Using Data Images for Outlier Detection

David J. Marchette and Jeffrey L. Solka

May 20, 2002

Abstract

The data image has been proposed as a method for visualizing high dimensional data. The idea is to map the data into an image, by using gray-scale (or color) values to indicate the magnitude of each variate. Thus, the image for a data set of size n and dimension d is a $d \times n$ image, where the columns correspond to observations and the rows to variates. We consider the application of this idea to the detection of outliers, providing a simple visualization technique that highlights outliers and clusters within the data.

Key Words: Data Image; Color histogram; Interpoint distance matrix; hierarchical clustering; Outlier detection.

1 Introduction

This paper considers a graphical technique for investigating outliers in high dimensional data. Outliers are, by definition, observations that are far from the rest of the data. This presupposes a definition of "far", which is provided by a user specified distance metric. The technique we propose utilizes the data image to visualize the interpoint distance matrix, with an ordering on the observations designed to highlight observations that are potential outliers.

One of the earliest published descriptions of the data image is Ling [1973]. This paper plotted the data image as a matrix of characters, using the fact that different characters contain different amounts of ink, to control the gray scale value of the point. This provides a very low-tech, but nonetheless useful display of the data. Also, as with most visualization techniques, Bertin [1967] describes a version of the data image. A more recent description of the technique can be found in Wegman [1990], where the data image is referred to as a "color histogram." See also Minnotte and West [1998] for discussion and examples of the data image.

Most of the examples in this paper will be two dimensional, so we can use simple scatter plots to illustrate the functioning of the data image approach. In addition to these pedagogical examples, we will illustrate the technique using a high dimensional data set gathered from an artificial olfactory device or "artificial nose".

An example of the data image is depicted in Figure 1. We have plotted the observations from the Setosa and Versicolor species of the well-known Iris data set. This is a 4 dimensional data set, and we have ordered the data so that the two species are separated. Above the plot is a dendogram, showing the clustering that the R routine "hclust" produces (see Ihaka and Gentleman [1996] for information on the R language). The observations are ordered according to this dendogram.

It is easy to see that the two classes are distinct, by considering the different intensities of the image in the variates. One can also see sub-class structure in the Versicolor, and some indication of sub-class structure in the Setosa. These subclasses are easier to see in the dendogram. However, it seems easier to make decisions on the number of distinct clusters using the data image in concert with the dendogram, rather than from the dendogram itself.

The ordering of the observations and/or the variates is an important aspect of the data image. This allows us to extract useful information about the data from the visual pattern of the image. An unsorted data image tends to look like a random image, from which any information gleaned is usually spurious.

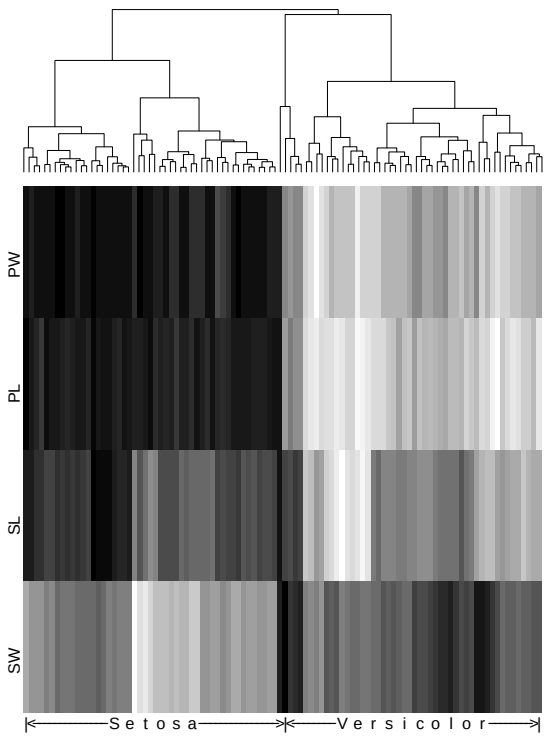

Figure 1: Data image for the Setosa and Versicolor species of irises. The variables are: sepal width and length and petal width and length, as indicated by the abbreviations.

2 Interpoint Distance Matrices

In order to define outliers, one must have a method for comparison of observations. Those observations that are "far" from, or incongruent with, the bulk of the data are considered outliers. In this paper, we will be concerned with data for which a distance has been defined, and we will use this distance to define "far".

We define the nxn interpoint distance matrix to be the matrix (a_{ij}) where the ijth entry is the distance between observations i and j and n is the number of data points. This presupposes a distance metric, and the choice of this metric is critical to the problem of outlier detection.

