Partial Bootstrap in CA : correction of the coordinates. Application to textual data

Ramón Álvarez[1], Mónica Bécue[2], Olga Valencia[3]

[1]Universidad de León. Departamento de Dirección y Economía de la Empresa. Facultad de CC. Económicas y Empresariales. Campus de Vegazana, s/n. 24071-León. España. Fax : 987 29 14 54. dderae@unileon.es

[2]Universidad Politécnica de Cataluña. Departamento de Estadística e Investigación Operativa. Campus Sur-Edificio U. C/ Pau Gargallo, 5. 08028-Barcelona. España. monica.becue@upc.es

[3]Universidad de Burgos. Departamento de Economía Aplicada. Facultad de CC. Económicas y Empresariales. Pl. Infanta Doña Elena, s/n. 09001 Burgos. España. Fax : 947 25 89 56. oval@ubu.es

Abstract

This paper studies the external stability of configurations issued from Correspondence Analysis of a lexical table by means of a Partial Bootstrap. As total inertia in the Bootstrapped tables is usually higher than total inertia in the original one, a correction of Bootstrap coordinates has been suggested. To build Bootstrap confidence regions for the position in factorial plans of every row and column, an algorithm based on a "peeling" of the convex hulls of scatter plots of bootstrapped points has been used.

Keywords : Correspondence Analysis, Partial Bootstrap, Lexical tables, open-ended questions.

1. Introduction

To include open and closed questions in questionnaires allows for combining the information provided by the free answers with the characteristics of the respondents. The aggregated lexical table, provided by cross-tabulating the words used in the free answers (rows) with the categories of respondents (columns) can be analysed by correspondence analysis (CA). The principal planes supplied by CA visualize the structure of the associations between groups of respondents and words.

However, it is necessary to verify the stability of the positions of both the words and the categories with respect to small perturbations in data. Partial Bootstrap (Lebart, 2004) provides a contribution to the stability study by drawing confidence regions for the coordinates of the rows (words) and of the columns (categories).

In this paper, we aim at pointing out that the bootstrap inertias are biased, with an expectancy greater than the original inertia. Therefore the obtained coordinates are miscalculated and we offer a way to correct these simulated coordinates. This correction leads to a "peeling" of the convex hulls and so to redrawn confidence regions.

In Section 2, we describe the type of stability that we consider and the facilities offered by the Bootstrap in the case of CA. Section 3 suggests a methodology for a correction of the Partial

Bootstrap coordinates. Section 4 presents an application of this methodology in textual data analysis.

2. Stability of the results of CA

According to the general notion (Gifi, 1990), CA results are stable if slight perturbations in the initial contingency table do not produce substantial alterations neither in the extracted principal axes nor in the configurations represented in the principal planes.

The analysis of the stability with respect to sampling fluctuations requires a real, abstract or simulated system of generation of samples. Discarded both a real drawing of samples, due to its usual non-viability in social sciences, and an abstract replication, because of its complexity in the case of CA, simulation of samples by means of a non-parametric bootstrap, i.e. without beforehand assumptions, is considered appropriate.

The utility of the Bootstrap to study the stability in the principal axes methods, particularly in CA, is underlined by Greenacre (1984), Meulman (1984), Ringrose (1992), Reiczigel (1996), Lebart, Morineau and Piron (2000), Tan *et al*. (2004) and Lebart (2004). Two main types of bootstrap resampling, with quite different objectives and degrees of complexity, can be distinguished : Total Bootstrap and Partial Bootstrap.

Total Bootstrap (TB) carries out a complete CA for each bootstrap table built. Coordinates, eigenvalues, eigenvectors, absolute contributions and relative contributions are obtained in each one of this CA analysis. All these elements allows for a study of the stability of the principal axes. However, complex problems arise from the direct comparison of statistics that belong to different subspaces (Álvarez, Bécue and Valencia, 2004).

