthy. BACON relays on a 'clean' starting set, while the stalactite and the hybrid algorithm perform many random restarts.

2.2 Probabilistic evaluations

All the methods presented in the preceding subsection perform at some stage of their work some evaluations based on statistical tests. In particular they evaluate the significance of the found outliers by calculating their Mahalanobis distances from a (robust) center of the analyzed data cloud.

The evaluated distances are supposed to have (at least asymptotically) a χ^2_p distribution, with p denoting the number of the considered variables.

Just this supposition is a week point in the mathematical approach to identify outliers. The presumption that Mahalanobis distances have a χ^2 distribution is valid only for data having multivariate normal distribution; which happens rarery in practice.

A remedium for this handicap (i.e. non–normality of the data) might be: to apply to the data some transformation, which would bring the observed distribution nearer to normality. This topic has been pursued since long; see e.g. Atkinson (1987), Velilla (1995), Riani and Atkinson (2000). However - it has to be said – that the transformed data may be difficult for interpretation.

2.3 The graphical approach

The idea of the graphical approach is to perform a kind of visualization of the data and look directly whether there are points which are much outstanding from the main bulk of the data. Widely in use are two-dimensional scatterplots constructed from observations on pairs of variables; also scatterplots constructed from principal coordinates.

Three-dimensional displays (of 3 variables) can be constructed using interactive graphics displaying views on a computer screen. Spin plots are a well known technique used for that purpose. Another technique for displaying multivariate data are scatterplot matrices. One of the earliest implementations of spin plots and scatterplot matrices appeared in XLispStat (Tierney, 1990). Today these techniques are implemented in most of statistical commercial packages.

Another technique, not so widely in use, is parallel coordinate plot. One early implementation of that method may be found in XLispStat (Tierney, 1990). A specific example illustrating suspected outliers by parallel coordinate plots is shown in Bartkowiak, Piotrowicz and Szustalewicz (1999), where also further references may be found.

A still more interesting technique is the grand tour (Asimov, 1985). It permits to obtain sequentially, as in a continuing movie, two-dimensional instances of views from (in) multivariate space.

Quite a different class of methods is yielded by the neural network methodology. Specially the Self-Organizing-Maps (SOMs) are very promissing in this branch of methods; see the papers by Morlini (1998) and Bartkowiak, Piotrowicz and Szustalewicz (1999) for some applications of self-organizing maps in finding outliers.

All the graphical methods, when implemented in an interactive computer environment, benefit much from such possibilities as linking objects and graphs, selecting, brushing and coloring, carried out in an interactive mode.

Using these additional possibilities detecting suspected outliers and verifying them may be carried out quite effectively.

3 The grand tour with concentration ellipses

3.1 The concept of the grand tour and one its implementation

The concept of the grand tour has appeared firstly in the paper by Asimov (1985). The idea was to obtain a sequential set of projections which would become dense in the manifold of all projections, thus permitting to obtain views which would contain possibly many (a dense set of) views of the data points contained in the analyzed data cloud. This goal may be realized in a variety of ways. Several proposals may be found in Asimov's paper.

The problem was elaborated in subsequent years by other researchers (for references, see e.g. Bartkowiak and Szustalewicz, 1997). Among others Tierney (1990) designed in XLispStat a simple procedure, **tour–plot**, for that purpose. The approach of Bartkowiak and Szustalewicz (1997) follows that of Tierney's, which means that their approach is based also on the concept of rotation.

Rotation may be considered both from a formal mathematical and from a purely geometrical point of view.

Mathematically rotation is considered as an algebraic 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 $\mathbf{X}_{n\times p}$ we obtain a transformed matrix $\mathbf{X}_{n\times p}^{transf} = \mathbf{X}_{n\times p}\mathbf{A}_{p\times p}$.

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 sees at any moment the projection of all the points onto a 2–dimensional plane perpendicular to the direction of his look, thus some points my be overshadowed by some others. To obtain different views the observer should either to move himself around the sphere, or to rotate the sphere. In both cases the observer has in view some dynamic projections of the data points onto some 2–dimensional planes.

The algorithm `out-tour` elaborated by Bartkowiak and Szustalewicz (1997) combines both the mathematical and the geometrical approach. They assume, that the observer is positioned in such way that he has in view the plane $< X_1, X_2 >$ spanned by the coordinate axes of the first two variables. The observer will see and perceive the points projected onto that plane. In practice the observer sees the plane on the screen of his computer and observes in such way what is going on when subsequent rotations are executed.

After rotating the sphere with the data points, but leaving the coordinate system unrotated, the fixed plane $< X_1, X_2 >$ will show the coordinates of the rotated points, and – in the same time, the projections of the rotated points onto $< X_1, X_2 >$. Practically, the 'old' content of this plane containing the projections of the points before the rotation will be cleared, and new content exhibiting projections of the rotated points will be redrawn. When the rotation (and the projections) are done in small steps, one may perceive the impression that points in the screen are moving.

Below we show main points of the algorithm.

Before beginning the actions we draw a scatterplot of the first two variables and a linked *count plot* exhibiting the values: $(i,\ count(i))$, $i = 0, 1, \ldots, (n-1)$, with $count(i) = 0$ for each i.

Next we repeat in turn cycles of actions composed of following steps :

1. The direction of the rotation and its angle α are established. Having these fixed the *rotation matrix* $\mathbf{A}$ is constructed.

2. The actual data matrix $\mathbf{X}$ is *transformed* yielding $\mathbf{X}^{(tr)}$:

$$\mathbf{X}^{(tr)} = \mathbf{X}\mathbf{A}. \tag{1}$$

3. The previous content of the exhibition in the observed scatterplot is cleared and a new content displaying the points $(x_{i1}^{(tr)}, x_{i2}^{(tr)}),\ i = 0, \ldots, n-1$, is drawn.

4. A *concentration ellipse* is superimposed onto the scatterplot displaying the projection of the transformed points. The role of this ellipse is to focus our attention on the points located far – in the Mahalanobis metric – from the data center.

