barcelona gse
graduate school of economics

A Simple Permutation Test for Clusteredness

Michael Greenacre

April 2011

Barcelona GSE Working Paper Series

Working Paper nº 555

A simple permutation test for clusteredness

Michael Greenacre

Department of Economics and Business
Universitat Pompeu Fabra
Ramon Trias Fargas, 25-27
08005 Barcelona
SPAIN
E-mail: michael.greenacre@upf.edu

Abstract: Hierarchical clustering is a popular method for finding structure in multivariate data, resulting in a binary tree constructed on the particular objects of the study, usually sampling units. The user faces the decision where to cut the binary tree in order to determine the number of clusters to interpret and there are various *ad hoc* rules for arriving at a decision. A simple permutation test is presented that diagnoses whether non-random levels of clustering are present in the set of objects and, if so, indicates the specific level at which the tree can be cut. The test is validated against random matrices to verify the type I error probability and a power study is performed on data sets with known clusteredness to study the type II error.

Keywords: Hierarchical clustering, distance, permutation test.

1. Introduction

Ascending hierarchical cluster analysis is one of the most commonly used multivariate methods in practice. A set of n objects – usually cases or other sample units – is described by a set of p variables, a dissimilarity measure is defined between the objects, and a method of clustering is chosen, for example, complete linkage or Ward clustering. The analysis proceeds in a stepwise fashion, from single objects to the complete set, to construct a binary tree, or dendrogram, which summarizes visually inter-object dissimilarities. A different approach is to apply a multidimensional scaling (MDS) to the inter-object dissimilarities, which aims to represent the objects as a map in a continuous space – usually a plane – so that inter-object distances in the map approximate the dissimilarities as closely as possible. Clustering and MDS reveal discrete and continuous structures respectively and complement each other: the discrete clusters of the objects are sought in the high-dimensional space of the variables whereas MDS looks for a few directions of spread in the same space, called principal axes, which represent the positions of the objects optimally in a low-dimensional display.

Faced with the dendrogram resulting from a hierarchical cluster analysis, the researcher has to make a decision where to cut the dendrogram to define the clusters. This is almost always done by the rule-of-thumb of looking for a large jump in the node heights, where there has been a large increase in the dissimilarity measure at that point to move to the next merging of the objects. The rationale is as follows: supposing that complete linkage clustering is performed, then cutting the dendrogram at a level d means that all inter-object dissimilarities within the clusters below this cut must be less than or equal to d. So if there is a large jump between nodes in the dissimilarity levels, say from d to $d+\Delta d$ for Δd relatively large, then cutting the dendrogram at d is more natural than at $d+\Delta d$ because of the much higher within-cluster variance that is implied by cutting at the higher level, where is just one cluster less. This decision about

where to cut the dendrogram, however, is not always clear – for example, there might be more than one obvious place where the dendrogram could be cut – so it would be helpful to have some further statistical guidelines to support this choice.

This article explains a simple permutation test to identify the dissimilarity level below which all the clusters can be considered non-random.

2. Context of the application

As an application, we use the data set on benthic (i.e., sea-bed) species abundances given by Greenacre (2007, 2010) – this is data set "benthos" available at www.multivariatestatistics.org. This data set consists of 13 marine samples of the sea-bed, 11 of which are taken at different locations, called sites, close to an oil-drilling platform in the North Sea, while the remaining two samples are from reference sites far enough away from the platform that they can be regarded as the natural environment for that region. Data n_{ij} (i=1,…,13; j=1,…,92) consist of abundance counts of 92 species in each sample – on average about 56 of these species are found in any one sample, so approximately 39% of the entries in the data matrix consists of zeros. One of the accepted ways of measuring biological distance between the samples for data such as these is the *chi-square distance* – see, for example, Greenacre (2007) – this is a weighted Euclidean distance computed on the relative abundances of the species in each sample:

$$d(i,i') = \sqrt{\sum_{j=1}^{p} \frac{(n_{ij}/n_{i+} - n_{i'j}/n_{i'+})^2}{n_{+j}/n_{++}}} \qquad (1)$$

where + indicates summation over the corresponding index. Having calculated the 13×13 distance matrix between the sites, a classical multidimensional scaling (MDS) of the distance matrix has a total of 12 dimensions, with distance variance decomposed along the dimensions shown in Figure 1. Two standard rules-of-thumb suggest that there are three dimensions worth

studying: the "elbow" rule* shows an elbow after the third dimension, while the "percentages above average" rule identifies three dimensions that exceed the average per dimension of 100/12 = 8.75%. Two views of the three-dimensional solution are given in Figure 2. The view of the first two principal axes shows three groups of sites, the two reference stations on the right, and three groups of sites on the left, with the sites closest to the pollution source at top left and those further away at bottom left. The third dimension separates out site 24, due to the unusually high abundance of one of the species at this site (75% of the species abundances in site 24 are due to one species, *Myriochele oculata*, whereas in the other sites this species accounts for much lower percentages of between 2% and 32%).

