Available online at www.sciencedirect.com
ScienceDirect
C. R. Palevol 8 (2009) 1–8
COMPTES RENDUS PALEVOL
http://france.elsevier.com/direct/PALEVO/

General Palaeontology

The application of Correspondence Analysis in palaeontology

Matthijs Freudenthal [a,*,b], Elvira Martín-Suárez [a], José Angel Gallardo [c], Antonio García-Alix Daroca [a], Raef Minwer-Barakat [a]

[a] *Departamento de Estratigrafía y Paleontología, Universidad de Granada, Avda. Fuentenueva s/n, 18071 Granada, Spain*
[b] *Nationaal Natuurhistorisch Museum, P.O. Box 9517, 2300 RA Leiden, The Netherlands*
[c] *Departamento de Estadística e I.O., Universidad de Granada, Avda. Fuentenueva s/n, 18071 Granada, Spain*

Received 26 September 2008; accepted after revision 4 November 2008
Available online 4 January 2009
Presented by Philippe Taquet

Abstract

Correspondence analysis (CA) is frequently used in the interpretation of palaeontological data, but little is known about the minimum requirements for a result to be valid. Far from being a fundamental mathematical study of CA, this paper aims to present a tool, which may serve to evaluate results obtained in (palaeontological) praxis. We created matrices of random data, grouped by matrix size and varying percentages of zero cells. Each matrix was submitted to CA. Per matrix group the minimum, mean and maximum percentages of total inertia were calculated for the first four axes. We compared these results with several real cases in vertebrate paleontology. Valid conclusions based on CA can only be drawn on percentages that are considerably higher than the axis percentages obtained from random matrices. ***To cite this article: M. Freudenthal et al., C. R. Palevol 8 (2009).***

Résumé

L'application de l'analyse des correspondances en paléontologie. Les données paléontologiques sont fréquemment interprétées par analyse des correspondances (CA), mais on connaît peu de choses à propos des expériences minimales que nécessite cette analyse pour en tirer des conclusions valables. Le but de ce travail n'est pas une étude mathématique fondamentale de CA, mais la présentation d'un instrument qui puisse servir pour évaluer les résultats obtenus dans la pratique paléontologique. Nous avons créé des matrices de contingence avec des valeurs aléatoires, groupées par dimensions et par pourcentages variables de zéro. Chaque matrice a été soumise à CA. Pour chaque groupe de matrices, nous avons calculé le minimum, la moyenne et le maximum des pourcentages d'inertie totale pour les quatre premiers axes. Ces résultats sont comparés avec plusieurs cas réels en paléontologie de vertébrés. Les conclusions basées sur CA ne sont valables que quand les pourcentages des premiers axes sont considérablement plus élevés que les pourcentages d'axe tirés des données aléatoires. ***Pour citer cet article : M. Freudenthal et al., C. R. Palevol 8 (2009).***

Keywords: Correspondence analysis; Vertebrate paleontology

Mots clés : Analyse des correspondances ; Paléontologie de Vertébrés

1. Introduction

Multivariate Data Analysis techniques are used to summarize the original data and present them in a

* Corresponding author.
E-mail address: mfreuden@ugr.es (M. Freudenthal).

1631-0683/$ – see front matter
doi:10.1016/j.crpv.2008.11.002

graphic form to facilitate their interpretation. They may be subdivided into two groups: classification methods and factorial methods. Correspondence analysis (CA) belongs to the latter group and searches to represent the original data in a space of fewer dimensions through calculations that essentially belong to linear algebra. It produces graphic representations, where the objects to be described are transformed into points on an axis or a plane, offering synthetic representations of wide groups of numeric values. Simple CA (just like other factorial reduction methods) substitutes a matrix by another one with fewer dimensions. One of its most interesting approaches is based on the general theory of Singular Value decomposition (SVD) of a matrix, which is the framework that many multivariant techniques have in common. From a geometric point of view it means calculating the subspace with less dimensions that best fits the data of the original matrix. This geometric adjustment uses an exact algebraic formula, which calculates the reduced matrix that has minimum distance to the original one.

