

Correspondence analysis and two-way clustering*

Antonio Ciampi¹, Ana González Marcos² and Manuel Castejón Limas²

¹McGill University, Montreal, Canada, ²University of León

Abstract

Correspondence analysis followed by clustering of both rows and columns of a data matrix is proposed as an approach to two-way clustering. The novelty of this contribution consists of: *i*) proposing a simple method for the selecting of the number of axes; *ii*) visualizing the data matrix as is done in micro-array analysis; *iii*) enhancing this representation by emphasizing those variables and those individuals which are 'well represented' in the subspace of the chosen axes. The approach is applied to a 'traditional' clustering problem: the classification of a group of psychiatric patients.

MSC: 62H25

Keywords: block clustering, selecting number of axes, data visualization.

1 Introduction

Cluster analysis is often introduced as the family of techniques aiming to describe and represent the structure of the pairwise dissimilarities amongst objects. Usually objects are observational units or variables. The dissimilarity between a pair of units is defined as a function of the values taken by *all* the variables on the two units. In a dual way, the dissimilarity between two variables is defined as a function of the values taken by the two variables on the set of *all* units. However, some modern clustering problems, such as those arising in micro-array analysis and text mining, pose a new challenge: not only to describe dissimilarities relationships among individuals and variables, but

* Address for correspondence: Antonio Ciampi, Department of Epidemiology and Statistics.
antonio.ciampi@mcgill.ca, dieago@unileon.es and diemcl@unileon.es.

Received: November 2003

Accepted: January 2005

also to discover groups of variables and of individuals such that the variables are useful in describing the dissimilarities amongst the individuals and vice versa. To this end, techniques known as two-way clustering and crossed classification clustering have been developed, with the aim of producing homogeneous blocks in the data matrix (Tibshirani *et al.* 1999).

Correspondence analysis (CA), as other biplot techniques, offers the remarkable feature of jointly representing individuals and variables. As a result of such analyses, not only does one gain insight in the relationship amongst individuals and amongst variables, but one can also find an indication of which variables are important in the description of which individuals (Gordon, 1999). It is therefore natural to develop clustering algorithms based on the coordinates of a CA, and indeed practitioners of “analyse des données” commonly advocated this well before the advent of micro-array and text mining.

More recently, in an early attempt to develop clustering methods for micro-array data Tibshirani *et al.* 1999 used the coordinates associated to the first vectors of the singular value decomposition of the data matrix to simultaneously rearrange its rows and columns. They eventually abandoned this approach to concentrate on block clustering. In this work we explore their early idea further. Instead of using only the first axis, we select a few important axes of a CA and apply clustering algorithms to the corresponding coordinates of both rows and columns. Then, instead of ordering rows and columns by the value of the respective first coordinates, we use one of the orderings of rows and columns induced by the classification thus obtained.

The motivation of this work does not come from micro-array or textual analysis, but from a more traditional type of data analysis problem arising from the area of criminal psychology. The problem and the data are briefly described in Section 2; the general approach is described in Section 3; Section 4 is devoted to the analysis of our data and finally we end with a brief discussion.

2 The data analysis problem: identifying ‘psychopaths’ in a data set

In criminal psychology, one is interested in identifying a type of individual that can be considered as especially dangerous, roughly corresponding to what is commonly termed ‘psychopath’. One of us (AC) was asked to help identify such a group in a data set consisting of the values taken by 20 variables on 404 patients living in an institution for offenders with recognized psychiatric problems. The source of the data cannot be disclosed. The 20 variables are the items of a standard questionnaire used to identify dangerous individuals. It is current practice to add the values of the 20 variables for each individual and classify an individual as dangerous if this sum exceeds a specific threshold. The expert psychologist was not satisfied with the standard approach, mainly because his intuition was that a ‘psychopath’ should be defined by only a few personality

variables, and in particular ‘glibness and superficial charm’ and a ‘grandiose sense of self-worth’.

The list of the variables with their meaning is given in Table 1. All variables are ordered categorical, ranging from 0 to 2, with 0 denoting absence, 1 moderate degree and 2 high degree of a certain characteristic. The variables are classified as *behavioural* or *personality related*; this is indicated in the table by the letters B and P respectively.

