{file s.tex January 20, 1999}

Paper proposed for IIS'99

SOME MODERN TECHNIQUES FOR VIEWING MULTIVARIATE DATA – A COMPARATIVE LOOK

Anna Bartkowiak and Adam Szustalewicz
Institute of Computer Science, University of Wrocław
Przesmyckiego 20, Wrocław 51-151 Poland
e-mail: aba, asz@ii.uni.wroc.pl

Abstract

We consider real data which are gathered typically in a two dimensional table $\mathbf{X}$ of size $n \times p$ with rows containing *individuals* and columns indicating *variables* measured on these individuals. The variables are measured on a continuous scale. Each row of this table may be considered as a *data point* located in the p-dimensional Euclidean space R^p.

Modern visualization techniques like *tour* plots, *augmented biplots*, *parallel coordinate* plots – combined with some other, auxiliary evaluations and using an appropriate GUI – permit to recognize the shape of the multivariate data cloud containing points representing subsequent rows from the data table $\mathbf{X}$, in particular to see outliers, if any.

We demonstrate the usefulness of these techniques considering the structure of earnings for some data extracted from the Eurostat New Cronos data base. The same data were recently analyzed by Morlini (NTTS'98) with the goal of detecting multivariate outliers by use of Kohonen networks. We compare ours with her results.

1 General state of the art, aim of the paper

Many data sets are really multidimensional, i.e. with dimension p greater than 3. The individual data vectors, denoted here as $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, may be considered as data points located in p-dimensional space R^p. When we consider n such points, so they may be imagined as a multivariate data cloud located in R^p.

Traditionally these data are visualized by using 2-dimensional (2D) graphs in the form of dot plots, histograms, box plots, scatter plots, profile plots representing uni- or bivariate relations.

With developments of computer facilities a 3-dimensional presentation of data in the form of spin plots was introduced.

Both 2D and 3D presentations use often some data reduction techniques (choice of representative variables or feature extraction like principal coordinates or canonical variates) which permit to project the analyzed data graphically to the 2- or 3-dimensional space.

Recently some techniques for visualizing truly multivariate data were proposed. We will consider here in more details two such techniques: the grand tour and parallel coordinate plot.

To explore efficiently multivariate data and find the shape of the multivariate data cloud located in R^p we need specific software allowing, a.o., for quick interaction, linkage of plots, selection and identifying data points, giving them specific symbols and colors, hiding and restoring indicated subgroups of data.

The aim of the paper is to illustrate and compare the mentioned above techniques for multivariate data visualization.

We think that specially useful is here the grand tour method. This method may be complemented by other techniques: parallel coordinates plots and sometimes by principal coordinates illustrated by a specially constructed spin plot.

2 Data used in the paper for comparison of methods

Fig.1. Boxplots from original and standardized data

We will illustrate our considerations by using a data set analyzed recently by Morlini [6]. The data were taken from the Eurostat New Cronos Data Base and can be presented in the form of the data matrix $\mathbf{X}_{n \times p}$, with $n = 158$ units of economical activities in Italy in 1995, each characterized by 5 variables:
$X1$ – "nbempl" –number of employees, $X2$ – "mwhour" – mean weekly hours, $X3$ – "mmearn" – mean gross monthly earnings, $X4$ – "cvarme" – coefficient of variation of monthly earnings, $X5$ – "bonus" – mean monthly value of periodic bonuses.

All the recorded values were measured in a real continuous scale with a very differentiated range, what can be seen in Figure 1, left plot. After standardizing the variables to mean=0 and standard deviation=1 one can present all the 5 distributions as boxplots with a reasonable common scale, what can be seen in Figure 1, right plot.

Both plots exhibited in Figure 1 show a high asymmetry in the distributions of all analyzed variables. To look in more detail, we show in Figure 2 (left plot) a histogram constructed for the variable $X1$="nbempl" exhibiting the highest skewness. After transforming the values by the *log* function the distribution of the transformed variable appears much more symmetric – what can be seen in Figure 2, right plot.

