ESTUDIOS DE ECONOMÍA APLICADA VOL. 30-1 2012 PÁGS. 1–28

A Bootstrap Procedure in the Context of Correspondence Analysis: Numerical Approach to Measure the Stability of Axes[*]

OLGA VALENCIA GARCÍA

Departamento de Economía Aplicada, Área de Métodos Cuantitativos para la Economía y la Empresa, UNIVERSIDAD DE BURGOS, ESPAÑA. E-mail: oval@ubu.es

RAMÓN ÁLVAREZ ESTEBAN

Departamento de Dirección y Economía de la Empresa. Área de Estadística e I.O., UNIVERSIDAD DE LEÓN, ESPAÑA E-mail: ramon.alvarez@unileon.es

ABSTRACT

Bootstrap procedures are commonly used to evaluate the stability of principal axes methods. In our research, attention is focussed on the particular case of Correspondence Analysis (CA). As in classical CA the metrics are induced by table margins and thus would vary over the replicated tables, we suggest here a specific Bootstrap procedure in which a generalised CA is performed on the replicated tables imposing as metrics those issued from the original table. The contribution of our paper is to provide a numerical procedure to quantify the stability of principal axes which is suitable for this context. We give a bounded measure of the stability of axes. Furthermore, as any fixed threshold handled to label an axis as stable or unstable would be arbitrary, our procedure is based on the comparison of real and randomly permuted data to determine stability thresholds. Computational results obtained on several data sets are offered.

Keywords: Correspondence Analysis, Principal Axes, Resampling, Bootstrap, Procrustes, Stability.

Un procedimiento Bootstrap en el contexto del Análisis de Correspondencias: Aproximación numérica para medir la estabilidad de los ejes

RESUMEN

Los procedimientos Bootstrap se utilizan habitualmente para evaluar la estabilidad de los métodos basados en ejes principales. En nuestra investigación, la atención se centra en el caso particular del Análisis de Correspondencias. Dado que en el Análisis de Correspondencias clásico las métricas están determinadas por las frecuencias marginales de las tablas y, por tanto, variarían a través de las tablas replicadas, sugerimos aquí un procedimiento Bootstrap específico en el que se lleva a cabo un Análisis de Correspondencias generalizado de las tablas replicadas imponiendo como métricas las procedentes de la tabla original. La contribución de nuestro trabajo es proporcionar un procedimiento numérico para cuantificar la estabilidad de los ejes principales que es adecuado para este contexto. Proponemos una medida acotada de la estabilidad de los ejes. Además, teniendo en cuenta que cualquier umbral fijo considerado para etiquetar un eje como estable o inestable sería arbitrario, nuestro procedimiento se basa en la comparación de datos reales y datos permutados aleatoriamente para determinar umbrales de estabilidad. Se presentan los resultados computacionales obtenidos en varios conjuntos de datos.

Palabras clave: Análisis de Correspondencias, ejes principales, remuestreo, Bootstrap, Procustes, estabilidad.

JEL Classification: C15

[*] *We thank Mónica Bécue-Bertaut (Polytechnic University of Catalonia, UPC) for their support and carefully reading versions of the manuscript.*

Artículo recibido en julio de 2011 y aceptado en noviembre de 2011
Artículo disponible en versión electrónica en la página www.revista-eea.net, ref. ə-30109

ISSN 1697-5731 (online) – ISSN 1133-3197 (print)

1. INTRODUCTION

Bootstrap procedures (Efron 1979; Efron and Tibshirani 1993) are commonly used to assess the stability of principal axes methods. Two strategies appear in literature. The strategy called Partial Bootstrap (PB) (Lebart 2006) projects Bootstrap rows and columns as supplementary elements onto the axes issued from the analysis of the original dataset. This just allows for assessing the stability of every row or column (Beran and Srivastava 1985; Gifi 1981; Greenacre 1984; Chateau and Lebart 1996; Álvarez *et al.* 2010). Alternatively, the so-called Total Bootstrap (TB) consists in performing the analysis on every replicated table. Row and column coordinates, eigenvalues and eigenvectors are computed for every bootstrap replication. Thus, the stability of the principal axes or the subspaces can be tackled (Diaconis and Efron 1983; Stauffer *et al.* 1985; Daudin *et al.* 1988; Lambert *et al.* 1990; Jackson 1993 and Peres-Neto *et al.* 2005).

In order to provide a right measure of stability by means of TB, the comparison of original and replicated axes becomes crucial. As Milan and Whittaker (1995) pointed out, a direct comparison can be misleading and an appropriate one requires eliminating the specific problems which arise from the conditions associated to the algorithm of eigen-decomposition used in principal axes methods. The constriction of orthogonality imposed to the eigenvectors as well as their ordination depending on the associated eigenvalues lead to over dispersion (and bias) of the bootstrap distributions of the statistics. In Milan & Whittaker's proposal, these problems are dealt with Procrustes-like rotations (Gower 1971; Gower 1975; Markus 1994), so the vectors of coordinates are computed for every replicated table and rotated to adjust, as far as possible, the initial configuration. The distributions of the rotated statistics are used to assess the stability of the row points (and/or column points) and the eigenvalues. This process avoids a too pessimistic evaluation of the stability of the principal axes, due to possible reflections, ordination and/or rotations of the axes (especially when the eigenvalues are only slightly different).

Following the same line of research, Linting *et al.* (2007) go more deeply into the study of the stability of nonlinear Principal Components Analysis (PCA) from a graphical perspective. They thoroughly examine the stability of several nonlinear PCA results, providing confidence intervals for the quantified variables and confidence ellipses for the eigenvalues, the component loadings and the person scores. Their study is focussed on the graphical representation of the bootstrap results for a two-dimensional solution. The confidence ellipses are constructed from a certain percentage of the bootstrap points closest to the centroid of the bootstrap cloud while retaining the orientation of the cloud in two dimensions. Hence, conclusions regarding the degree of stability of different results are based on the relative sizes of their respective ellipses.

Within a similar strategy that combines a Total Bootstrap with Procrustes rotations, we propose here to extend this approach to the particular case of Correspondence Analysis (CA). However, our attention is focussed on the use of TB with the purpose of measuring the stability of CA principal axes and therefore a numerical approach is developed.

CA is a particular case of principal axes methods (Benzécri *et al.* 1973; Lebart *et al.* 1984; Greenacre, 1984 and Escofier, 2003). The best validation criterion in CA consists in verifying the stability of the pattern as well (Lebart *et al.* 2000). Lebart (1976) and O'Neill (1978) have analysed the asymptotic eigenvalue distribution. Escofier and Le Roux (1972) and Bénasséni (1993) consider perturbational aspects in correspondence analysis. Delta method has been developed by Gifi (1981). Pearson test statistic can be used to determine the eigenvalues significantly different from zero (Gilula and Haberman, 1986). However, this test can be only applied on contingency tables issued from a simple random sample. Whenever this or other requirements are not fulfilled, other approaches like bootstrap resampling have to be adopted.

Concerning bootstrap techniques and CA, some authors have resorted to PB just to build confidence regions for row and column categories (Ringrose 1992; Greenacre 1984; Lebart *et al.* 2000; Lebart 2004a and Lebart 2004b). Other studies have applied TB to determine meaningful eigenvalues (Reiczigel 1996) or contributions of row and column categories (Tan *et al.* 2004).

In this paper, we suggest a numerical procedure to quantify the stability of principal axes in this context, i.e., taking into account both the specific features of CA and the above-mentioned problems concerning the combination of Bootstrap with principal axes methods.

As in CA the metrics are induced by table margins and thus vary through the replicated tables, a generalised CA is performed on the resampled tables imposing as metrics those issued from the original table. By means of this particular TB, including also Procrustes corrections, we provide a bounded measure of the stability of every principal axis. Furthermore, in order to assess the computed stability measures, we propose to avoid arbitrary thresholds by means of comparisons with hazard.

Section 2 recalls the principles of CA with imposed metrics (Escofier 1984, 1987, 2003; Escofier and Pagès 1983) used to analyze the bootstrapped samples. Section 3 sets out in detail the methodology proposed to quantify the stability of principal axes, supported by a small example. Finally, in Section 4, we offer and discuss the results obtained on several data sets.

2. CORRESPONDENCE ANALYSIS AND GENERALISED CORRESPONDENCE ANALYSIS

2.1. General framework for principal axes methods

The different principal axes methods can be presented as particular variants of Principal Component Analysis (PCA) (Tenenhaus and Young 1985) using a data matrix X and two diagonal metric matrices, $\mathbf{D_r}$ the row weighting matrix (and metric in the column space), and $\mathbf{D_c}$ the column weighting matrix (metric in the row space).

PCA(X, $\mathbf{D_r}$, $\mathbf{D_c}$) consists in identifying the principal axes of row and column clouds and computing the coordinates matrices **F** and **G** of, respectively, rows and columns coordinates on these axes. $\mathbf{X^tD_rXD_c}$ is diagonalised and thus the column eigenvectors matrix **U** (orthonormal) and the diagonal eigenvalue matrix $\mathbf{\Lambda}$ are computed. Let **V** be the row eigenvector matrix, $\mathbf{V=\Lambda^{-1/2}F}$. The matrices of coordinates **F** and **G** are computed in the following way:

$$\mathbf{F = XD_cU} \qquad \mathbf{G = U\Lambda^{1/2}} \tag{1}$$

We use F_s to denote the s-column of **F**, v_s the s-column of **V**, $\mathrm{G_s}$ the s-column of G, u_s the s-column of **U** and λ_s the s-diagonal term of $\mathbf{\Lambda}$. We have v_s = standardized F_s, u_s = standardized G_s, where $\mathrm{Var}(F_s)= \mathrm{Var}(G_s)= \lambda_s$

2.2. Classical CA

Classical CA applied to proportion table **P** with general term p_{ij} is equivalent to PCA (**X**, $\mathbf{D_r}$, $\mathbf{D_c}$) where **X** has for general term:

$$x_{ij} = (p_{ij} - p_{i.}p_{.j})/(p_{i.}p_{.j}) \tag{2}$$

$\mathbf{D_r}$ is a diagonal matrix with general term $p_{i.}$ $\{i = 1,\ldots, I\}$ and $\mathbf{D_c}$ is a diagonal matrix with general term $p_{.j}$ $\{j= 1, \ldots, J\}$(Escofier and Pagès 1998).

