Outliers – which genuine and which spurious

Anna Bartkowiak and Adam Szustalewicz

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

Summary

The grand tour is a dynamic graphics method permitting to recognize visually the shape of a multivariate data cloud and the degree of outlyingness (closeness) – of some indicated data vectors – from (to) the bulk of the data. The suspected outliers – obtained by some analytical methods like those in (Atkinson 1994, Rocke & Woodruff 1996, Hadi & Velleman 1998) or the count plot from an implementation of a grand tour (Bartkowiak & Szustalewicz 1997) – usually need additional verification stage.
We feel the need to subdivide the investigated set of points into some subgroups similar with respect to their location in the feature space. We propose to do it using cluster analysis based on similarities evaluated from angular distances between the points. We apply the agglomerative complete linkage method. When based on properly defined matrix of distances the proposed method permits to find bundles of data vectors anchored in the (median) center of the data. We can then say exactly how large is the maximal angle between pairs of vectors belonging to the same bundle.
The procedure is illustrated with the *Milk container* data, analyzed formerly, a.o. by Atkinson (1994) and Muruzábal & Muñoz (1997).

Keywords: Exploratory data analysis, The Grand tour, Suspected outliers, Clustering by angular distances, Dynamic graphics

1 Introduction – Motivation

There is no need to argue how important is the identification of outliers hidden in the data. There are several analytical methods which permit quite surely to find out and indicate for data vectors suspected to be atypical, see e.g. Atkinson & Mulira (1993), Hadi & Velleman (1998), Rocke & Woodruff (1996). However, these methods are based, more or less, on some significance testing (e.g. of Mahalanobis distances assumed to have the χ^2 distribution) which in principle needs the assumption of normality of the data. Also, these methods yield only a global index (characterization) of the outlyingness of the identified data points, and do not explain specifically, why the indicated data points were pointed to as 'atypical'.

The grand tour method performing sequentially rotations and projections of the data cloud may be specifically adapted for identifying the outliers (Bartkowiak & Szustalewicz 1997, Bartkowiak & Szustalewicz 1998). One may then identify not only the atypical points hidden in the data, but also "see" the mutual position of the points of interest and find out, how far they are located each from the other.

Suppose that by some methods we got a set S of points *suspected* to be outliers. Such a set may be obtained, e.g. from a grand tour method (Bartkowiak & Szustalewicz 1997), which produces count plots like that visible in Figure 1 – obtained when analyzing the *milk* data.

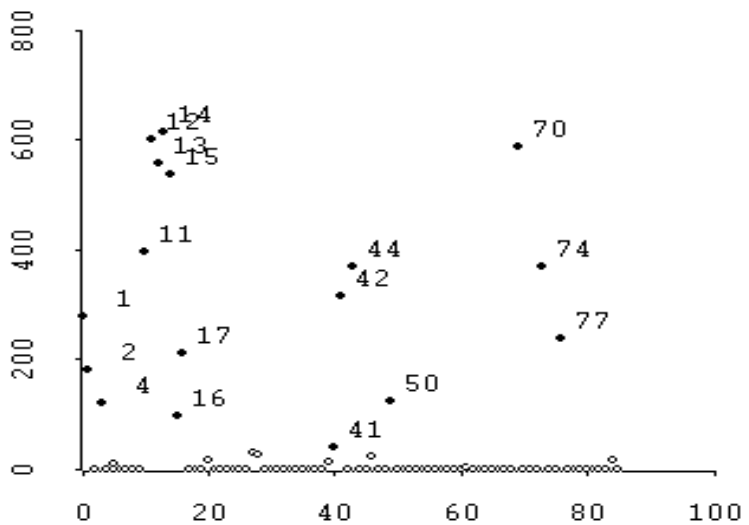

Figure 1. Count plot for the milk *data, obtained when running a grand tour. The highlighted seventeen points are suspected to be outliers*

We see in Figure 1 that there are 17 points which appeared quite frequently outside the concentration ellipse generated by the grand tour method. Let us call this set by S. Of course, this set could be obtained by a different method.

Are the points in S truly atypical observations? We would like to verify once more, what is the position of the points from S in relation to other, unsuspected points.

