Quaderni di Statistica
Vol. 2, 2000

Peeling multivariate data sets: a new approach

Giovanni C. Porzio
Dipartimento di Economia e Territorio, Università degli Studi di Cassino
E-mail: porzio@eco.unicas.it

Giancarlo Ragozini
Dipartimento di Scienze Statistiche, Università degli Studi di Napoli Federico II
E-mail: giragoz@unina.it

Summary: Peeling a data set consists of identifying its successive layers, from the outermost to the innermost. It is used for many purposes in multivariate data analysis, including data ordering, trimming, outlier detection, robust estimation of location, correlation and probability contours. However, existing peeling procedures, mainly based on the convex hull idea, require a remarkable computational effort (many convex hulls have to be computed), and can fail with respect to their own goals in presence of irregularities at the boundary of the data region. To overcome these drawbacks, noting that in practice peeling procedures are essentially used to identify the bulk of the data, we propose a new peeling approach that splits the observations in just two layers. In particular, the method distinguishes between $inner$ and $outer$ data, identifying as $outer$ those observations that lie closer to the boundary region than to the remainder (the $inner$ data). It exploits the first outer convex hull and a quasi-clustering procedure: observations close to the boundary region are clusterized through appropriate radial projections around the convex hull vertices.

Key words: Convex hull, Partial ordering, $Outer$ and $inner$ data.

1. Introduction

The aim of a peeling procedure consists of identifying nested layers of a data set, assigning to each observation an index which considers the proximity of that point with respect to the outside of the data swarm. To determine such an index, an assigned shape with minimum volume that contains the data is computed, and observations lying on its border take index value one. Then, the procedure is iterated on the remaining observations yielding a sequence of k shapes. Data points lying on the border of the $i-th$ shape ($i = 1, ..., k$) will take index value i (Green, 1981).

Although given geometric shapes can be considered, such as rectangles (Nath, 1971), ellipses (Silverman and Titterington, 1980), or circles (Daniels, 1952), the main peeling approach is based on the convex hull of the data (i.e. the smallest convex set containing them). Indeed, usually it is not possible to make assumptions on the data set shape, and hence the use of a more flexible and conservative figure such as the convex hull is more appropriate (Brooks *et al.*, 1988).

Convex hull peeling is used in different settings in statistics. Considering the analogy between the convex hull vertices and the extremes of an univariate set, it has been applied to order multivariate data (Barnett, 1976; Eddy, 1981). Eddy and Hartigan (1977) used it to estimate probability density contours, while Bebbington (1978) exploited the peeling to trim bivariate data sets in order to obtain a robust estimate of the correlation coefficient. Recently, a bivariate boxplot based on this kind of method has been proposed by Zani *et al.* (1998), while Liu *et al.* (1999) considered convex hull peeling within the framework of multivariate analysis by data depth.

However, this peeling approach has some drawbacks, related to the computational effort (Petitjean and Saporta, 1992) and the effectiveness of the procedure (Donoho and Gasko, 1992).

As in the practice of data analysis a complete convex hull peeling may be not required, in this paper we propose a suitable peeling approach that identify only two layers, distinguishing between $inner$ and $outer$

data. The method allows at the same time to correctely identify the bulk of the data, even under irregular data structure, and to avoid intensive computations.

The paper is organized as follows. In section 2 we discuss convex hull peeling along with its main drawbacks, while section 3 proposes a distinction between $outer$ and $inner$ data, giving foundations for our approach. Sections 4 and 5 illustrate our proposal along with some examples, while final remarks follow in section 6.

2. Convex hull peeling

Convex hull peeling consists of computing iteratively the nested hulls of the set, from the outermost to the innermost. At each step, the convex hull vertices are deleted, the convex hull of the remainder is constructed and new vertices are identified. Each convex hull in the nested sequence defines a deeper layer, from the outer to the inner. As an illustration, in Figure 1 we present a complete convex hull peeling for a 50 observation data set from a bivariate normal distribution.

Some drawbacks affect convex hull peeling. First, the computational effort for a complete peeling of the data is heavy, especially when dimensions increase (Petitjean and Saporta, 1992). To overcome this drawback, it has been proposed to consider only few nested layers, or to peel projected data. However, both approaches could be misleading. The first one is uneffective in the case of multiple outliers, while projections could hide the real structure of the data.