Table 1:

LABEL	NAME	B/P	MEAN
PCL1	Glibness, superficial charm	P	0.1782
PCL2	Grandiose sense of self worth	P	0.3441
PCL3	Need for stimulation, proness to boredom	P	0.6361
PCL4	Pathological lying	P	0.3936
PCL5	Conning, manipulative	P	0.6386
PCL6	Lack of guilt or remorse	P	1.0594
PCL7	Shallow affect	P	0.5544
PCL8	Callous, lack of empathy	P	0.6015
PCL9	Parasitic life-style	B	0.8490
PCL10	Poor behavioural controls	B	0.8837
PCL11	Promiscuous sexual behaviour	B	0.6188
PCL12	Early behaviour problems	B	0.7005
PCL13	Lack of realistic long term plans	P	0.8218
PCL14	Impulsivity	P	0.7995
PCL15	Irresponsibility	B	0.9084
PCL16	Failure to accept respons for actions	P	0.9604
PCL17	Many short-term marital relationships	B	0.3663
PCL18	Juvenile delinquency	B	1.1188
PCL19	Revocation of conditional release	B	1.0544
PCL20	Criminal versatility	B	1.2747

For instance, PCL1 is classified as personality related (P); it is a number between 0 and 2 according to the level of ‘glibness, superficial charm’ shown by the subject in a video taped interview.

From the analyst’s point of view, some specific aspects of the problem can be identified: *a)* data are recorded on a 3-point ordered scale; *b)* some kind of variable selection should be useful in separating signal from noise; *c)* perhaps some individuals may provide noise and removing them might result in a crisper classification.

The method developed aims at dealing with these aspects. We are looking for clusters, in particular for one ‘stable’ cluster or ‘taxon’ that can be easily identified. There is an additional difficulty: this cluster is probably a rather small one, as, fortunately, psychopaths are rare, even in a population detained for crimes.

3 Clustering individuals by the coordinates of a correspondence analysis

We will consider data in the form of a two-way contingency table and the problem of clustering the rows and the columns of this table. Extensions to data matrices where rows are observational units and columns continuous or categorical ordered variables are easy and will be discussed later. The chi-squared distance between rows and columns is perhaps the most natural dissimilarity that can be defined between pairs of rows and columns. It is well known that if we apply (simple) correspondence analysis (CA) to our table, we obtain a new set of coordinates for both rows and columns, in which the chi-squared distance becomes the classical Euclidean distance, or Pythagorean distance. Therefore, the chi-square distance between two objects calculated directly from the contingency table is equivalent to the Pythagorean distance between the two objects as represented in the factor space of the CA. But the factor space has the interesting property of decomposing the total inertia of the data so that the first axis is associated to the greatest proportion of the total inertia, the second axis to the second largest proportion and so on. In other words, CA defines a sequence of subspaces containing an increasing proportion of the total inertia. It is at the root of the practice of CA that taking only a few axes often clarifies the relationship amongst objects, separating, in some sense, interesting ‘signal’ from the uninteresting ‘noise’.

All this is well known and applied in current data analysis practice (Greenacre, 1984, 1993). Here we propose a few tools for taking maximum advantage of this approach.

3.1 *Selecting the number of axes of the CA*

One fundamental problem in CA is the identification of the ‘important’ dimensions. In practice, the selection is done informally, by studying the axes corresponding to the first few eigenvalues of the CA and retaining those which are interpretable. However, when CA is the first step of an algorithm as in our case, it is important to have a slightly formalized selection procedure. We propose one based on the following reasoning. Suppose we had a good idea of the distribution of the eigenvalues of a family of contingency tables similar to ours but generated under the independence assumption. Then if the eigenvalues of our data appear to follow this distribution, we may conclude that none of the axes contains interesting information. On the other hand, if the first k axes of our data matrix contain information and the others do not, then we would expect that the first k eigenvalues markedly differ from the behaviour of the first k eigenvalues of matrices generated under the independence assumption. More formally, this leads to the following rule:

1. Perform a CA of the data matrix and draw the scree plot of the ordered eigenvalues of the CA.

2. Generate an artificial data matrix by randomly permuting the responses (rows) of each variable (column).
3. Perform the CA of the generated data and superimpose the scree plot of its eigenvalues to the graph obtained in 1.
4. Repeat 2 to 3 a number of times.
5. Look at the graph containing all the scree plots. If the graph of the real data is indistinguishable from the band formed by the simulated scree plots, or if the former is consistently below the latter, conclude that there is no structure in the data, *i.e.* the two variables defining the contingency table are independent. Otherwise, identify the point of intersection of the real data's scree plot with the band of simulated scree plots and conclude that the number of interesting axes is the largest one at the left of the abscissa of the intersection.

We remark that this rule can also be seen as a formalization of the elbow rule, which is very popular in factorial analyses.