   The concentration ellipse is given by the equation

$$(\mathbf{x} - \mathbf{m})\mathbf{S}^{-1}(\mathbf{x} - \mathbf{m})^T \leq \chi^2_{2,\beta} \tag{2}$$

   where: $\mathbf{x} = (x_1, x_2)$, $\mathbf{m} = (m_1, m_2)$ are the coordinates and their mean (median), $\mathbf{S}$ is the covariance matrix of $\mathbf{x}$, and $\chi^2_{2,\beta}$ is the quantile of order β in the chi–square distribution with 2 degrees of freedom.

   item 5

   The values of the variable count are updated: for points i $(i = 0, 1, \ldots, (n - 1)$ described by a vector $\mathbf{x}$ not satisfying (2) we sunstitute: $COUNT(i) := COUNT(i) + 1$. Next the content of the count plot is redrawn.

More detailed description of the algorithm out-tour may be found in (Bartkowiak & Szustalewicz, 1998), where also the derivation of the formula for the rotation matrix $\mathbf{A}$ is given. The same paper shows also results from simulations having as aim to find out how many cycles are needed to obtain a reasonable uniformity of the projections.

In Figure 1 we show three plots illustrating the idea of the grand tour – performed for the LIEB125 data (more details on this subset is presented in Subsection 4.3). The plots exhibit results obtained after the first cycle of the algorithm. In the left plot of that figure the scatterplot of the values $(x_{i1}^{(tr)}, x_{i2}^{(tr)}),\ i = 0, \ldots, n-1$, after the first rotation is exhibited. The middle plot of Figure 1 contains the concentration ellipse constructed on the basis of the displayed two-dimensional points. One may notice here that the points no. 3, 4, 42, 124 have appeared outside the concentration ellipse. Comparing this plot with that on the left one may state how much the concentration ellipse helps in fixing some outlying points.

The outlyingness of the four points can be also stated when looking at the count plot located at the right side of that figure.

Let us underline that the concentration ellipse and the count plot play a different role: The concentration ellipse (redrawn after each rotation and projection) shows, how much the given points are outstanding in the actual projection; the count

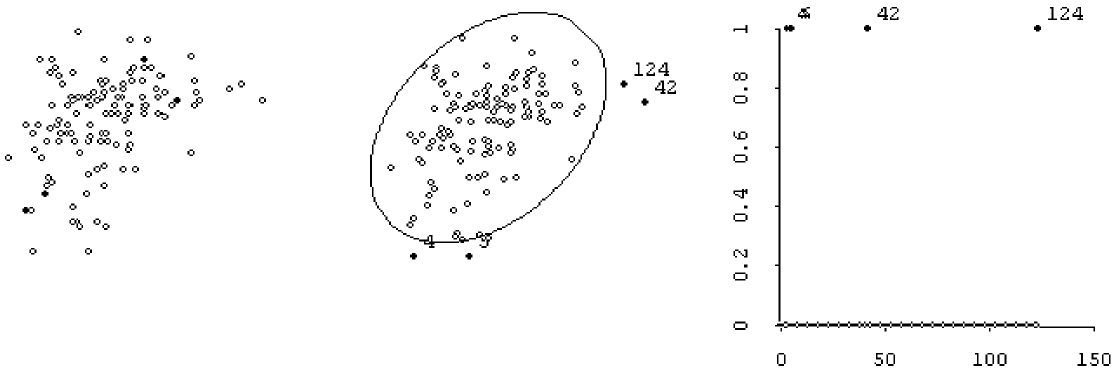

Figure 1: *The idea of the grand tour. View after first rotation. Left: Scatterplot of transformed values of the first two columns of the data matrix. To gain a clearer presentation, the axes of the scatterplots both in this and the middle plot are suppressed. Middle: The same, but with the 95 % ellipse of concentration superimposed and with identification of points located outside the ellipse. Right: Linked count plot, indicating that points no. 3, 4, 42, 124 were found for one time outside the concentration ellipse*

plot (actualized after each projection) shows how many times, i.e., in how many projections all the analyzed data points were outstanding so far.

For the situation exhibited in Figure 1, the count plot shows that the points no. 3, 4, 42, 134 have appeared once (for one time) beyond the ellipse border (this is o.k., because there was only one rotation performed so far).

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

After performing several hundreds 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, i.e. as representing data vectors atypical in magnitude or interdependence structure of their components.

In Figure 2 we show two instances of views produced by the grand tour. One can see that the data contains several clearly outstanding data points and some others that appear at the border of the main bulk of the data.

Looking at the count plots one may notice, that quite a large part of the data points has *zero* counts, which means that these points have never appeared outside the borders of the concentration ellipse. We propose to consider these points as constituting a 'clean' subset.

The grand tour has been successful applied to many sets of data. It has been very helpful in showing what kind of data we deal with. It has also clearly identified some outliers in several benchmarks and several real data sets. For the later it has been observed that the differentiation between the main bulk of data and some peripheral

