Statistica, anno LXII, n 1, 2002

IDENTIFYING MULTIVARIATE OUTLIERS - A MEDICAL EXAMPLE

Adam Szustalewicz

1. INTRODUCTION

One can say that an outlier is an atypical observation - it means - not matching the pattern suggested by the majority of observations.

Such incorrect data may distort the model - estimated relationships among the considered traits (variables). In extreme cases, specially in medical considerations, false diagnoses can be made. Therefore before starting the proper data analysis one should always check the considered data set for outliers.

This may be specially difficult for multivariate data: most of the proposed methods start with calculation of the covariance matrix which can be seriously changed by the influence of erroneous observations hidden in the data set. Thus, obtaining a clean set of only typical observations for further considerations may be not easy this way.

Our data are true medical data which have been taken from Wroclaw University of Medicine. These can be presented in the form of a data matrix X,, $_p$ with $n = 125$ individuals, each characterized by $p = 9$ variables. The data will be briefly described in section 2.

Next, in section *3* we describe a convenient technique which allows the visualization of multidimensional data by a simple two dimensional representation. This is *parallelcoordinates plot* which permits to watch all the data points in one picture and to judge bp eye if the considered set of multivariate data may contain outliers or not.

In the next section we present a method permitting to distinguish some of the data points as outliers; a method, which does not use the covariance matrix, thus is not influented by atypical data. The method - grand tour - comes originally from Asimov (1985). Our implementation of his idea is described in details in Bartkowiak and Szustalewicz (1997, 1998). The method permits to watch the whole cloud of multivariate data points 'from different sides' and basing on this to designate some observations as outliers. The method grand tour will be shortly presented in section 4.

Next we verify once more what are the positions of the suspected to be outliers points in relation to the main group of points. In section 5 we describe the conception of *angular distance* treated as a measure of neighbourness of two data points.

Applying then cluster analysis to the distance matrix we obtain a subdivision of all the suspected points into several subsets. The elements of the same subgroup should be located 'on the same side' of the main bulk of data. It can be easily verified by running grand tour once more with highlighted points of the same subset.

Some summary and conclusions are presented in section 6.

2. DATA – A MEDICAL EXAMPLE

We apply our considerations to the true medical data obtained from the ambulatory of the Department of Internal Medicine and Allergology, Wrocław University of Medicine (kindness of prof. J. Liebhart).

The data matrix X of size n x p contains observed data for $n = 125$ individuals - patients with obturative disorders (Obturation).

For each patient p = 9 traits 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-expiratorp Flow Time.

The set of p = 9 observations for one patient will be in the following called *data point* and denoted as $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ for the i-th individual.

Geometrically $\mathbf{x}_i$ means a point located in the p-dimensional Euclidean space RP.

For the convenience in next considerations, we shift the origin of coordinates to the data gravity center which means practically that we subtract from the coordinates of data points the means $\mu_1, \ldots, \mu_p$ of the corresponding variables:

$$x_{ij} := x_{ij} - \mu_j \qquad \text{for } j = 1, \ldots, \text{p} .$$

From now on our data set is centered at the origin of the coordinate system.

The data hare been previously analyzed by Bartkowiak [3]. Applying the methods of principal components and robustified Mahalanobis distances the author has found 12 data points identified as outliers. These are: 2, 43, 80, 117, 20, 42, 48, 49, 53, 68, 88, 91 in the Lisp numeration, *i.e.* starting from 0.

3. THE PARALLEL COORDINATES PLOT

The *parallel coordinates* plot was introduced in 1985 by Inselberg. A detailed descriptions of this technique may be found in Inselberg (1996) and Bartkowialc (1997). This is a convenient method permitting to see the values of all separate variables (coordinates) for every individual (data point) from the R^P space.

The variable axes are drawn as p vertical, equidistant lines. All the values of every variable for the whole data set are plotted onto the same axis. Each data point $\mathbf{x}_i = (x_{,,}, \ldots, x_{ip})$ is illustrated as a horizontal (p - 1)-segment line crossing the succeeding axes at the values of corresponding coordinates (analyzed variables) $x_{i1}, \ldots, x_{ip}$.