Besides, we note that the existing peeling procedures could fail with respect to their own goals. With irregular data structures, such as clustered data along the boundary or marked asimmetry, they assign the same index layer both to observations lying on the periphery of the data region and to others belonging to its bulk (see also Donoho and Gasko, 1992). To show such an effect, we consider the *ad hoc* simulated data set displayed in Figure 2.

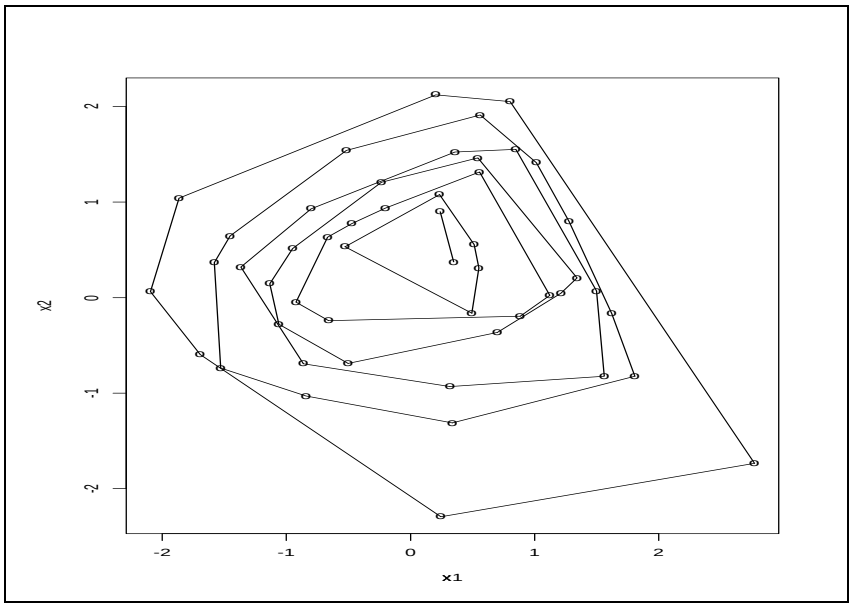

Figure 1: *Simulated data set. Complete convex hull peeling.*

The data set consists of 200 observations generated from a mixture of two normal distributions:

$$(1-\alpha)\ \Phi(x|\mu_1,\Sigma)\ \ +\ \ \alpha\ \Phi(x|\mu_2,\Sigma),$$

with

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 0.5 \\ -5 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

The mixture parameter α was set to 0.10, and the second distribution was shifted in mean in order to obtain a small well-separated subsample.

According to the classical peeling procedure, the first seven nested convex hulls were superimposed to the data set. We note that the same index layer (e.g. the 6-th) is assigned to observations belonging both to the periphery and to the core of the main structure of the data. As a consequence, if the identification of the main structure of the data is the issue, the procedure will fail. If in addition, according to the practice,

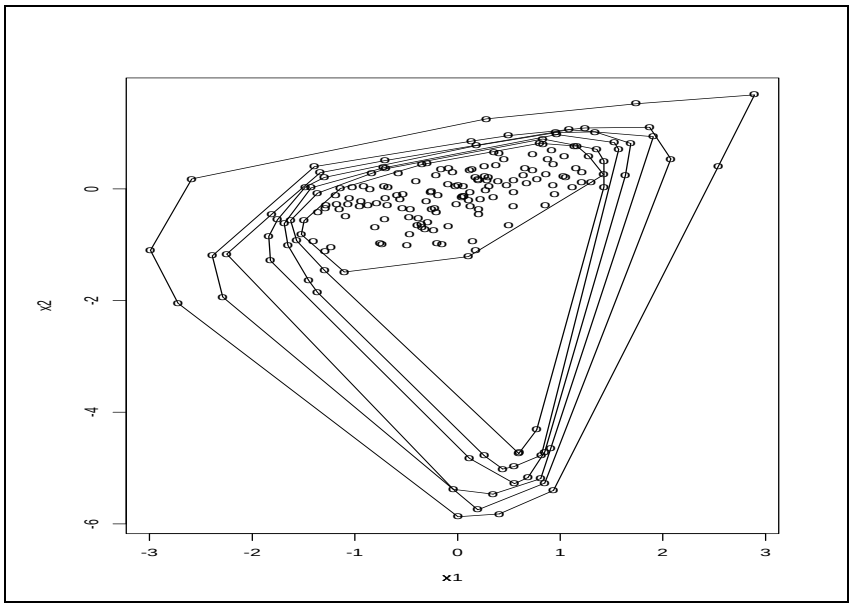

Figure 2: *The classical convex hull peeling when a cluster of data lies on the periphery of the scatter. The first seven layers.*

observations lying on the most outer convex hull layers are removed, it will be caused a loss of information.