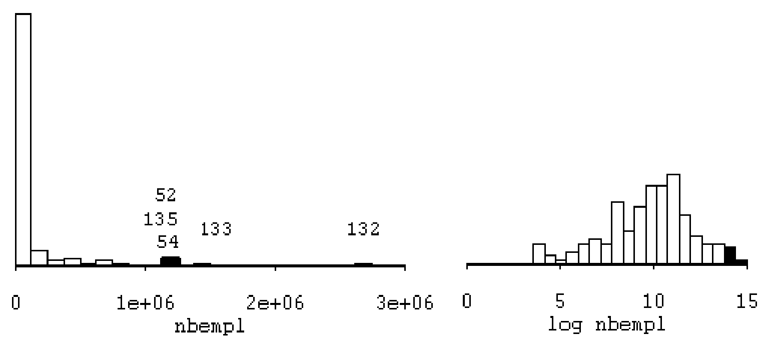

Fig. 2. Histogram of the variable X1="nempl" of the analyzed data.
Left: original values. Right: values after the log *transformation*

In the following we will analyze the raw data, without transformations reducing their skewness.

3 The grand tour

3.1 Nomination of outliers

The grand tour, originally proposed by Asimov, was adapted by Bartkowiak & Szustalewicz [3] specifically for detecting and identifying outliers. It works as a proceeding sequence of rotations and projections of the considered data points (see [3] for further description and references).

The grand tour was performed on standardized variables.

One snapshot of the projections visible in the computer screen at time t_k is shown in Figure 3. The momentary view at time t_k consists of a scatterplot with an concentration ellipse (left view) and an indexplot (called *count plot*) exhibiting counts telling how many times the given point for time $t < t_k$was notified outside the ellipse (right view).

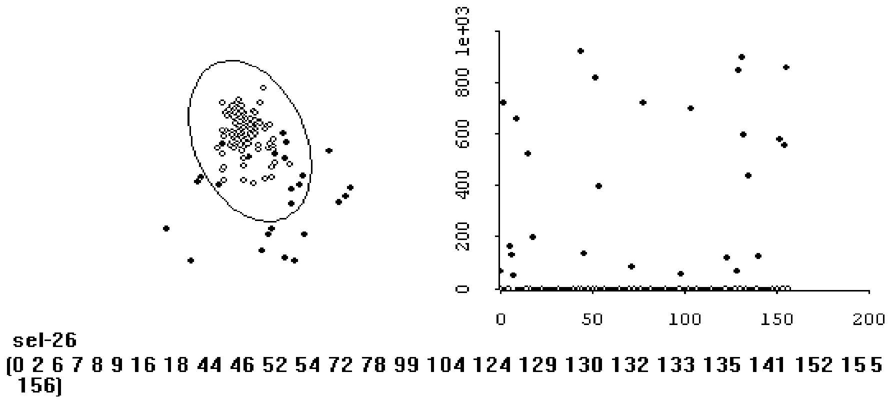

Fig. 3. One snapshot from the grand tour with ordinary concentration ellipses.
Left: A momentary projection seen on the screen containing the point projections and a 95% ordinary concentration ellipse.
Right: Corresponding count plot indicating how many times each data point was found – till present t_k – outside the concentration ellipse.
26 points with the highest counts are highlighted

The data points appearing outside the concentration ellipse are nominated as suspected outliers. Proceeding this way we have nominated 26 data-points (the set *sel-26*, marked in Figure 3 by filled circles) as possible outliers.

These are (the numeration starts from 0): { 0 2 6 7 8 9 16 18 44 46 52 54 72 78 99 104 124 129 130 132 133 135 141 152 155 156}.

All the nominated points have high counts indicated in the count plot.