3.2 Visualization of the data matrix

Consider a $R \times C$ contingency table with elements n_{ij} , $i = 1 \dots R$, $j = 1 \dots C$. A visual representation of such data is obtained by associating different colors to segments of the range of the n_{ij} 's and drawing a picture in the form of $R \times C$ grid with colors replacing the n_{ij} 's. This is exactly what is done in micro-array analysis, but it would be applicable to any range of data, even negative data. Normally this picture is rarely useful, unless the rows and the columns have been previously permuted in an appropriate way, aiming to extract information. What we propose here is to cluster the points representing rows and columns in the (reduced) factor space with Euclidean distance by a hierarchical clustering algorithm (e.g. Ward). Then the rows and columns of the picture can be rearranged using (one of) the ordering(s) induced by the clustering. The hierarchical trees can also be drawn in the picture, imitating again what is current practice in micro-array analysis.

Obviously this representation is applicable to any data matrix with non-negative entries, and is especially useful for a cases \times variables rectangular matrix with variables which are measured in the same units or which have been preliminarily scaled and centred.

3.3 Taking out poorly represented variables and/or cases

As is well known, CA provides useful aids to interpretation, among which the quality of the representation of each object on each object factorial axis: this is defined as the square of the cosines of the angle that the object forms with the object axis. By summing

these quantities for the first k axes, one obtains the quality of the representation of the object in the object factorial subspace spanned by the first k axes.

In our approach, we propose to cluster both variables and individuals using only their coordinates on a few chosen axes. This was motivated by the aim of decreasing noise. Now, it may be useful to further reduce noise by representing graphically only the individuals and the variables that are well represented on the subspace spanned by the chosen axes. In this paper we distinguish well-represented objects from poorly represented objects by defining as well-represented an object which is better represented in the factorial subspace than outside of it (quality of the representation on the factorial plane $> 50\%$). As we will demonstrate in the next section, it is useful to remove the individuals and/or the poorly represented variables, and to repeat the analysis with the sub-matrix of the original data matrix consisting of only the well-represented variables and/or individuals.

3.4 Applying the approach to other types of data

CA, initially developed for contingency tables, is formally applicable to any data matrix with non-negative entries. In particular it is applicable to a data matrix of the form cases \times variables as long as the variables can only assume non-negative values. Because of the *distributional invariance* property of the chi-square distance, on which CA is based, the application is particularly well justified if the object of the analysis is to study profiles. An example where CA is useful and relevant is the case of nutrition data, when one wishes to find dietary patterns (clusters) based on the proportion of each nutrient that an individual absorbs rather than on absolute values. Thus two individuals are considered similar if they have a very close profile, i.e. if they absorb similar proportions of each nutrient regardless of the total amount, which can be quite different.

Another well-justified application of CA to situations other than the two-way contingency table is to cases \times variables data matrices when the variables are ordinal, with levels represented by non-negative integers. Here, however, one needs to use the artifice of doubling the number of variables, as explained for example in (Greenacre, 1993). Thus for each ordinal variable X_i taking values between, say, 0 and p_i , one creates its non-negative 'double', $p_i - X_i$, and performs CA on the data matrix consisting of all variables and their 'doubles'.

4 Application to our data

The first step of our analysis was to double the variables, i.e. create for each PCLi its double 2-PCLi. The resulting 404×40 matrix was then treated by CA. The graph of the eigenvalues is given in Figure 1. The graph also contains a few graphs based on simulated matrices with random permutations of the responses of each variable.

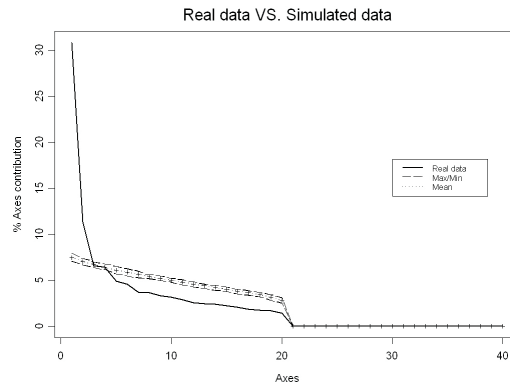


Figure 1: Real data VS. Simulated data.

The graph of the real matrix and the band of graphs of the simulated matrices, intersect at a point corresponding to three factorial axes. Before this point the graph of the real matrix is above the band, and lies below it after the intersection. We chose therefore the three-dimensional subspace spanned by the first three factors to represent both individuals and variables. We applied Ward’s clustering algorithm with the Euclidean distance to the coordinates of both subject-points and variable-points in this subspace. To verify that the choice of three axes is a sensible one, we plotted our original data matrix (doubled variables are omitted) with both rows and columns ordered according to the clustering for varying number of axes. This is shown in Figure 2.