To overcome the above drawbacks, we propose a new peeling approach, based on a distinction between $outer$ and $inner$ data.

3. $Outer$ *and* $inner$ *observations: a new peeling approach*

In order to analize a data set structure, a fast and broad ordering could be as useful as a complete one. In the univariate case, Tukey (1977, Chap.2) proposed a sort of partial order, distinguishing among far out, outside, adjacent observations and inner values. In the multivariate case, we propose to distinguish among $outer$ and $inner$ observations, and we provide the corresponding tool to allocate the observations according to

such a distinction.

We intend as $outer$ observations those far from the rest of the data scatter, closer to its boundary then to the bulk (in some sense they correspond to the far out and the outside observations of Tukey's univariate case). As a counterpart, we call $inner$ observations those that are not $outer$.

Accordingly, to split the data in two layers (the $outer$ and $inner$ ones), we propose a new peeling approach that combines the main idea of the classical convex hull peeling procedure with an *ad hoc* quasi-clustering method.

We identify as $outer$ observations the convex hull vertices along with their closest data points. The $inner$ layer is then defined as the complement to the $outer$ one with respect to the whole data set.

The convex hull vertices are included in the first layer as they are the most extreme observations by definition (Barnett, 1976). Then we clusterize around them all the observations lying in a neighbourhood of the boundary region. In particular, points closer to the vertices than to the rest of the data will be included in the $outer$ layer.

To identify which are the closest observations, we consider for each vertex the distances to the remaining points along a radial projection. Then, along each of these directions, we look at the univariate ordering of the points from the closest to the furthest. The presence of $gaps$, if any, in these univariate orderings will point out some empty space in the data structure, splitting the $outer$ observations from the $inner$ ones.

This kind of peeling allows us to strip out all the $outer$ observations in a single run. In this way we obtain a relevant computational gain, as we compute a single convex hull instead of a nested series. Consequently, it is more feasible to perform this peeling procedure in the complete variable space without relying on projections.

Finally, we note that our method is able to point out particular structures, such as clusters of data, on the boundary of the data region. On the other hand, in their absence, the $outer$ layer will consist essentially of the convex hull vertices.

4. Clustering around vertices: the algorithm

To clarify how our peeling method works, we present in detail the algorithm, providing a graphical illustration of some of its steps.

Given a data set of n points in p dimensions $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the peeling algorithm works as follows.

Step 1: Construct the convex hull of $\mathbf{X}$, $\mathcal{CH}(\mathbf{X})$, and let $\mathbf{V} = \{\mathbf{x}_{i^*}\}$, with $i^* \in I^*$, be the set of v vertices of $\mathcal{CH}(\mathbf{X})$;

Step 2: Compute the $v \times n$ distance matrix $\mathcal{D}(\mathbf{V}, \mathbf{X}) = \{d(i^*, j)\}$, $i^* \in I^*$, $j = 1, \ldots, n$, with $d(i^*, j) = \|\mathbf{x}_{i^*} - \mathbf{x}_j\|_2$ the distance between the i^*-th vertex and the j-th data point. Let $d(i^*, \cdot)$ be the row vector of $\mathcal{D}(\mathbf{V}, \mathbf{X})$, which collect the distances between $\mathbf{x}_{i^*}$ and $\mathbf{x}_j \in \mathbf{X}$, $j = 1, ..., n$ (note that it is not required to compute the complete distance matrix);

Step 3: For each vertex i^*, sort in ascending order the $d(i^*, \cdot)$ vector; call $d_{(k)}(i^*, \cdot)$ the k-th element in the sorted sequence. Then compute the first difference of the sorted distance vectors $\left[d_{(k+1)}(i^*, j) - d_{(k)}(i^*, j')\right]$, for $k = 1, ..., n-1$. For each vector compute the maximum of such first differences, and let k_{i^*} be the corresponding k:

$$k_{i^*} = \arg\max_k \left[d_{(k+1)}(i^*, j) - d_{(k)}(i^*, j')\right].$$

For each vertex $\mathbf{x}_{i^*}$, such a maximum identifies a possible gap in the data structure, and hence a cluster, if any, of $outer$ observations around the vertex can be selected. Let us denote with C_{i^*} such a cluster for the i^* vertex, $i^* \in I^*$.