CA has been developed by Benzécri since 1964 and published in French in 1972 [1] and later in English [2]. Since the 1980s, and evidently closely related with the development of informatics, the abstract mathematical approach by the Benzécri school has been transformed into matrix notation, e.g. by Greenacre [6], suitable for its use in computer programs. One might say that the French school emphasized the probabilistic model, whereas in the Anglo-Saxon school the exploratory model prevailed.

Greenacre [7] stated: "An important aspect of CA which distinguishes it from more conventional statistical methods is that it is not a confirmatory technique, trying to prove a hypothesis, but rather an exploratory technique, trying to reveal the data content". This is primarily achieved by graphical representations of the distributions on the first axes, which allow an easy access to the data and permit to formulate hypotheses. Such hypotheses can then be tested formally by conventional statistical methods.

Apart from the graphical representations, fundamental data in CA are the total inertia of the data matrix and the percentages of the total inertia covered by each of the axes. The higher the values obtained on the first axes, the easier it will be to interpret the results; and when these values are low, it is very difficult, if not impossible, to formulate a good hypothesis.

CA is widely used in fields as different as sociology, economy, linguistics, ecology, medicine and psychology and it is being used ever more frequently in the analysis of palaeontological data. When we tried to apply it to our own data tables and compare it with classical palaeontological methods, we realized that little is known about the minimum percentages required for a result to be useful. CA always produces a result, sometimes better, sometimes worse, but there seems to be no instrument to decide whether a result should be accepted or rejected.

Far from being a fundamental mathematical study of CA, this article aims to present a tool that may serve to evaluate results obtained in (palaeontological) praxis. We achieved this by creating a large number of data tables with random data and submitting these to the program PAST [8]. These tables are grouped according to:

- table size, the number of cells in the table;
- percentage of zero cells, the sum of "absence" and "missing data" cells;
- data range, the range between the smallest and the largest value in the table.

For each group of random tables the mean and the range of the values of the first four axes of CA were calculated, and it became clear that the results of CA are strongly influenced by the above mentioned group criteria. The values obtained are given in Tables 1 and 2, which may be used to evaluate the results of real cases. When a real case does not score considerably higher than the corresponding random table, one should conclude that CA does not give a useful result. Of course, this does not mean that the subsequent analysis of the data leads to incorrect results; it only means such an analysis is not supported by CA.

2. Methods

We created 20 files with random data, imitating taxon by locality matrices, through a program written in Visual Basic. The random numbers were created by means of the random generator of Visual Basic, using the decimal fractions only and placing one digit in each cell. In the resulting matrices of 20 rows by 10 columns the numbers 0 to 9 occur in more or less equal frequencies, and no structure whatsoever is expected to exist in such a matrix. These 20 matrices are represented in Table 1 on the line "standard".

Inside each matrix the contributions of the individual cells to the matrix sum vary only moderately, so we created a derivate of each matrix by multiplying certain values by 10, in order to get a distribution of several high values and many low values per row and column, a situation that is probably more similar to a real taxon

Table 1
Percentages for the first axis, the sum of the first two axes, the sum of the first four axes, and the inertia of correspondence analysis over 330 random matrices.
Tableau 1
Pourcentages pour le premier axe, la somme des deux premiers axes, la somme des quatre premiers axes et l'inertie de l'analyse des correspondances sur 330 matrices aléatoires.