Partial Bootstrap (PB) projects rows or columns of the bootstrapped tables on the subspace issued from CA of the original table, as illustrative or supplementary elements. This way, and contrary to the TB, the aim at the study of PB is only the analysis of the stability of the configurations visualized through the principal planes and presents less complexity than TB. That is one of the reasons why PB it is used more frequently in the case of the principal axes methods, specially concerning CA. However, the increase of the total inertia of the Bootstrap samples with respect to the original sample can cause an effect of "expansion" of the coordinates obtained by means of PB. This effect has to be studied and corrected.

3. Partial Bootstrap and Bootstrap confidence regions

The generation through Bootstrap of B replicated contingency tables is carried out by considering the relative frequencies of the cells of the original lexical table as the estimations of the probabilities corresponding to a multinomial model. Under this model, samples with the same size of observations can be reproduced.

Partial Bootstrap (Greenacre, 1984) uses the principal subspace issued from CA of the original table as a reference space for all the simulated tables. On this reference subspace, the rows and columns of every Bootstrap table are projected as supplementary elements. Therefore, B simulated coordinates are calculated for each element (row or column). When a principal plane is considered, a scatterplot of the Bootstrap points informs on the stability of every element in this plane.

3.1. Bias of the bootstrap coordinates

Starting from the empirical analysis of different types of contingency tables, it has been able to check that the mean of total inertia of the bootstrap tables may be sensibly bigger than the original inertia.

A lexical table is a special kind of contingency table more prone to present problems of instability due to its nature. Indeed, it is a sparse matrix, with a notable difference between the number of rows (textual units) and the number of columns (categories of a closed question or individual answer to a classical textual document), cells with very low frequency (even zero), and where the categories with very low marginal frequencies (especially in the textual units) are plentiful.

For this type of contingency tables the total inertia of the bootstrap tables is considerably bigger than the total inertia of the original one. This means that the sum of bootstrap eigenvalues is bigger than the sum of original ones, and most of the bootstrap eigenvalues is also bigger than the corresponding original eigenvalues.

If the first bootstrap eigenvalue is bigger than the first original eigenvalue the bootstrap coordinates are dilated (bigger than original coordinates) in this axis.

This way, in the case of CA, partial bootstrap leads to overvalue coordinates, with an expectancy greater that the original coordinates. As the dilation degree is different for each factor, it is necessary to carry out a specific correction for each one of them if the objective is to achieve an appropriate representation.

Concerning the inertia, in the context of a Partial Bootstrap, we use the term “pseudo-inertia” since the total inertia and the principal inertias are computed from the coordinates of supplementary elements.

The total inertia obtained from a CA for the original table where *n* designates the number of rows, *p* the number of columns and *q* the number of axes (minimum between *n* and *p* minus one) is :

$$I_{TOTAL} = \sum_{\alpha=1}^{q} \lambda_{\alpha}$$

The part of the inertia corresponding to the original axis α :

$$I_{\alpha} = \lambda_{\alpha} = \sum_{i=1}^{n} f_{i.} \psi_{\alpha(i)}^{2} = \sum_{j=1}^{p} f_{.j} \varphi_{\alpha(j)}^{2}$$

In a similar way, for every axis, we calculate the pseudo-inertia of the *b*-th Bootstrap row and column clouds, by taking into account the Bootstrap coordinates and frequencies :

$$\lambda_{\alpha b}^{PBrows} = \sum_{i=1}^{n} f_{i.b} \psi_{\alpha b(i)}^{+2} \text{, and } \lambda_{\alpha b}^{PBcol} = \sum_{j=1}^{p} f_{.jb} \varphi_{\alpha b(j)}^{+2}$$

So, the total pseudo-inertia is :

$$I_{b}^{PBrows} = \sum_{\alpha=1}^{q} \lambda_{\alpha b}^{PBrows} \text{ or } I_{b}^{PBcol} = \sum_{\alpha=1}^{q} \lambda_{\alpha b}^{PBcol}$$

This empirical study focuses attention on denote the increase of the inertia in the case of the replicated analyses leads to the hypothesis that the Partial Bootstrap coordinates are overvalued. To obtained replicated configurations with the same expectancy than the original configuration, we suggests a methodology for correcting the bootstrap coordinates and, consequently, the achievement of corrected confidence regions for every point of the configuration.