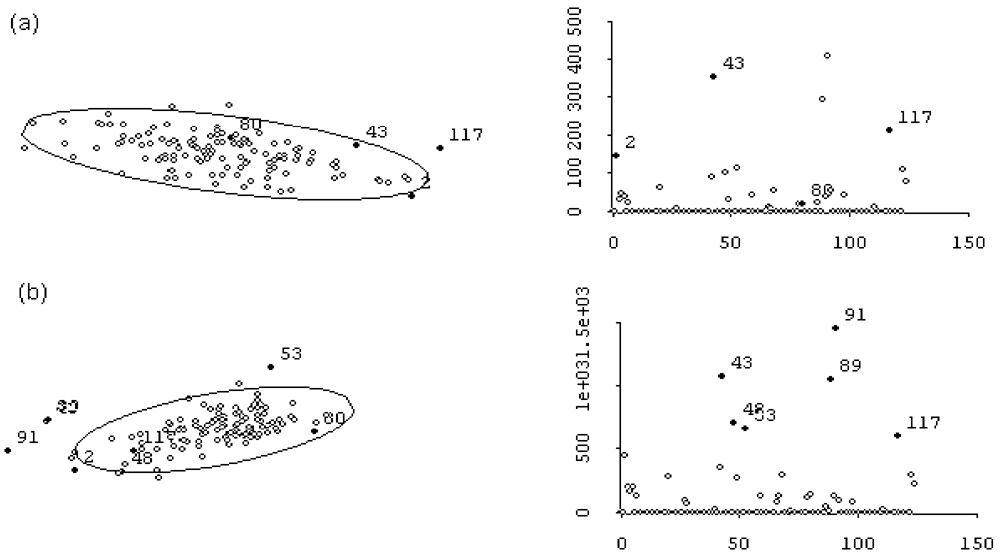

Figure 2: *Two views obtained when running the grand tour for the LIEB125 data. Left exhibits: ordinary concentration ellipses; right exhibits: corresponding count plots. Please notice, that most of the points in the count plots have zero counts, which means that they were always notified in the interior of the concentration ellipse*

observations is not so sharp and other criteria are needed to identify finally which observed data vectors are outliers.

What concerns the concentration ellipse, it has been stated, that the outliers hidden in the data may inflate the covariance matrix. The concentration ellipse evaluated from inflated covariances appears then too large and cannot sort out the outliers.

To get the outliers outside the borders of the ellipse – one has to diminish the confidence parameter β, say to $\beta = 0.90$ or $\beta = 0.75$, which puts the user in an uncomfortable situation. To copy with it, we propose to construct a robust concentration ellipse.

3.2 Construction of a robust concentration ellipse

Because in the out-tour algorithm the construction of concentration ellipses is carried out many times (hundreds or thousands), we need a relatively fast algorithm for calculation of the parameters and drawing the ellipse. Certainly, we can not use for that purpose extensively iterative algorithms, which need many iterations to attain convergence.

After analyzing several algorithms we have chosen an old method described in Gnanadesikan & Kettenring (1972), also Gnanadesikan (1997). The method calculates the elements of a covariance matrix pairwise, using sums and differences of the considered variables. The following identity (called in the following *DSD, Difference of Sum and Difference* is exploited:

$$cov(X_k, X_l) = 0.25[var(X_k + X_l) - var(X_k - X_l)],$$

with the symbol $var(\xi)$ denoting the variance of the random variable ξ.

The point is that the variances $var(X_k+X_l)$ and $var(X_k-X_l)$ can be estimated by a robust method, using e.g., the MAD statistics, i.e. the *M*edian of *A*bsolute *D*eviations from the median:

$$\widehat{var}\xi = MAD(\xi)/0.675 \tag{3}$$

with ξ denoting generally the computed variable.

In the paper we have simplified the calculations of the MAD by taking one half of the InterQuartile Range (IQR) of the computed variable. Thus we have applied the following formula for robust estimation of the variance of the variable ξ:

$$\widehat{var}\xi = IQR(\xi)/1.35 \tag{4}$$

The method explained above is very quick, however it may yield negative definite covariance matrices not suitable for calculation of $\mathbf{S}^{-1}$, needed for the concentration ellipse (see formula 2). A remedy for this would be to find a supplement Δ to $\mathbf{S}$, which makes $\mathbf{S} + \mathbf{\Delta}$ a positive definite matrix. In (Bartkowiak and Ziętak, 2000) it is shown how to find an appropriate $\mathbf{\Delta}$.

Meanwhile, we may state the following: Our ellipse is constructed from values x_{i1}, x_{i2} which were obtained as linear combinations of the original observations. By the central limit law – even if the original observations are coming from non-normal distributions, their linear combination might be quite near to the normal variate. This means practically, that we may be permitted to use use not so much sophisticated robust methods, and we may have hope that these less sophisticated methods will work quite efficiently when running the grand tour.

3.3 Search for a clean subset of data vectors

As is pointed out in the paper by Billor et al., 2000, the outlier detection methods provide the analyst with a set of proposed outliers, which next can be corrected (if identifiable errors are the cause) or taken out from the body of the data for a separate analysis. The remaining data then more nearly satisfy homogeneity assumptions and can be more safely analyzed with standard methods.

So the problem is to obtain a homogeneous data set not containing suspected outliers. Hadi (1992) calls such set a *clean* data set. Some recent algorithms (e.g. the general BACON algorithm, Billor et al., 2000; the forward search combined with the *fan plot*, Riani and Atkinson, 2000) start from a small 'clean' data set and then, after adding to that set a number of the remaining observations, perform a kind of testing, whether the data set with the added observations can stll be considered as a 'clean' data set.

The crucial step in this procedure is to find a really clean data set, not containing any masked outliers. Some analytical methods for finding such a basic clean subset are suggested by Hadi (1992) and Billor *et al.* (2000). The basic subset is usually small and next it is augmented by a kind of forward search.

We make another proposal: we find a clean subset from the results obtained by the grand tour method.

Let us look once more at the results of the grand tour displayed in the count plots visible in Figure 2. Remind that the ordinate (y) in the count plot tells us how many times a given point (identified by its id-number marked in the x-axis) has appeared outside the concentration ellipse.

Figure 2 was obtained for the LIEB125 data containing values for $n = 125$ patients characterized by $p = 9$ variables each. Looking at the count plots we see that:

- There is a dozen or more of points which have occurred frequently outside the concentration ellipse – thus these points are suspected to be outliers.
- A good deal of the points has zero frequency count: these are points which appeared always inside the concentration ellipse.

The proposal is: Take as the clean data set those data vectors which have *zero* frequency count in the count plot.