Figure 2: Clustering with the correspondence analysis coordinates: Euclidean distance.

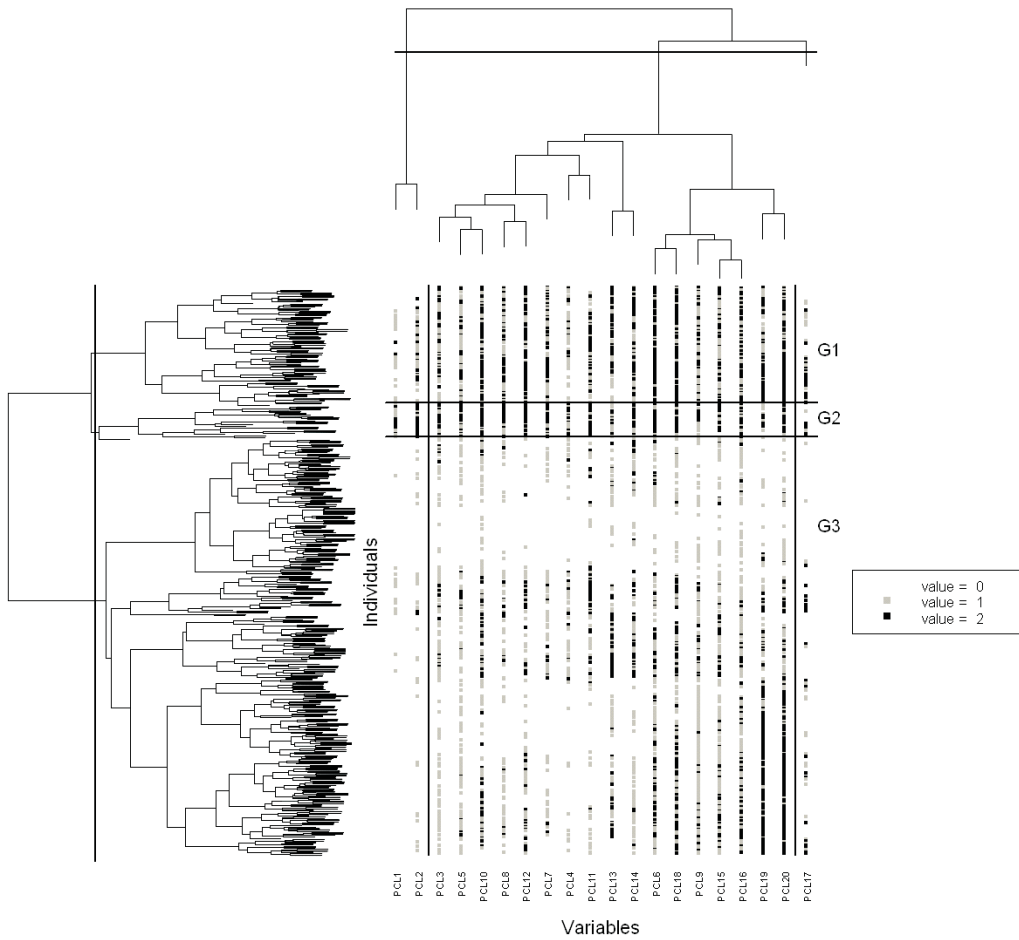


Figure 3: Clustering with the correspondence analysis coordinates: 3 groups.

It seems that taking a higher number of axes does not improve the general pattern obtained with three axes. On the other hand, three axes give a much neater picture than one axis only, which would have been the choice of [Tibshirani *et al.* 1999]. Figure 3 is a larger picture of the data matrix with rows and columns ordered according to clustering obtained from the three axes choice; in it we also show the hierarchical classification trees of both columns and rows, with tentative ‘cuts’ yielding three clusters of variables and three clusters of individuals. Note that two branches come together at the distance between the two clusters being merged.

Interestingly, the clustering of variables places two personality characteristics (PCL1 and PCL2) in one cluster and all the remaining variables in the other with the exception of one behavioural characteristic (PCL17), which remains isolated. A description of this first clustering of subjects is given in Figure 4.