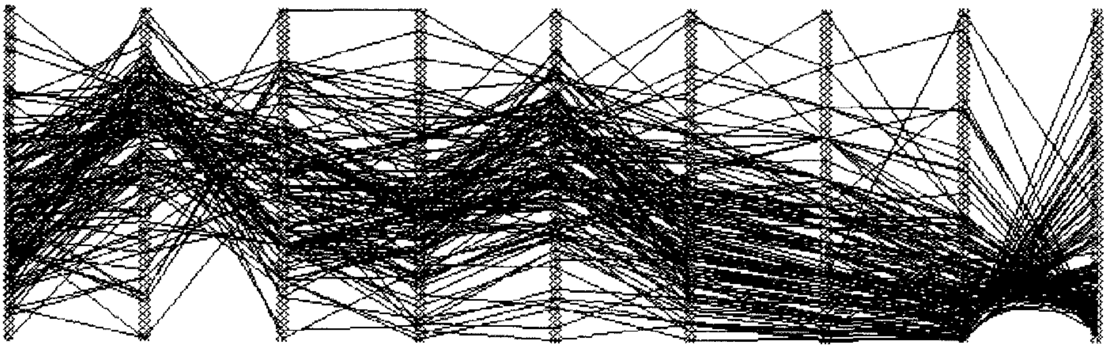

Figure 1 – Parallel coordinates plot for the analyzed data.

Particular data points may be selected (highlighted) by clicking with a mouse the respective segment lines. One can see then, what arc the values of the succeeding variables for the selected individuals. The atypical patterns of the selected segment lines may suggest their misfit to the main set of data points and one may judge by eye then whether the analyzed set of data contains some atypical points or not.

In figure 1 we see the illustration of our data set obtained this way. This is a composition of 125 differently changing segment lines representating analyzed individuals. We may notice a main, wide band of almost similar lines, and many others – irregular, with greater values of their coordinates $x_{.j}$ lor $j = 4, ..., 8$. There arc also some lines with the second segment ($[x_{.2}, x_{.3}]$) sloping strongly up, i.e. atypically to the rest of data.

4. THE GRAND TOUR – NOMINATION OF OUTLIERS

The first idea of the method for visualizing multivariate data was oryginally proposed by Asimov (1985) and further elaborated by other researchers (for references see Kartkowiak and Szustalewicz, 1997) a.o. also by Tierney (1994), who implemented a procedure in Lisp-Stat. The method was specifically adapted for detecting multivariate outliers by Hartkowiak and Szustalewicz, 1997, 1998. The approach of these authors develops that of Tierney.

The concept is simple and follows the physical interpretation of viewing a set of data points in three-dimensional Euclidean space R'. According to our assumption we may imagine the considered set of data points as centered at the origin of the coordinate system and located in some closed sphere with the center at the origin as well.

Observing the set of points from beyond the sphere is equivalent to watching the projections of the points onto a plane perpendicular to the direction of the observer's look. Observing the given set of points from different sides is equivalent to the rotating ol the whole sphere with the contents round the origin in randomly chosen directions and each time watching the projections of the points onto the same plane. Without loss of generality this may be the plane $< x_1, x_2 >$ spanned by the first two coordinate axes.

Algebraically, for real p-dimensional data stored in the subsequent rows of ma-

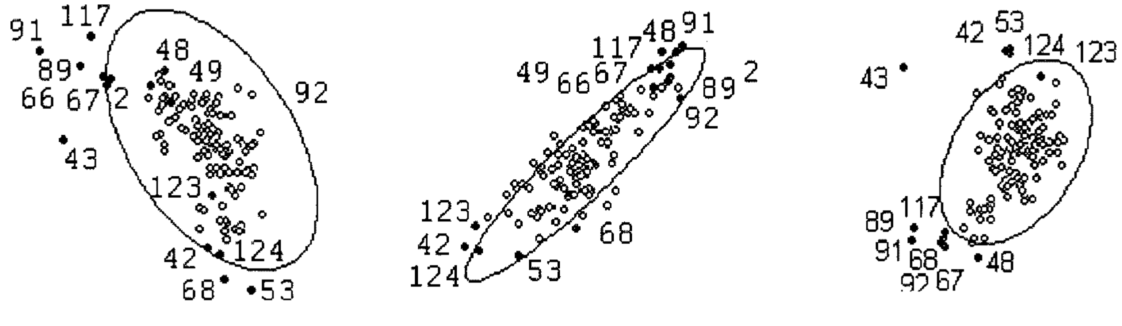

Figure 2 - Three exemplary snapshots from the grand tour projections. A 99% robustified concentration ellipse is superimposed on every of them.