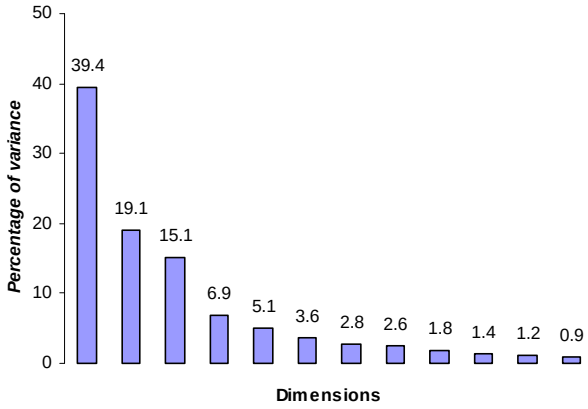

Figure 1: Scree plot of the percentages of variance along the 12 dimensions of the multidimensional scaling (MDS).

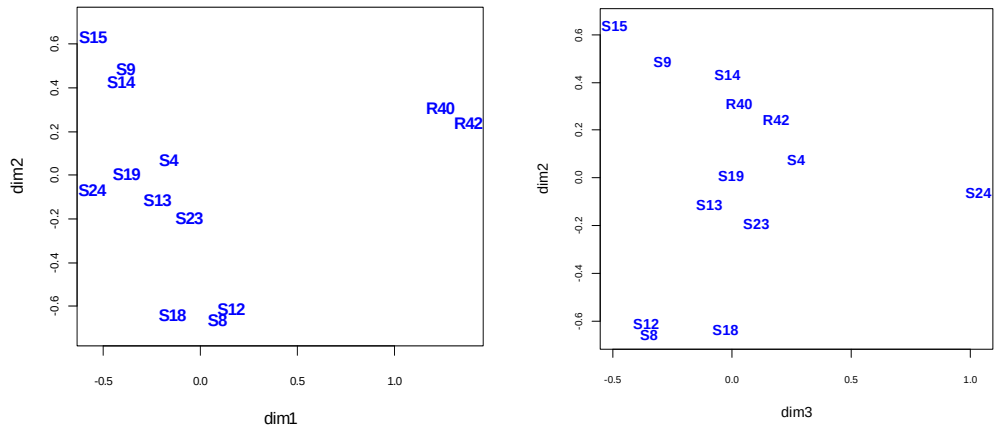

Figure 2: Two projections of the three-dimensional MDS solution, showing dimensions 1 and 2 on the left, and dimensions 3 and 2 on the right.

* The elbow rule in MDS is analogous to the "look for a big jump" rule in hierarchical cluster analysis for cutting the dendrogram, described in the introduction.

The above MDS solution of the sites clearly suggests that they form groups. A hierarchical cluster analysis performed on the same distance matrix gives the dendrogram in Figure 3. Since the chi-square distance is a weighted Euclidean distance, Ward clustering is an appropriate clustering option.

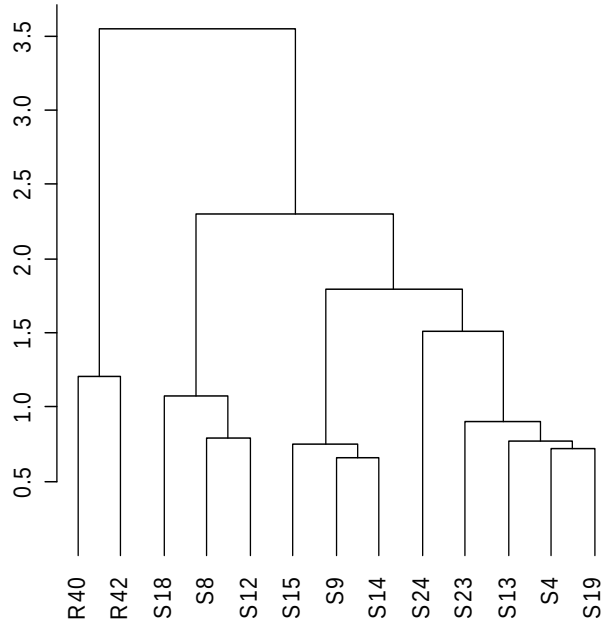

Figure 3: Dendrogram of the chi-square distances between sites, using Ward clustering.

The dendrogram shows the same four major groups seen in the first projection of Figure 2, as well as site 24 possibly separating from the group of sites 23, 13, 4 and 19, that was seen along the third axis in the second projection.

Now the question arises as to the statistical significance of these results, whether they are non-random: a cluster analysis would reveal groupings in any data set, even in many random ones. Figure 4, for example, shows a cluster analysis based on the same data set where the data for each of the 92 species have been randomly permuted – now, for example, that unusually high value which was in site 24 will pop up in any of the other sites (in this case it was randomly allocated to site 23). This dendrogram now clearly suggests two clusters consisting of several sites in each and two "outliers", sites 23 and 24 forming two isolated clusters of one site each.