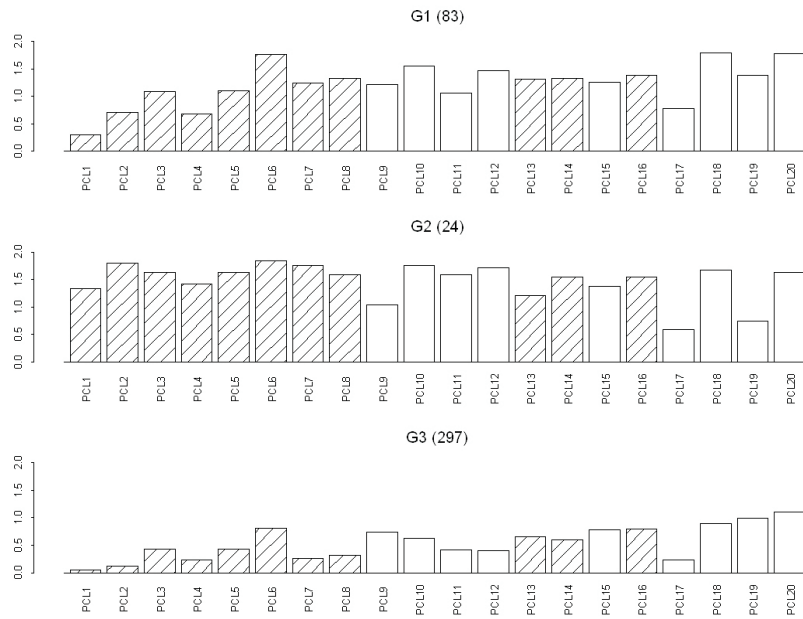


Figure 4: Cluster description.

One might interpret group 2 (24 patients) as including the ‘psychopaths’, since the individuals in this group have high values of PCL1 and PCL2; they also have high value of other personality and behavioural variables. Group 3 includes the majority (297 patients), characterized by generally low levels of all variables. Group 1 (83 patients) is similar to group 2 but some variables, and in particular PCL1 and PCL2, are not as high on the average. Notice that if we had cut the tree so as to have two clusters, we would not have seen the difference between group 2 and group 1. On the other hand, we have also looked at finer cuts up to the eight-cluster partition (data not shown) and found the following features. Group 1 remains a distinct entity even in the eight-cluster solution. Group 2 splits only once, at the 4 cluster cut, into two sub-clusters of 23 and 1 patients respectively, with the isolated patient being characterized by having low values for the behavioural variables. On the other hand, group 3 is the one that appears less stable, splitting into up to 5 sub-clusters.

Next, as explained in Section 2.3, we proceeded to an elimination of variables and individuals and applied our algorithm to what is left of the data matrix. We decided to keep only the rows and columns with quality of the representation on the 3-dimensional factorial subspace greater than 50%. We are left with a 171×8 matrix. The 8 remaining variables are 5 of the 11 personality variables (PCL1, PCL2, PCL7, PCL8, and PCL13) and 3 of the 9 behavioural variables (PCL18, PCL19, PCL20). This is an indication that personality variables in general, and PCL1 and PCL2 in particular, are more relevant to the goal of identifying clusters in the whole data set.

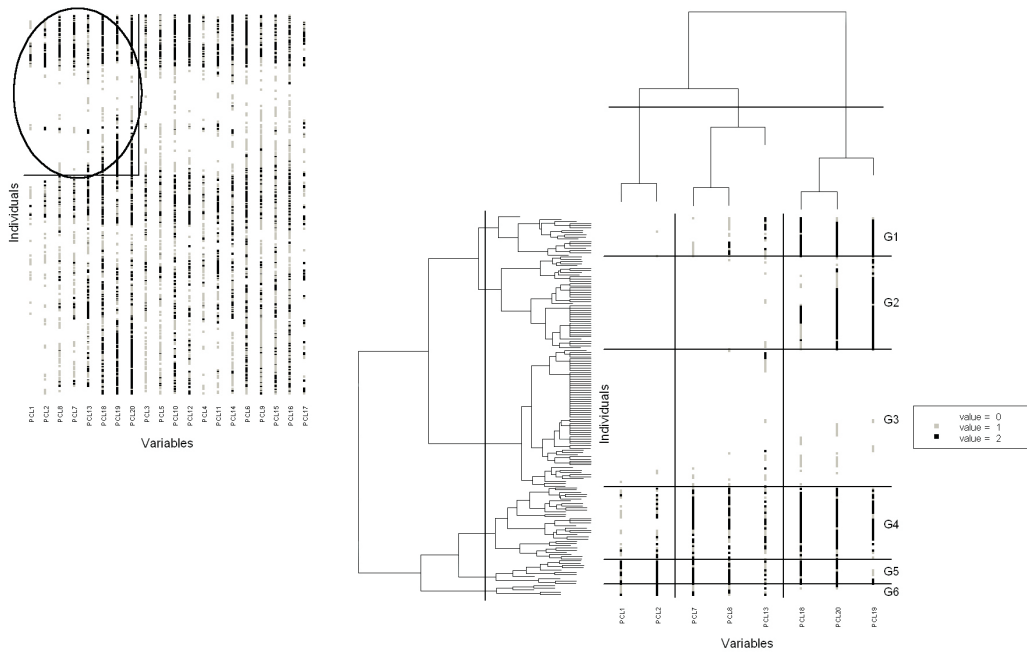


Figure 5: Clustering with the correspondence analysis coordinates of the representative variables and individuals.