2.3. CA with imposed metrics

Escofier (2003) has proposed to perform CA with imposed metrics, $\mathbf{Q_I}$ and $\mathbf{Q_J}$ other than those issued from the margins. The output of this generalised CA is obtained by performing a PCA on table **X** with general term

$$x_{ij} = (p_{ij} - p_{i.}p_{.j})/(q_{i.}q_{.j}) \quad \sum_{i=1}^{I} q_i = 1 \quad \sum_{j=1}^{J} q_j = 1 \tag{3}$$

in metrics $\mathbf{Q_I}$ and $\mathbf{Q_J}$. CA with imposed metrics preserves the duality property between rows and columns. Transition formulae allow crossing from **F** coordinates to **G** coordinates replacing $\mathbf{D_c}$ by $\mathbf{Q_J}$ and $\mathbf{D_r}$ by $\mathbf{Q_I}$ into the general framework transition formulae (section 2.1).

3. METHODOLOGY

3.1. Outline of the methodology through a small example

We suppose the observation of two categorical variables on a random sample composed of 97 individuals which leads to the original data set. The first bootstrap resample is obtained by the well-known Non-parametric Bootstrap which replicates the original extraction method for the sample. This gives rise to a data set with the same grand total (97 individuals) but different marginal frequencies (Table 1). Figure 1 illustrates the processing.

Table 1
Outline of the methodology through a small example: original and bootstrap data sets

Original data set					Bootstrap data set				
	C1	**C2**	**C3**	**Marginal**		**C1**	**C2**	**C3**	**Marginal**
R1	6	2	12	20	R1	4	1	10	15
R2	5	1	4	10	R2	8	1	2	11
R3	14	7	16	37	R3	13	3	19	35
R4	10	8	12	30	R3	10	9	17	36
Marginal	35	18	44	97	Marginal	35	14	48	97

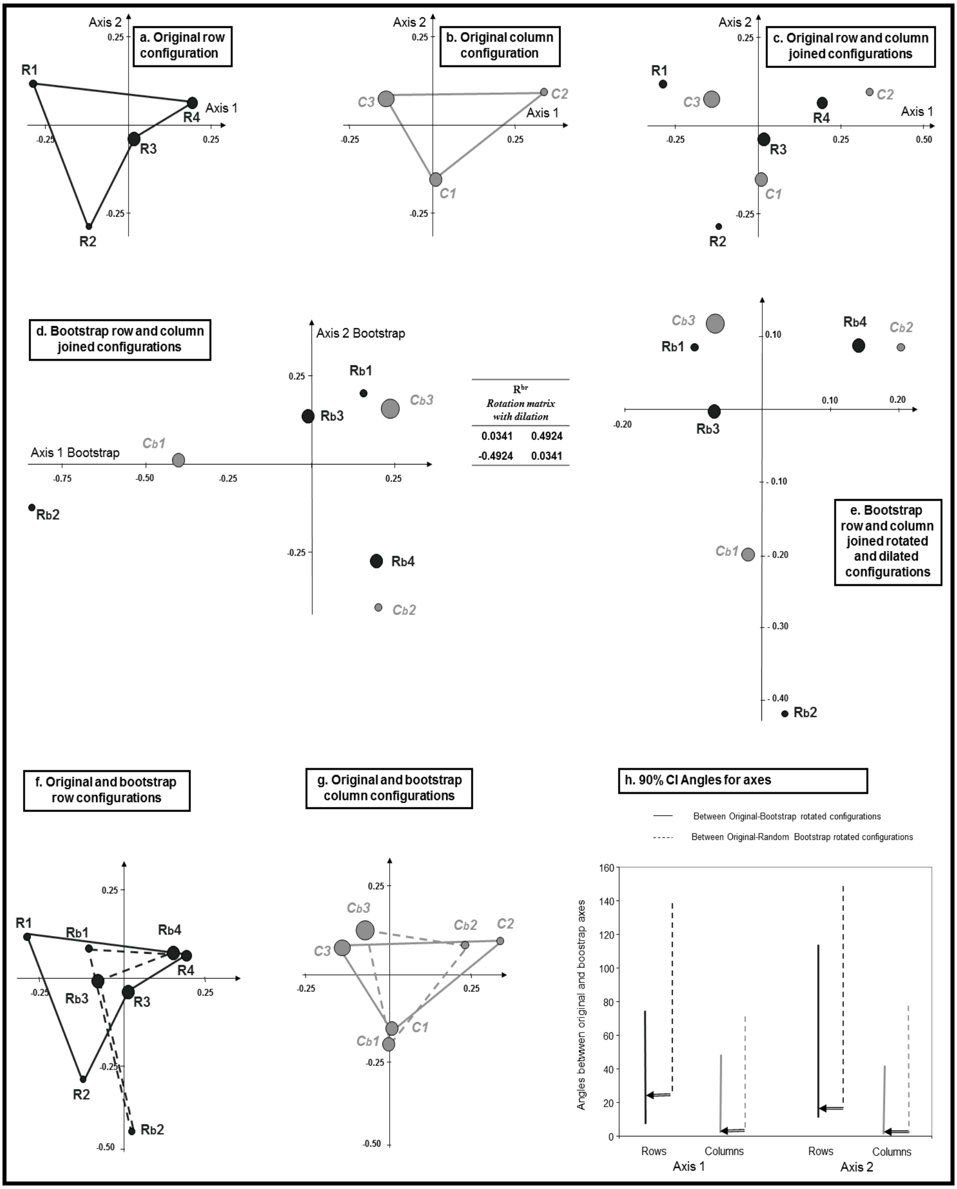
Source: Own elaboration.

A classical CA is conducted on the original data set. A CA with imposed metrics has been carried out on the bootstrap resample (see 3.2 for details). The imposed metrics are those issued from the original data set. The coordinates of the rows and the columns of both the original and the bootstrap data sets are represented in the same Euclidean space (Figs. 1a, 1b and 1c). However, the direct comparison among either original coordinates or principal axes (Fig. 1c) and their respective bootstrap (Fig. 1d) needs a previous adjustment in order to remove the apparent differences. This adjustment, performed by means of orthogonal Procrustes rotations, will be crucial in the study of stability suggested in this paper. Figure 1e depicts the new bootstrap coordinates of rotated points which can be compared jointly with the original joined configuration (Fig. 1c) or overlapping the original rows (Fig. 1f) or columns (Fig. 1g).

Angles between original vectors of coordinates and their respective bootstrap are computed (see 3.2 for details), but these measurements include the apparent differences. Once Procrustes rotations are applied, angles between original and bootstrap principal axes give a more suitable measure of the stability of principal axes.

Figure 1

Outline of the methodology through a small example



(a) Original row configuration; (b) Original column configuration; (c) Original row and column joined configuration; (d) Bootstrap row and column joined configuration; (e) Bootstrap row and column joined rotated configuration; (f) Original and bootstrap row configurations; (g) Original and bootstrap column configurations; (h) 90% CI angles between original and rotated bootstrap axes from real data (solid line) and angles between original and rotated bootstrap axes from randomly permuted data (broken line).

Source: Own elaboration.

If this bootstrap procedure is carried out repeatedly enough, these angles may be an appropriate index of principal axes stability (solid lines in Figure 1h). However, interpreting the magnitude of these angles in terms of stability is somehow complex as any threshold could be arbitrary. Accordingly, the angles computed from real data have been compared with those calculated from randomly permuted data (broken lines in Figure 1h).

3.2. Stability of CA axes through Total Bootstrap

In this section, we indicate the rationale of the processing that is used to quantify the stability of CA principal axes and eventually to select the stable subspace either for representing the data or for further analyses such as clustering.

3.2.1. CA applied to the original table

Classical CA results **F**, **G**, **U** and **V** are computed performing classical CA on the original proportion matrix.

3.2.2. CA applied to the replicated table

B Bootstrap samples are extracted replicating the original sampling method by the well-known Non-parametric Bootstrap. As above-mentioned, the bootstrapped tables have the same grand total as the original table but, in general, different marginal frequencies.

Generalised CA are conducted on these *B* resamples, imposing as metrics those of the original table, i.e. $\mathbf{Q_I}=\mathbf{D_r}$ and $\mathbf{Q_J}=\mathbf{D_c}$, the diagonal matrices with the row and column marginal proportions of the original data set. So the output of this generalised CA is obtained by performing a PCA on each table $\mathbf{X^b}$ with general term

$$x_{ij}^{b} = (p_{ij}^{b} - p_{i.}^{b} p_{.j}^{b}) \big/ (p_{i.} p_{.j})$$

thereby obtaining $\mathbf{F^b}$, $\mathbf{G^b}$, $\mathbf{U^b}$ and $\mathbf{V^b}$.

Thus, the rows (respectively, the columns) of both the initial and the replicated tables are located in the same metric space.