4 Practical examples

4.1 The Hawkins–Bradu–Kass data

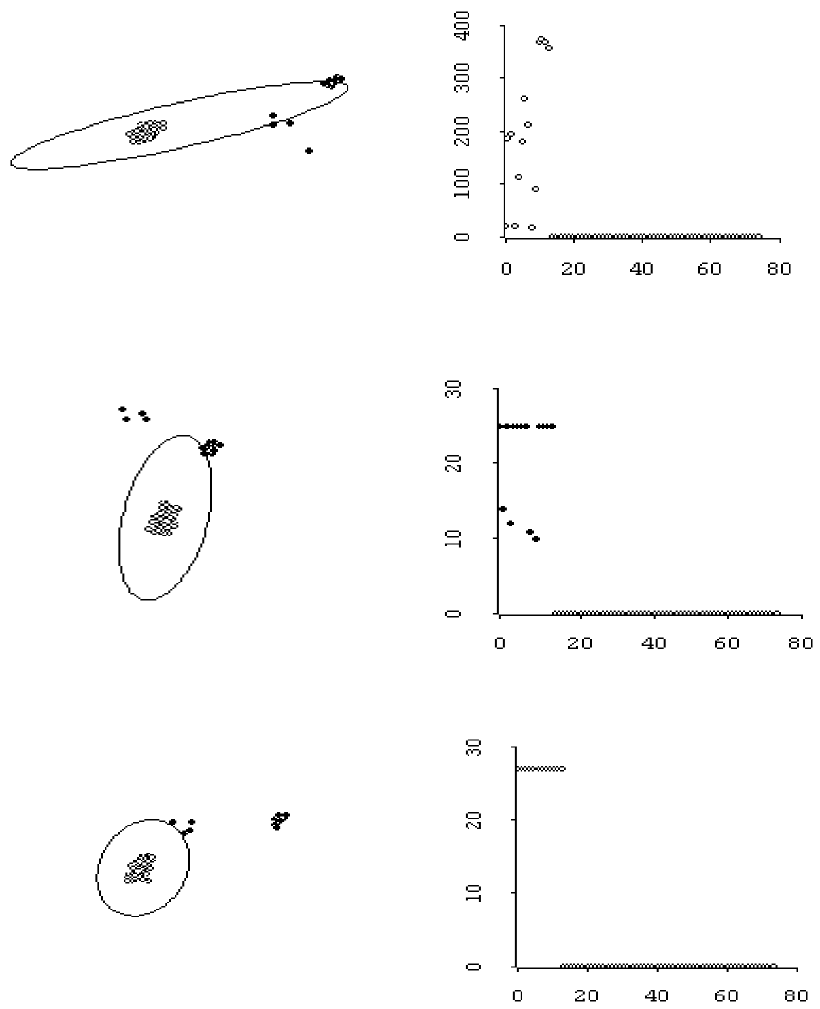

Figure 3: *Some views obtained when running the grand tour for the Bradu-Hawkins-Kass data. It can be clearly seen that the data set is composed from three clusters. Three variants of concentration ellipses were applied in separate runs: ordinary (top), positioned in median centre (middle), and using robust estimators (bottom). All 3 ellipses were obtained using the same confidence parameter* $\beta = 0.95$

These are artificial (synthetic) data constructed by Hawkins, Bradu, Kass (1984) with the special purpose to show a situation where truly influential data points are not detected by classical diagnostic – because of a masking effect. The data consists of n=75 cases (data vectors) each characterized by 3 explanatory and 1 predictor variable.

The source of the data may be found in the original paper by Hawkins, Bradu and Kass (1984), also in the book by Rousseeuw & Leroy (1987).

It provides a good example of a masking effect. The first 10 observations are bad leverage points – very influential for the fitted regression. They are masked by observations no. 11, 12, 13, 14, which are good leverage points. The data were analyzed a.o. by Hadi (1992), Atkinson & Mulira (1993), Hadi & Simonoff (1997), Hadi & Velleman (1997), Chatterjee, Hadi & Mächler (1997), Rocke & Woodruff (1996).

The out-tour with ordinary concentration ellipses evaluated for the confidence parameter $\beta = 0.95$ permits at once to see, what kind of outliers was implanted, which is shown in Figure 3, top plot. None the less, it can be seen, that the implanted outliers have influenced (enlarged) the ellipse. The formal identification of the outliers was possible after performing about 100 projections.

A better performance is achieved when positioning the ellipse in the median center of the projections. This is shown in Figure 3, middle plot.

Constructing a robust concentration ellipse outperforms the former two methods. The implanted data vectors are recognized at once.

All the three ellipses exhibited in Figure 3 were obtained using the same confidence parameter $\beta = 0.95$.

The points no. 15 – 74 points have never appeared outside the concentration ellipse and constitute, without doubts, a clean data subset. This is indicated in all count plots obtained when using any of the three variants of the concentration ellipse.

Summarizing:

- In this example the ordinary concentration ellipse permits – in a longer run – to identify the implanted 14 outliers. None the less, the other two variants (ellipse positioned in median center and constructed by the robust DSD method) do it much quicker.

- The points 15 – 74 (in the numeration starting from zero) constitute a clean subset.

4.2 The modified wood gravity data

The data consist of $n = 20$ observations with $p = 5$ explanatory variables and one response. The source for this example may be found a.o. in the book by Rousseeuw and Leroy (1987), where also the origin and history of the data is described. The primary task was to compute the regression of wood specific gravity (Y) in dependence on five anatomical factors called X_1, X_2, X_3, X_4, X_5. Four of the data vectors representing samples no. 4, 6, 8, 19 were specially contaminated and represent regressional outliers. Neither the hat matrix nor any other classical diagnostic is able to detect this fact, because the outliers are susceptible to a masking effect.

The data set was analyzed, a.o., by Atkinson (1994), Atkinson & Mulira (1993), Hadi & Simonoff (1997), Rocke & Woodruff (1996),

Applying the out-tour algorithm with ordinary ellipses is not helpful in identifying the outliers. The concentration ellipse constructed from ordinary covariances gets inflated by the implanted outliers. We can see in the projections that the implanted outliers constitute a separate cluster; none the less – when using ordinary ellipses – we are not able to identify them by the formal algorithm of the ordinary grand tour presented in Subsection 3.1.