		1 axis			2 axes			4 axes			Sum eigenvalues		
	n	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
20 × 10													
Standard	20	21.19	25.29	30.73	40.36	45.23	52.69	68.75	73.70	79.13	0.30	0.37	0.44
Adapt	20	19.03	22.72	27.60	36.08	42.03	48.11	64.23	71.56	79.96	1.02	1.17	1.45
Zero	20	20.94	25.56	28.90	41.09	45.17	48.65	68.62	72.62	79.41	0.73	0.83	0.98
Pattern	20	21.77	27.52	34.71	40.15	49.07	57.39	68.19	76.25	81.94	0.56	0.83	1.04
30 × 15													
Standard	20	15.40	17.61	20.87	29.08	32.38	37.81	52.09	55.29	60.15	0.33	0.37	0.41
Adapt	20	14.00	16.37	18.37	26.49	30.18	33.17	48.50	52.90	56.60	0.38	1.29	1.52
Zero	20	14.16	16.96	20.70	27.01	31.46	35.95	50.27	54.30	59.85	0.81	0.85	0.93
Pattern	20	17.23	22.40	27.08	31.54	39.99	44.63	55.79	64.14	69.80	0.56	0.86	0.99
40 × 30													
Standard	20	9.46	10.83	12.43	18.02	20.16	21.76	33.14	36.12	37.98	0.34	0.37	0.41
Adapt	20	8.93	10.26	11.71	17.55	19.41	21.62	32.76	35.03	37.95	0.41	1.39	1.54
Zero	20	9.46	10.52	11.36	18.24	19.97	21.37	33.57	36.02	38.36	0.84	0.87	0.89
Pattern	20	11.37	17.28	22.11	22.35	31.83	38.94	40.07	51.21	57.88	0.56	0.84	0.98
98 × 30													
Standard	20	7.14	7.77	8.62	13.85	14.87	16.21	26.47	27.49	28.86	0.36	0.38	0.40
Adapt	20	6.83	7.27	7.92	13.17	14.05	14.96	25.02	26.29	27.23	1.42	1.48	1.55
Zero	20	7.04	7.55	8.26	13.94	14.53	15.59	26.17	27.10	28.56	0.87	0.89	0.91
Pattern	20	9.06	16.42	18.87	17.80	31.09	34.39	33.09	49.18	52.31	0.61	0.90	0.96
60 × 50													
Standard	10	6.51	6.82	7.20	12.67	13.06	13.51	23.62	24.38	25.26	0.36	0.38	0.39
14 × 14													
Standard	10	21.08	25.21	29.88	40.42	44.70	52.83	66.98	71.03	77.63	0.27	0.34	0.41

by locality array. These 20 matrices are represented in Table 1 on the line "adapt".

A second set of derivate matrices was created by substituting randomly the contents of about 35% of the cells by 0. These 20 matrices are represented in Table 1 on the line "zero". The line "pattern" indicates matrices with added zeros, created by an algorithm that introduces some kind of a pattern in the data. For details see the chapter "The influence of zeros in a matrix".

Each matrix was submitted to CA using PAST version 1.71 [8]. Per matrix group (standard, adapted, zero and pattern) the minimum, mean and maximum percentages were calculated for the first axis, the sum of the first two axes, and the sum of the first four axes (Table 1).

Then, the entire process was repeated for matrices of 30 rows by 15 columns, matrices of 40 rows by 30 columns and matrices of 98 rows by 30 columns, resulting in a total of 320 matrices.

Table 2
Percentages for the first axis, the sum of the first two axes, and the sum of the first four axes of correspondence analysis over 20 random 35 × 30 matrices with 70% zeros (A), and idem with multiplication of certain values (B).
Tableau 2
Pourcentages pour le premier axe, la somme des deux premiers axes et la somme des quatre premiers axes de l'analyse des correspondances sur 20 matrices aléatoires 35 × 30 avec 70 % de zéros (A), et idem avec multiplication de certaines valeurs (B).

	1 axis			2 axes			4 axes			Sum eigenvalues		
	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
A	13.90	15.90	17.75	25.40	28.75	31.23	43.63	47.89	50.78	2.77	3.10	3.45
B	12.01	14.11	17.55	22.85	25.67	29.84	41.66	44.17	47.49	3.58	3.92	4.56

The results were first calculated over five and 10 matrices in each group; between the results for five and 10 matrices there were some important differences; the results for 10 and 20 matrices are very similar, so we may assume a sample of 20 matrices is sufficiently reliable. In a few cases we analyzed up to 60 equally sized matrices and this confirmed that 20 is a sufficiently large number.

This procedure allows an analysis of the correlation between matrix size and axis percentages in CA. However, another factor in CA is the inertia, or sum of the eigenvalues. In our random files the total inertia rarely exceeds 1.0, which is considered to be a low value, caused by the fact that the majority of the values range between 0 and 9, and higher values are scarce in the adapted matrices, and absent in the other two groups.