Figures 2 and 3 illustrates this issue. The data all lie on a line, except for one observation that lies off the line, as depicted in Figure 2. Two of the observations are depicted using different symbols, a triangle and a plus, and these are two points that are obvious candidates for outliers.

One comment on the display of the data images is in order. For interpoint distance matrices we will usually use a color map that depicts a zero value as white, with larger values going toward black in the display. We have found that this usually produces pictures that are lighter than they otherwise would be. In a few cases we find that reversing the colormap produces a slightly better display. The colormap used can be inferred from the values along the diagonal, which always correspond to a distance of zero.

The observations are ordered according to a hierarchical clustering of the distances. Thus, each column of the inner point distance matrix is treated as an n dimensional vector, and these vectors are clustered via the R routine "hclust" (or any other clustering method one wishes to use; see for example Hartigan [1975]). This ordering places observations that are far from the rest of the data together, which makes outlier detection a relatively easy visual task.

Note that we could just as easily use the ordering defined by the clustering of the original data, instead of the interpoint distances. The advantage of using the interpoint distances to order the data image is that observations

that are the same distance away from all the others appear close in the image. Thus outliers are grouped roughly according to how "outlying" they are.

In a data image of an interpoint distance matrix, one looks for a "v" or "+" of dark (large distances). These correspond to points that are far from the rest of the points. The "v" or "+" will have a vertex (or crossing point) of light, corresponding to the observations that are close to the potential outlier. In the case of a singleton outlier, this will be seen as a single white square at the vertex. This is illustrated in Figure 3.

The data image of the Euclidean interpoint distance matrix for these data (Figure 3, left) shows one potential outlier, which corresponds to the triangle. This observation is far from all the others. However, since it lies on the line defined by (all but one) of the other observations, one might not want to consider it an outlier. Also, the obvious outlier depicted by the "+" symbol is not detectable in this image (it is shown as the right most observation in the image).

Since the data are (nearly all) linear, it might make sense to try a different metric. The data image on the right of Figure 3 depicts the data image for the Mahalanobis distance. This shows clearly the two outliers as the "v" in the upper right corner. This example illustrates quite clearly that the distance chosen is critical to the detection of outliers. In fact, it is critical to the *definition* of outliers. In some applications, the observation denoted by the triangle might not be considered an outlier at all, since it lies on the line defined by the rest of the data (minus the "+" outlier). On the other hand, it is clearly quite separated from the other observations, and so should probably be flagged as suspicious. Note that it looks essentially the same as the "+" outlier. A further comment is that the Mahalanobis distance was computed using all the data. Recomputing using all but the outliers would result in a singular matrix, in this trivial example, due to the perfect linearity of the data.

3 Examples

In this section we present several examples of outlier detection using the data image. The first few are taken from the book Rousseeuw and Leroy [1987], and thus provide a comparison with standard outlier detection techniques. In each case distance is computed wtih repect to the standard Euclidean distance metric unless otherwise specified.

Figure 4 shows the relationship between body weight and brain weight for a number of living and extinct animals (the values for the extinct animals are estimated from their skull size and skeletons). Five outliers are numbered on the scatter plot, and their corresponding columns are numbered on the data image.

The data image clearly shows the outliers, and further shows the subclusters corresponding to observations 2 and 3 (diplodocus and triceratops) and 4 and 5 (Asia and Africa elephant). There is considerable overplotting in the scatter plot. This overplotting corresponds to the large near-white box in the data image, depicting a block of observations that are all close together.

The next example is the stackloss data set from Brownlee [1965]. Rousseeuw and Leroy [1987] states that most researchers conclude that observations 1, 3, 4 and 21 are outliers, and that some would add observation 2 to this list. Figure 5 shows the data image on the interpoint distance matrix and the pairs plot of the data.

The data image shows three main outliers, corresponding to observations 1,2,3. If one stretches one's imagination, one might call the two observations in the last two columns outliers. Since these correspond to observations 4 and 21, this is the complete set of outliers reported by Rousseeau. Realistically, however, this approach seems to have only detected three of the five outliers.

Most of the analysis on these data is concerned with the prediction of the stackloss (the last variate) via the other variables, and most researchers have investigated linear regression to perform this prediction. Thus, it seems appropriate to consider the Mahalanobis distance. Figure 6 depicts the interpoint distance matrix for the stackloss data using Mahalanobis distance. In this case only observation 21 shows up (weakly) as a potential outlier in the first column. The next five columns correspond to observations 17, 4, 2, 1, and 3, but these are not obvious candidates, according to this plot. In fact, the evidence for observation 21 as an outlier is quite shaky in this plot.