However the robust concentration ellipse constructed by the DSD method, with confidence parameter $\beta = 0.95$, is able to identify the implanted outliers quite easy and without doubts. Only this situation is illustrated here.

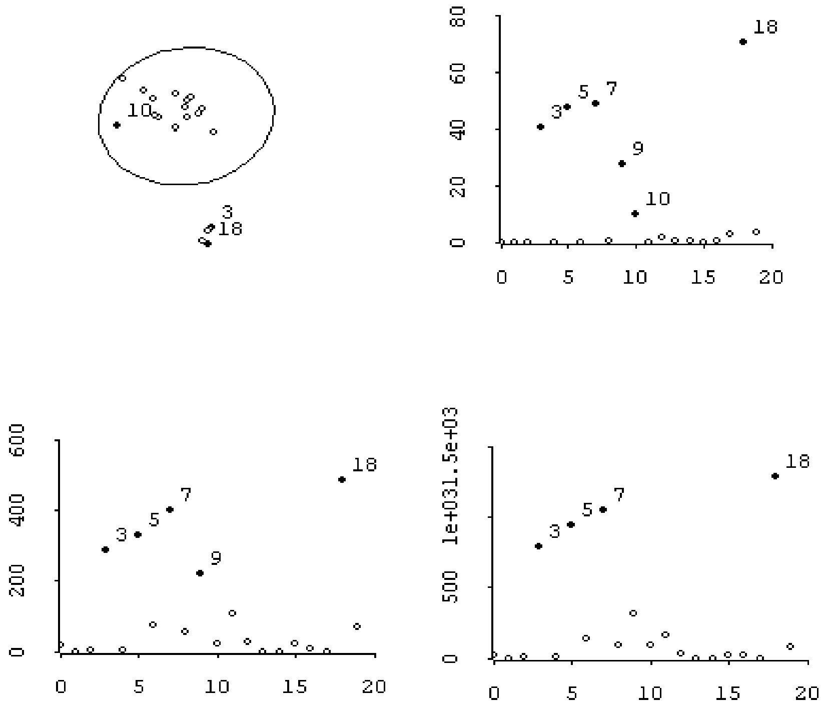

Figure 4: *Some views obtained when running the grand tour for the modified wood gravity data. Robust concentration ellipses with confidence $\beta = 0.95$ were drawed. The outliers can be identified without doubt. However, due to the shape of the data cluster, also other (i.e. non implanted) data points appear quite often outside the borders of the ellipse, which is recorded by the count plot. There are only 6 points left for constituting a clean subset of the data*

In Figure 4 we show views captured when running the grand tour and using robust concentration ellipse with confidence parameter $\beta = 0.95$. The implanted outliers can be seen in the projections at once. However, due to the shape of the data cluster, also other (i.e. non implanted) data points appear quite often outside the borders of the ellipse, which is recorded by the count plot. In particular, in Figure 4, top right, one may see that the count plot has recorded quite often also the points no. 9 and 10. This continues in subsequent projections, what can be seen in the exhibits bottom left and bottom right of the same figure. None the less, it can be notified, that the implanted outliers have a pronounced frequency in the count plots.

Only one point (out of the analyzed 20 points) has zero frequency count, which means that it was always notified as belonging to the interior of the confidence ellipse. Further 5 points could be indicated as having very small frequency counts. Thus – in that case – , and allowing also for very small frequency counts in the count plot, the clean subset could be constituted from 6 data points only: no. 1, and no. 2, 4, 13, 14, 17.

Summarizing:

- This specially prepared data example needs robust concentration ellipses to identify the implanted outliers.
- A clean subset of the data could be constituted by the data vectors no. 1, 2, 4, 13, 14 and 17.

4.3 The LIEB125 data

The data were recorded by prof. Liebhart from the Medical Academy of Wrocław. We use here only a part of the data, namely a matrix $\mathbf{X}$ of size $n \times p = 125 \times 9$ containing values for $n = 125$ patients. The patients have been diagnosed as having pulmonary malfunctioning named *obstruction* or *obturation*.

For each patient $p = 9$ variables are considered: (1) RV, Residual Volume, (2) Age, (3) Height, (4) VC, Vital Capacity, (5) VC%, percentage of due VC, called also predicted normalized vital capacity, (6) FEV1, forced expiratory volume in the first second, (7) FEF, forced expiratory flow at the level of 0.2–1.2 VC, (8) MMFR, maximal mid-expiratory flow rate, (9) MMFT, maximal mid–expiratory flow time.

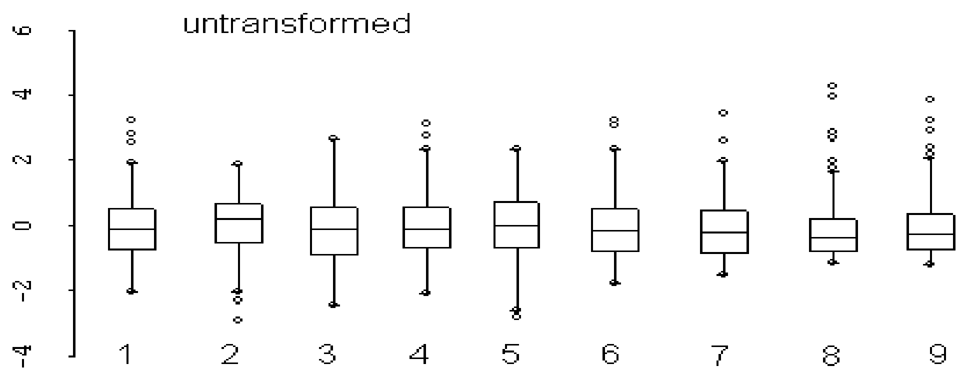

Figure 5: *LIEB125 data. Boxplots constructed from standardized values for nine variables. Standardization means subtracting from each column of the data matrix its mean and next divide the result by the column standard deviation*

The set of 9 observations for one patient will be in the following called also data vector or data point, and denoted – for the ith patient – as $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$.