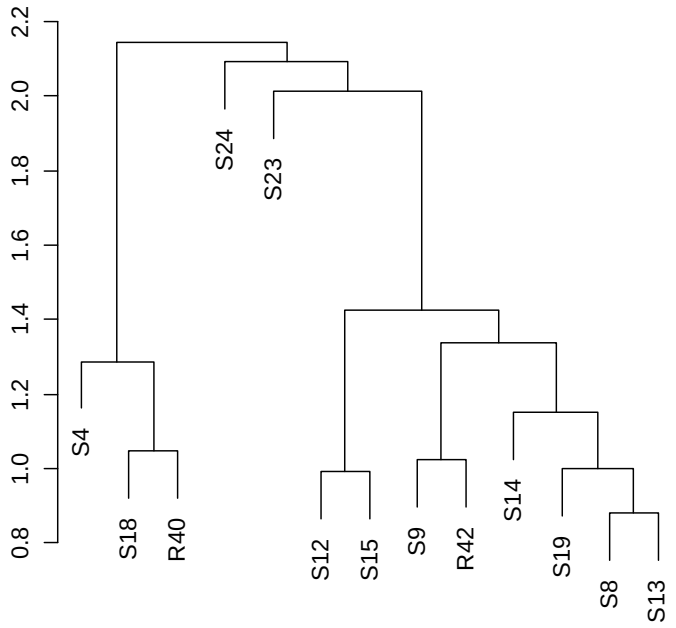

Figure 4: Dendrogram of the chi-square distances between sites, after random permutation of the values for each species.

The remainder of this article describes a permutation test which will show that the groups identified in Figure 3 are not random, whereas those in Figure 4 are (as one would expect) compatible with randomness.

3. Permutation test

We use the terminology “inferior nodes” for those in the lower part of the dendrogram (formed initially), “superior nodes” for those in the upper part (formed later), and “top node” for the one that merges all the objects (i.e., the last node in the ascending hierarchical process). Figure 5 shows three additional dendrograms (a, b and c), based on further random permutations of the species data across the sites as well as the original dendrogram (d) of Figure 3. This last dendrogram, based on the actual unpermuted data, distinguishes itself from the others by having higher levels for the superior nodes, especially the topmost node, while the inferior nodes have levels generally below those of the dendrograms based on permuted data. This observation is simply saying that clusteredness consists in having groups with within-group distances small and between-group distances high. Taking one of the dendrograms a, b or c of Figure 5, choosing a

level somewhere in the middle of the vertical scale and then stretching the nodes above this level upwards, and especially pushing the nodes below this level downwards to make a gap between the node levels is essentially making the groups more similar internally and further apart. Such a level would then be where one would want to cut the dendrogram to arrive at a classification of the sites.

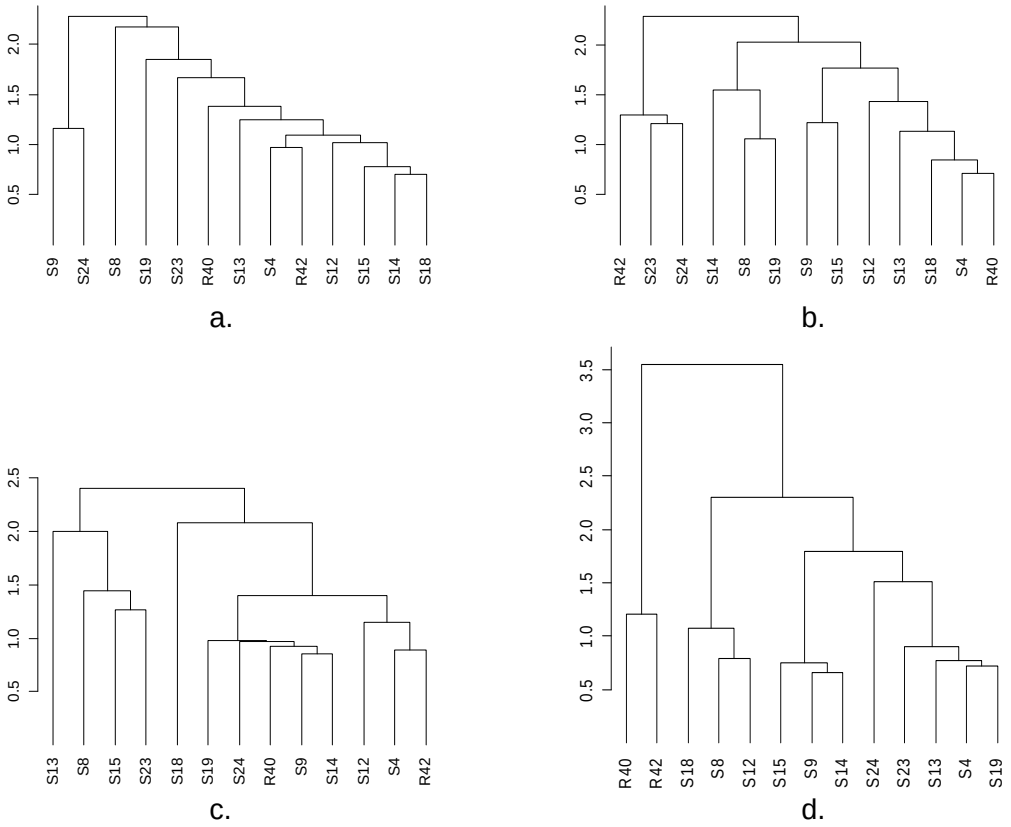

Figure 5: Dendrograms resulting from three data sets based on random permutations of the data (a.-c.) and the dendrogram obtained from the original data (d.) – this last one is identical to Figure 3..