This could be done in the following way:

1. Using angular distance as a measure of similarity of two data points, subdivide the set S into several subsets. The subdivision may be obtained from a dendrogram produced by cluster analysis based on a properly defined distance matrix. Points belonging to one subset should be located on the same 'side' of the data cloud.

2. Take in turn each of the subset and run the grand tour with the points belonging to the watched subset being highlighted. You will see directly, how much the highlighted points are distant from the main data cloud.

The procedure is illustrated by considering the *milk container* data, analyzed formerly, a.o. by Atkinson (1994) and by Muruzábal & Muñoz (1974).

For the calculations we use own software written in Turbo Pascal (for clustering) and XLisp-Stat (for the dynamic grand tour).

2 Clustering by angular distances

Principles of cluster analysis may be found, a.o., in (Gnanadesikan 1987).

For our purpose we have firstly to define the angular distances.

Let $\vec{\mathbf{x}} = (x_1, \ldots, x_p)$ and $\vec{\mathbf{y}} = (y_1, \ldots, y_p)$ be two vectors anchored at $\mathbf{0}$ (the median of the entire data set). These vectors may be also imagined as data points located in the feature space $I\!R^p$. The *angle* α between these two vectors can be obtained by the well known formula for the scalar product $(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ of these two vectors:

$$(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = ||\vec{\mathbf{x}}||\ ||\vec{\mathbf{y}}||\ \cos(\alpha_{x,y}).$$

The value $\cos(\alpha_{x,y})$ appearing in the formula above will be called *the angular distance* between the vectors $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$.

Before beginning the proper calculations we transform the suspected data vectors by subtracting from their components the corresponding medians evaluated from the entire data set.

The angular distance defined for the transformed data vectors provides a measure of similarity between the data vectors (data points): if two of them are

anchored at 0 and have a small distance, then they are close each other and are located in $I\!R^p$ at the same 'side' of the data cloud.

The measure $s_{x,y}$ of similarity between two (transformed) data vectors is defined by us as:

$$s_{x,y} = [1 + \cos(\alpha_{x,y})] \times 0.5. \tag{1}$$

Cluster analysis based on the matrix of similarities between the data vectors yields a dendrogram – like that shown in Figure 2. The branches of such dendrogram permit to establish bundles of similar vectors.

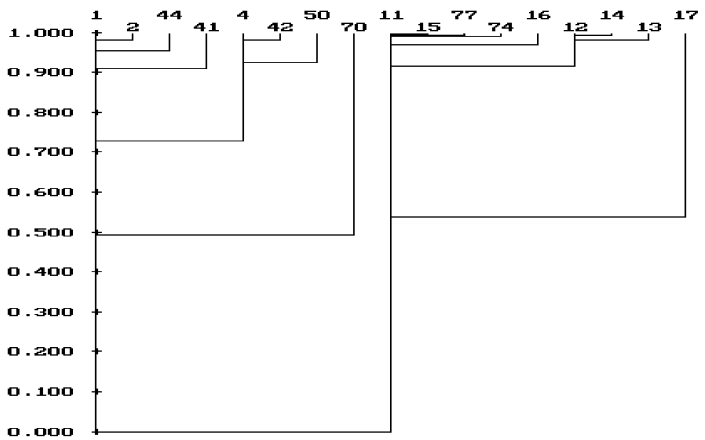

Figure 2. Dendrogram based on similarities defined in formula (1), obtained by the complete linkage *method – for seventeen suspected outliers found in the milk container data. According to the branches of the dendrogram we may subdivide the entire set into five subsets*

3 Application to the *Milk container* data

The data comprise observations of $p = 8$ variables for $n = 86$ milk containers. Description and history of the data may be found in (Atkinson 1994), who has identified in the data 17 outliers. One of the points (no. 70) is supposed to contain two typing errors. Two others (no. 63 and no. 64) are identical.

To detect outliers, Atkinson has used the proposed by him *stalactite plots*.

The data set was also analyzed with the same goal by Muruzábal and Muñoz (1974); they have used for this purpose the so called *self organizing maps* (a neural network method proposed by T. Kohonen).

We have analyzed the data using our grand tour method (Bartkowiak & Szustalewicz 1997, Bartkowiak & Szustalewicz 1998) with the results, that the following 17 data vectors were indicated as outliers:

1, 2, 4, 11, 12, 13, 14, 15, 16, 17, 41, 42, 44, 50, 70, 74, 77.