Geometrically a data vector $\mathbf{x}_i$ means a point located in the p-dimensional space of the analyzed variables. Using the expression "p–dimensional data point" we have in mind a data point in R^p given by p coordinates.

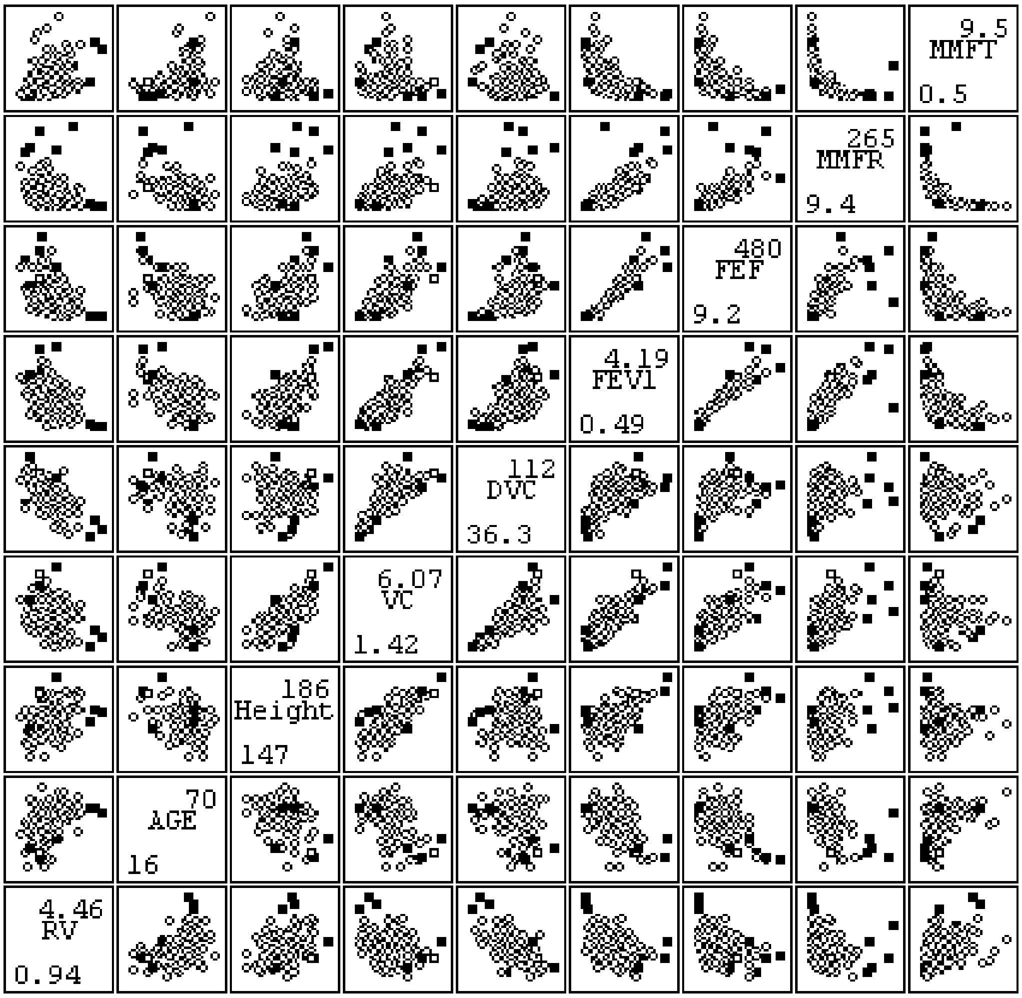

Figure 6: *Scatterplot matrix for the LIEB125 data with marked points no. 2, 42, 43, 48, 89, 91, 117, 123, 124, which appeared most frequently outside the concentration ellipse. One can see here, that generally the selected points held an extreme position.*

The data have been previously analysed by Bartkowiak, Liebhart and Szustalewicz (1997) in the context of finding interrelationships between the considered variables.

To illustrate the shape of the distributions of the investigated variables, we present in Figure 1 boxplots constructed from standardized values of the considered variables (we had to take standardized values, i.e. subtract the mean, and divide by the respective standard deviation, to make the boxplots presentable in one figure).

Looking at the plot one can see, that some of the variables exhibit a high degree of asymmetry. Several isolated points located outside the upper Tukey's fence (i.e. cases beyond $Q_3 + 1.5 \times IQR$) are clearly visible in the boxplots.

The scatterplot matrix for the analyzed data is shown in Figure 6. It contains the scatterplots of all pairs of the considered variables. Clearly the distributions for some pairs of variables are not normal.

To learn more about the shape of the data and identify outliers, if any, we have applied to this data the grand tour method.

In Figure 2 we have shown already two momentary views of the projection $R^9 \to R^2$, the respective plots exhibit the obtained ellipses of concentration and linked to them count plots.

The ellipses of concentration in Figure 2 are ordinary ellipses.

Observing the projections, and also the count plots, we see that the data constitute *in fact* one cloud of data, which is quite irregular at its borders. One may notice several points which appear at more outstanding position, but the gap between these points and the remaining ones is not big.

Next we have repeated the run of the grand tour with other setup: (a) the ellipses were positioned at the medians and (b) the ellipses were based on robust estimators as explained in Section 3.2 . Both variants gave similar results. In Figure 7 we show count plots obtained by these methods. Generally they look alike. There is no sharp border between the data points found at outstanding positions.