trix X, $_{\times p}$ the procedure described above is equivalent to the multiplication of the matrix X by a suitably constructed orthogonal matrix $A_{p \times p}$:

$$X^{(new)} = X^{(old)}A$$

Very important is ensuring that after carrying out a sufficiently long sequence of rotations and projections we have seen the detected data set from really many directions and we could notice the majority of characteristic internal patterns of the data. For details see Bartkowiak and Szustalewicz (1998).

A scatterplot of projections $(x_{i1}^{(new)}, x_{i2}^{(new)})$ is presented alter each rotation. We superimpose a concentration ellipse on the scatterplot to focus our attention on the points located far (in the sense of Mahalanobis metric) from the data center. Some exemplary projections seen on the screen are presented in figure 2.

A linked count-plot is opened in the screen to the comfort of the user as well. There is a graphical counter, which shows for every data point how many times this point was located beyond the concentration ellipse. After performing several

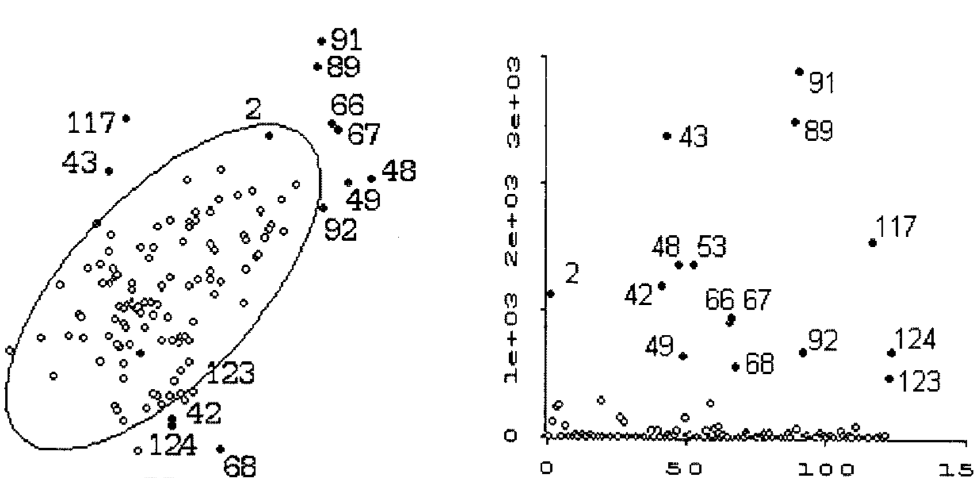

Figure 3 - A snapshot from the grand tour. Left: An exemplary point projections with 99% concentration ellipse superimposed, Right: A linked count-plot illustrating how many times (till the time of the snapshot on the left hand side) each data point was notified beyond the ellipse. The data points are enumerated according to the Lisp-Stat from 0 to 124.

hundreds of rotations it is very easy to notice the indexes of all individuals which may be nominated as outliers.

5. CLUSTERING BY ANGULAR DISTANCE

The nominated outliers are noticed on the snapshots from the grand tour as grouped into some smaller subsets.

This suggests that the elements of such a subgroup are really lying close one to each other in the R^p space, and that the subsets may be located 'at various sides' of the main bulk of data.

To verify the neighbourness of the suspected points we may use the *complete linkage method* (see a.o. Gnanadesikan, 1997). We apply the method to the matrix of *angular measure of the distance between data points* $s(\mathbf{x}_i$, xj) described later.

As a result we obtain a dendrogram with branches containing the observed in grand tour subsets of close to each other individuals (in the sense of angular distance).