Since we have posited observations 1–3 as outliers from Figure 5, it makes sense to eliminate them from the calculation of the covariance matrix used in the Mahalanobis distance. Figure 7 depicts the resulting interpoint distance matrix. In this case we see that the first four columns correspond to apparent outliers (these are observations 1–4) and the "+" in the interior of the plot is another potential outlier (observation 21).

As this example has demonstrated, it is important to determine the appropriate metric for computing the interpoint distance matrix, that correctly takes the geometry of interest into account.

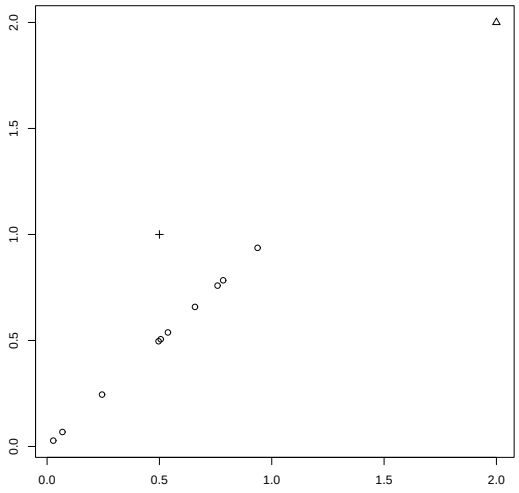

Figure 2: Scatter plot of data lying on a line, with one outlier off the line.

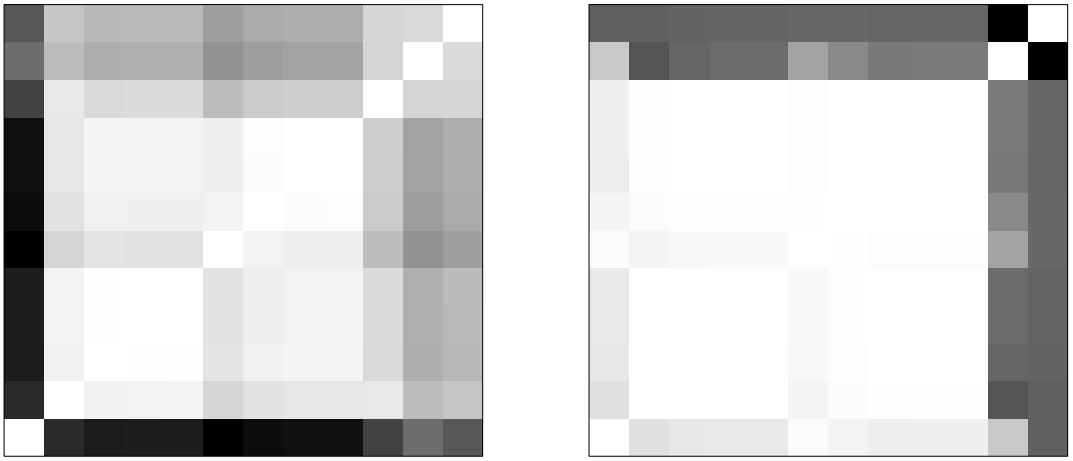

Figure 3: Data images of the interpoint distance matrix using Euclidean distance(left) and Mahalanobis distance (right) for the data depicted in Figure 2. The dark "v" in the lower left and upper right corners of the plots are indicative of potential outliers.