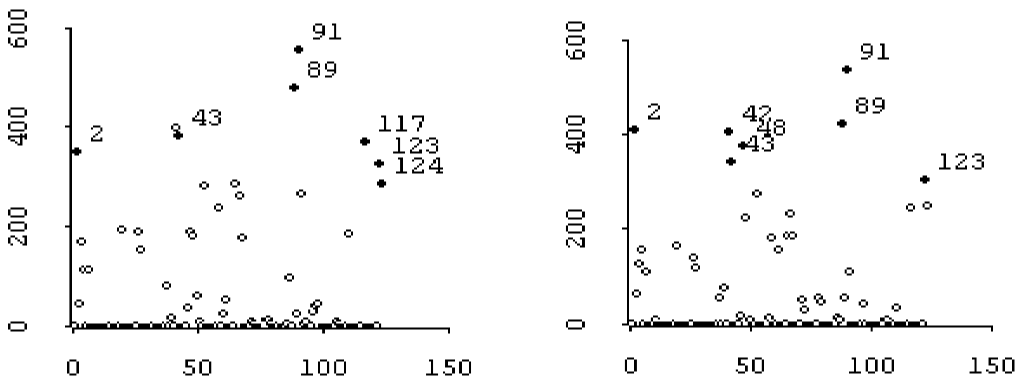

Figure 7: *Count plots for the LIEB125 data. Left: Robust concentration ellipses were used for recording outstanding points. Right: Ordinary concentration ellipses positioned at the medians were used to gather the points for the plot.*

On the basis of the count plots we have selected the following set of data points which have appeared most frequently outside the concentration ellipse: 2 42 43 48 89 91 117 123 124 (remind, the numeration of the analyzed data points starts from 0).

To see, what kind of points are these, we may look at the scatterplot matrix shown in Figure 6.

One may notice, that the selected points are mainly extreme points. There is no clear gap between these points and the main bulk of data. To decide whether these are truly outliers, we should use another criterion. In the present situation we may say only, that these are *suspected* outliers.

Concerning the clean subset: There are 67 data points which have never appeared beyond the border of the concentration ellipse – thus these points could be nominated as constituting a clean subset.

It may be also noticed that a large part of the bi–variate distributions exhibited in the scatterplot matrix does not seem to be distributed normally – thus probabilistic considerations based on normality assumptions may be doubtful.

Summarizing:

- For this real data set, gathered in an medical ambulatory environment, applying ordinary concentration ellipses is sufficient to learn about the shape of the data and identify some suspected outliers.
- The same ordinary concentration ellipses yield a clean subset composed of 67 data vectors.

Analysis of transformed data

It is known, that the scale of the variables may be responsible for the fact, that some data points appear – or do not appear – as outliers. Riani and Atkinson (2000, page 385) say directly: 'Outliers in one transformed scale may not be outliers in another scale'.

To bring the distributions nearer to normality, we have been looking for a transformation (Box–Cox transformation) which would yield the distributions of the variables from the LIEB125 data appearing more normal. Although there are specialistic methods for that purpose (see, e.g. Atkinson, 1994; Velilla, 1995; Riani and Atkinson, 2000) – we have done it by interactive graphics performed by the function `bcfun`, a slight modification of the function `bcdemo` offered by XLispStat (Tierney, 1990). As a result of this investigation we came to conclusion, that the Box-Cox transformation with a suitably chosen parameter λ helps to reduce the pronounced asymmetry of variables no. 2, 4, 6, 7, 8 and 9. The values of the appropriately chosen parameter λ are shown in the table below ($\lambda = 0$ means the logarithmic transformation):

Variable	(2)	(4)	(6)	(7)	(8)	(9)
	Age	VC	FEV1	FEF	MMFR	MMFT
λ	2	0.4	0.4	0.4	0	0

The transformations were applied to the raw (i.e. unstandardized) data. After doing that we have stated that the distributions became more symmetric and regular.

Next we have rerun the grand tour on these transformed data.

We obtained very similar results. The formerly (i.e. for the untransformed data) most frequently notified outstanding points appeared also to be frequently notified for the transformed data, albeit sometimes with a changed frequency of counts.

Figure 8 shows directly two indexplots exhibiting the frequency counts recorded for the untransformed and transformed data. One can see that, in principle, the notified points are the same in both plots, only the frequencies vary for some points. Thus, we may infer, that it was not the stated asymmetry of the distributions of the analyzed variables which made some points appearing at outlying positions.

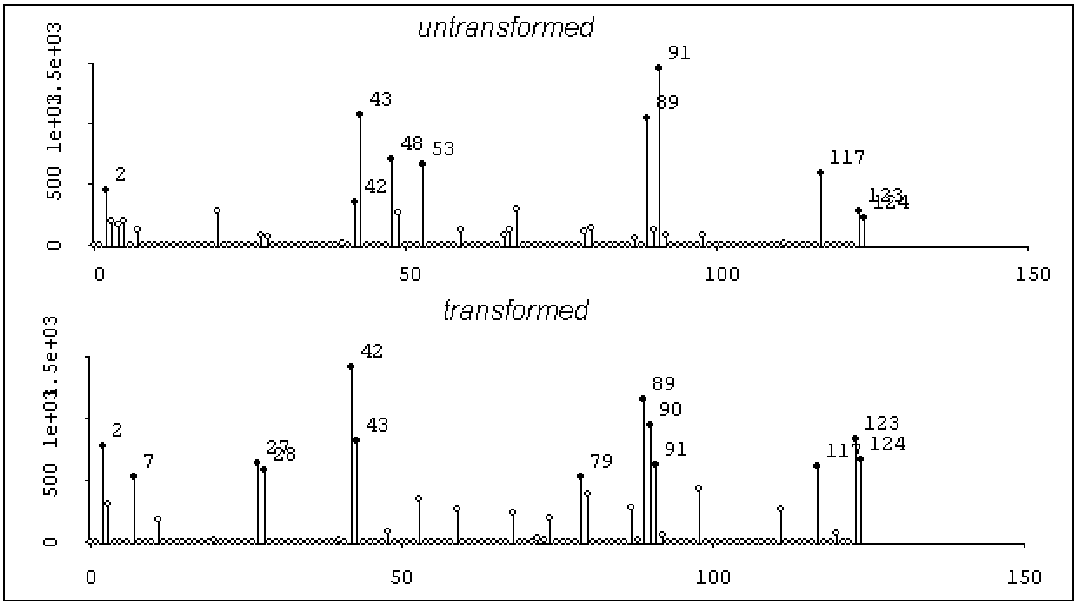

Figure 8: *Outliers detected in the original (upper plot) and transformed data (bottom plot). In the horizontal axis the id-numbers for the points–patients no. 0 through 124 are shown. The vertical spikes denote, how frequently the given point was notified as falling beyond the concentration ellipse.*

Looking at these plots one can see, that the points appearing in both plots with the highest frequency are the same ones.