Thus, when looking at the dendrogram in Figure 3 (same as Figure 5d) the question is whether one can identify a level where the nodes below this level are significantly lower than one might expect by chance. At the same time, one might consider the topmost node and ask whether a random version of the data could have led to a dendrogram with a higher topmost node. This is the basis of the permutation test proposed here.

The test is designed as follows:

1. The dendrogram is constructed on n objects using the original data, and with the distance/dissimilarity function and clustering method chosen by the user. (In our application, n was 13, the data were the relative abundances, the distance was chi-square, and the method was Ward.)

2. The n–1 node heights are stored in the first row of a matrix **H**: nodes are typically numbered from 1 (the lowest level) to n–1 (the highest, or top node).

3. Then for a large number N–1of times (we assume the typical value of 999 times, i.e. N=1000) do the following: Form random permutations of the data matrix by randomly shuffling the values for each variable across the cases, perform the same cluster analysis and store the node heights in rows 2 to N of **H**.

4. For each of the n – 1 nodes, count how many of the heights computed on all permutations, including the original one, are smaller than or equal to those computed for the original data; i.e., for each column j of **H**, count how many satisfy the condition $h_{ij} \leq h_{1j}$ (i = 1,…, N), storing these counts in the vector $\mathbf{f} = [\, f_1 \;\; f_2 \; \ldots \; f_{n-1} \,]$. Divide this count by N to obtain estimates of p-values $p_1, \ldots, p_{n-1}$ in the vector $\mathbf{p} = (1/N)\,\mathbf{f}$.

5. Significant clustering is indicated by a value of an inferior node that is well below 0.05 (the aspect of multiple testing and lowering the significance level is discussed in Section 8). This indicates a possible level for cutting the dendrogram. If there are several possibilities, choose the highest node, cut the dendrogram and then repeat the permutation test on each of the resultant clusters to check for possible subclusters.

6. For the top node, which had count f_{n-1}, the count of how many heights are greater than or equal to the one computed for the original data, is $N - f_{n-1}+1$. Dividing this count by N

gives the probability that a random dendrogram gives a top-node height greater than or equal to the observed one. This can be regarded as a test for overall separation of the points, compared to what one would expect by chance.

For example, in the case of Figure 3, the set of counts for N=1000 was 4, 1, 1, 1, 1, 1, 1, 2, 323, 749, 996, 1000, giving p-values of 0.004, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001, 0.002, 0.323, 0.749, 0.996, 1. The natural threshold is thus between nodes 8 and 9, which implies five clusters, including the single one formed by site 24 (Figure 6).

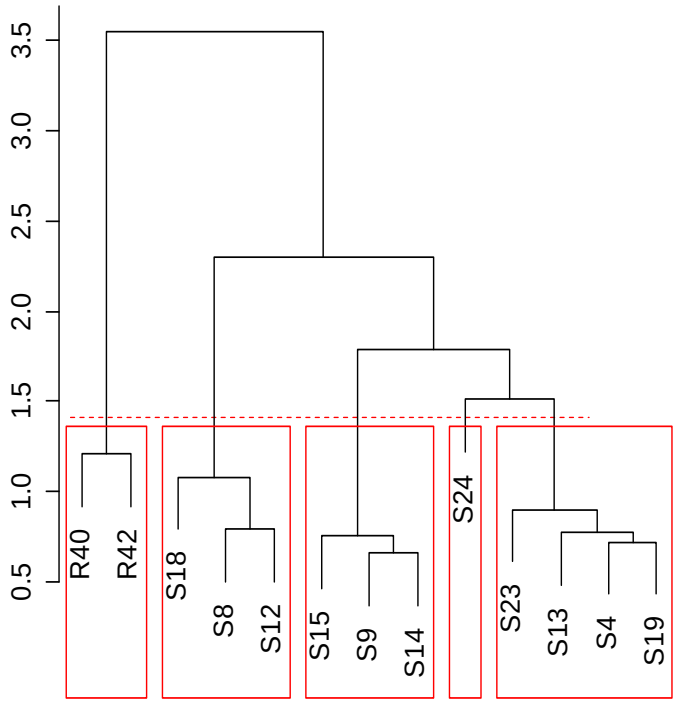

Figure 6: Level for significant clustering determined by the node-height permutation test, implying five clusters: (R40,R42), (S18,S8,S12), (S15,S9,S14), (S24) and (S23,S13,S4,S19).

In this example all the node heights below this threshold are lower than one would expect by chance, with p-values very low. As might be expected by simple inspection of each of the five clusters implied by the above cutting of the dendrogram, none show further subclustering when the same permutation test is applied to each of them (only those with more than two objects can be tested). As far as the top node is concerned, the p-value for its "highness" is 0.001, because all the node heights of the randomly permuted matrices give top nodes lower than this value.