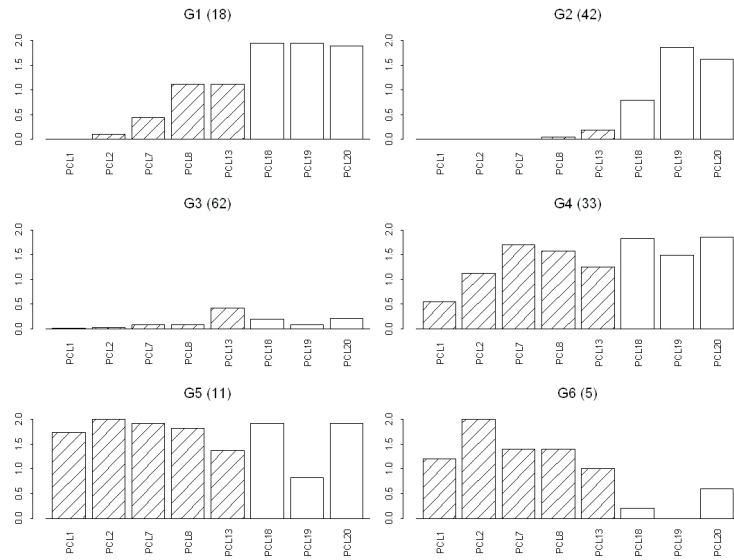


Figure 6: Cluster description.

Applying our algorithm to the resulting 171×8 data matrix, we obtained the following results. Our modified scree plot approach (not shown here) suggested choosing two factorial axes to represent both individuals and variables. The left side of Figure 5 shows the original data matrix, but now the rows and columns that are poorly represented in the three-dimensional factorial subspace, are shown at the margin. The right portion of the figure shows the row- and column-clustering for the data matrix with the poorly represented rows and columns taken out. A tentative cut of the two trees suggests six clusters of subjects and three clusters of variables.

These clusters are described in Figure 6.

Now it is group 5 which can be seen as consisting of ‘psychopaths’. This group has high values of PCL1 and PCL2, but also of many other personality and behavioural variables. Interestingly, the smaller cluster of five individuals consists of individuals with high levels of PCL1 and PCL2 and low levels of the behavioural variables.

Next, given the emphasis of the expert on finding clusters of *subjects*, we re-introduced the poorly representative individuals as supplementary and repeated the clustering for the rows. Figure 7 shows the results of the clustering with the well-represented columns and the entire set of rows. Again, we cut the dendrogram of the rows to obtain six clusters of subjects.