Generally, the nominated outliers may be located at various 'sides' of the multivariate data cloud $\subset R^p$. The proximity of some points noticed in the momentary projections does not mean that these points are really close in the R^p space. To assess their proximity we may use the method of complete clustering based on angular distances between the points (for a definition of angular distance see [4]). As a result we obtain then a dendrogram, which typically exhibits some branches containing 'close' points (the closeness is measured by the angular distances among these data points).

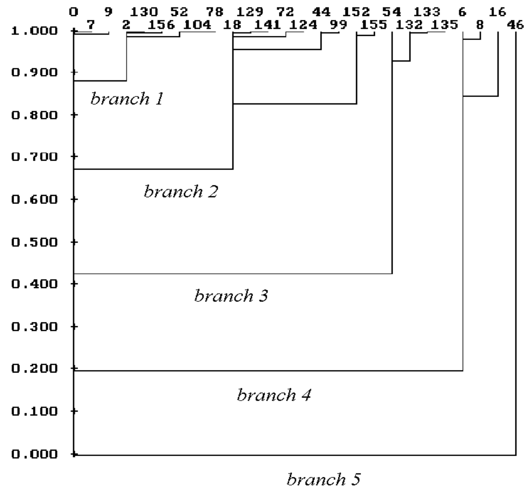

Fig.4. Proximities among nominated outliers established on the base of angular distances by use of the complete linkage clustering

For the nominated 26 outliers we have obtained the dendrogram exhibited in Figure 4. One may see there 5 branches containing the following points:
branch 1: {0 7 9 2 130 156 52 104 78}, branch 2: {18 129 141 72 124 44 99 152 155}, branch 3: {54 132 133 135}, branch 4: {6 8 16}, branch 5: {46}.

3.2 Confirmation – by repeated runs with highlighted groups of points belonging to the watched branches

After obtaining the branches we start to run the grand tour again with selection (highlighting) of points belonging only to one branch and eventually with increased confidence coefficient of the superimposed ellipse.

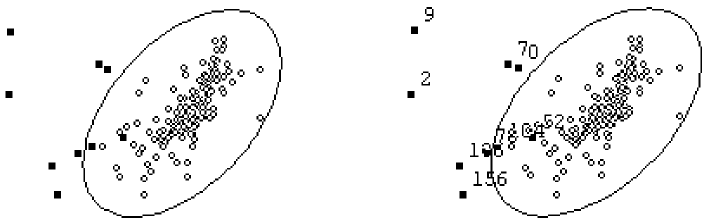

Fig. 5. Repeated grand tour. Watching points belonging to branch 1.
Left: Points from branch 1 selected, but unlabelled. Right: The points selected are labelled

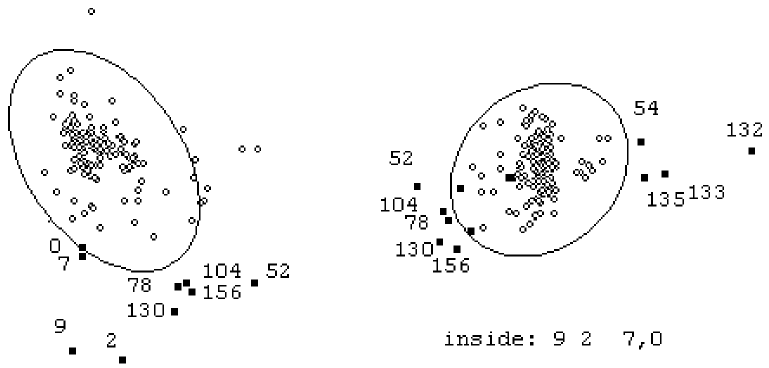

Fig. 6. Repeated grand tour. Left: Watching points belonging to branch 1, another situation. Right: Watching simultaneously points belonging to branch 1 and branch 3; inside the ellipse are – highlighted but not labelled – points {9, 2}, and one on the top of the other, points {7, 0}

In Figure 5 we show one snapshot of the grand tour when watching specially points belonging to branch 1 (Figure 5, left: points unnumbered, right: the same view, but with labelled points).