The coordinates $(x_{i1}, ..., x_{ip})$ of our data points may be treated as components of vectors $\{\vec{\mathbf{x}}_i\}$ anchored at the origin of the coordinate system: $\vec{\mathbf{x}}_i = [x_{i1}, ..., x_{ip}]$. We may evaluate the value of the cosine of the angle $\alpha_{i,j}$ between such two vectors $\vec{\mathbf{x}}_i$ and $\vec{\mathbf{x}}_j$ from the definition of the scalar product

$$(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = | \vec{\mathbf{x}}_i \| \, \| \vec{\mathbf{x}}_j \| \cos \alpha_{i,j}$$

where $\| \cdot \|$ denotes the Euclidean vector length.

Transforming now the obtained value of the trigonometric function onto the interval [0; 1] we introduce the desired *measure of angular distance s* between our data points (individuals) $\mathbf{x}_i$, $\mathbf{x}_j$

$$s(\mathbf{x}_i, \mathbf{x}_j) = 0.5(1 + \cos \alpha_{i,j}) \ . \tag{1}$$

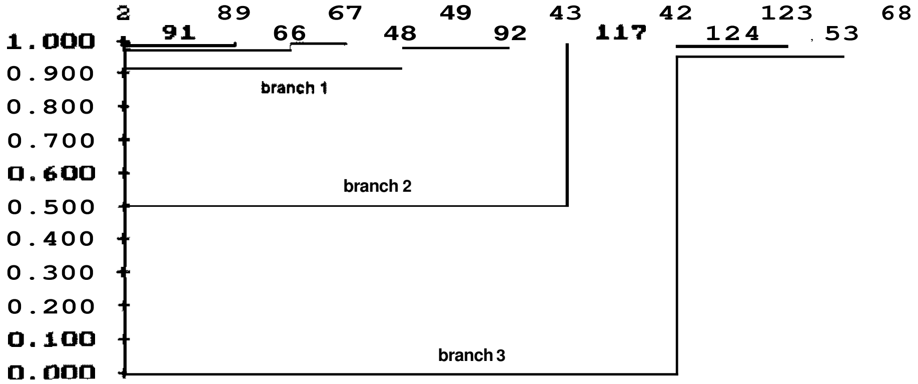

Figure 4 – Dendrogram obtained for the 15 data points which were notified by the grand tour method mostly outside the 99% concentration ellipse. The measure of their neighbourness is defined by the formula (1).

As we see in figure *4*, the set of data has been divided into three separate subsets (called 'branches' in the picture) of the points lying very close one to each other. These subsets are: $S1 = \{2, 48, 49, 66, 67, 89, 91, 92\}$, $S2 = \{43, 117)$ and $S3 = \{42, 53, 68, 123, 124)$ enumerated according to the Lisp-Stat from 0 to *124*.

The angular distance from the branch *1* to *3* is almost maximal. This suggests, that these branches ought to be located on 'nearly opposite sides' of the main bulk of data.

Branch *2* lies angularly midway between branches 1 and *3*. Then, its elements {43, 117) - as possibly outliers - should be located anywhere on the 'equator' if to treat the branches 1 and 3 as the 'poles' of the sphere containing the analyxed set of data points.

6. CONFIRMATION OF THE OUTLYINGNESS OF THE POINTS NOMINATED BY THE GRAND TOUR METHOD

After obtaining the branches from the clustering method we start to run the grand tour again with highlighting of points belonging to separate branches.

Three snapshots from such watching are shown in figure 2. We see there that branches *1* and 3 are really located on 'nearly opposite sides' of the main bulk of data, as this was suggested in the previous section. The smallest branch 2, particularly the point *43*, is located in an orthogonal direction to the direction defined by branches 1 and 3. The elements 43 and 117 of branch *2* are probably not to much close by and the best snapshot with the view of branch *2* is shown in the left part of figure *3*.

Let us look now at the parallel coordinates plot of the suspected to be outliers 15 data points. In the left part of figure 5 we see an entaglement of segment lines obtaining the minimal or maximal values of the respective variables. The lines are numbered and we may distinguish three bands of more similar segment lines appropriate to the obtained branches.

In the right part of this figure we see the 'cleaned' set of data, i.e. after removing all the suspected data points. The corresponding parallel coordinates plot looks much better than before, but as we may judge by eye - the set is not 'clean' yet.