The respective count plot exhibiting these data points is shown in Figure 1.

Comparing our set of suspected outliers with that found by Atkinson, we state that 15 points are identical, and 2 are exclusive different: our list contains additionally points {4, 11}, while Atkinson's list contains {3, 75}. Yet let us say that Atkinson analyzed 85 points only (with no. 70 omitted).

Muruzábal and Muñoz have identified only points {70} and {74, 12, 13, 14, 15}, and eventually {11, 77}.

Taking into account the similarities between the 17 indicated vectors and carrying out an agglomerative cluster analysis using the complete linkage method – we find from the dendrogram shown in Figure 2, that the entire set of 17 outliers may be subdivided into following subgroups:

S1={1, 2, 44, 41}, S2={4, 42, 50}, S3={70}, S4={11, 15, 77, 74, 16, 12, 14, 13} and S5={17}.

Subsets S1 and S2 are relatively close. Subset S4 is located rather opposite to subsets S1 and S2. Subsets S3 and S5 contain only one data point each.

Repeated analysis by the grand tour method confirms these statements. Some snapshots from the grand tour projections are shown in Figure 3 and Figure 4. They show also the positions of points no. 4 and no. 11 from our list.

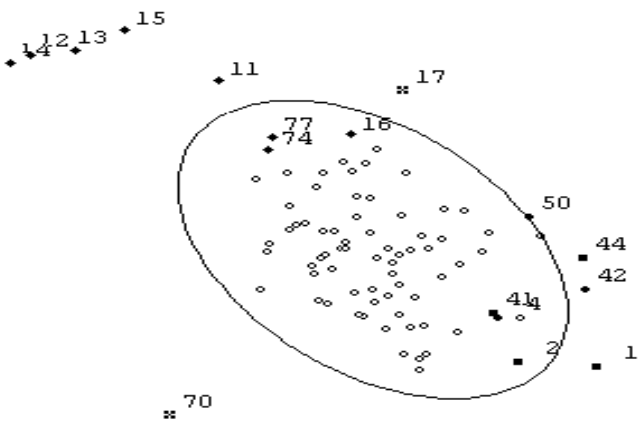

Figure 3. One exemplary projection obtained by the grand tour method with the suspected outliers highlighted

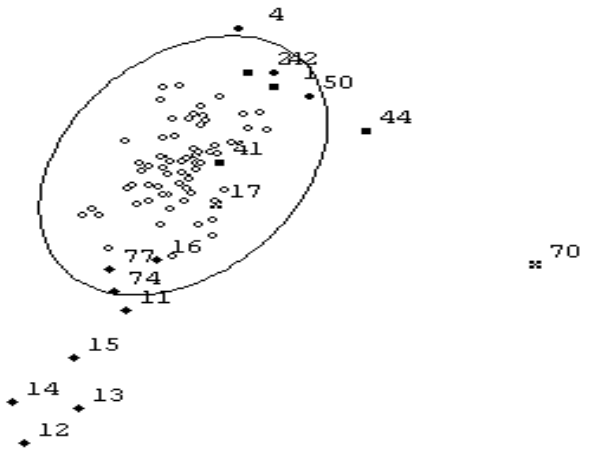

Figure 4. Another projection obtained by the grand tour method

References

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 outliers', *Statistics and Computing* **3**, 27–35.

Bartkowiak, A. & Szustalewicz, A. (1997), 'Detecting outliers by a grand tour', *Machine Graphics & Vision* **6**(4), 486–505.

Bartkowiak, A. & Szustalewicz, A. (1998), 'Watching steps of a grand tour', *Manuscript*, pp. 1–28, submitted.

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

Hadi, A.S. & Velleman, P. (1998), 'BACON: Blocked adaptive computationally efficient outlier nominators', *Manuscript*, pp. 1–23, submitted.

Muruzábal, J. & Muñoz, J. (1997), 'On the visualization of outliers via self-organizing maps', *Journal of Computational and Graphical Statistics* **6**, 355–382.

Rocke, D.M. & Woodruff, D.L. (1996), 'Identification of outliers in multivariate data', *Journal of the American Statistical Association* **91**, 1047–1061.