Thus there is not only clusteredness in the samples but also a significantly high dispersion. As a side comment, the cutting of the dendrogram at this point is consistent in this application with the "cut where there is a big jump" rule-of-thumb described in the introduction: Figure 7 shows the 11 differences in the 12 successive nodes of Figure 6 – the average difference is 0.262 and the cutpoint between nodes 8 and 9 is the first that is above this average (reminiscent of the "elbow" rule in MDS mentioned previously).

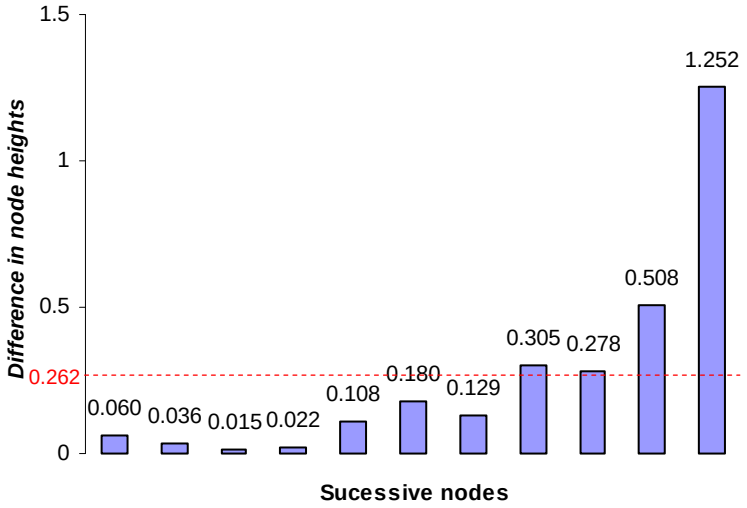

Figure 7: Differences in node heights between successive nodes (2-1, 3-2, …, 12-11) for the dendrogram of Figure 6. Average difference is 0.262 and the node difference 9-8 is the first one that exceeds this value.

We now compare this situation with what happens if we apply the same permutation test to the dendrogram shown in Figure 4, which was based on a randomly generated matrix but which exhibited apparent clustering. The corresponding *p*-values for the nodes are:

0.664, 0.827, 0.458, 0.199, 0.047, 0.124, 0.370, 0.155, 0.072, 0.998, 0.849, 0.142

which show no evidence of significant clusteredness or dispersion, as would be expected. The fact that there is a single *p*-value of 0.047 would not be taken seriously, since there is the issue of multiple testing across the 11 nodes, and so significance levels should be lowered accordingly (see discussion in Section 8).

4. Verifying the type I error and significance level of the test

In order to check that the test leads to the correct probability of a type I error, we generated 1000 data matrices from random permutations as described above. For each of these we computed the node heights based on their hierarchical clustering and then the node heights based on a further 999 permutations, that is we repeated the permutation test on 1000 random variations of the data. Choosing a significance level of $\alpha = 0.05$ we counted how many times, out of the 1000 tests, the *p*-values at each node were less than or equal to 0.05, and expressed this as a proportion out of 1000. The results were as follows for the 12 nodes:

0.069, 0.057, 0.052, 0.045, 0.062, 0.058 0.049, 0.045, 0.038, 0.046, 0.047, 0.050

which is consistent with the significance level of 0.05. Similarly, counting how many were equal to 0.001, which is the lowest value one can get based on 1000 permutations, the results were:

0.001, 0.003, 0.001, 0.001, 0.001, 0.001, 0.002, 0.003, 0.001, 0.002, 0.002, 0.001

also consistent with the significance level (in this case, one must obtain a value of 0.001 or higher).

5. Power study

In order to investigate the power of the proposed permutation test, a series of data sets were generated from a pair of bivariate normal distributions with increasing separation: $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, where:

$$\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 = \delta \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{for } \delta \text{ varying from 0 to 3 in steps of 0.1.}$$

For each value of δ random samples of size 20 were generated from each of the pair of distributions, and this was repeated 100 times. For each combined sample of 40 objects a Ward

clustering was performed and our permutation test (involving N=1000 permutations, as described previously) was applied to the second node from the top, which splits the 40 objects into two clusters. Counting how many times out of a 100 the p-value was below 0.05 gives an estimate of the power of the test to detect two clusters, and this value is plotted on the left in Figure 7 as well as a smoothed version. Again one can see that at the null hypothesis ($\delta = 0$) the power is equal to the significance level 0.05. The rise in power is slight until 1 standard deviation difference in the means and then increases rapidly, reaching a probability of 0.97 for $\delta = 3$. For each of the 100 data sets and the two clusters identified at the second highest node, the misclassification rate was also computed, and this is shown on the right in Figure 7. The decrease in the misclassification rate is steady, reaching an average of about 1 in 40 for $\delta = 3$.