In Figure 6, left, we show another snapshot from watching branch 1. In the same figure, right, we show a view for a different situation: here branches 1 and 3 were watched together. The location of the two groups in the watched projections is nearly opposite, what could be expected when looking at the dendrogram.

4 Biplot in 3 dimensions

Another technique for visualizing multivariate data constructs – by a SVD decomposition of the data matrix $\mathbf{X}$ – new features, called *principal coordinates*, which are typically visualized in the form of a 2D plot, called biplot (see [2], [7]).

For our data the new features $PC1, \ldots, PC5$ – calculated for standardized data – reproduce cumulatively the following percentage of the total inertia of the data: PC1 – 46.90, PC2 – 67.31, PC3 – 85.02, PC4 – 94.59, PC5 – 100.00.

Looking at the above percentages one can see that the representation by a classical biplot is not fair: it covers only a part (67.31%) of the mutual relations among the analyzed data points.

A much better representation – of 85.02% – can be obtained when drawing a biplot in 3 dimensions [2]. Such a plot reproduces the following percentages of the inertias of individual standardized variables (denoted below by the letter V):
$V1$ – 0.98, $V2$ – 0.77, $V3$ – 0.81, $V4$ – 0.89, $V5$ – 0.81.

Fig. 7. Spinplot exhibiting principal coordinates $PC1 = X$, $PC2 = Z$, $PC3 = Y$ for data points from **branches 1** *(squares)* **and 3** *(diamonds)*

Figures 7 and 8 show three-dimensional biplots (realized as spinplots based on the first three principal coordinates) with points from branches 1+3 (Figure 7) and 2+3 (Figure 8) highlighted.

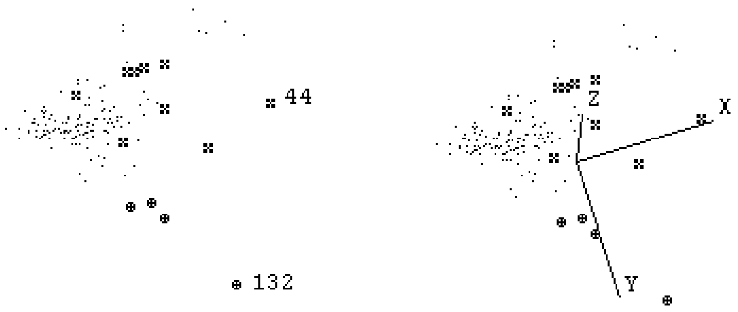

Fig. 8. Spinplot exhibiting principal coordinates $PC1 = X$, $PC2 = Z$, $PC3 = Y$ for data points from **branches 2** *(symbol: x)* **and 3** *(symbol: $\oplus$)*

Again, these plots confirm our former results, i.e. confirm the fact that points belonging to one branch are close, and points belonging to different branches are distant in the considered space of the first three principal coordinates.

5 Parallel coordinate plot

Parallel coordinate plot is a special technique permitting to see for the data points (individual data vectors) the values of their variables (see [1] and [5] for further references).

A parallel coordinate plot drawn for our data is shown in Figure 9.

Each data vector is represented there as $p-1$ horizontally connected line segments.

The p vertical equidistant lines represent the axes of the analyzed variables ('variable axes').

Individual data points can be selected (highlighted, or given special color) by clicking with a mouse some segment belonging to a representation of the chosen data vector.