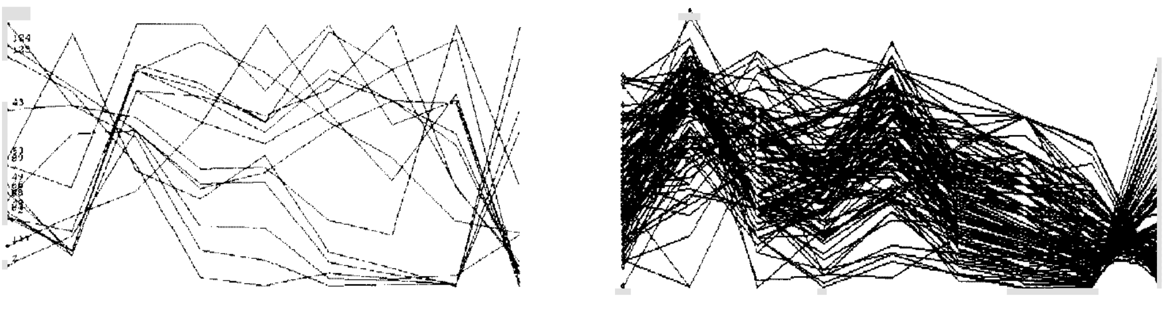

Figure 5 - Parallel coordinates plots. Left: for the data points nominated as outliers by the grand tour method. Right: for the cleaned set of data points.

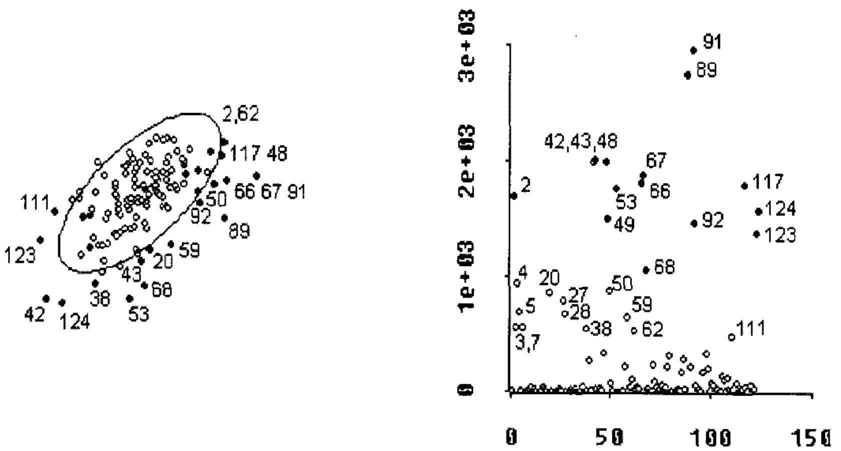

Figure 6 - *A* snapshot from the grand tour applying the 95% ellipse of concentracion. Left: An exemplary point projections with concentration ellipse superimposed. Right: A linked count-plot illustrating how many times each data point was notified beyond the ellipse.

There are several segment lines outstanding in the pattern character from the main band of lines.

The easiest way to obtain a 'cleaner' set of data is to repeat the grand tour method with a weakened selection criterion for suspecting outliers. We use now a concentration ellipse covering 95% of all the points (instead of *99%* previously).

The repeated grand tour procedure nominated as outliers 12 additional individuals: *(3, 4,* 5, *7,* 20, *27, 28, 38, 50,* 59, *62,* 111}. (The numeration of individuals starts from *0.)* As we see in the count-plot in figure 6, there is no well-marked boundary separating all the suspected outliers from the rest of data.

Looking at the dendrogram in figure 7 we see the extended about *3* points branch 1. It connects *11* points now. *S1'* = {2, *3, 48, 49, 50, 62, 66, 67,* 89, *91,* 92}, spreaded as close as previously in the space.

The second branch did not change. It consists of the same two points *S2'* = {43, 117}.