Therefore, we created new random matrices, in which we randomly multiplied certain cells by stepwise increasing factors, incrementing the total inertia to values over 6.0. We analyzed the correlation of increasing inertia with decreasing values of the first axis of CA.

Finally, we analyzed a set of matrices with 70% zeros, as frequently found in palaeontological practice.

3. Analysis of the correlation matrix size/axis values

Since the matrices contain random data that, in principle, present no correspondence (except maybe for some fortuitous case), one has to admit that conclusions based on real cases with similar axis values are invalid. Valid conclusions based on the axis percentages of CA can only be drawn on values that are considerably larger than the results obtained from the random matrices.

The size of the array is strongly related with the percentages obtained on the first axes (Fig. 1). The percentages of the first four axes in a large matrix are considerably smaller than in a 20×10 matrix. In real cases, the threshold from where results may be considered to be useful should be placed much higher in small arrays. Apart from that, one should consider whether —independent of the array size— results of less than about 70% for the first four axes are useful.

The values obtained for the "adapted" matrices are constantly lower than those for the corresponding "standard" matrices. Assuming that the method of adaptation used did not introduce structure into the matrix (and there is no reason to believe it did), one must conclude that a relatively small number of cells with much higher values than the majority has a considerable influence on the results obtained. Ter Braak [3] suggested logarithmic transformation for such matrices. We tried this, and in some cases the result for the first axis in the log-transformed matrices was lower than in the original matrix, but in quite some cases considerably higher values were found. Apparently, logarithmic transformation produces unpredictable results.

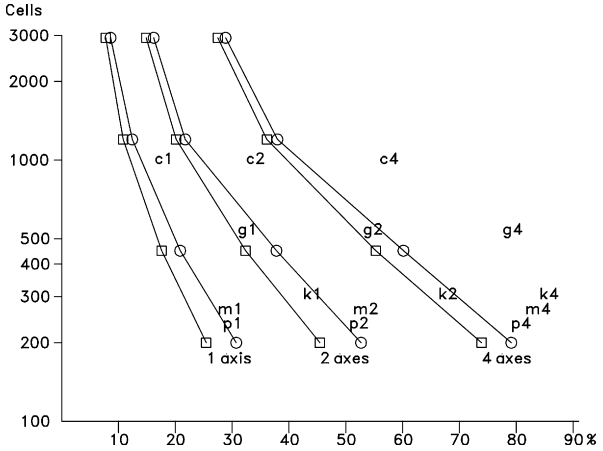

Fig. 1. Correlation between matrix size (number of cells) and percentage of the eigenvalues for the first axis, the sum of the first two axes, and the sum of the first four axes. Squares represent the means for 20 random matrices, circles represent the maximum values found. C, g, k, m, and p are the positions of one, two and four axes in real matrices.
Fig. 1. Corrélation entre la taille de la matrice (nombre de cellules) et le pourcentage de valeurs propres pour le premier axe, la somme des deux premiers axes et la somme des quatre premiers axes. Les carrés représentent les moyennes pour 20 matrices aléatoires, les cercles représentent les valeurs maximales trouvées. C, g, k, m, et p sont les positions de un, deux, trois et quatre axes dans les matrices réelles.

Plotting the first two axes of CA may give diffuse distributions, or there are some groups and isolated points, for both the standard and the adapted matrices. In the largest matrices no groups or outlying points can be recognized.

On Fig. 1 the mean values and the maxima of the first four axes of the standard matrices are plotted against the number of cells. The correlation between the number of cells and the percentages of the axes is evident. The letters **c**, **g**, **k**, **m** and **p** on Fig. 1 represent the position of real data taken from the literature that will be discussed afterwards.

For the decision whether the results of CA for a real matrix are meaningful, we have to take as a threshold the maximum value found in our random data, for the corresponding matrix size, plus a certain margin. In practice this means that the results of a medium-sized matrix (500–1000 cells) should be considered insufficient when the value of the first axis is below 25–30%, or the sum of the first two axes is below 35–45%. In large matrices the threshold should be chosen at 20% (one axis) and 30% (two axes). For small matrices the thresholds are around

40 and 60%, and these percentages must be considered conservative estimates.