3.2.3. Rotation of the replicated set of points

We have to compare the configuration of points $\mathbf{C^b}$ issued from the generalised CA applied to the replicated table and the original configuration **C** issued from CA applied to the original table. Both configurations of points have *I*+*J* rows and *K* columns, where *K* is the number of principal axes in CA, *K*=min(*I*,*J*)-1. In our case all configurations are centred.

Let D be the weighted vector ($\mathbf{D_r}$ and $\mathbf{D_c}$ diagonals). The square distance S^2 between both configurations is computed as:

$$S^2 = \sum_{h=1}^{I+J} \sum_{k=1}^{K} D_h (C_{hk} - C_{hk}^b)^2 \qquad (4)$$

This total distance can be split into row distance and column distance. Differences between original and bootstrap coordinates can be explained by *apparent* and *real* variability. In CA, apparent variability derives from the application of bootstrap to a method that incorporates an eigenvalue/eigenvector decomposetion. On the contrary, the real variability is merely due to resampling fluctuations.

Procrustes rotation is used to remove the apparent variability looking for minimizing S^2. This rotation maintains the internal relationships among the points. A generalised orthogonal rotation (Krzanowski 1999; Ten Berge and Bekker 1993; Ten Berge 2006) can be applied either to only one set of points (row or column set) or to both sets but considered as a whole (Milan and Whitakker 1995). In the case of CA, where rows and columns play a symmetric role, the second option seems more suitable. Generalised orthogonal rotation includes reflections (180º rotation) and exchanges (90º rotation) as particular cases (Gower and Dijksterhuis 2004).

The rotation matrix $\mathbf{R^{br}}$ (Table 2) is computed as $\mathbf{R^{br}} = \mathbf{UV^t}$, where $\mathbf{U\Sigma V^t}$ is the singular value decomposition of the product matrix $\mathbf{C^{bt}DC}$. Thus: $\mathbf{F^{br}} = \mathbf{F^b}\ \mathbf{R^{br}}$, $\mathbf{G^{br}} = \mathbf{G^b}\ \mathbf{R^{br}}$, $\mathbf{V^{br}} = \mathbf{V^b}\ \mathbf{R^{br}}$ and $\mathbf{U^{br}} = \mathbf{U^b}\ \mathbf{R^{br}}$.

Table 2
Outline of the methodology through a small example:
Rotation matrix between bootstrap and rotated configuration without and with dilation

Rotation matrix without dilation		Rotation matrix with dilation	
0.0691	0.9976	0.0341	0.4924
-0.9976	0.0691	0.4924	0.0341

Source: Own elaboration.

3.2.4. Dilation

A uniform dilation, that is, a stretching or shrinking of all the points by a constant (homothethy) can be also performed after the rotation step (Table 2). Correlations between principal coordinates are not changed by dilations. However, the sum of the weighted square distances between original and bootstrap coordinates (S^2) diminishes.

3.2.5. Assessing stability of the axes

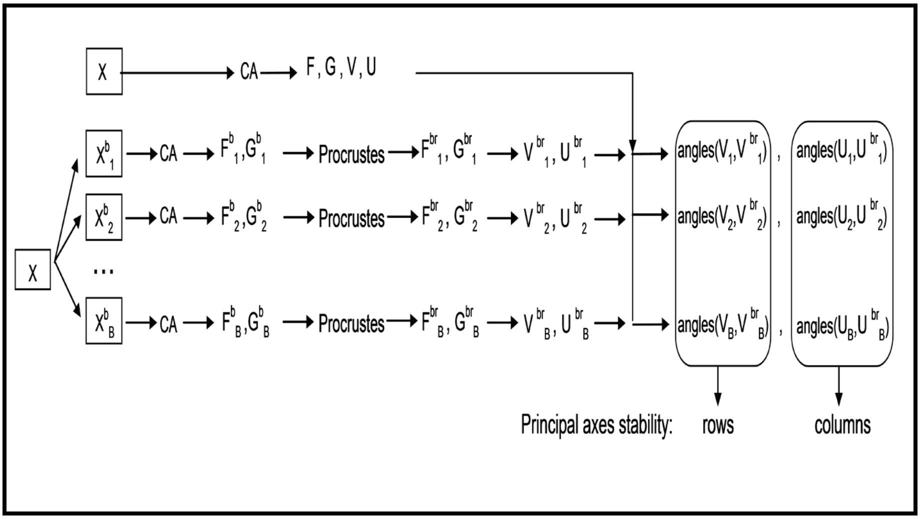
A series of indicators can be used for measuring the global similarity between original and replicated configurations (Schönemann and Carrol 1970; Abdi 2007; Holmes, 2008). In our study, the similarity between original and replicated coordinates vectors with the same rank is measured by Pearson's correlation *Corr* of principal coordinates F_s and F^b_s or, equivalently, between v_s and v^b_s. The correlation is usually expressed through the corresponding angle:

$$\mathrm{Corr}(F_s, F_s^b) = v_s^{\mathrm{t}} D_I v_s^b = \cos(\theta) \qquad (5)$$

Thus, for every dimension, we have the bootstrap distribution of the angle between the original and the rotated replicated principal axes with the same rank (Fig. 2 summarises the processing). We have computed the mean, the median, the 5th percentile (P_5) and the 95th percentile (P_{95}) of these distributions.

For our original table, 1000 bootstrap simulations have been carried out. The mean of the angles between original and bootstrap principal axes are not near 0º (solid lines in Figure 1h). Following this criterion, every principal axis is assessed as *Stable* (S) if it shows a high degree of stability before (bP) and after (aP) Procrustes rotation, *Apparently unstable but actually stable* (AU-S) when it presents a high degree of instability bP but not aP rotation (apparent variability), and *Unstable* (U) if the principal axis behaves as unstable both bP and aP.

Figure 2
Measure of stability of principal axes by means of Bootstrap and Procrustes


X → CA → F, G, V, U
X^b_1 → CA → F^b_1,G^b_1 → Procrustes → F^br_1, G^br_1 → V^br_1, U^br_1 → angles(V_1,V^br_1) , angles(U_1,U^br_1)
X^b_2 → CA → F^b_2,G^b_2 → Procrustes → F^br_2, G^br_2 → V^br_2, U^br_2 → angles(V_2,V^br_2) , angles(U_2,U^br_2)
X
…
X^b_B → CA → F^b_B,G^b_B → Procrustes → F^br_B, G^br_B → V^br_B, U^br_B → angles(V_B,V^br_B) , angles(U_B,U^br_B)
Principal axes stability: rows columns

Source: Own elaboration.

3.3. Comparison with hazard to determine a threshold for the angles

Interpreting the magnitude of these angles in terms of stability is somehow complex as any threshold could be arbitrary, depending on the data set analysed. Random matrices $\mathbf{X}_\mathbf{i}^{\mathbf{b\ random}}$ are also generated permuting **X** initial table cells in each *i* bootstrap sample (Fig. 1h). The distributions of the rotated statistics can be compared with these hazard distributions.

Angles computed for the data sets analysed are compared with those obtained from the random bootstrap. Being $(P_5^{real}, P_{95}^{real})$ the 90% CI of angles computed with real data by our suggested bootstrap procedure and $(P_5^{rand}, P_{95}^{rand})$, the 90% CI of angles computed with randomly permuted data by this procedure, a principal axis is labelled unstable if $(P_5^{real}, P_{95}^{real})$ overlaps with $(P_5^{rand}, P_{95}^{rand})$ (Fig. 1h).

4. RESULTS ON SEVERAL SAMPLES

4.1. Data sets and results

In order to check the performance of the proposed methodology, a series of data sets has been studied We have selected four frequency tables with different characteristics regarding the number of categories, the magnitude of cell frequencies, the presence of very low cell frequencies, the existence of empty cells and the grand total of the table (see Appendix A).

Tables 3 to 6 show the outcome obtained on the data sets analyzed. These tables display just angles computed from real data. The figures in bold indicate that instability holds according to the hazard criterion (comparison with randomly permuted data) included in the bootstrap procedure proposed.

In order to check this criterion for setting thresholds of instability, Tables 3b to 6b exhibit the results of real and randomly permuted data, both computed by means of bootstrap after Procrustes rotation. Figures in italics point out the 90% confidence intervals that are overlapped, i.e. the cases where $(P_5^{real}, P_{95}^{real})$ overlaps with $(P_5^{rand}, P_{95}^{rand})$, that revealing instability of the corresponding axes.

Regarding the issue of the imposed metrics, as noted above, we emphasize that the reason why these metrics are included in the procedure is to enable the rows (respectively, the columns) of both the initial and the replicated tables be located in the same Euclidian space. Actually, the resampled data sets simply have the same distribution as the original one, without imposing any restrictions on the marginal frequencies.

For ease of comparison between the proposed procedure with imposed metrics and the equivalent standard procedure (without imposed metrics), we have included all the computing for bootstrap after Procrustes rotation in

Appendix B. The calculations have been made from the same resampled tables in both cases. As one can see, the results on the stability of axes are similar in the data sets considered, i.e., both procedures point out the same unstable axes. So the imposed metrics does not lead to lower dispersion or any bias on results of the proposed procedure, but instead to just a more appropriate calculation of the similarity between every original and bootstrapped axis.

Finally, Pearson test statistic has been conducted in every data set, even if the data are not drawn from a simple random sample. Appendix C provides the results of this eigenvalue significance test as well as the eigenvalues and their corresponding percentages of total inertia of the raw data sets examined.