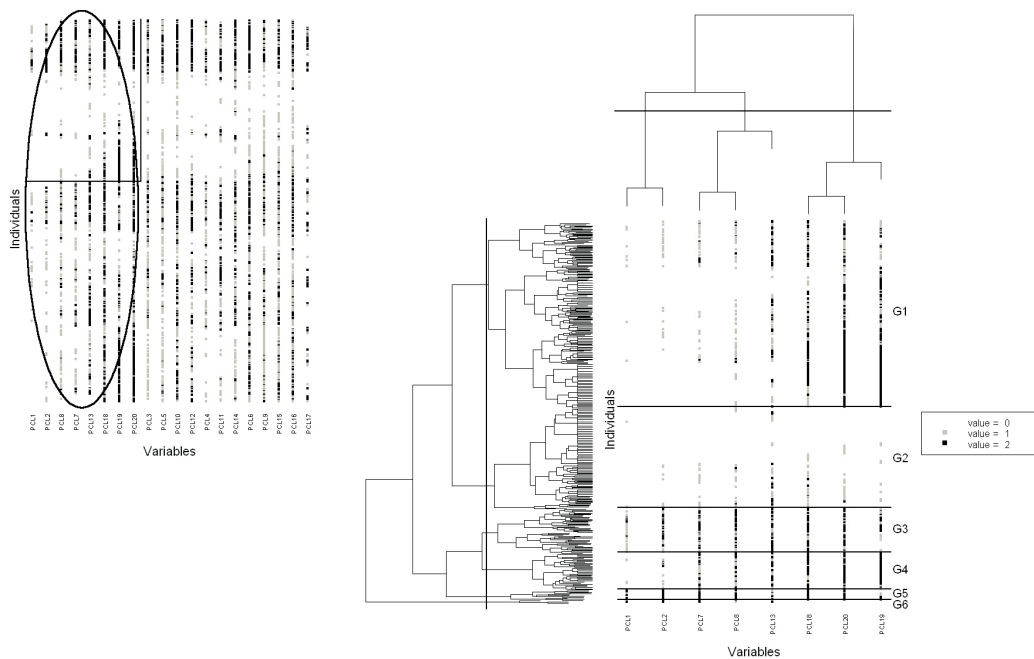


Figure 7: Clustering with the correspondence analysis coordinates of the representative variables and individuals: poorly represented individuals added as supplementary.

Comparing these clusters, which are described in Figure 8, with those shown in Figure 6, we observe nearly the same profiles.

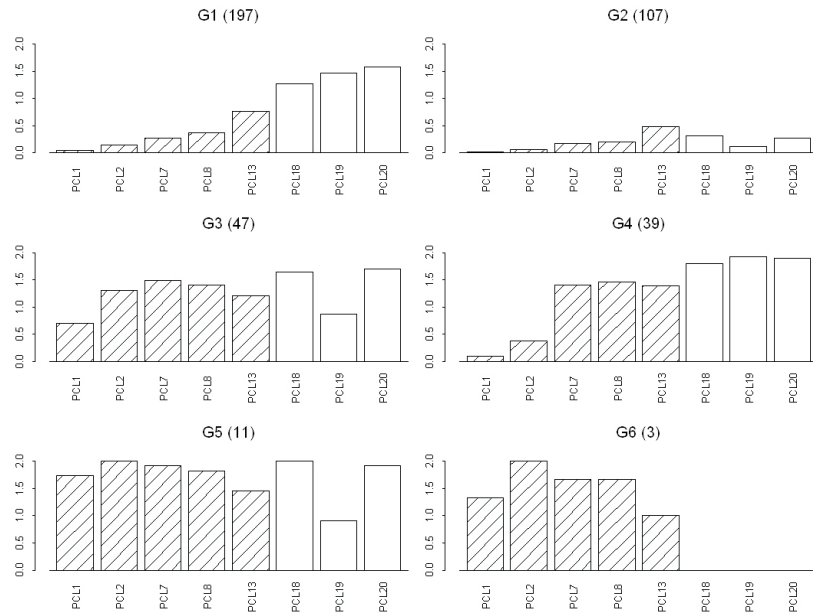


Figure 8: Cluster description.

The differences are indeed minor. We have verified that the group of ‘psychopaths’ consists of the same 11 patients in both cases (group 5). The small group of ‘psychopaths’ with a normal behaviour (group 6 in both clusterings), consists of 3 of the 5 individuals of the earlier clustering (the other 2 patients join group 3 of the new clustering). We have also considered coarser solutions and found that the group of the ‘psychopath’ with normal behaviour is clearly distinguishable from the rest even for the two-cluster cut, while the cluster of the ‘psychopaths’ appears at the four-cluster cut.

For these data it appears that removing variables and/or individuals that are poorly represented in the reduced subspace of the ‘important’ first factors, results in a sharper classification. Moreover, the results of the analysis correspond to the expert’s original intuition: psychopaths are an identifiable ‘taxon’, but they are better identified by the personality variables than by the behavioural variables, and, in particular, by PCL1 and PCL2. Our method succeeds in correctly identifying the important variables and obtaining a rather clear hierarchical classification of our patient group. An interesting and somewhat surprising result is the identification of the very small subgroups consisting of individuals with high values of the personality variables PCL1 and PCL2 and low values of the behavioural variables (group 6). Indeed this led our expert to comment that these individuals are ‘psychopaths’ that appear quite normal in their daily

social behaviour and specialize in crimes that are cleverly disguised. He added that the reason why there are so few of them in our psychiatric prison sample, might be that they rarely get caught! Thus group 6 can be considered as a variant of the typical ‘psychopath’.

5 Discussion

In this work we have shown by example how CA can be used as a powerful tool in clustering, particularly in two-way clustering, where clustering of both rows and columns (observational units and variables) is of interest. The basic idea consists of first obtaining a representation of both units and variables as points in a subspace of the factor space identified by the CA. Next, a standard hierarchical clustering algorithm is applied to the points of this subspace.