An additional problem with big matrices with low axis values is the number of axes to be interpreted. Interpretation of more than four axes is practically impossible. So, probably no valid conclusions can be drawn when the first four axes cover less than 70% of the inertia, independent of the size of the matrix.

On the other hand, this does not mean that there is no correlation in a matrix with low axis values. It only means that in such cases no valid conclusions can be based on CA, and visual inspection of the matrix may be more fruitful.

4. The shape of the matrix

We took the number of cells as a measure for the matrix size, but the relation between the number of columns and rows has some influence too: as a general rule square tables give lower axis values than oblong tables; e.g. in ten 50×60 standard matrices the value of the first four axes are constantly lower than in the 98×30 tables, which have practically the same number of cells. In the same way the values for 14×14 tables are slightly lower than for 20×10 tables (Table 1). When one decreases the value of one of the dimensions of the table, maintaining more or less the same number of cells, the values of the first four axes of CA will increase until —of course— reaching 100% in matrices with only five columns or five rows. In such matrices one should analyze only one or two axes, and these should give very high values in order to be meaningful. In small matrices at least one of the dimensions is necessarily small and that is one of the reasons why they show higher axis values than large matrices.

5. Analysis of the correlation inertia/axis value

When discussing the results with Dr Casanovas-Vilar (Sabadell) the question arose whether the total inertia could influence the results. In our standard and zero matrices the values of the cells vary between 0 and 9; in real matrices the differences between cells are usually much greater, resulting in a greater inertia. We therefore refined the method of multiplication applied to the adapted matrices, and found that there is some correlation between total inertia and the percentages found for the first axis, depending on the size of the matrix (Fig. 2).

For each of the classes of 1200, 450 and 200 cells we created matrices in which we increased the inertia through six steps of multiplications of cell values by increasing factors, returning to the original matrix

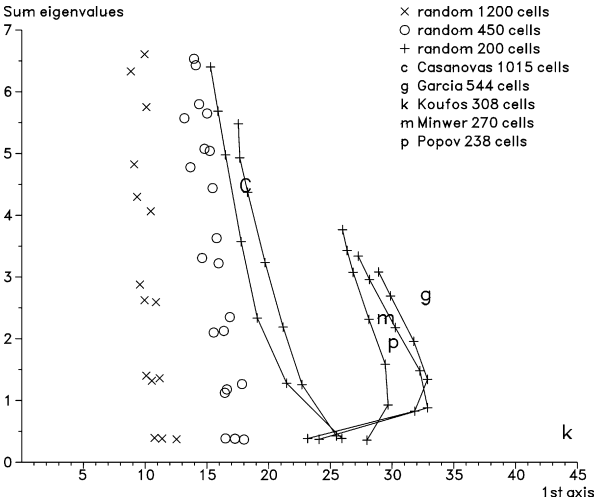

Fig. 2. Correlation between the sum of the eigenvalues (inertia) and the percentage found for the first axis of CA for various matrix sizes.
Fig. 2. Corrélation entre la somme des valeurs propres (inertie) et le pourcentage trouvé pour le premier axe de CA pour des tailles variées de matrices.

after each step. We did this in two different ways. In the first algorithm we multiplied the same cells in these six steps. This does not change the structure of the matrix, it merely stretches the range of the values. In the second algorithm, in each step we randomly chose the cells to be multiplied, creating six matrices with different structures. There were no important differences in the results of these two methods.

For the resulting 20×10 matrices the variability is very great and the points are distributed in an irregular way, but higher values on the first axes seem to be correlated with lower eigenvalues, though in some matrices the opposite is the case. On Fig. 2 the consecutive multiplication steps for each 20×10 matrix are connected, so one can see that in several cases the first step causes a strong increase of the percentage of the first axis; after that there is a negative correlation.

For large matrices there is practically no correlation, the points plot on an almost vertical line.

6. The influence of zeros in a matrix

Creating random matrices with a high percentage of zeros is not easy, and several algorithms were rejected because they apparently introduced rhythmic sequences in the matrix, often recognized by the fact that the first two axes gave practically the same values. When we analyzed graphic representations of such matrices (coloring the background of zero cells [Fig. 3]), we observed in quite some cases diagonal or V-shaped patterns of zeros.