4.1.1. Data set 1. Fisher (1940)

Fisher (1940) analyses 5,387 schoolchildren from Scotland classified according to hair colour and eye colour, leading to a 5 x 4 contingency table with two low cell frequencies but no empty cells (Table 7 Appendix A). This is a Maung's compilation of Tocher's data (1908) with special district grouping of 502,155 children.

Table 3 shows that there are not stable axes before Procrustes rotation. Rotation allows for seeing that the variability of the first and second principal axis is mostly apparent. The third principal axis is really unstable, specifically in the case of the row set. This example shows that small mean angles (17.5º and 4.5º) are not enough to ensure stability. Pearson test statistic, α=5%, points out the first two eigenvalues different to zero.

Table 3
Stability performance of principal axes of data set 1
(Fisher, 1940)

Principal Axes	Before Procrustes				After Procrustes				Stability		
	Mean	Median	P_5	P_{95}	Mean	Median	P_5	P_{95}	S	AU-S	U
v_1	20.4	3.4	1.4	**176.6**	3.1	2.9	1.2	5.5		x	
v_2	87.7	12.4	2.7	**177.4**	5.5	5.1	1.9	10.1		x	
v_3	70.2	21.5	3.6	**175.0**	17.5	11.6	2.9	**53.8**			x
u_1	19.7	2.6	0.7	**177.4**	2.2	2.1	0.6	4.5		x	
u_2	87.4	10.5	1.4	**178.6**	4.1	3.6	0.9	8.6		x	
u_3	68.0	6.3	1.1	**178.6**	4.5	3.4	0.9	**11.4**			x

Figures in bold point out that instability holds according to the criterion suggested.

Source: Own elaboration.

Table 3b
Comparison between real and randomly permuted data. Bootstrap aP. Data set 1
(Fisher, 1940)

Principal Axes	Randomly permuted data				Real data			
	Mean	Median	P_5	P_{95}	Mean	Median	P_5	P_{95}
v_1	75,5	72,9	40,2	114,7	3.1	2.9	1.2	5.5
v_2	78,7	79,7	39,8	108,7	5.5	5.1	1.9	10.1
v_3	84,4	83,3	***29,4***	***148,9***	17.5	11.6	***2.9***	***53.8***
u_1	45,6	44,4	12,3	84,4	2.2	2.1	0.6	4.5
u_2	43,3	41,8	10,6	80,6	4.1	3.6	0.9	8.6
u_3	31,7	28,9	***9,0***	***60,9***	4.5	3.4	***0.9***	***11.4***

Figures in bold italics point out real and random 90% CI which are overlapped.

Source: Own elaboration.

4.1.2.Data set 2. Escofier y Pagès (1992)

It is a cross-tabulation of 271,096 French female school leavers who abandoned the educational system in 1973 and found a job with nine different types of job (rows) and eight educational levels (columns) (Table 8 Appendix A). Cells frequencies are very high (three figure numbers) but there are several empty cells resulting from the actual incompatibility between some jobs and some educational levels (e.g. engineers and no schooling at all), that is, from the existence of structural zeros. This gives rise to well-defined row and column categories and thus, to distances between profiles markedly determined by the presence or absence of categories for both rows and columns.

Table 4 with the results of the applied methodology indicates that the instability of the structure of the data set 2 is revealed apparent after Procrustes (very low angles). In this data set, axes 4 to 7 have very small eigenvalues and account each one for less than 1% of the total inertia (see Appendix C) but they should be considered meaningful dimensions. Pearson test statistic, α=5%, shows all the eigenvalues different to zero.

4.1.3.Data set 3. Abascal y Grande (2005)

A table with the frequencies of association between thirteen brands of milk and eleven milk attributes, according to the responses of a small sample of consumers but 4,533 evaluations (Table 9 Appendix A). This is a convenience sampling when cases or individuals are the evaluations. Some cell frequencies are low but no empty cells exist.

Table 4
Stability performance of principal axes of data set 2
(Escofier y Pagès, 1992)

Principal	Before Procrustes				After Procrustes				Stability		
Axes	Mean	Median	P_5	P_{95}	Mean	Median	P_5	P_{95}	S	AU-S	U
v_1	21.7	0.8	0.3	**179.3**	0.5	0.5	0.3	0.9		x	
v_2	7.9	1.2	0.6	2.3	0.7	0.7	0.3	1.3	x		
v_3	1.0	0.9	0.5	1.6	0.7	0.7	0.4	1.2	x		
v_4	35.3	3.4	1.4	**177.7**	2.1	2.0	0.9	3.4		x	
v_5	70.0	11.1	2.7	**175.8**	2.7	2.6	1.3	4.1		x	
v_6	86.5	17.9	3.9	**175.8**	3.3	3.2	1.4	5.7		x	
v_7	98.0	172.3	2.7	**177.4**	3.3	3.3	1.6	5.3		x	
u_1	21.7	0.8	0.3	**179.3**	0.5	0.4	0.2	0.7		x	
u_2	7.8	1.1	0.5	2.2	0.6	0.6	0.3	1.1	x		
u_3	0.9	0.9	0.4	1.5	0.6	0.6	0.3	1.0	x		
u_4	35.6	3.9	1.5	**177.4**	2.4	2.2	0.9	4.4		x	
u_5	69.9	10.9	2.8	**175.8**	2.2	2.1	0.9	3.7		x	
u_6	86.5	18.9	3.6	**176.1**	2.6	2.5	1.1	4.3		x	
u_7	98.1	172.6	2.5	**177.8**	2.2	2.0	0.8	4.2		x	

Figures in bold point out that instability holds according to the criterion suggested.

Source: Own elaboration.

Table 4b
Comparison between real and randomly permuted data, Bootstrap aP. data set 2
(Escofier y Pagès, 1992)

Principal Axes	Randomly permuted data				Real data			
	Mean	Median	P_5	P_{95}	Mean	Median	P_5	P_{95}
v_1	74,0	74,5	49,7	97,3	0.5	0.5	0.3	0.9
v_2	73,9	74,6	47,8	98,3	0.7	0.7	0.3	1.3
v_3	80,9	82,5	61,2	94,3	0.7	0.7	0.4	1.2
v_4	59,3	57,5	30,5	94,7	2.1	2.0	0.9	3.4
v_5	61,9	59,6	33,6	99,1	2.7	2.6	1.3	4.1
v_6	65,6	61,9	39,2	104,9	3.3	3.2	1.4	5.7
v_7	68,3	67,6	34,8	104,9	3.3	3.3	1.6	5.3
u_1	62,6	63,0	38,8	85,7	0.5	0.4	0.2	0.7
u_2	57,8	57,2	31,1	86,0	0.6	0.6	0.3	1.1
u_3	60,8	62,0	39,3	78,6	0.6	0.6	0.3	1.0
u_4	60,2	56,1	33,9	99,7	2.4	2.2	0.9	4.4
u_5	56,5	53,0	27,0	95,6	2.2	2.1	0.9	3.7
u_6	56,7	54,8	29,5	88,1	2.6	2.5	1.1	4.3
u_7	55,0	54,7	28,2	85,4	2.2	2.0	0.8	4.2

Figures in bold italics point out real and random 90% CI which are overlapped.

Source: Own elaboration.

Table 5 shows that according to a direct analysis, all the principal axes should be considered rather unstable. However, the rotation points out stability in the first principal axis and some instability in the second one whereas the rest has a striking and increasing instability. In fact, the similarity of cell frequencies results in a weak structure of associations between row and column categories. Therefore, just associations between brands of milk and milk attributes unveiled by the first factorial plan should be considered valid although with some caution regarding the second axis using the 90% CI. Only the first axis is stable using $P_{2.5}$, $P_{97.5}$ and Pearson test statistic, α=5%.

Table 5
Stability performance of principal axes of data set 3
(Abascal y Grande, 2005)

Principal	Before Procrustes				After Procrustes				Stability		
Axes	Mean	Median	P_5	P_{95}	Mean	Median	P_5	P_{95}	S	AU-S	U
v_1	101.3	161.9	10.7	**170.1**	13.2	13.0	8.5	18.3		x	
v_2	90.9	97.5	27.3	**152.2**	28.6	28.0	18.0	**40.8**			x
v_3	89.7	89.7	53.2	**128.7**	41.0	40.1	26.7	**58.9**			x
v_4	88.1	88.3	52.1	**122.2**	43.5	42.2	26.1	**63.6**			x
v_5	90.3	90.0	58.3	**121.5**	46.9	45.9	27.8	**67.8**			x
v_6	89.4	89.3	58.5	**121.1**	48.8	47.8	29.8	**71.1**			x
v_7	89.8	89.7	59.7	**119.9**	54.5	52.5	31.4	**84.8**			x
v_8	89.0	89.4	57.0	**119.6**	57.8	56.3	32.7	**87.2**			x
v_9	90.4	90.3	56.9	**121.8**	64.9	64.1	34.6	**100.0**			x
v_{10}	89.8	90.0	57.5	**122.1**	75.0	72.5	39.5	**115.7**			x
u_1	101.1	162.9	9.4	**170.5**	12.0	11.8	7.7	17.1		x	
u_2	91.0	98.8	24.7	**155.5**	25.5	25.3	14.8	**36.9**			x
u_3	89.8	89.6	48.3	**130.3**	34.0	33.4	19.3	**49.8**			x
u_4	89.4	89.7	52.8	**126.7**	35.9	35.1	21.0	**53.8**			x
u_5	90.4	89.9	53.8	**127.0**	36.7	35.8	20.6	**55.2**			x
u_6	89.6	89.3	54.3	**125.1**	37.7	36.9	21.8	**58.0**			x
u_7	90.0	89.3	53.5	**125.3**	39.8	38.7	20.4	**60.8**			x
u_8	88.8	88.4	52.2	**125.5**	38.5	37.9	19.8	**60.5**			x
u_9	90.1	90.4	51.2	**127.0**	37.5	36.3	19.7	**59.9**			x
u_{10}	89.9	90.9	47.5	**128.4**	36.6	35.4	20.2	**56.2**			x

Figures in bold point out that instability holds according to the criterion suggested.