As an illustration, Figure 3 shows some radial projections with respect to the vertex $\mathbf{x}_{i^*}$, the corresponding univariate ordering (on the right side of the figure), the gap, and the cluster C_{i^*} around $\mathbf{x}_{i^*}$ for an artificial bivariate data set with different kind of $outer$ observations created $ad\ hoc$.

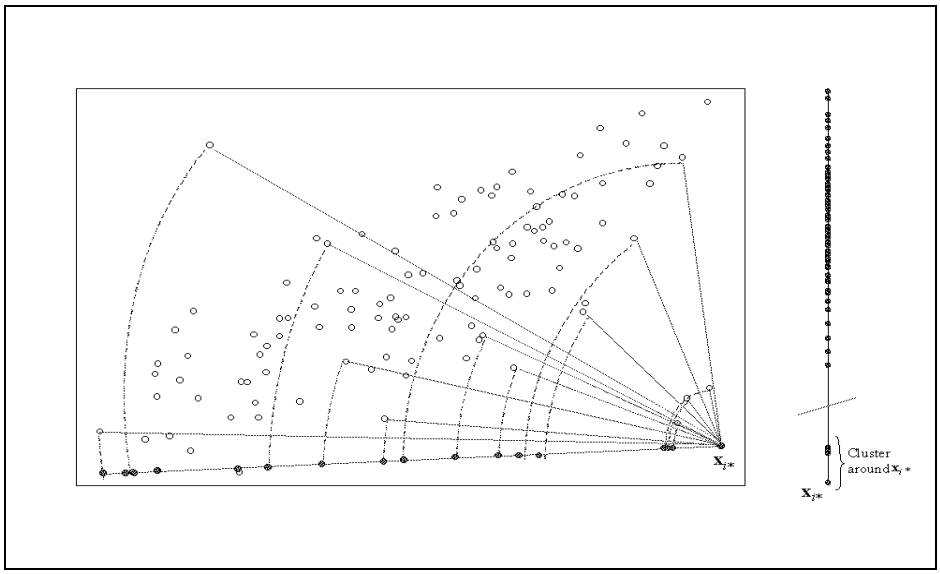

Figure 3: *Artificial data set. Step* 3 *of the proposed procedure. The radial projections with respect to the vertex* $\mathbf{x}_{i^*}$ *(left), along with the univariate ordering, the gap, and the cluster* C_{i^*} *(right).*

Step 4: Once k_{i^*} has been identified, determine the clusters C_{i^*} of *outer* observations according to the following rules.

i) if $k_{i^*} = 1$, set $C_{i^*} = \{\mathbf{x}_{i^*}\}$: in this case the maximum gap occurs next to the vertex $\mathbf{x}_{i^*}$, and hence the latter is an *outer* observation lying far from the others, (see for instance obs. 1 in Figure 4) or on the border near the core of the data (obs. 31 and 112);

ii) if $1 < k_{i^*} < \frac{1}{2}n$, set $C_{i^*} = \left\{\mathbf{x}_j : d_{(k)}(i^*, j) \leq d_{(k^*)}(i^*, j')\right\}$: in this case the maximun gap occurs before the median position, the vertex $\mathbf{x}_{i^*}$ is the outermost of an *outer* group (e. g. groups around obs. 4 and 42), and all the observations before the gap, $\mathbf{x}_{i^*}$ included, belong to the *outer* layer;

iii) if $k_{i^*} \geq \frac{1}{2}n$, set $C_{i^*} = \{\mathbf{x}_{i^*}\}$: the maximum gap occurs after the median position in the sequence of the sorted distances, usually