This basic idea, as recognized in the introduction, is far from new. Indeed, data analysts commonly use it, in spite of lack of a strong theoretical foundation. However, in view of recent theoretical work, the idea acquires a new strength: roughly speaking, it appears that if there are clusters, then CA is the best representation to discover them (Caussinus, H. & Ruiz-Gazen, A. (2003), Caussinus, H. & Ruiz-Gazen, A. (1995)). Furthermore, recent work in unsupervised machine learning (Bengio *et al.* (2003), Ng *et al.* (2002)) seems to indicate that ‘spectral’ data reduction algorithms applied to the matrix of the pairwise distances between points, provide impressive results in retrieving ‘unusual’ cluster shapes. This is not the same as the basic idea developed in this work, since in our case the spectral decomposition is applied to the data matrix and not the distance matrix (see, however, Greenacre (2000)). Nevertheless, the connection is intriguing. In any case, further theoretical work along the two lines of research mentioned above seems to be highly promising for providing a deeper theoretical justification to the common practice of applying clustering algorithms to reduced data.

The novel contributions of this work are: *i*) proposing a simple method for selecting the number of axes previous to clustering; *ii*) proposing a visualization of the data matrix which generalizes the one current in micro-array analysis; *iii*) enhancing this visualization by emphasizing those variables and those observational units which are ‘well represented’ in the subspace of the chosen axes. Each of these contributions is grounded more on intuition than on theoretical results. Also, in this paper we have simply presented the ideas and their motivation, illustrating them by the analysis of a non-trivial problem. Each of these contributions should be considered as themes for further research.

The problem of selecting the dimension of the subspace on which to represent the data is all pervasive in data reduction and model building. Many approaches have been proposed and ours is just one within the family of computational intensive proposals. The same approach can be applied to the problem of selecting the number of clusters

or, equivalently for hierarchical clustering algorithms, the level at which to cut the dendrogram. We have, in our case, preferred to not propose a single cut, in keeping with the exploratory aim of our analysis.

The visualization of the data matrix with the aid of CA and clustering may be improved at many levels. Our priority, however, is to extend the approach to the representation of multiple categorical variables, starting from some version of multiple correspondence analysis. As for the enhancement of the visualization by emphasizing the well represented objects, we have already outlined some possibilities that, we feel, deserve to be explored. For instance, one could start by using only the first axis (as in Tibshirani et al., 1999), obtain a visualization of the data, and then pull out the objects that are not well represented. The next step would be to repeat the same approach starting from the second factorial axis, and so on: one obtains as many classification schemes as there are 'important' axes, and each such scheme applies to a subset of individuals and variables, with possibly overlapping subsets.

6 Acknowledgments

This work has been partially funded with a research scholarship granted by the State Secretary of Education and Universities of the Spanish Ministry of Education, Culture and Sport. The authors want to recognize the hospitality of the Department of Epidemiology and Statistics members during the visits of A. González and M. Castejón to McGill University between 2002 and 2004.

7 References

- Bengio, Y., Vincent, P., Paiement, J-F., Delalleau, O., Ouimet, M., and Le Roux, N. (2003). 'Spectral clustering and Kernel PCA are learning eigenfunctions'. Technical Report 1239, Département d'Informatique et Recherche Opérationnelle, Université de Montréal.
<http://www.iro.umontreal.ca/~lisa/publications.html>.
- Caussinus, H. & Ruiz-Gazen, A. (1995). Metrics for finding typical structures by means of principal component analysis. In *Data Science and its Applications*, Y. Escoufier & C. Hayashi (eds), Tokyo: Academic Press, 177-192.
- Caussinus, H. & Ruiz-Gazen, A. (2003). Which structures do generalized principal component analysis display? The case of multiple correspondence analysis. To appear in *Multiple Correspondence Analysis and Related Methods* (eds. Michael Greenacre and Jörg Blasius), London: Chapman & Hall, 2006.
- Gordon, A.D. (1999). *Classification*, 2nd Edition, London: Chapman & Hall.
- Greenacre, M.J. (1984). *Theory and Application of Correspondence Analysis*, London: Academic Press.
- Greenacre, M. (1993). *Correspondence Analysis in Practice*. London: Academic Press.
- Greenacre, M. (2000). Correspondence analysis of a square symmetric matrix. *Applied Statistics*, 49, 297-310.

- Ng, A. Y., Jordan, M. I., and Y. Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, T. Dietterich, S. Becker and Z. Ghahramani (eds.), volume 14. Cambridge MA: MIT Press.
- Tibshirani, R., Hastie T., Eisen, M., Ross, D., Botstein, D., and Brown, P. (1999). Clustering methods for the analysis of DNA microarray data. *Technical report*, Department of Statistics, Stanford University. <http://www-stat.stanford.edu/~tibs/lab/publications.html>.