Source: Own elaboration.

Table 5b
Comparison between real and randomly permuted data, Bootstrap aP. Data set 3
(Abascal y Grande, 2005)

Principal	Randomly permuted data				Real data			
Axes	Mean	Median	P_5	P_{95}	Mean	Median	P_5	P_{95}
v_1	66,0	65,7	45,5	87,1	13.2	13.0	8.5	18.3
v_2	64,8	64,3	43,4	87,3	28.6	28.0	18.0	40.8
v_3	60,5	60,2	***38,1***	***86,2***	41.0	40.1	***26.7***	***58.9***
v_4	67,4	66,8	***44,1***	***90,8***	43.5	42.2	***26.1***	***63.6***
v_5	68,2	68,0	***45,5***	***90,5***	46.9	45.9	***27.8***	***67.8***
v_6	71,2	71,3	***49,6***	***93,0***	48.8	47.8	***29.8***	***71.1***
v_7	70,5	70,2	***48,3***	***95,7***	54.5	52.5	***31.4***	***84.8***
v_8	68,4	67,2	***45,7***	***94,7***	57.8	56.3	***32.7***	***87.2***
v_9	66,9	65,3	***40,6***	***98,3***	64.9	64.1	***34.6***	***100.0***
v_{10}	74,1	72,4	***46,5***	***108,6***	75.0	72.5	***39.5***	***115.7***
u_1	51,4	50,6	32,8	72,8	12.0	11.8	7.7	17.1
u_2	51,9	51,4	***32,4***	***73,5***	25.5	25.3	***14.8***	***36.9***
u_3	54,3	54,2	***33,6***	***76,3***	34.0	33.4	***19.3***	***49.8***
u_4	49,6	49,1	***31,4***	***70,8***	35.9	35.1	***21.0***	***53.8***
u_5	49,1	48,9	***29,5***	***71,3***	36.7	35.8	***20.6***	***55.2***
u_6	46,9	46,1	***29,4***	***66,4***	37.7	36.9	***21.8***	***58.0***
u_7	48,1	46,9	***30,1***	***69,5***	39.8	38.7	***20.4***	***60.8***
u_8	50,2	49,7	***30,5***	***72,3***	38.5	37.9	***19.8***	***60.5***
u_9	51,9	52,1	***32,2***	***72,7***	37.5	36.3	***19.7***	***59.9***
u_{10}	51,1	50,6	***34,0***	***69,9***	36.6	35.4	***20.2***	***56.2***

Figures in bold italics point out real and random 90% CI which are overlapped.

Source: Own elaboration.

4.1.4. Data set 4. Valencia (2006)

It is a contingency table from the national survey on Labour Force in Spain (EPA) for the second quarter of 1999. The 68,959 frequencies represent working population on 18 Spanish regions (rows) and 16 economic sectors (columns) (Table 10 Appendix A). High cell frequencies together with low cell frequencies exist and also a small number of empty cells.

Table 6 shows that the principal axes seem to be highly unstable. Having applied the adjustment proposed, just the last four axes should be labelled as unstable, so variability initially observed in axes 1 to 11 is to some extent apparent. Anyway, the degree of stability is considerably lower in axes 9 to 11, with potential angles around 30º (P_{95}) or even higher in the eleventh axis. The first 10 axes are stable using a 95% confidence interval and Pearson test statistic, α=5%.

Table 6
Stability performance of principal axes of data set 4
(Valencia, 2006)

Principal	Before Procrustes				After Procrustes				Stability		
Axes	Mean	Median	P_5	P_{95}	Mean	Median	P_5	P_{95}	S	AU-S	U
v_1	85.4	13.0	4.5	**175.4**	4.3	4.3	3.1	5.7		x	
v_2	86.3	14.0	5.7	**174.1**	5.1	5.0	3.6	6.5		x	
v_3	79.5	40.9	9.5	**169.7**	6.8	6.7	4.6	9.3		x	
v_4	94.6	131.6	11.1	**169.8**	7.4	7.3	5.3	9.9		x	
v_5	97.0	121.9	13.3	**168.2**	9.8	9.7	6.3	14.1		x	
v_6	89.8	85.1	14.6	**165.9**	10.9	10.8	7.7	14.6		x	
v_7	91.9	128.2	16.6	**162.6**	14.2	14.1	9.7	19.7		x	
v_8	95.4	115.7	21.5	**158.5**	16.7	16.4	10.9	23.4		x	
v_9	94.5	108.8	24.7	**154.7**	18.9	18.5	12.3	26.6		x	
v_{10}	91.2	92.3	32.1	**149.1**	23.5	23.2	14.3	34.7		x	
v_{11}	89.4	88.7	37.2	**142.5**	27.7	27.1	16.9	40.5		x	
v_{12}	92.0	92.9	39.2	**140.5**	31.6	30.7	18.8	**47.8**			x
v_{13}	89.1	89.4	42.8	**134.9**	41.4	39.4	21.6	**70.7**			x
v_{14}	90.5	90.4	47.5	**133.9**	46.9	44.6	24.5	**78.1**			x
v_{15}	91.1	90.1	47.8	**133.6**	69.8	67.4	28.5	**123.2**			x
u_1	85.4	13.2	4.3	**175.6**	4.1	4.1	2.9	5.6		x	
u_2	86.3	14.2	5.6	**174.5**	4.9	4.9	3.4	6.7		x	
u_3	79.5	41.1	9.6	**169.6**	6.9	6.8	4.7	9.3		x	
u_4	94.6	131.3	11.3	**170.0**	7.3	7.3	4.9	10.1		x	
u_5	97.1	120.5	13.0	**168.4**	9.1	8.9	6.0	12.7		x	
u_6	89.7	84.2	13.8	**166.3**	10.2	10.0	6.8	14.1		x	
u_7	92.0	129.0	15.9	**163.8**	12.8	12.6	8.2	17.7		x	
u_8	95.3	115.1	21.0	**159.2**	14.5	14.2	9.2	20.4		x	
u_9	94.6	109.9	23.0	**155.9**	16.7	16.3	10.3	24.4		x	
u_{10}	90.9	94.5	31.2	**150.0**	20.8	20.2	12.9	30.6		x	
u_{11}	89.3	89.4	35.1	**146.0**	21.5	21.0	11.8	33.5		x	
u_{12}	92.1	92.5	36.1	**143.5**	24.3	23.6	12.7	**38.9**			x
u_{13}	89.3	90.2	36.1	**142.3**	29.0	27.0	13.8	**48.9**			x
u_{14}	90.4	91.0	36.2	**143.4**	29.4	27.7	13.8	**54.8**			x
u_{15}	89.7	91.0	30.7	**149.3**	22.1	20.7	10.4	**39.1**			x

Figures in bold point out that instability holds according to the criterion suggested.

Source: Own elaboration.

Table 6b

Comparison between real and randomly permuted data, Bootstrap aP. Data set 4
(Valencia, 2006)

Principal	**Randomly permuted data**				**Real data**			
Axes	**Mean**	**Median**	P_5	P_{95}	**Mean**	**Median**	P_5	P_{95}
v_1	67,6	67,8	53,8	80,2	4.3	4.3	3.1	5.7
v_2	64,8	65,3	49,1	79,4	5.1	5.0	3.6	6.5
v_3	74,4	74,5	56,0	91,8	6.8	6.7	4.6	9.3
v_4	67,0	66,6	49,9	84,4	7.4	7.3	5.3	9.9
v_5	73,7	74,4	49,5	94,3	9.8	9.7	6.3	14.1
v_6	54,3	53,3	39,1	73,1	10.9	10.8	7.7	14.6
v_7	59,4	59,2	44,6	75,4	14.2	14.1	9.7	19.7
v_8	62,6	62,0	46,6	79,9	16.7	16.4	10.9	23.4
v_9	61,8	61,5	44,2	79,2	18.9	18.5	12.3	26.6
v_{10}	69,3	69,6	49,3	87,9	23.5	23.2	14.3	34.7
v_{11}	62,4	61,4	42,8	84,1	27.7	27.1	16.9	40.5
v_{12}	66,3	65,2	***47,5***	***87,3***	31.6	30.7	***18.8***	***47.8***
v_{13}	72,0	71,8	***52,9***	***92,2***	41.4	39.4	***21.6***	***70.7***
v_{14}	71,8	71,0	***52,0***	***94,5***	46.9	44.6	***24.5***	***78.1***
v_{15}	75,0	74,1	***51,9***	***101,1***	69.8	67.4	***28.5***	***123.2***
u_1	70,6	70,8	57,3	82,8	4.1	4.1	2.9	5.6
u_2	68,8	67,9	55,8	83,2	4.9	4.9	3.4	6.7
u_3	50,0	48,4	37,1	67,6	6.9	6.8	4.7	9.3
u_4	60,3	59,5	45,2	77,0	7.3	7.3	4.9	10.1
u_5	40,6	38,3	24,4	64,3	9.1	8.9	6.0	12.7
u_6	70,1	70,2	52,5	88,8	10.2	10.0	6.8	14.1
u_7	68,8	68,9	53,4	84,2	12.8	12.6	8.2	17.7
u_8	62,2	62,3	45,8	79,6	14.5	14.2	9.2	20.4
u_9	65,0	65,6	45,2	82,9	16.7	16.3	10.3	24.4
u_{10}	52,8	52,6	36,4	71,0	20.8	20.2	12.9	30.6
u_{11}	57,2	56,3	39,9	77,8	21.5	21.0	11.8	33.5
u_{12}	55,6	55,3	***38,5***	***73,9***	24.3	23.6	***12.7***	***38.9***
u_{13}	55,1	54,6	***38,3***	***73,1***	29.0	27.0	***13.8***	***48.9***
u_{14}	54,1	54,0	***38,9***	***70,7***	29.4	27.7	***13.8***	***54.8***
u_{15}	53,2	52,9	***38,7***	***69,3***	22.1	20.7	***10.4***	***39.1***

Figures in bold italics point out real and random 90% CI which are overlapped.