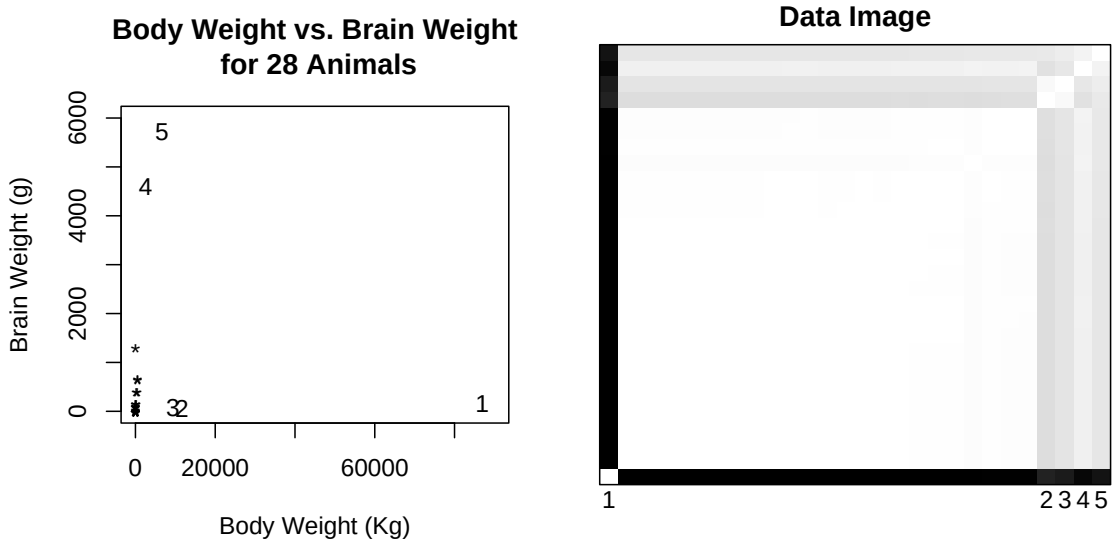

Figure 4: Body and brain weight from Rousseeuw and Leroy [1987], page 57, originally from Jerison [1973] and Weisberg [1980]. The five outliers are numbered in both plots, and correspond to: brachiosaurus, diplodocus, triceratops, Asia elephant and Africa elephant.

Figure 8 depicts the artificial data set originally from Hawkins et al. [1984]. The data image of the interpoint distance matrix has a set of unusual points in the upper right hand corner: the first block of 10 observations correspond to the first ten observations in the data, that Rousseeau calls "bad leverage points", and the last four columns correspond to the next four observations which Rousseeau calls "good leverage points". That is, their x values are outlying, but their y values fit the underlying model well. These correspond roughly in character to the observation denoted by a triangle in Figure 2. Note that the ordering allows us to see the clustering of these two groups quite distinctly. This informs the user that the outliers are not homogeneous, and gives the user the opportunity to explore the two groups separately.

An example that is a little different from the others presented so far is the one depicted in Figure 9. A variant of this example was suggested by Dan Carr. In this case, the obvious outlier, in the center of the ellipse, will not show up as an outlier in any linear projection of the data. Figure 10 depicts the data image for these data for two distance metrics. Note that the outlier stands out quite clearly as a "+" (or cross) in both images, although one could argue that it is slightly more obvious in the Mahalanobis distance case. The uniform gray value for the arms of the "+" indicates that this observation is equidistant from all the other observations, a property that makes it stand out from the other observations.

The detection of the central point in Figure 10 is not specific to two dimensional data. To illustrate this, we drew data uniformly on a 5 dimensional sphere (Figure 11). As can be seen, there is a distinct "+" in the data image of the interpoint distance matrix, corresponding to the central outlier.

Another interesting data set is the Tufts artificial nose dataset Priebe [2001]. The artificial nose is a sensor made up of 19 chemically doped fibers. For each "sniff", the fibers react to the odorant by a characteristic fluorescence pattern, and this fluorescence is measured at two frequencies for 60 time steps. Thus, each observations consists of $19 * 2 * 60 = 2280$ values. These could be viewed as points in a very high dimensional space, or, more properly, as 38 functions sampled at 60 points.

There are 1112 observations in the data set, consisting of 760 observations of the ground water contaminant TCE in several concentrations and mixed with several confuser chemicals, and 352 observations of the confusers (and air) with no TCE present. The data image for the interpoint distance matrix of the TCE present data is depicted in Figure 12.

In Figure 12, it is clear that there is a cluster of outliers in the lower corner. These correspond to TCE in