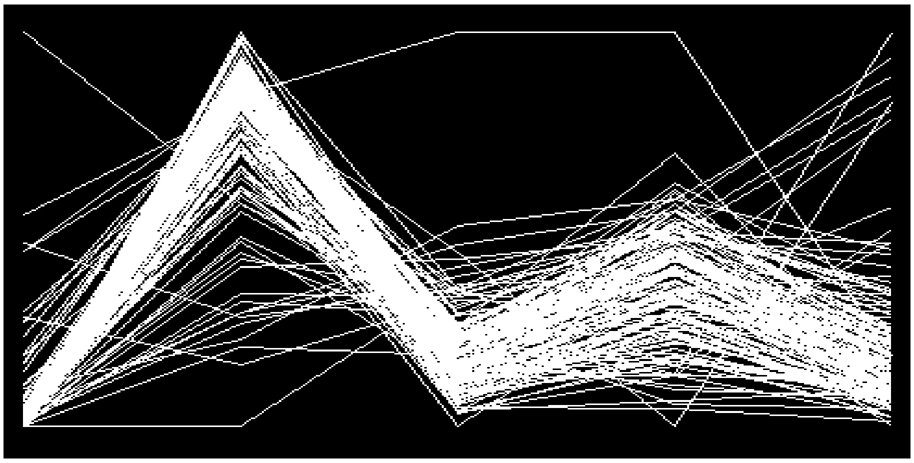

Fig. 9. Parallel coordinate plot for the analyzed data

Of course, we can select also a group of data vectors. Then, looking at the position of the crossings of the line segments with the variable axes, one can see, which are the values of the variables for the selected individuals. In particular, one can see whether these values differ much, or not at all, from the values exhibited by the not selected data vectors.

The parallel plot needs operating with colors; it is different to present its valor using only gray shades or different patterns in black and white colors.

In any case, looking at the plot shown in Figure 9 one can see there the main bulk of the 4-segment lines representing the majority of the analyzed data vectors, and some atypical ones belonging to the *sel-26* set. The variable axes drawn originally in red are not visible in that plot.

6 Final remarks and conclusions

The considered techniques: grand tour, parallel coordinate plot, spinplot based on principal coordinates – when used interactively in a dynamic graphics mode – are very useful in visualizing multivariate data. They permit to recognize the shape of the analyzed data and identifying atypical observations, if any.

It seems that the methods complement each other. The most universal, when properly operated, seems to be the grand tour.

The parallel coordinate plot, when using alone, needs a special skill to recognize the shape of the data cloud. However, it is very useful in assessing the nominated outliers.

All these methods need a special software allowing for properly designed dynamic graphics. In fact, none of the present available software packages meets fully all the needs of the data analysts.

For all our presentations in the paper we have used XLisp-Stat.

We did not consider Kohonen networks and Sammon's self organizing maps SOM's which are used for mapping points from a multivariate space R^p to a 2D or 3D representations. Our impression is, that the statistical methods presented above are much flexible and give more information about the mutual position of points in a multivariate data space. The results obtained by using Kohonen network and SOMs show many similar patterns to the results presented here for the considered data.

References

[1] Alexakis A., Deftereos M., Samiotakis Y., Visualization techniques in statistics. The parallel coordinates visualizer (PARCOVI), 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, 19–24.

[2] Bartkowiak A., Szustalewicz A., Some issues connected with a 3D representation of multivariate data points, *Machine Graphics* & *Vision* 5.4, 1996, 563–578.

[3] Bartkowiak A., Szustalewicz A., Detecting multivariate outliers by a grand tour, *Machine Graphics* & *Vision* 6.4, 1997, 487–505.

[4] Bartkowiak A., Szustalewicz A., Outliers – which genuine and which spurious, In: *NTTS'98*, as in [1], 61–66.

[5] Inselberg A., Parallel Coordinates; a guide for the perplexed Plot, *Hot Topics Sect. of IEEE Comp. Soc. Conf. on Visualization '96*, San Francisco, Proc., 35–38.

[6] Morlini I., Multivariate outlier detection with Kohonen networks: an useful tool for routine exploration of large data sets. In: *NTTS'98*, as in [1], 345–350.

[7] Pagnotta S., Interactive graphics for principal component analysis, In: *NTTS'98*, as in [1], 369–374.