Source: Own elaboration.

4.2. Some remarks about stability results

Some authors have indicated that instability in principal axes methods is rather linked to the presence of low cell frequencies. This feature may be crucial for the study of the stability of individual variables or categories (Markus, 1994; Linting, Meulman, Groenen y Van der Koojj, 2007). Regarding CA, as far as we are concerned, it seems that the mere existence of low cell frequencies does not necessarily mean instability of the principal axes. Anyway, the combined effect of the magnitude of cell fre-quencies, the similarity among them and their location within the data matrix, definitely influence the stability.

From the empirical study carried out, other insights could be mentioned:

- When cell frequencies are not extremely low but there are small differences among them, as in data set 3 (Abascal y Grande, 2005), instability may come up even along the first axes.
- A small eigenvalue (inertia) of an axis does not necessarily mean an unstable axis if a suitable procedure to remove apparent variability (such as Procrustes-like rotations) has been applied. For instance, this takes place in the last axes of data set 2, which show a high degree of stability.
- A rather low average angle for a principal axis does not automatically imply stability of this axis (as the third axis in data set 1). It seems more convenient to regard the distribution of the computed angles (percentiles) and also to avoid comparisons with rigid thresholds which do not take into account the features of the data matrix.
- To appoint an axis stable or unstable requires the use of thresholds. Although these thresholds are set up in a non-arbitrary way, stability is still a matter of degree and therefore the consultation of numerical measurements (e.g. angles) seems necessary.
- There are many applied economics studies whose methodologies are based on bootstrap resampling (Del Hoyo, Llorente y Rivero, 2011; Miguel y Olave, 2002). When, as in the example of data set 4 (Valencia, 2006), categorical variables together with simple or multiple correspondence analysis are employed (as in Rodriguez, 2005), a strategy to determine the stability of patterns seems quite useful.

5. DISCUSSION

Assessing the stability of patterns is crucial in principal axes methods to avoid the rejection of useful information and the interpretation of meaningless dimensions. The main objective of this paper has been to suggest a methodology which determines the degree of stability of principal axes in CA, bearing in mind that an interpretable but unstable principal axis should not be kept in mind as indicator of the latent or underlying structure of the data.

The Bootstrap methodology seems to be a useful tool for this purpose but it has to be adapted in some ways. First of all, in the case of CA, a specific problem arises from the different metrics of the replicated tables, which lead us to propose to perform CA with imposed metrics (the metrics from the original data set) on the bootstrapped tables. Thus, the original and replicated clouds of point categories are located in the same metric space. Secondly, the direct comparison of some statistics such as principal axes issued from TB may lead to pessimistic evaluations of stability. So, Procrustes rotations of the bootstrapped configurations are applied to remove the apparent variability of the replicated configurations. In our case, the bootstrapped principal axes are reconstructed from rotated configurations to obtain the measure of stability of each one.

Finally, since the stability is a matter of degree, any approach to qualify an axis as stable or unstable may be arbitrary. In our proposal, the comparison of a measure (e.g. angles) with the same measure obtained from randomly permuted data is considered a more suitable approach than any fixed threshold for the average of angles.

Although regarding the matter of selection of axes, our methodology offers quite similar results to Pearson test statistic, it can be applied for non random samples and complex designs.

As far as we know, the purpose of both obtaining a bounded measure of the stability of every principal axis through a Bootstrap procedure and labelling each axis as stable or unstable by means of a non-arbitrary threshold had not been tackled so far in the literature on bootstrapping CA.

Future research could be aimed at deepening in our findings by means of simulations studies with several data sets. These would allow analysing the individual effects on stability derived from low frequencies, empty cells, similar magnitude of cell frequencies, low total inertia and a high difference between the number of row and column categories.

REFERENCES

ABASCAL, E. y GRANDE, I. (2005). *Análisis de encuestas.* Madrid: ESIC.

ABDI, H. (2007). “RV coefficient and congruence coefficient”. En Salkind N.J. (ed.): *Encyclopedia of Measurement and Statistics* (pp. 849-853). Thousand Oaks (CA): Sage

AL-IBRAHIM, A.H. y AL-KANDARI, N.M. (2008). “Stability of principal components”. *Computational Statistics, 23 (1), pp.153-171.*

ÁLVAREZ, R.; VALENCIA, O. y BÉCUE-BERTAUT, M. (2010). "Assessing the Stability of Supplementary Elements on Principal Axes Maps Through Bootstrap Resampling. Contribution to Interpretation in Textual Analysis". En Skiadas,C.H.(ed.): *Advances in Data Analysis. Statistics for Industry and Technology* (pp. 3-11). Boston: Springer/ Birkhäuser.

BENASSENI, J. (1993). "Perturbational aspects in correspondance analysis". *Computional Statistics & Data Analysis,* 15, pp. 393-410.

BENZECRI, J.P. (1973). *L'analyse des données. Tome 2: L´Analyse des Correspondances.* Paris: Dunod.

BERAN, R.B. y SRIVASTAVA, M.S. (1985). "Bootstrap tests and confidence regions for functions of a covariance matrix". *Annals of Statistics,* 13, pp. 95–115.

CHATEAU F. y LEBART, L. (1996). "Assessing sample variability in the visualization techniques related to principal component analysis: bootstrap and alternative simulation methods". En Prats, A. (ed.): *XII Symposium on Computational Statistics COMPSTAT96* (pp. 205-210). Heidelberg:,Physica Verlag.

DAUDIN, J.; BUBY,C. y TRECOURT, P. (1988). "Stability of principal component analysis studied by the bootstrap method". *Statistics,* 19 (2), pp. 241-258.

DEL HOYO, J.; LLORENTE, G. y RIVERO, C. (2011). "Consumo de electricidad y producto interior bruto. Relación dinámica y estabilidad". *Estudios de Economía Aplicada,* 29(2), pp. 473-492.

DIACONIS, P. y EFRON, B. (1983). "Computer-intensive methods in statistics". *Scientific American,* 248, pp. 96-108.

EFRON, B. (1979). "Bootstrap methods: Another look at the Jackknife". *Annals of Statistics,* 7, pp. 1-26.

EFRON, B. y TIBSHIRANI, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

ESCOFIER, B. (1984). *Analyse factorielle en référence à un modèle. Application à l'analyse de tableaux d'échange.* En RR-0337, INRIA00076220, Rennes, 1984.

ESCOFIER, B. (1987). *Notice d'utilisation du programme d'analyse factorielle en référence à un modèle.* En RR-0616, INRIA 00075938, Rennes, 1987.

ESCOFIER, B. (2003). *Analyse des correspondances: Recherches au cœur de l'analyse des données.* Rennes: Presses universitaires de Rennes.

ESCOFIER, B. y LE ROUX, B. (1972). "Etude de trois problèmes de stabilité en analyse factorielle". En *Publications de l'Institut de Statistique de Paris,* 21 Fasc.(3-4) (parution 1973).

ESCOFIER, B. y PAGÈS, J. (1983). "Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractérisation de vins rouges du Val de Loire". *Révue de Statistique Appliquée,* 31(2), pp. 4-59.

ESCOFIER, B. y PAGÈS, J. (1992). *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación.* Bilbao: Servicio Editorial Universidad del País Vasco.

ESCOFIER, B. y PAGÈS, J. (1998). *Analyse factorielles simples et multiples* Paris: Dunod.

FISHER, R. A. (1940). "The precision of discriminant functions" en Annals of Eugenics, 10, pp.422-429. Permanent link to this item: http://hdl.handle.net/2440/15240

GIFI, A. (1981). *Nonlinear multivariate analysis.* Chichester: Wiley.

GILULA, Z. y HABERMAN, S.J. (1986). "Canonical analysis of contingency tables by maximum likelihood". *Journal of the American statistical association*, 81, pp. 780-788.

GOWER, J.C. (1971). "Statistical methods for comparing different multivariate analyses of the same data". En Hodson, J.R.; Kendall, D.G. y Tautu, P. (eds.): *Mathematics in the Archaeological and Historical Sciences* (pp. 138-149). Edinburgh: University Press.

GOWER, J.C. (1975). "Procrustes Analysis". *Psichometrika,* 40(1), pp. 33-51.

GOWER J.C. y DIKSTERHUIS, G.B. (2004). *Procrustes problems.* Oxford, New York: Oxford University Press.

GREENACRE, M. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.

HOLMES, S. (2008*). "Multivariate data analysis: The French way".* En Nolan, D. y Speed, T. (eds.): *Probability and Statistics: Essays in Honor of David A. Freedman* (pp.219-233). Beachwood, Ohio, USA: Institute of Mathematical Statistics.

JACKSON, D.A. (1993). "Stopping rules in principal component analysis. A comparison of heuristical and statistical approaches". *Ecology,* 74 (8), pp. 2204-2214.