5 Conclusions and final remarks

The method of the grand tour, with the proposed modification of constructing a robust concentration ellipse is a powerful tool for detecting suspected outliers in the data. The method may yield both suspected outliers and a clean data set. However – it has to be said – the method detects efficiently outliers from data which have approximately normal or elliptical shape.

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 further analysis of the found outliers – when clustering them by angular distances – is shown in Bartkowiak & Szustalewicz (2000).

Acknowledgements

The work was partially sponsored from the Polish State Committee for Scientific Research (KBN) grant no. 8 T11C 031 16.

The author wishes express her thanks to prof. Jerzy Liebhart from the Medical Academy of Wrocław, for lending the data analysed in this paper.

References

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

Atkinson A.C. (1994), *Plots, Transformations and Regression.* 2nd Edition. Oxford Science Publications, Clarendon Press, Oxford U.K.

Atkinson, A.C. (1994), Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association,* **89**, 1329–1339.

Atkinson A.C., Mulira H.M. (1993), The stalactite plot for the detection of multivariate outliers. *Statistics and Computing*, **3**, 27–35.

Barnett V., Lewis T. (1994), *Outliers in Statistical Data.* 3rd Edition. Wiley, Chichester.

Bartkowiak A. (1997), Some basics for detecting multivariate outliers in regressional context. *Biocybernetics and Biomedical Engineering*, **17**, 57–83.

Bartkowiak A., Szustalewicz A. (2000), Outliers – finding and classifying which genuine and which spurious. *Computational Statistics*, **15**, no. 1, 3–12.

Bartkowiak A., Ziętak K. (2000), Correcting possible non–positiveness of a covariance matrix estimated elementwise in a robust way. In: J. Sołdek, J. Pejaś, Eds, *Advanced Computer Systems, ACS'2000, Proceedings, Szczecin – Poland – 23–25 October 2000*, Fac. of Computer Science & Information Systems, Technical University of Szczecin, Informa, 91–96.

Bartkowiak A., Piotrowicz P., & Szustalewicz A. (1999), Detecting outliers in multivariate data by use of Kohonen's self–organizing maps. In: J. Sołdek, J. Pejaś, Eds, *Advanced Computer Systems, ACS'1999, Sixth International Conference, Szczecin November 1999, Proceedings*, Fac. of Computer Science & Information Systems, Technical University of Szczecin, Informa, 321–328.

Bartkowiak A., Szustalewicz A. (1998), Watching steps of a grand tour implementation. *Machine Graphics & Vision*, **7**, no. 3, 655–680.

Bartkowiak A., Liebhart J., and Szustalewicz A., (1997), Visualizing the correlation structure of some spirometric data by a biplot extended to 3 dimensions. *Biocybernetics & Biomedical Engineering.* **17**, 85–99.

Bartkowiak A., Szustalewicz A. (1997), Detecting multivariate outliers by a grand tour. *Machine Graphics & Vision*, **6**, no. 4, 487–505.

Billor N., Hadi A.S. and Velleman P.F. (2000), BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, **34**, no. 3, 279–298.

Campbell N.A. (1980), Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics*, **29**, 231–237.

Chatterjee, S., Hadi, A.S. and Mächler M. (1997), *A Re-weighted Least Squares Method for Robust Regression Estimation and Outlier Detection.* Manuscript 1-17. Paper presented at ISI in Istanbul.

Gnanadesikan, R.: (1997), *Methods for Statistical Data Analysis of Multivariate Observations.* 2nd Edition (1st edition 1977), Wiley, New York.

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

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. (1992), Identifying multiple outliers in multivariate data. *J. R. Statist. Soc., B*, **54**, no. 3, 761–771.

Hadi A.S., Simonoff J.S. (1997), A more robust identifier for regression data. *Bull. of the Int. Statistical Institute, ISI'97, 51th Session at Istanbul.* 281–282.

Hadi A.S., Velleman P.F. (1997), BACON: blocked adaptive computationally efficient outlier nominators. Paper presented at the *Joint Statistical Meeting in Anaheim.* Manuscript 1–23.

Hawkins D.M. (1994), The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics and Data Analysis*, **17**, 197–212.

Hawkins D.M., Bradu D., Kass G.V. (1984), Location of several outliers in multiple regression using elemental subsets. *Technometrics*, **26**, 197–208.

Morlini I. (1998), Multivariate outlier detection with Kohonen networks: an useful tool for routine exploration of large data sets. In: Ph. Nanopoulos, P. Garonna, C. Lauro (Eds), *NTTS'98 International Seminar on New Techniques and Technologies for Statistics, Sorrento, Italy 4-6 November 1998*, Contributed Papers, 345–350.

Riani M., &Atkinson A.C. (2000), Robust diagnostic data analysis: Transformations in regression. *Technometrics*, **42**, 384–394.

Rocke D.M., Woodruff D.L. (1996), Identification of outliers in multivariate data. *JASA*, **91**, Nb 435, 1047–1061.

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

Rousseeuw P.J., van Zomeren B. (1990), Unmasking multivariate outliers and leverage points. *JASA*, **85**, 633–639.

Rousseeuw P.J., Leroy A. (1987), *Robust regression and outliers detection.* Wiley, New York.

Tierney L. (1990), *Lisp–Stat, an Object–Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley, New York.

Velilla S. (1995), Diagnostics and robust estimation in multivariate data transformations. *JASA*, **90**, 945–951.

Woodruff D.L., Rocke D.M. (1993), Heuristic search algorithms for the minimum volume ellipsoid. *Journal of Computational and Graphical Statistics*, **2**, 69–95.