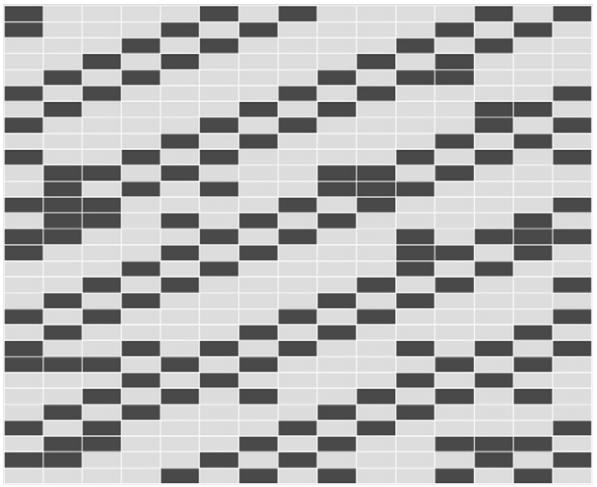

Fig. 3. Diagonal distribution of zero cells. Light grey: cells with values from 1 to 9; dark grey: cells with 0.
Fig. 3. Distribution diagonale des cellules zéro. Gris clair : cellules avec valeurs de 1 à 9 ; gris foncé : cellules zéro.

These matrices gave very high percentages for the first axes, in comparison with the standard matrices they were derived from (see line "pattern" in Table 1), but in fact they are no longer random matrices. Before executing CA on a real matrix one should analyze it, to make sure that there is no meaningless accidental pattern that would influence the results of CA.

Ter Braak ([3], table 5.3) noted the influence of diagonal structures on CA. A hidden diagonal structure may become visible by reordering the rows (and maybe that is what CA does, because randomly reordering the rows has no influence on the axis values); however, in a real matrix where the rows are placed in stratigraphic order, such reordering would be senseless because it approaches information that by nature is separated.

The rejected algorithm revealed a second problem: the lines "pattern" in Table 1 refer to matrices with 35% zeros that form a pattern. The values for the CA axes are much higher than in the matrices they were derived from. We repeated the same procedure substituting with "1" and "2". In these cases the axis values do not deviate significantly from those of the standard matrices. Apparently a pattern formed by zeros has a much greater influence on CA results than a pattern formed by a nonzero value.

The matrices produced by the correct algorithm, not introducing a pattern, are represented on the lines "zero" in Table 1. They are not significantly different from the matrices they were derived from.

The matrices of Casanovas-Vilar and Agusti [4], García-Alix et al. [5] and Minwer-Barakat [10] have about 70% of zeros. Line A in Table 2 lists the values obtained for 20 random 35×30 matrices with 70% zeros. Their inertia varies between 2.8 and 3.5, mean 3.1. Apparently the high proportion of zeros produces a very great variability, and some very high values for the first axis. We repeated this for matrices, where randomly chosen cells were multiplied to obtain greater total differences between cells. Their inertia varies between 3.6 and 4.6, mean 3.9, but the maxima obtained for the axes do not differ substantially from the previous case. The results are given in Table 2, line B.

7. Absence/presence matrices

Sometimes CA is applied to absence/presence matrices that contain only zeros and ones. We transformed five random matrices of each size group to such absence/presence matrices, and found nearly always an increase of the value of the first axis of CA. In the 20×10 matrices the greatest increase found was from 24.0 to 41.3%, and since we tried only five matrices greater increases are certainly possible. In the 30×15 arrays the increase was about 5%, with one exception: from 17.9 to 27.2%. In the larger matrices only slight increases were found. In the 30×35 matrices with 70% zeros the maximum increase found was from 14.0 to 21.5%. Moreover, whereas the original matrices normally give a diffuse plot, these random absence/presence matrices tend to show clear groupings in the plot of the first two axes.

When applying CA to absence/presence matrices, one must consider a higher threshold to decide whether the results are meaningful.