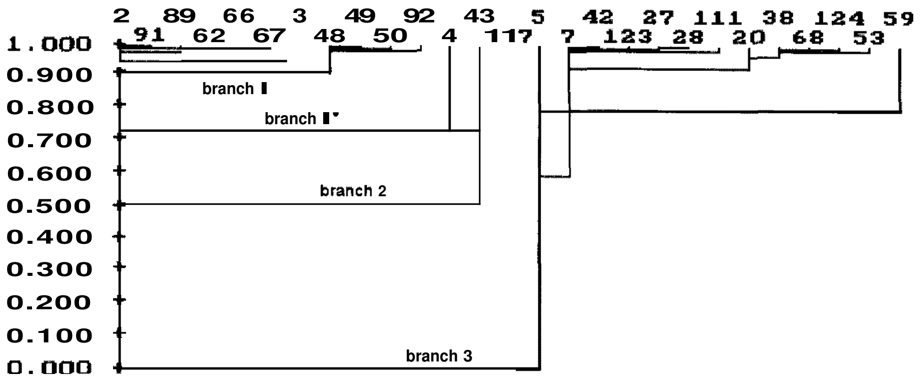

Figure 7 - Dendrogram obtained for the 27 data points which were notified by the grand tour method mostly outside the 95% concentration ellipse.

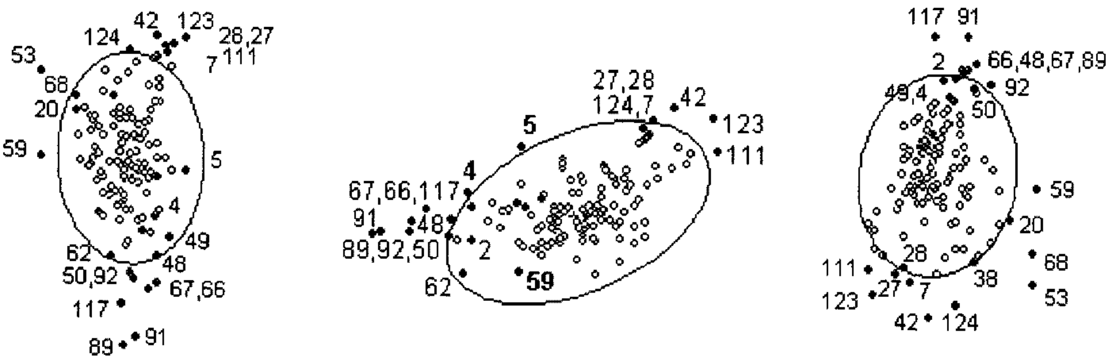

Figure 8 – Three exemplary snapshots from the grand tour projections with superimposed a 95% concentration ellipses.

A new isolated point with number 4 appeared between these two branches. This is notified mostly nearer by the points of branch 1.

Branch 1 is almost three times greater than previously. It consists of 13 points now, $S3' = \{5, 7, 20, 27, 28, 38, 42, 51, 59, 68, 111, 123, 124)$. Two of them - with numbers 5 and 59 are lying wide apart and from the others. We see that on the dendrogram in figure 7 and on the snapshot in figure 8.

The parallel coordinates plot of the additionally cleaned set of data points is presented in figure 9.

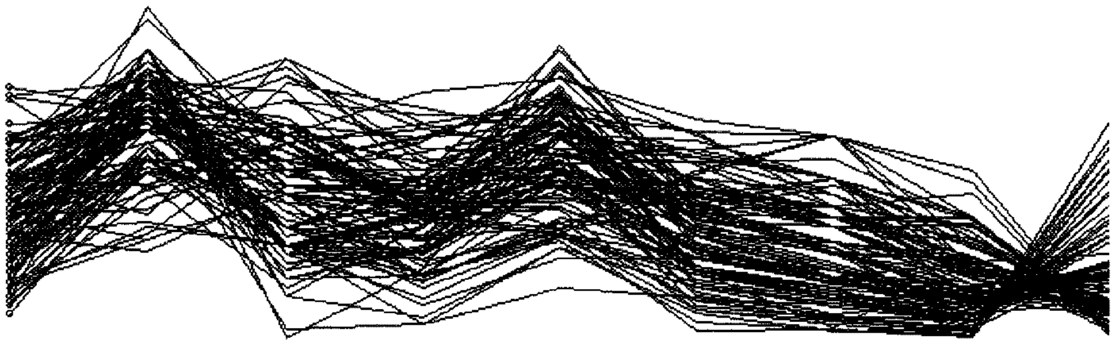

Figure 9 Parallel coordinates plot for the additionally cleaned set of data points