KRZANOWSKI, W.J. (1999). *Principles of multivariate analysis. A user's perspective.* Belfast: Oxford Statistical Science Series.

LAMBERT, Z.V.; WILDT, A.R. y DURAND, R.M. (1990). "Assessing sampling variation relative to number-factors criteria". *Educational and Psychological Measurement,* 50, pp. 33-48.

LEBART, L. (1976). "The significance of eigenvalues issued form correspondence analysis". *Proceedings in Computational Statistics,.* COMPSTAT (pp. 38-45). Vienna: Physica Verlag.

LEBART, L. (2004a). "Validité des visualisations de données textuelles. Le poids des mots". En Purnelle, G.; Fairon, C. y Dister, A. (eds.): *Proceedings of the 7th International Conference on Textual Data Statistical Analysis, JADT 2004* (pp. 708-715). Belgique: Presses Universitaires de Louvain.

LEBART, L. (2004b). “Validation techniques in text mining (with applications to the processing of open-ended questions)”. En Sirmakessis. S. (ed): *Text mining and its applications* (pp. 169-178). Berlin, Heidelberg: Springer.

LEBART, L. (2006). “Validation techniques in multiple correspondence analysis”. En Greenacre, M.J. y Blasius, J. (eds.): *Multiple Correspondence Analysis and Related Methods* (pp.179-195). London: Chapman and Hall.

LEBART, L.; MORINEAU, A. y PIRON, M. (2000). *Statistique exploratoire multidimensionnelle (3ª ed.).* Paris: Dunod.

LEBART, L.; MORINEAU, A. y WARWICK, K. (1984). *Multivariate descriptive statistical analysis.* New York: Wiley.

LINTING, M.; MEULMAN, J.J.; GROENEN, P.J.F. y VAN DER KOOJJ, A.J. (2007). “Stability of Nonlinear Principal Components Analysis: An empirical study using the balanced bootstrap”. *Psychological Methods,* 12(3), pp. 359-379.

MARKUS, M. (1994). *Bootstrap confidence regions in nonlinear multivariate analysis.* Leiden: DSWO Press.

MIGUEL, J.A. y OLAVE, P. (2002). “Avances recientes en métodos Bootstrap para procesos ARCH. Una aplicación en el mercado español de valores”. *Estudios de Economía Aplicada,* 20(2), pp. 487-498.

MILAN, L. y WHITTAKER, J. (1995). “Application of the parametric bootstrap to models that incorporate a singular value decomposition”. *Applied Statistics,* 44(1), pp. 31-49.

O’NEILL, M.E. (1978). “Asymptotic distributions of the canonical correlations from contingency tables”. *Australian Journal of Statistics,* 20(1), pp. 75-82.

PERES-NETO, P.R.; JACKSON, D,A. y SOMERS, K.M. (2005). “How many principal components? Stopping rules for determining the number of non-trivial axes revisited”. *Computational Statistics and Data Analysis,* 49(4), pp. 974-997.

REICZIGEL, J. (1996). “Bootstrap tests in correspondence analysis”. *Applied Stochastic Models and Data Analysis,* 12, pp. 107-117.

RINGROSE, T.J. (1992). “Bootstrapping and correspondence analysis in Archaeology”. *Journal of Archaeological Science,* 19, pp. 615-629.

RODRIGUEZ, O.M. (2005). “El crédito commercial en las pymes canarias desde una perspectiva multivariante”. *Estudios de Economía Aplicada*, 23 (3), pp. 773-816.

SCHÖNEMANN, P.H. y CARROL, R.M. (1970). “Fitting one matrix to another under choice of a central dilation and a rigid motion”. *Psychometrika,* 35(2), pp. 245-255.

STAUFFER, D.F.; GARTON, E.O. y STEINHORST, R.K. (1985). “A comparison of principal components from real and random data”. *Ecology,* 66(6), pp. 1693-1698.

TAN, Q.; BRUSGAARD, K.; KRUSE, T.A.; OAKELEY, E.; HEMMINGS, B.; BECK-NIELSEN, H.; HANSEN, L,. y GASTER. M. (2004) “Correspondence analysis of microarray time-course data in case-control design”. *Journal of Biomedical Informatics,* 37, pp. 358-365.

TEN BERGE, J.M.F. (2006). “The rigid orthogonal procrustes rotation problem”*. Psychometrika,* 71(1), pp. 201-205.

TEN BERGE, J.M.F. y BEKKER, P.A. (1993). “The isotropic scaling problem in Procrustes Analys”. *Computational. Statististics & Data Analysis,* 16, pp. 201-204.

TENENHAUS, M. y YOUNG, F.W. (1985). “An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data”. *Psychometrika*, 50(1), pp. 91-119.

TOCHER, F. (1908). “Pigmentation survey of school children in Scotland”. *Biometrika*, 6, pp. 30-235.

VALENCIA, O. (2006). *Estudio de la estabilidad de los métodos factoriales mediante procedimientos de remuestreo: Aplicación al Análisis de Correspondencias de tablas léxicas.* León: Universidad de León, Secretariado de Publicaciones.

Appendix A: Data sets

Table 7
Data set 1 Hair colour/ Eye colour

	Light	Blue	Medium	Dark
Fair	688	326	343	98
Red	116	38	84	48
Medium	584	241	909	403
Dark	188	110	412	681
Black	4	3	26	85

Source: Fisher (1940).

Table 8
Data set 2. Type of job/Educational level, Females

	Without studies	BEPC	BEP/ CAP	BAC general	BAC technical	DEUG/ ENT	DUT/BTS/ health	Higher
Farmers	5089	1212	1166	0	0	0	0	0
Engineers	0	0	0	316	0	0	304	1033
Technicians	281	0	320	320	283	0	683	0
Qualified workers	7470	1859	4017	1752	657	0	285	0
Not qualified workers	29997	4334	4538	1882	0	0	0	0
Senior management	0	0	0	2236	595	911	569	6788
Middle management	1577	1806	4549	17063	875	4152	15731	3991
Qualified employees	21616	19915	32452	16137	5865	1256	3332	1286
Not qualified employees	19849	7325	6484	5111	898	294	635	0

Source: Escofier y Pagès (1992).

Table 9
Data set 3. Brands of milk/Milk attributes

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
PASCUAL	29	76	41	53	71	60	70	62	65	58	61
PULEVA	25	54	34	43	47	49	54	52	47	44	39
ASTURIANA	41	57	44	56	62	48	56	46	46	43	38
KAIKU	40	61	42	37	50	41	51	45	41	36	38
RAM	25	33	26	25	35	29	41	34	34	33	26
CELTA	33	39	35	34	36	33	37	32	34	31	36
PRESIDENT	15	26	24	26	28	24	29	26	26	26	20
BONMILK	49	40	36	13	39	11	41	28	29	32	26
DIA	51	28	29	6	34	8	37	24	29	24	21
CAPRABO	31	24	17	4	24	9	27	25	23	22	21
CARREFOUR	33	23	14	7	23	11	29	22	23	22	21
LAUKI	19	19	13	2	17	2	17	17	18	18	15
RENY-PICOT	16	16	11	2	16	2	16	16	16	16	15

Source: Abascal y Grande (2005).

Table 10
Data set 4. Spanish regions/Economic sectors

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
R1	1121	32	36	1320	73	1198	1854	672	515	245	570	747	745	582	349	344
R2	244	0	22	706	31	271	525	165	159	98	193	234	188	235	105	62
R3	161	8	65	276	11	185	295	102	80	33	88	128	115	91	66	46
R4	35	5	4	212	19	235	353	350	161	43	113	122	68	97	82	32
R5	219	14	6	252	31	484	834	503	248	55	249	355	254	172	170	93
R6	91	17	10	304	12	174	221	87	73	22	61	75	64	67	63	34
R7	673	0	72	1166	45	728	1059	387	356	165	352	609	452	447	238	165
R8	538	1	14	957	27	694	814	277	239	120	183	362	312	290	160	117
R9	370	13	23	1938	49	851	1296	500	439	200	612	335	459	451	353	240
R10	292	6	12	1507	38	609	1156	323	292	140	364	272	300	290	225	182
R11	341	0	32	194	29	349	452	163	85	41	105	209	162	166	94	58
R12	661	147	37	793	18	512	741	221	228	105	203	314	256	236	157	117
R13	36	0	11	637	34	350	568	204	328	195	463	388	235	221	184	133
R14	246	6	5	342	9	219	442	82	90	44	91	106	119	112	44	46
R15	98	0	3	454	7	135	197	71	73	41	97	87	111	117	55	37
R16	76	12	6	1139	21	351	544	217	208	91	275	184	245	229	150	157
R17	88	0	0	312	3	88	144	44	21	30	54	50	41	55	22	11
R18	1	1	4	18	5	23	93	30	33	8	26	166	40	23	24	5

Source: Valencia (2006).