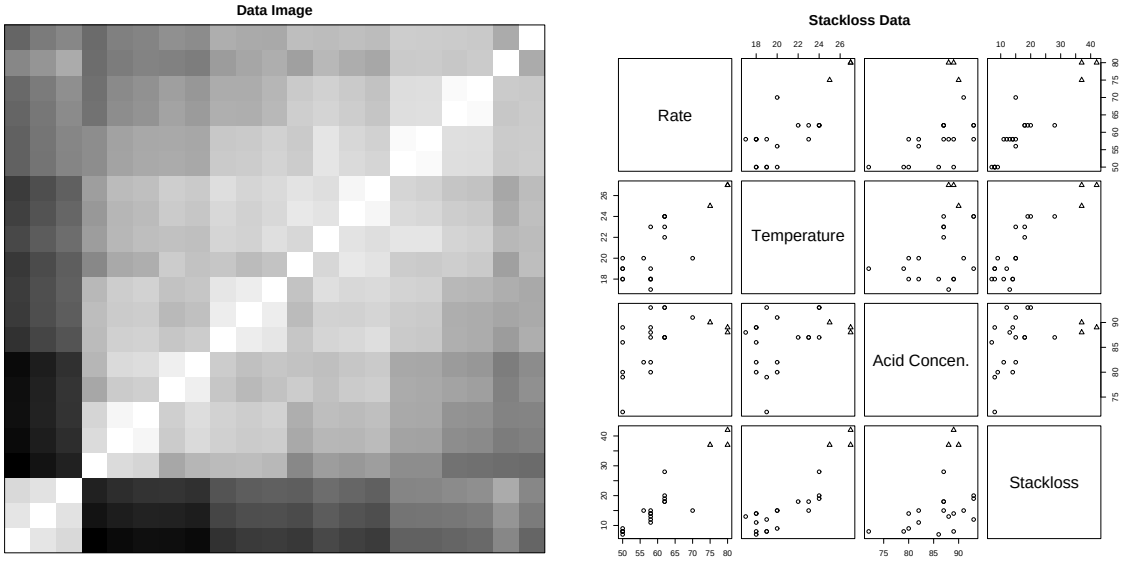

Figure 5: Stackloss data from Rousseeuw and Leroy [1987], page 76, originally Brownlee [1965]. The three outliers detected in the data image are depicted with triangles in the pairs plot.

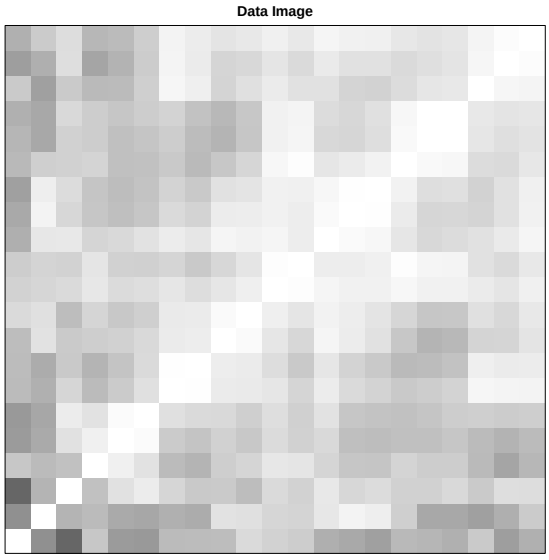

Figure 6: Data image for the Mahalanobis distance matrix for the stackloss data of Figure 5.

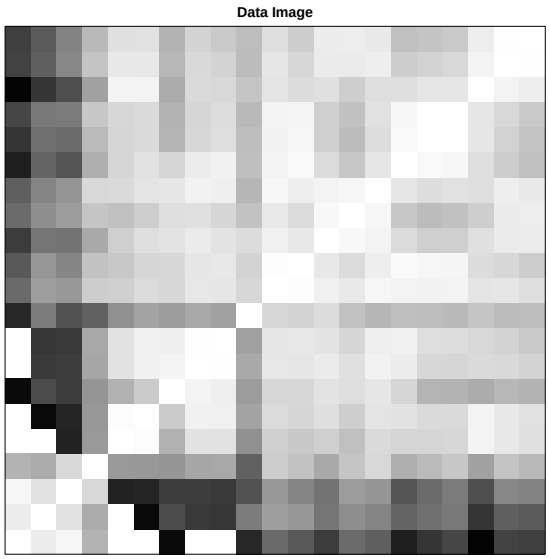

Figure 7: Data image for the Mahalanobis distance matrix for the stackloss data of Figure 5, where the covariance in the Mahalanobis calculation is constructed using observations 4–21.

chloroform. Further analysis has determined that the chloroform observations are indeed quite different from all the other observations. The other "v"s and "+"s in the Figure are more complicated, and have not yet been fully characterized. The upper right corner seems to be primarily observations with low concentrations of TCE, but these are not pure, and further data analysis is necessary to fully characterize these.

4 Discussion

The data image is a powerful tool for the display and analysis of high dimensional data. Utilizing it to display the interpoint distance matrix allows one to detect outliers visually, regardless of the dimensionality of the data.

The choice of metric is critical to the detection of outliers. We have seen examples where metrics other than Euclidean are appropriate, and this choice is very problem dependent. The selection of appropriate metrics for a task is a difficult one, which is often performed in an ad hoc manner by the analyst. The technique used in this paper presupposes the choice of the metric, and so the utility of the technique is dependent on the user's ability to select the metric appropriately. Work on metric selection is ongoing; see for example Priebe et al. [2000].

One issue that has not been addressed is the one of large sample sizes. The data image, as presented here, cannot handle more observations than the number of pixels available to display the data. One potential solution is to use the cluster tree to selectively collapse points to their cluster center. For example, in Figure 4 the observations corresponding to the large square of light could have been collapsed into a smaller region, since their variability in interpoint distances is small compared to those of the outliers. Thus, a measure of the variability of within cluster distances could be used to bin the data. Another solution, for moderate sized datasets, would be to save the data image as an image, and use tiling and panning techniques developed by the image processing community to traverse large images.