8. Comparison of CA and Principal Components Analysis (PCA)

Apart from CA we submitted our 20×10 and 98×30 matrices to PCA. The results, represented in Table 3, are quite similar to the results of CA. This may mean that the same considerations presented here for CA apply to PCA too.

9. Comparison with real data matrices

We compared the results of our random data with real matrices taken from Casanovas-Vilar and Agustí [4], García-Alix et al. [5], Koufos [9], Minwer-Barakat [10] and Popov [11].

On Fig. 1, c1, c2 and c4 represent the first axis, the sum of the first two axes and the sum of the first four axes of the Casanovas matrix (1015 cells). C1 is very close to

Table 3
Results of PCA for 20 × 10 and 98 × 30 matrices over the same data as in Table 1.
Tableau 3
Résultats des PCA pour 20 × 10 et 98 × 30 matrices sur les mêmes données que dans le Tableau 1.

20 × 10		1 axis			2 axes			4 axes			Total inertia		
	n	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
Standard	20	19.33	23.02	27.90	37.50	42.06	49.13	65.51	70.26	75.46	75.84	82.32	92.63
Adapt	20	21.34	25.90	33.02	37.32	45.59	54.89	63.55	73.15	79.58	1358.20	1943.44	2636.69
Zero	20	21.27	24.52	29.50	38.67	42.56	47.79	67.40	70.37	76.20	90.27	100.91	113.73
Pattern	20	20.09	27.41	33.59	38.79	48.35	56.16	65.54	74.24	81.65	88.03	100.61	115.20
98 × 30													
Standard	20	6.86	7.50	8.45	13.47	14.38	15.68	25.69	26.68	28.28	231.59	241.41	251.54
Adapt	20	6.95	7.76	8.49	13.70	14.85	16.04	26.33	27.55	29.57	5405.45	5890.76	6219.45
Zero	20	6.90	7.46	8.20	13.61	14.36	15.56	25.70	26.70	28.13	286.26	298.27	305.92
Pattern	20	9.16	16.59	18.97	17.75	31.15	34.44	32.84	49.18	52.49	281.19	297.50	311.27

the range of random matrices, c2 and c4 are outside the range of random matrices.

The c on Fig. 2 also represents the Casanovas matrix. The value of the first axis: 18.1, with a sum of the eigenvalues of 4.48 (pers. comm. Dr Casanovas-Vilar), scores better than our random matrices. By interpolation one may estimate the maximum of the first axis value for a 1000 cell random matrix to be between 11 and 12 (in a few trials we found a maximum of 11.0). It is difficult to say whether the difference between 12 and 18 is large enough to conclude that the Casanovas matrix is outside the zone of random matrices. On the other hand, the sum of the first four axes: 57.72, is so low that the interpretation may easily be incorrect.

The Casanovas matrix contains about 70% zeros, so we created 20 random matrices of the same size (30 × 35) with 70% zeros and 20 matrices with 70% zeros and multiplication of values (Table 2), using the algorithm that does not introduce a pattern. In both cases the sum of the eigenvalues is comparable to the value found by Casanovas-Vilar and Agusti [4], and the maximum values obtained for the axes come so close to the values in the Casanovas matrix that one must conclude that the latter are not significantly different from a random result.

On Fig. 1, g1, g2 and g4 represent the axis data of the García matrix [5]. The values obtained are well above the limits calculated from the random files, but we could not make a useful interpretation of these results, and think that in this case they are fortuitous. The same goes for m1, m2 and m4 of the Minwer matrix [10]. On Fig. 2 both these matrices fall outside the range of random matrices in view of the number of cells they contain.

On Fig. 1, k1, k2 and k4 are the data for the Koufos matrix [9]. The values are very high, and their interpretation by Koufos appears to have a sound basis. This is confirmed on Fig. 2, in spite of the low value of the inertia.

The Popov matrix (p on Fig. 1) is a 34 × 7 absence/presence matrix (238 cells). On Fig. 1 and Fig. 2 it falls within the range of the standard random matrices. As said before, we found a value of 41% for the first axis of CA in a 20 × 10 absence/presence matrix, which is considerably larger than the 30.1% found by Popov [11]. The conclusions of Popov may be perfectly correct, but they cannot be inferred from the results of CA.