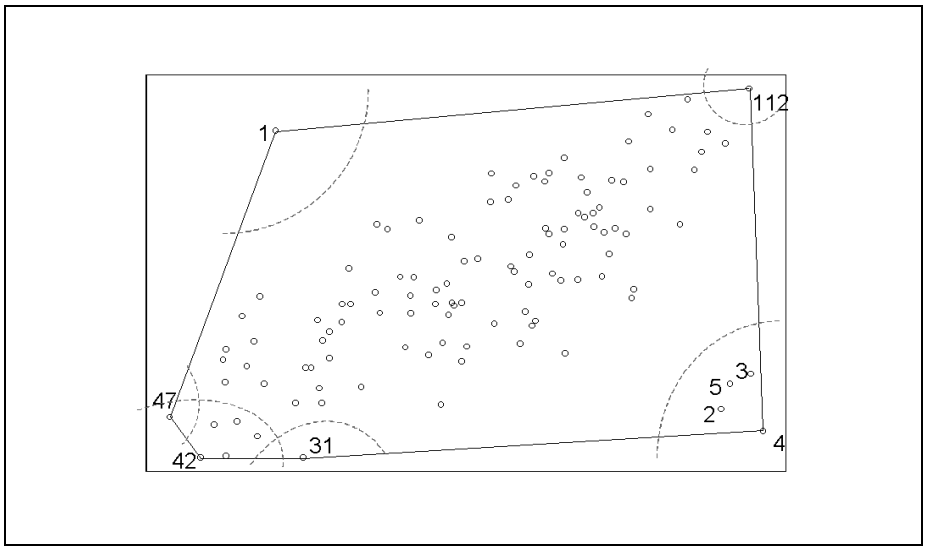

Figure 4: *Artificial data set as in Figure 3. Step* 4 *of the proposed peeling procedure: clusters of outer data with respect to the convex hull vertices.*

near the end of the sequence (e.g. obs. 47 has its maximum gap with respect to obs. 112), and $\mathbf{x}_{i^*}$ is a single *outer* observation as in i).

This last step yields hence a set of clusters, each of them containing some *outer* observations. The *outer* layer $\mathcal{O}$ is obtained as the union of such clusters: $\mathcal{O} = \bigcup_{i^* \in I^*} \{C_{i^*}\}$. The *inner* layer $\mathcal{I}$ will be its complement: $\mathcal{I} = \bar{\mathcal{O}}$.

5. Some illustrative examples

In this section, we illustrate our method when the data have a regular structure, and when such a structure presents either clustered observations along the boundary or outliers. For visual enhancement, for each data

set we compute also the convex hull of the observations belonging to the $inner$ layer, and we highlight such an $inner$ region through a shaded area in the plots.

To illustrate the method when data do not present irregularities, we generated 500 observations from a normal bivariate distribution. The corresponding scatterplot is displayed in Figure 5, along with the shaded area corresponding to the $inner$ layer.

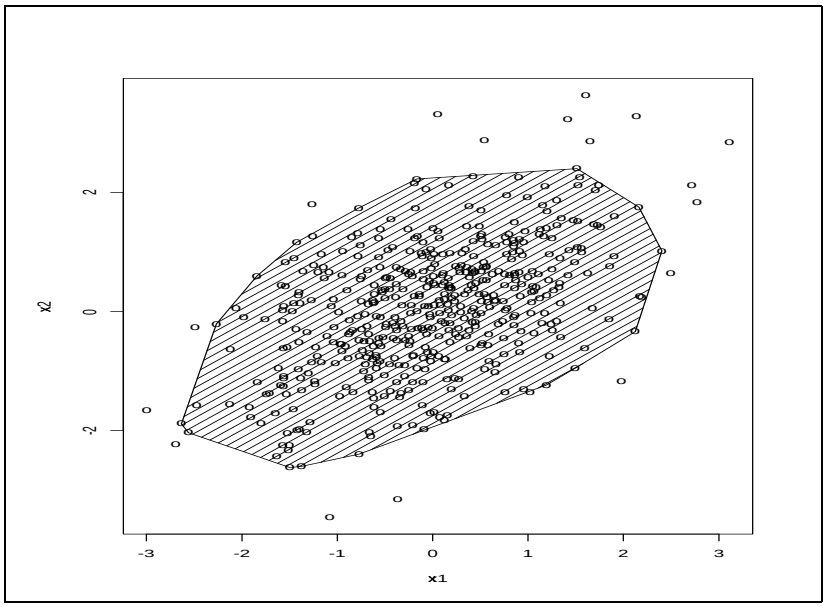

Figure 5: *Normal data scatter. The proposed peeling approach. The shaded area in the plot corresponds to the $inner$ layer. The remaining data are $outer$ observations.*

We note that the $outer$ layer includes just the observations lying on the first convex hull and few others thinly scattered. In other words, in such a case, our approach yields results in agreement with the classical peeling procedure.