The data image of the interpoint distance matrix is redundant, since the interpoint distance matrix is symmetric. It would be nice to better utilize the extra space available due to this redundancy. One possibility would be to use a different distance metric in the other half of the image.

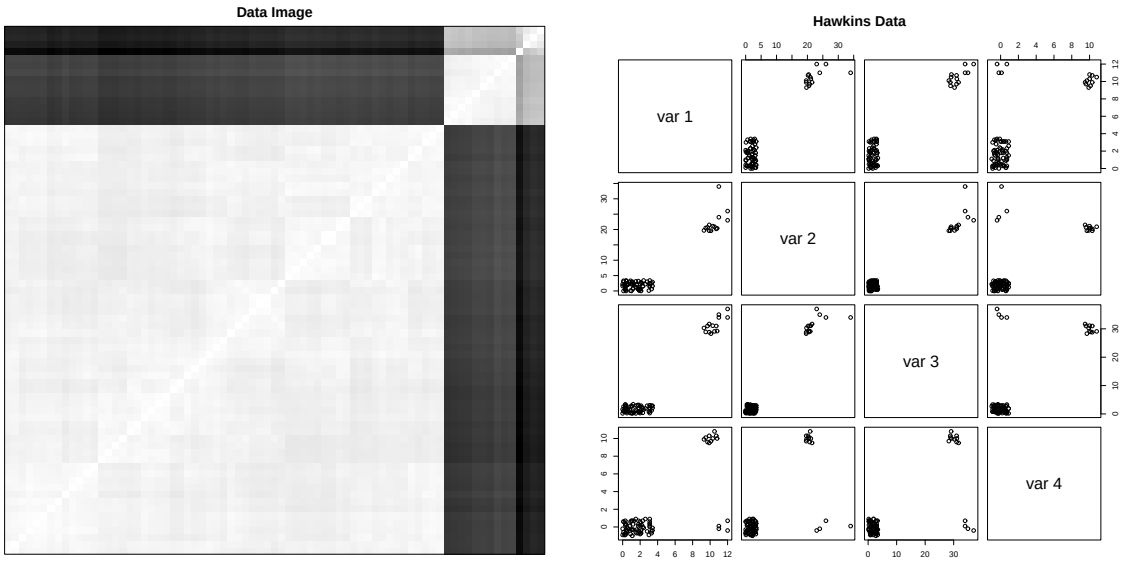

Figure 8: An artificial data set, Rousseeuw and Leroy [1987], page 94, originally from Hawkins et al. [1984].

References

J. Bertin. *Sémiologie Graphique.* Editions Gauthier-Villars, Paris, 1967. (English translation by W. J. Berg as *Semiology of Graphics*, University of Wisconsin Press, Madison, 1983).

K. A. Brownlee. *Statistical Theory and Methodology in Science and Engineering.* John Wiley & Sons, New York, 1965.

J. A. Hartigan. *Clustering Algorithms.* John Wiley & Sons, New York, 1975.

D. M. Hawkins, D. Bradu, and G. V. Kass. Location of several outliers in multiple regression data using elemental sets. *Technometrics*, 26:197–208, 1984.

R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *IEEE Transactions on Information Theory*, 5:299–314, 1996.

H. J. Jerison. *Evolution of the Brain and Intelligence.* Academic Press, New York, 1973.

R. F. Ling. A computer generated aid for cluster analysis. *Communications of the ACM*, 16(6):355–361, 1973.

M. Minnotte and W. West. The data image: a tool for exploring high dimensional data sets. In *1998 Proceedings of the ASA Section on Statistical Graphics*, pages 25–33, 1998.

C. E. Priebe. Olfactory classification via interpoint distance analysis. *IEEE PAMI*, 23:404–413, 2001.

C. E. Priebe, D. J. Marchette, and J. L. Solka. On the selection of distance for a high-dimensional classification problem. In *American Statistical Association, 2000 Proceedings of the Statistical Computing Section and Section on Statistical Graphics*, pages 58–63, 2000.

P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* John Wiley & Sons, New York, 1987.

E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *JASA*, 85:664–675, 1990.