10. Conclusions

Conclusions obtained from CA cannot be evaluated correctly when the total inertia and the matrix size are not given. In publications of the results of CA this information should be available. Another indispensable datum is the percentage of zero cells.

CA is doubtlessly a useful technique. But, the high values obtained from random matrices demonstrate that one should be careful when using CA for analyzing real data matrices; the obtained values should be well above the threshold values presented here. This is especially true when these matrices are not very big, or when they contain a high percentage of zeros. Simple visual inspection of a data matrix is then probably more reliable.

For the decision whether the numerical results of CA for a real matrix are meaningful, we have to consider a threshold based on the results from the random matrices. For small matrices (up to 500 cells) the first axis should represent at least 40% of the inertia, and the first two axes should sum 60%. In a medium-sized matrix (500–1000 cells) these limits are 30 and 40%, respectively. In large matrices the thresholds should be chosen at 20% (one

axis) and 30% (two axes). In such cases, however, serious problems arise, because one has to interpret too many axes, and, probably, one should refrain from using CA when the sum of the first four axes is less then 70%. For matrices with many zeros and for absence/presence matrices the thresholds are higher than stated before.

It is necessary to check whether a matrix contains an accidental hidden diagonal structure, which may result in a high but meaningless value for the first axis.

Acknowledgements

We thank Dr Ø. Hammer (Oslo) for valuable information about CA. Dr I. Casanovas-Vilar (Sabadell) contributed substantially to this article through a highly appreciated email discussion. Dr Koufos (Thessalonica) kindly provided us with his original data table. This study was realized in the framework of the project Consolider-Ingenio 2010, CSD2006-00041.

References

[1] J.P. Benzécri, Pratique de l'analyse des données, Analyse des correspondances, vol. 2, Dunod, Paris, 1972, p. 424.
[2] J.P. Benzécri, Correspondence Analysis Handbook, M. Dekker, New York, 1992, p. 688.
[3] C.J.F. Ter Braak, Ordination, in: R.H. Jongman, C.J.F. ter Braak, O.F.R. van Tongeren (Eds.), Data Analysis in Community Ecology, Cambridge University Press, 1995, pp. 91–173.
[4] I. Casanovas-Vilar, J. Agustí, Ecogeographical stability and climate forcing in the Late Miocene (Vallesian) rodent record of Spain, Palaeogeogr. Palaeoclimatol. Palaeoecol. 248 (2007) 169–189.
[5] A. García-Alix, R. Minwer-Barakat, E. Martín-Suárez, M. Freudenthal, J.M. Martín, Late Miocene–Early Pliocene climatic evolution of the Granada Basin (southern Spain) deduced from the paleoecology of the micromammal associations, Palaeogeogr. Palaeoclimatol. Palaeoecol. 265 (2008) 214–225.
[6] M. Greenacre, Theory and Applications of Correspondence Analysis, Academic Press, 1984, p. 364.
[7] M. Greenacre, Correspondence analysis in medical research, Stat. Methods Med. Res. 1 (1992) 97–117.
[8] Ø. Hammer, D.A.T. Harper, P.D. Ryan, PAST: Paleontological Statistics Software Package for Education and Data Analysis, Palaeontol. Electron. 4 (2001), 9p. http://palaeo-electronica.org/2001_1/past/issue1_01.htm (website accessed on 11th December 2008).
[9] G.D. Koufos, Palaeoecology and chronology of the Vallesian (Late Miocene) in the Eastern Mediterranean region, Palaeogeogr. Palaeoclimatol. Palaeoecol. 234 (2006) 127–145.
[10] R. Minwer-Barakat, Roedores e insectívoros del Turoliense superior y Plioceno del sector central de la cuenca de Guadix, Doctoral Thesis, Universidad de Granada, 2006, 548 p.
[11] V.V. Popov, Pliocene small mammals (Mammalia, Lipotyphla, Chiroptera, Lagomorpha, Rodentia) from Muselievo (North Bulgaria), Geodiversitas 26 (2004) 403–491.