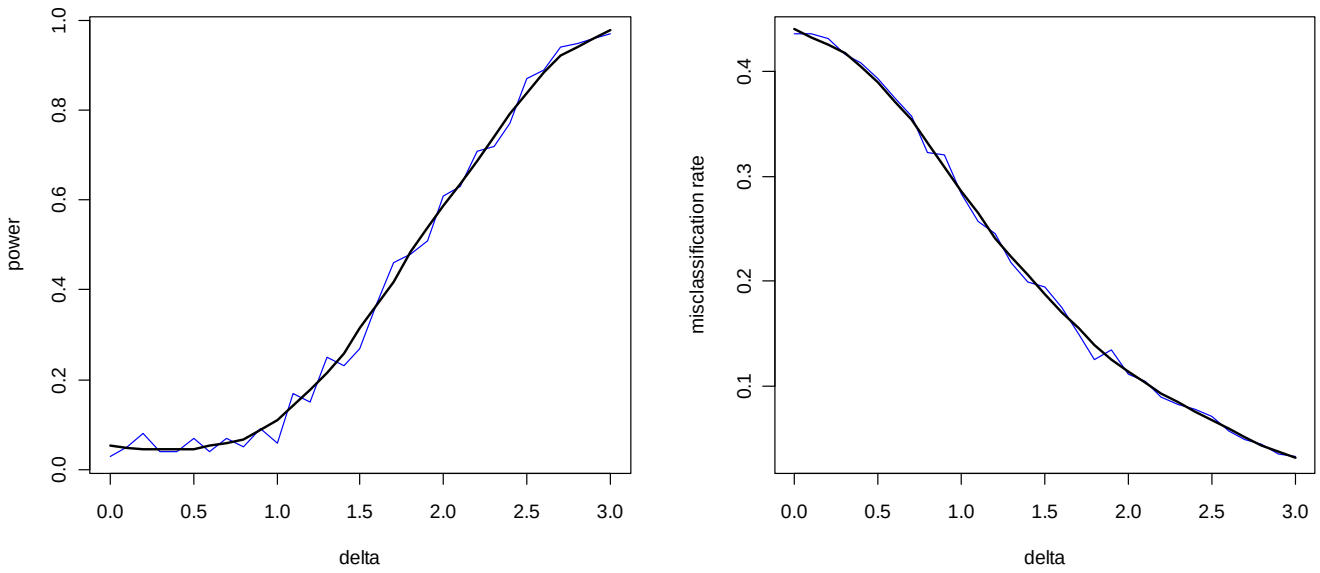

Figure 8: Power function of the permutation test for samples from two bivariate normal distributions with increasing separation between their means, as well as the misclassification rate for the two-cluster solutions.

6. Comparison with previous research

Park et al. (2009) published a procedure for determining significance of clusters based on a permutation test at each node where two clusters, one of which contains at least two objects, are merged. The test is based on comparing the within-cluster structure of the observed data at that

node with those obtained by permuting the cluster membership of the objects merged at each node. In this approach which is essentially a permutation test associated with MANOVA, the set of data associated with each object is conserved and the objects are simply randomly reassigned to one of the two clusters at that node. The result is shown in Figure 9, with estimated *p*-values at each of the nodes tested. Three nodes show significant splitting, thus four clusters are implied by this approach.

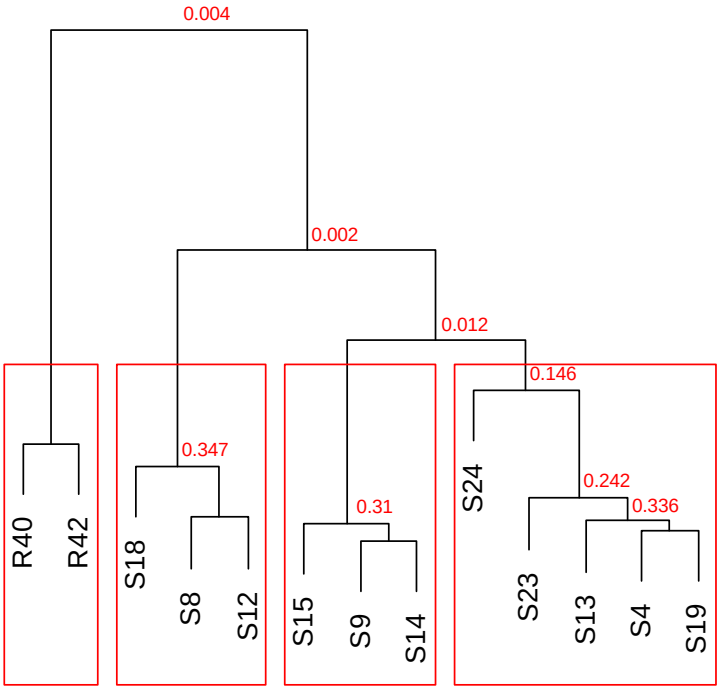

Figure 9: Original cluster analysis of Figure 3, with *p*-values at each node computed by the method of Park et al. (2009). The result implies four clusters.