S. Weisberg. *Applied Linear Regression.* John Wiley & Sons, New York, 1980.

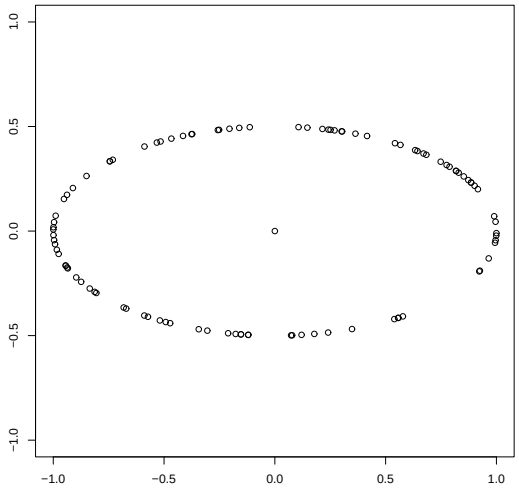

Figure 9: An artificial elliptical data set.

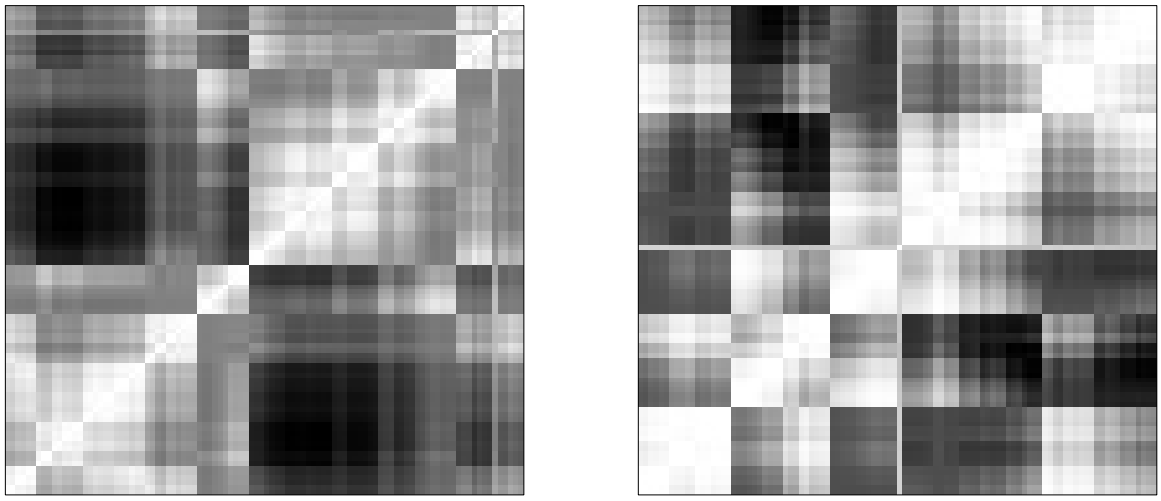

Figure 10: Data images for the Euclidean (left) and Mahalanobis (right) distance matrices for the data in Figure 9.