Institute of Computer Science
University of Wrocław, Poland

ADAM SZUSTALEWICZ

REFERENCES

A. ALEXAKIS, M. DELTEREOS, Y. SAMIOTAKIS (1998), *Visualization techniques in statistics. The parallel coordinates visualizer (PARCOVI)*, in: PH. NANOPOULOS, P. CARONNA, C. LAURO (eds.), *NTTS'98 International Seminar on New Techniques and Technologies for Statistics*, Sorrento, Italy 4-6 November 1998, Contributed Papers, pp. 19-21.

D. ASIMOV (1985), *The grand tour, a tool for viewing multidimensional data*, "SIAM Journal *on* Scientific and Statistical Computing", 6.1, pp. 128-143.

A. BARTKOWIAK (1997), *Some basics for detecting multivariate outliers in regressional context,* "Biocybernetics and Biometrical Engineering", 17, pp. 57-83.

A. BARTKOWIAK (1999), *Identifying multivariate outliers by dynamic graphics - as applied to some medical data,* Paper presented at ICASMES'97, August 12-13, Ankara, printed in Applied Statistical Science IV, e d ~M. Ahsanullah, F. Yildirim, Nova Science Publishers Inc. New York.

A. BARTKOWIAK, A. SZUSTALEWICZ (1997), *Detecting multivariate outliers by the grand tour,* "Machine Graphics & Vision", 6, pp. 487-505.

A. BARTKOWIAK, A. SZUSTALEWICZ (1998), *Watchings steps of a grand tour implementation,* "Machine Graphics & Vision", 7, pp. 655-680.

R. GNANADESIKAN (1997), *Methods for statistical data analysis of multivariate observations,* Wiley, New York.

A. INSELBERG (1996), *Parallel coordinates; a guide for the perplexed Plot,* "Hot Topics Sect, of IEEE Comp. Soc. Conf. on Visualization '96", San Francisco, Proc., pp. 35-38.

L. TIERNEY (1994), *LISP-STAT, an object-oriented environment for statistical computing and dynamic graphics,* Wiley, New York.

RIASSUNTO

L'identificazione di outlier multivariati: un esempio medico

Molti insiemi di dati, e in particolare quelli di tipo medico, sono costituiti da una matrice X,, $_p$ che contiene i valori di p variabili osservati su *n* individui.

Abbiamo a che fare con i valori di $p = 9$ caratteristiche, quali l'età, l'altezza e altre variabili spirometriche quali RV, VC, VC%, FEVI, FEF, ... osservate per $n = 125$ pazienti. Tale tavola può essere interpretata come una nube di *n* punti nello spazio euclideo *p*-dimensionale R". I dati analizzati contengono outlier in dimensione e struttura. Quest'ultimo tipo in particolare non può essere individuato quando ogni variabile viene considerata separatamente.

Si dimostra l'utilità di moderni metodi di visualizzazione per insiemi di dati multivariati, quali *grand tour with a count plot* (Bartkowiak and Szustalewicz. 1997), che individua un insieme di punti potenzialmente outlier; successivamente il metodo del legarne completo (basato su distanze angolari) e il grafico a coordinate parallele confermano i risultati ottenuti.

SUMMARY

Identifying multivariate outliers - m medical example

Many data sets, especially medical data, consist of a two-dimensional table $\mathbf{X}_{n \times p}$ containing *p* variables measured for every of *n* individuals.

We are concerned with values of $p = 9$ traits, such as *Age, Height* and other spirotnetric variables like RV, VC. VC%, FEV1, FEF, ... recorded for $n = 125$ patients. Such table can be interpreted as a cloud of *n* points in the *p*-dimensional Euclidean space R^p.

The analysed data contain outliers both in size and structure. Especially the last type could riot be detected when considering each variable individually.

We demonstrate the usefulness of modern visualization methods for multivariate data, as *grand tour with a count plot* (Bartkowiak and Szustalewicz, 1397) which finds a set of points suspected to be outliers, and then, the *complete linkage method* (based on *angular distances)* and *parallel coordinate plot* - which additionally confirm the obtained results.