However, this approach has a problem with clusters merging few objects, which is clearly seen here. For example, when a node merges one object with a cluster of two (there are three such nodes in Figure 8), there are only three possible ways of re-assigning the labels, one of which is the actual situation, so the *p*-value is theoretically 1/3 since it is impossible, by construction, that another permutation gives a lower within-cluster variance measure – so the randomization process is just estimating the probability of 0.333, as 0.347, 0.310 and 0.336 at the respective three nodes (the margin of error for estimating 0.333 with a sample size of 1000 is

approximately ±0.030). Similarly, where S23 merges with the cluster of three, the theoretical probability would be 0.25, estimated here as 0.242. At the node where S24 merges with the cluster of four sites, there are five ways to re-assign labels, so the theoretical probability is 0.2, estimated as 0.012. So this test would never be able to separate S24 at that node, simply because of the number of objects involved. The same phenomenon is apparent in Figure 2(a) of Park et al (2009), where almost all the *p*-values at the lower nodes can be determined theoretically just from the numbers of objects in each cluster. Even when the two merged clusters consist of several objects, the *p*-value can be reliably estimated from the numbers in each cluster. For example, the *p*-value of 0.012 in Figure 8 at the node merging (S15,S9,S14) and (S24,S23,S13,S4,S19) must be an estimate of the probability 1/56 = 0.018, where $56 = \binom{8}{3}$, the number of ways of choosing 3 out of a set of 8.

7. Clusteredness versus dispersedness

A simple example illustrates that the lack of clusteredness does not mean that there is no interesting structure in a multivariate data set. The data set "bioenv" used in Greenacre (2010), also available online from www.multivariatestatistics.org, is an artificial data set of abundances of 5 species (a,b,c,d,e) at 30 sites (S1,S2,…,S30), where the abundances are related to two additional environmental variables, pollution and depth. A hierarchical clustering of the sites, also based on the chi-square distances between sites, is shown in Figure 10 – the impression of at least two clusters is clearly given by this result. However, the permutation test gives the following set of *p*-values, from the lowest to highest nodes:

0.859,0.952,0.903,0.864,0.855,0.863,0.874,0.871,0.911,0.880,0.897,0.858,0.838,0.838,0.862

0.888,0.900,0.887,0.880,0.961,0.933,0.936,0.958,0.923,0.865,0.873,0.882,0.838,0.978

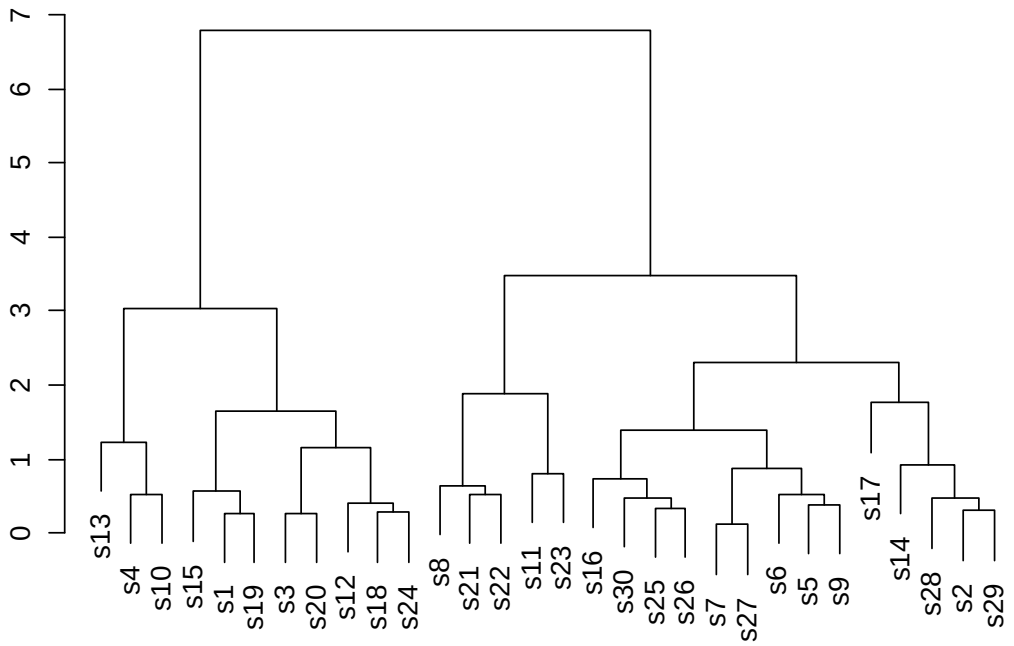

Figure 10: Cluster analysis of 30 sites of data set "bioenv", based on the relative abundances of five species, the chi-square distance and Ward clustering.

There is absolutely no sign of any clusteredness – none of the nodes have levels low enough to be significant. The topmost node has a *p*-value for "highness" equal to (1000-978+1)/1000 = 0.023, indicating significant variation within the whole data set, but with no clusters. The MDS of the chi-square distances is shown in Figure 11, onto which have been regressed the five species and the two environmental variables as a biplot (see, for example, Greenacre (2010: chap. 2) – this is essentially a correspondence analysis of the abundance data, except that the sites have been equally weighted whereas in correspondence analysis they would be weighted proportionally to their marginal totals.