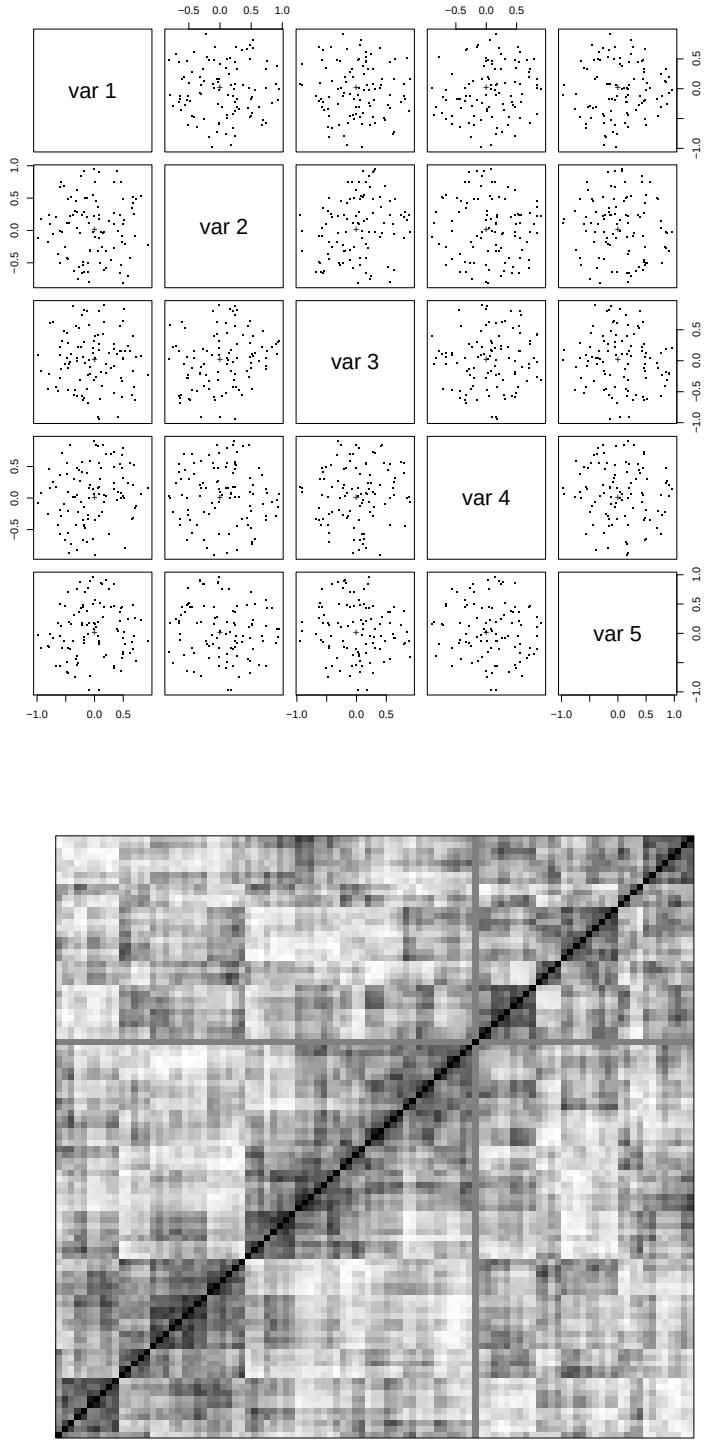

Figure 11: Pairs plot and data image of the interpoint distance matrix for a 5 dimensional data set. 100 observations were drawn uniformly on the 5 dimensional sphere, and one observation (indicated by a "+" in the pairs plot) was placed at the origin.

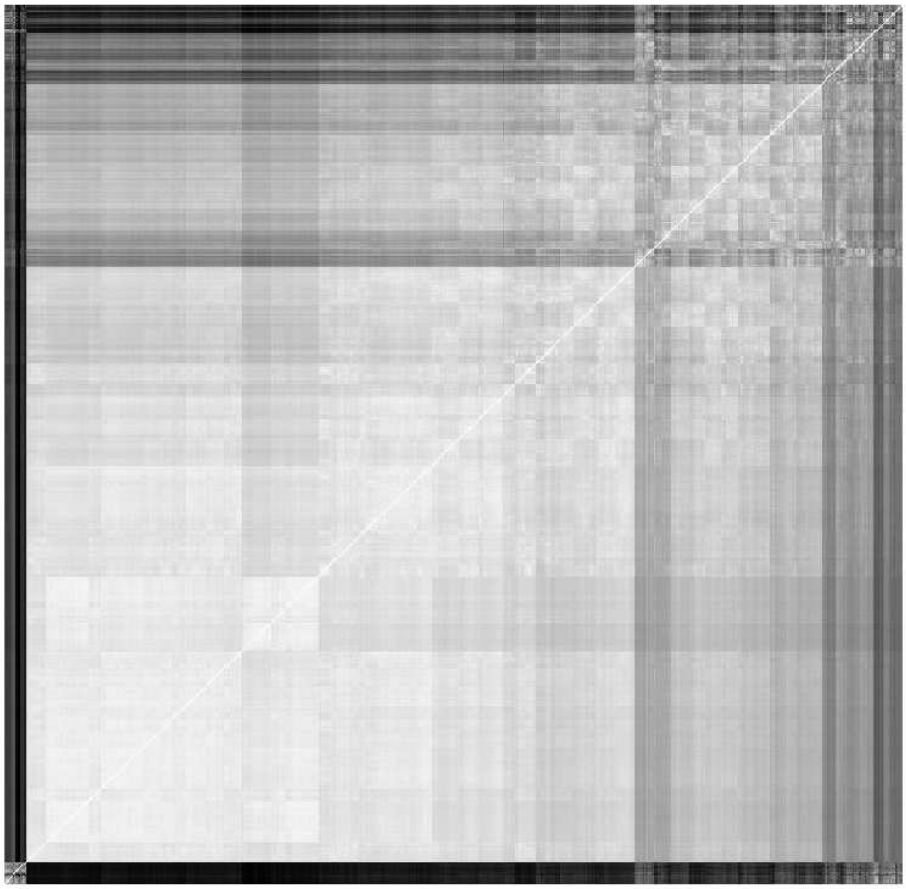

Figure 12: Artificial nose data, TCE present. The outliers are TCE in chloroform.