Appendix B: Comparison between results without and with imposed metrics, both computed by means of Bootstrap after Procrustes rotation

Tabla 11
Bootstrap without and with imposed metrics

Data Sets	Principal Axes	Without imposed metrics				With imposed metrics			
		Mean	Median	P_5	P_{95}	Mean	Median	P_5	P_{95}
Data set 1	v_1	2,2	2,1	0,84	4,03	3.1	2.9	1.2	5.5
	v_2	4,9	4,6	1,52	9,48	5.5	5.1	1.9	10.1
	v_3	18,5	11,3	3,00	56,95	17.5	11.6	2.9	53.8
	u_1	1,6	1,5	0,47	3,21	2.2	2.1	0.6	4.5
	u_2	3,4	2,9	0,73	8,07	4.1	3.6	0.9	8.6
	u_3	3,6	3,2	0,84	8,23	4.5	3.4	0.9	11.4
Data set 2	v_1	0.3	0.3	0.1	0.5	0.5	0.5	0.3	0.9
	v_2	0.5	0.5	0.2	1.0	0.7	0.7	0.3	1.3
	v_3	0.6	0.6	0.3	1.0	0.7	0.7	0.4	1.2
	v_4	1.7	1.6	0.8	2.9	2.1	2.0	0.9	3.4
	v_5	2.2	2.1	0.9	4.0	2.7	2.6	1.3	4.1
	v_6	2.7	2.6	1.1	4.8	3.3	3.2	1.4	5.7
	v_7	3.0	2.9	1.2	5.2	3.3	3.3	1.6	5.3
	u_1	0.3	0.3	0.1	0.4	0.5	0.4	0.2	0.7
	u_2	0.5	0.5	0.2	0.9	0.6	0.6	0.3	1.1
	u_3	0.5	0.5	0.3	0.9	0.6	0.6	0.3	1.0
	u_4	1.1	1.0	0.5	2.0	2.4	2.2	0.9	4.4
	u_5	1.2	1.2	0.6	1.9	2.2	2.1	0.9	3.7
	u_6	1.3	1.3	0.6	2.3	2.6	2.5	1.1	4.3
	u_7	1.1	1.0	0.5	2.0	2.2	2.0	0.8	4.2
Data set 3	v_1	11,5	11,4	7,6	16,0	13.2	13.0	8.5	18.3
	v_2	24,9	24,3	15,7	36,7	28.6	28.0	18.0	40.8
	v_3	39,7	38,6	23,3	60,4	41.0	40.1	26.7	58.9
	v_4	41,1	39,3	23,4	62,1	43.5	42.2	26.1	63.6
	v_5	45,1	44,0	25,8	69,3	46.9	45.9	27.8	67.8
	v_6	45,3	44,2	26,6	67,9	48.8	47.8	29.8	71.1
	v_7	55,7	53,8	30,0	86,5	54.5	52.5	31.4	84.8
	v_8	63,6	62,4	35,9	97,0	57.8	56.3	32.7	87.2
	v_9	75,1	75,0	40,7	113,1	64.9	64.1	34.6	100.0
	v_{10}	83,2	82,8	48,4	121,6	75.0	72.5	39.5	115.7
	u_1	10,0	9,9	6,3	14,4	12.0	11.8	7.7	17.1
	u_2	19,8	19,4	11,5	29,3	25.5	25.3	14.8	36.9
	u_3	23,3	22,8	12,9	35,2	34.0	33.4	19.3	49.8
	u_4	26,7	26,0	14,6	41,6	35.9	35.1	21.0	53.8
	u_5	27,5	26,6	15,6	43,1	36.7	35.8	20.6	55.2
	u_6	31,1	29,8	17,6	48,8	37.7	36.9	21.8	58.0
	u_7	32,8	31,0	18,2	52,5	39.8	38.7	20.4	60.8
	u_8	30,2	29,7	17,7	44,4	38.5	37.9	19.8	60.5
	u_9	33,1	32,4	19,5	48,7	37.5	36.3	19.7	59.9
	u_{10}	39,7	39,1	22,9	60,6	36.6	35.4	20.2	56.2

Tabla 11 (continue)
Bootstrap without and with imposed metrics

Data Sets	Principal Axes	Without imposed metrics				With imposed metrics			
		Mean	Median	P_5	P_{95}	Mean	Median	P_5	P_{95}
Data set 4	v_1	3.8	3.7	2.6	5.0	4.3	4.3	3.1	5.7
	v_2	4.2	4.1	2.9	5.4	5.1	5.0	3.6	6.5
	v_3	5.3	5.3	3.4	7.2	6.8	6.7	4.6	9.3
	v_4	5.8	5.8	4.0	7.9	7.4	7.3	5.3	9.9
	v_5	8.1	7.9	5.2	11.6	9.8	9.7	6.3	14.1
	v_6	9.2	9.0	6.2	12.7	10.9	10.8	7.7	14.6
	v_7	12.1	12.0	8.1	16.5	14.2	14.1	9.7	19.7
	v_8	13.8	13.6	8.7	20.0	16.7	16.4	10.9	23.4
	v_9	16.1	15.6	10.1	23.5	18.9	18.5	12.3	26.6
	v_{10}	19.8	19.2	11.2	30.4	23.5	23.2	14.3	34.7
	v_{11}	24.2	23.1	14.3	37.4	27.7	27.1	16.9	40.5
	v_{12}	27.2	26.1	15.5	43.2	31.6	30.7	18.8	47.8
	v_{13}	37.9	35.4	18.0	68.0	41.4	39.4	21.6	70.7
	v_{14}	46.1	43.5	21.8	80.2	46.9	44.6	24.5	78.1
	v_{15}	71.0	67.6	29.6	125.4	69.8	67.4	28.5	123.2
	u_1	3.3	3.3	2.2	4.4	4.1	4.1	2.9	5.6
	u_2	3.7	3.7	2.6	5.1	4.9	4.9	3.4	6.7
	u_3	5.2	5.1	3.4	7.0	6.9	6.8	4.7	9.3
	u_4	5.5	5.4	3.6	7.6	7.3	7.3	4.9	10.1
	u_5	7.5	7.3	4.8	11.0	9.1	8.9	6.0	12.7
	u_6	7.4	7.2	4.7	10.8	10.2	10.0	6.8	14.1
	u_7	9.3	9.1	5.8	13.6	12.8	12.6	8.2	17.7
	u_8	10.7	10.6	6.5	15.6	14.5	14.2	9.2	20.4
	u_9	12.5	12.3	7.4	19.3	16.7	16.3	10.3	24.4
	u_{10}	16.3	15.7	9.5	24.1	20.8	20.2	12.9	30.6
	u_{11}	17.0	16.3	9.7	27.6	21.5	21.0	11.8	33.5
	u_{12}	19.8	18.6	10.5	31.9	24.3	23.6	12.7	38.9
	u_{13}	25.4	23.2	12.0	46.0	29.0	27.0	13.8	48.9
	u_{14}	28.2	25.4	13.6	53.8	29.4	27.7	13.8	54.8
	u_{15}	38.2	31.1	17.1	97.4	22.1	20.7	10.4	39.1

Source: Own elaboration.

Appendix C: Eigenvalues and Eigenvalue significance Test of the data sets analysed

Table 12

Eigenvalues and eigenvalues significance tests

Data Sets	Eigenvalues and percentages of variance			Eigenvalue significance Test			
	Eigenvalue	Percent	Cum. Percent	H_k	Chi-squared	d.f	p-value
Data set 1	0,19924	86,6%	86,6%		1240,0	12	0,00000
	0,03009	13,1%	99,6%	1	166,7	6	0,00000
	0,00086	0,4%	100,0%	**2**	4,6	2	**0,09876**
Data set 2	0,52970	58,2%	58,2%	0	246759,5	56	0,00000
	0,23983	26,3%	84,5%	1	103160,4	42	0,00000
	0,12667	13,9%	98,5%	2	38144,5	30	0,00000
	0,00815	0,9%	99,4%	3	3803,9	20	0,00000
	0,00283	0,3%	99,7%	4	1594,4	12	0,00000
	0,00211	0,2%	99,9%	5	826,4	6	0,00000
	0,00093	0,1%	100,0%	6	253,4	2	0,00000
Data set 3	0,03986	79,6%	79,6%	0	227,1	120	0,00000
	0,00590	11,8%	91,3%	**1**	46,4	99	**1,00000**
	0,00146	2,9%	94,2%	2	19,7	80	1,00000
	0,00110	2,2%	96,4%	3	13,1	63	1,00000
	0,00076	1,5%	97,9%	4	8,1	48	1,00000
	0,00057	1,1%	99,1%	5	4,7	35	1,00000
	0,00024	0,5%	99,6%	6	2,1	24	1,00000
	0,00016	0,3%	99,9%	7	1,0	15	1,00000
	0,00005	0,1%	100,0%	8	0,3	8	0,99999
	0,00001	0,0%	100,0%	9	0,0	3	0,99763
Data set 4	0,03847	35,8%	35,8%	0	7406,6	255	0,00000
	0,02826	26,3%	62,1%	1	4753,7	224	0,00000
	0,01313	12,2%	74,4%	2	2804,8	195	0,00000
	0,01164	10,8%	85,2%	3	1899,3	168	0,00000
	0,00575	5,3%	90,5%	4	1096,7	143	0,00000
	0,00447	4,2%	94,7%	5	700,5	120	0,00000
	0,00218	2,0%	96,7%	6	392,1	99	0,00000
	0,00141	1,3%	98,0%	7	242,0	80	0,00000
	0,00096	0,9%	98,9%	8	144,5	63	0,00000
	0,00051	0,5%	99,4%	9	78,5	48	0,00354
	0,00033	0,3%	99,7%	**10**	43,5	35	**0,15228**
	0,00020	0,2%	99,9%	11	21,1	24	0,63481
	0,00006	0,1%	100,0%	12	7,3	15	0,94784
	0,00004	0,0%	100,0%	13	2,9	8	0,94165
	0,00001	0,0%	100,0%	14	0,4	3	0,94792

H_k : only k non-zero eigenvalues. Under the hypothesis H_k, the test statistic is asymptotically distributed like a chi-square with (p-k-1)(q-k-1) degrees of freedom. Figures in bold point out the k accepted number of significant eigenvalues (α=0,05).In data set 2, all eigenvalues are significant.

Source: Own elaboration.