To allow for the presence of unusual structures, we consider again the data set displayed in Figure 2 (200 observations generated from a mixture

of two normal distributions). For these data, we recall that the classical peeling procedure fails, combining $inner$ and $outer$ data up to the 6^{th} layer (Figure 2). On the other hand, our proposed approach correctly selects as $outer$ the 20 clustered observations plus few more (Figure 6). That is, the method identify properly the $outer$ and $inner$ layer, providing in addition a computational gain.

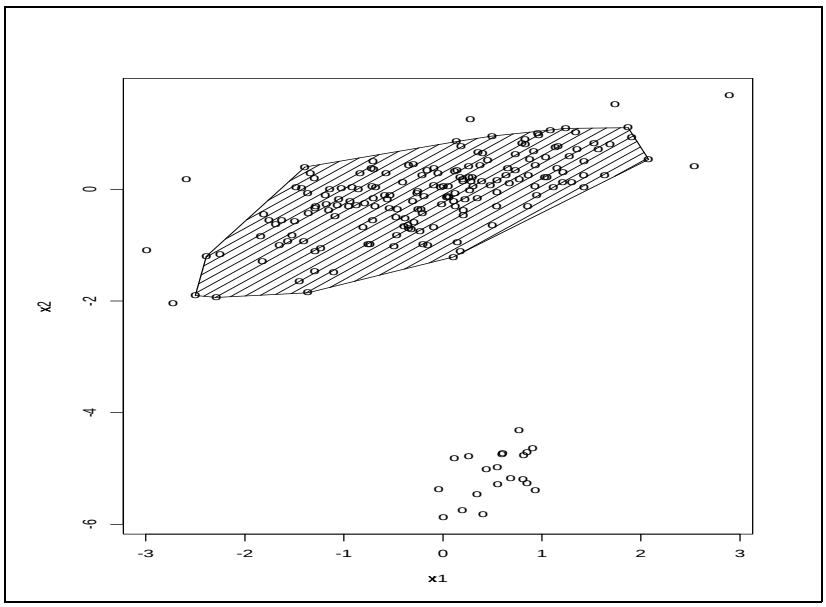

Figure 6: *Data set as in Figure 2. A cluster of data lie along the periphery of the scatter. The proposed peeling approach. The shaded area in the plot corresponds to the $inner$ layer. The remaining data are $outer$ observations.*

Finally, we consider the Brain and Body weight data (Rousseeuw and Leroy, 1987; pag. 57), a well known data set in outlier analysis, referring to the logarithm of the body and brain weight of 28 species. Rousseeuw and van Zomeren (1990) in a regression setting identified as outliers 5 observations (6, 16, 25, 14, and 17) using the minimum volume ellipsoid. Figure 7 shows the scatterplot of the data along with the shaded $inner$

region according to our procedure. The three clustered outliers on the right, and the other two atypical observations belong to the *outer* layer, although as expected this latter includes other *outer* observations as well.

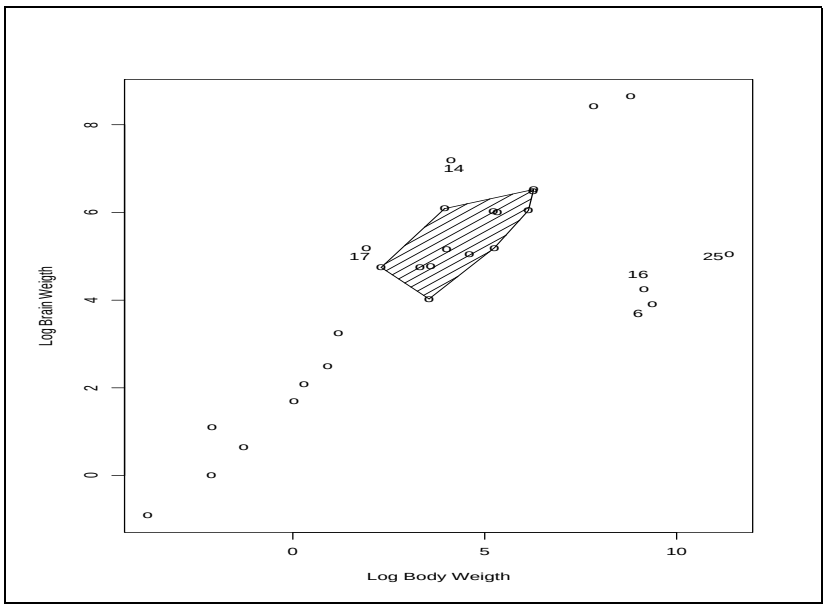

Figure 7: *Brain and Body data. The proposed peeling approach. The shaded area in the plot corresponds to the inner layer. The remaining data are outer observations.*

It is worth noting that, in this example, many data points are identified as *outer* observations. This is due to the sparseness and the regression nature of the data. Indeed, our procedure looks for a dense core of the data scatter, and does not consider the remoteness from a possible regression line.