A permutation test can be performed on the parts of variance on each dimension, equal to 52.4%, 22.0%, 16.2% and 9.4%. The data are permuted in the same way as before and the percentages of variance computed for each permutation. The *p*-values are computed for each dimension by counting how many of the percentages are greater than or equal to those in the original analysis – these turn out to be (from first to fourth dimension) 0.015, 0.966, 0.892, 0.851, showing that the

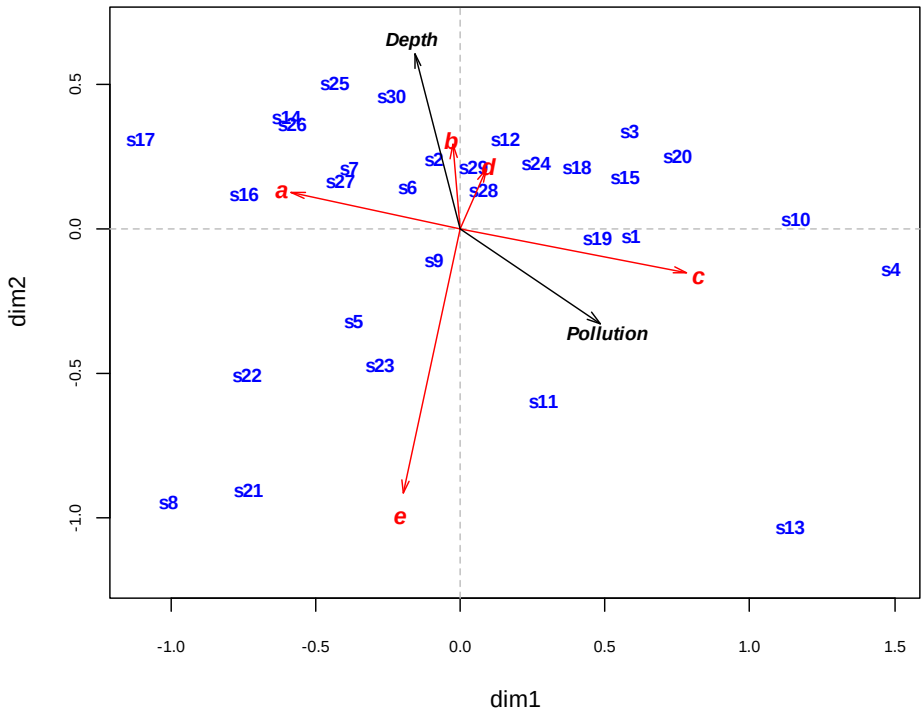

Figure 11: MDS of chi-square distances between sites of data set "bioenv", showing five species and two environmental variables regressed on the two dimensions. The percentage of explained variance on the two dimensions is 52.4%+22.0%=74.4%

first dimension is the only one that is significantly non-random. This agrees with the rule of thumb that chooses dimensions with variance larger than the average (more than 25% of the variance in this four-dimensional case).

The conclusion about the structure in this "bioenv" data set is that there are significant associations amongst the species that emerge as the first dimension of the MDS, correlated with the pollution variable, but that there are no significant clusters of the sites.

8. Discussion and conclusion

Most of the literature on identifying significant clustering has concentrated on making tests of the separateness of the clusters merged at each node (for example, Gordon 1994, Park 1999). Our approach here has been simply to test whether there are nodes with levels lower than would be expected from dendrograms constructed on random permutations of

the data. Cases will cluster because they have similar description vectors across the variables, hence in order to generate null distributions each variable's data are randomly permuted across the cases (i.e., columnwise permutations in the usual data format).

Clusters of n objects generate n–1 nodes, each of which is being tested by our procedure. From the results of Section 4 it is clear that for 21 objects, for example, we would tend to find a significantly low node for random data if the significance level of 0.05 is used. Hence the significance level must be lowered accordingly, the most extreme example of which is to use the Bonferroni correction and work at a significance level of $0.05/(n–1)$. Notice that in the power study of Section 67, where the level at only one node was tested, the significance level could be maintained at 0.05.

All computations were performed using the R package (R Development Core Team 2010). The R script is available from the author.

Acknowledgments

This research has been supported by the Fundación BBVA, Madrid, Spain. Partial support of Spanish Ministry of Education and Science grants MTM2008-00642 and MTM2009-09063 is also acknowledged.

References

Gordon, A.D. (1994). Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis*, **18**, 561-581.

Greenacre, M. J. (2007). *Correspondence Analysis in Practice. Second Edition.* London: Chapman & Hall / CRC Press. Published in Spanish translation by the BBVA Foundation, Madrid, 2008, and freely downloadable from URL http://www.fbbva.es/TLFU/tlfu/ing/publicaciones/fichalibro/index.jsp?codigo=300

Greenacre, M. J. (2010). *Biplots in Practice.* Madrid: BBVA Foundation.

Park, P.J., Manjourides, J., Bonetti, M., and Pagano, M. (1999). A permutation test for determining significance of clusters with applications to spatial and gene expression data. *Computational Statistics and Data Analysis*, **53**, 4290–4300.

R Development Core Team (2010). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org