7. Final remarks

Our approach can be viewed as a resistant partial peeling method, and can be applied in an iterative and interactive way as exploratory and diagnostic tool. It can be used to identify and compare the main structure of different groups in cluster analysis. Moreover, it may represent a first step in robust analysis and trimming procedures.

In addition, we stress that $outer$ data are not necessarily outliers. However, our procedure seems to be particularly attractive to deal with such occurence in the data. Indeed, outliers will be naturally included in what we have called the $outer$ layer. In particular, our peeling method should be well suited to deal with the masking effect, that occurs when outlier clusters hide themselves to the single case diagnostic. In fact, our approach allows to identify candidate outlier clusters in neighbourhoods of the boundary region. This property allows also to avoid the combinatorial size computations required by the classical block-deletion procedures (see also Porzio and Ragozini, 2000).

Finally, it has to be noted that, as any convex hull based procedure, our peeling approach is thought to deal with convex shape data sets. This notwithstanding, we believe that little deviation from convexity should not compromise its performance. In any case, although our approach is particularly effective with heavy and sparse tail data set, we recommend to iterate the procedure if unduly complex structures are suspected.

Further developments of this work could involve the possibility of considering more layers through an iterative use of our peeling approach. In this way a finer ordering of the whole data set could be obtained.

Acknowledgment: Our work was supported by funds from the MURST.

References

Barnett, V. (1976) The ordering of multivariate data (with discussion), *Journal of Royal Statistical Society A*, 139, 318-354.

Bebbington, A.C. (1978) A method of bivariate trimming for robust estimation of the correlation coefficient, *Applied Statistics*, 27, 221-226.

Brooks, D. G., Carroll, S. S., Verdini, W. A. (1988) Characterizing the Domain of Regression Model, *The American Statistician*, 42, 187-190.

Daniels, H. E. (1952) The covering circle of a sample from a circular normal distribution, *Biometrika*, 39, 137-143.

Donoho, D.L., Gasko, M. (1992) Breakdown properties of location estimates based on halfspace depth and projected outlyingness, *Annals of Statistics*, 20, 1803-1827.

Eddy, W.F. (1981) Comment on: Graphics for Multivariate Two-Sample Problem (Friedman J.H., Rafsky, L.C.), *Journal of the American Statistical Association*, 76, 287-289.

Eddy, W.F., Hartigan, J.A. (1977) Uniform convergence of the empirical distribution function over convex sets, *Annals of Statistics*, 5, 370-374.

Green, P.J. (1981) Peeling bivariate data, in: *Interpreting Multivariate Data*, V. Barnett Ed., John Wiley, New York, 3-19.

Liu, R.Y., Parelius, J.M., Singh, K. (1999) Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference, *The Annals of Statistics*, 27, 783-858.

Nath, G. B. (1971) Estimation in truncate bivariate normal distributions, *Applied Statistics*, 20, 313-319.

Petitjean, P., Saporta G. (1992) On the Performance of Peeling Algorithms, *Applied Stochastic Models and Data Analysis*, 9, 91-98.

Porzio, G.C., Ragozini G. (2000) Exploring the Periphery of Data Scatters: Are There Outliers?, in: *Data Analysis, Classification, and Related Methods*, H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, M. Schader Ed., Springer-Verlag, Heidelberg, 235-240.

Rousseeuw, P.J., Leroy, A. (1987) *Robust Regression and Outlier Detection*, John Wiley, New York.

Rousseeuw, P.J., van Zomeren, B.C. (1990) Unmasking multivariate outliers and leverage points (with discussion), *Journal of the American Statistical Association*, 85, 633-651.

Silverman, B. W. and Titterington, D. M. (1980), Minimum covering ellipses, *SIAM Journal on Scientific and Statistical Computing*, 1, 401-409.

Tukey, J.W. (1977) *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

Zani, S., Riani, M., Corbellini, A. (1998) Robust bivariate boxplots and multiple outlier detection, *Computational Statistics and Data Analysis*, 28, 257-270.