3.2. Methodology for correcting the Partial Bootstrap coordinates

In order to compute the corrected coordinates, we assume that the ratio between the pseudo-inertia on the axis α, $I_{\alpha b}^{PBrows}$, and the original principal inertia on this axis, $I_\alpha=\lambda_\alpha$, has to be equal to the ratio between the contribution of the row *i* to the inertia of the axis α, $I_{\alpha b(i)}^{PB}$, and this corrected inertia, $I_{\alpha b(i)}^{PBc}$:

$$\frac{I_{\alpha b}^{PBrows}}{I_\alpha} = \frac{I_{\alpha b(i)}^{PB}}{I_{\alpha b(i)}^{PBc}}$$

From :

$$f_{i.b}\psi_{\alpha b(i)}^{PBc\ 2} = I_{\alpha b(i)}^{PBc}$$

we deduce that the corrected coordinate of row *i* on the axis α, $\psi_{\alpha b(i)}^{PBc}$, is given by :

$$\psi_{\alpha b(i)}^{PBc} = \psi_{\alpha b(i)}^{+}\sqrt{\frac{I_\alpha}{I_{\alpha b}^{PBrows}}}$$

In an analogous way, we calculate the corrected coordinate of the column *j* on the axis α :

$$\varphi_{\alpha b(j)}^{PBc} = \varphi_{\alpha b(j)}^{+}\sqrt{\frac{I_\alpha}{I_{\alpha b}^{PBcol}}}$$

We could consider this reasoning too restrictive because a different correction is performed for a coordinate on the axis α in each bootstrap sample. Another alternative is to carry out the correction starting from a global adjustment : the mean of the total pseudo-inertia for all the bootstrap samples. The new coordinates will be then :

$$\psi_{\alpha b(i)}^{PBc} = \psi_{\alpha b(i)}^{+}\sqrt{\frac{I_\alpha}{\bar{I}_\alpha^{PBrows}}} \quad \text{and} \quad \varphi_{\alpha b(j)}^{PBc} = \varphi_{\alpha b(j)}^{+}\sqrt{\frac{I_\alpha}{\bar{I}_\alpha^{PBcol}}}$$

3.3. Bootstrap confidence regions

From the scatterplots of the corrected PB coordinates, it is possible to construct bidimensional nonparametric confidence regions by means of an algorithm based on the “peeling” of the corresponding convex hulls. The convex hull of a set of points on a given plane is the smallest convex polygon that envelops them. “Peeling” a convex hull consists in discarding the points that constitute the vertices of the convex hull. Afterwards, the convex hull of the remaining points is built up. This process is repeated until the elimination of a specific percentage δ% of

the points. This process means to discard the δ% most extreme values of the empirical distribution, i.e. suppresses the most extreme points of a bidimensional distribution.

However, this procedure will not allow, in general, for obtaining a polygon that exactly encompasses (1-δ)% of the points of the scatterplot. For example, if we intend to elaborate a 90% Bootstrap confidence region, it is possible that the k-th convex hull includes 92% of the points, whereas the next convex hull, the (k+1)-th, contains only 87%.

Therefore, it is necessary to perform a second "peeling" procedure in order to obtain the exact required percentage of points. As Markus (1994) proposes, the procedure of the second "peeling" consists of discarding points of the k-th convex hull by using an elimination criterion, until the desired percentage of points (1-δ)% is reached. The elimination criterion that we suggest is based on an iterative procedure : successively, every point of the k-th convex hull is discarded and the corresponding area of the convex hull of the remaining points is computed. The point leading to the smallest area when discarded is suppressed. This process is repeated until removing the required number of points. Figure 1 shows the "peeling" procedure. The result is a convex hull that contains exactly the desired percentage of points and constitutes a (1-δ)% nonparametric Bootstrap confidence region for the location on the principal plane of the corresponding element.

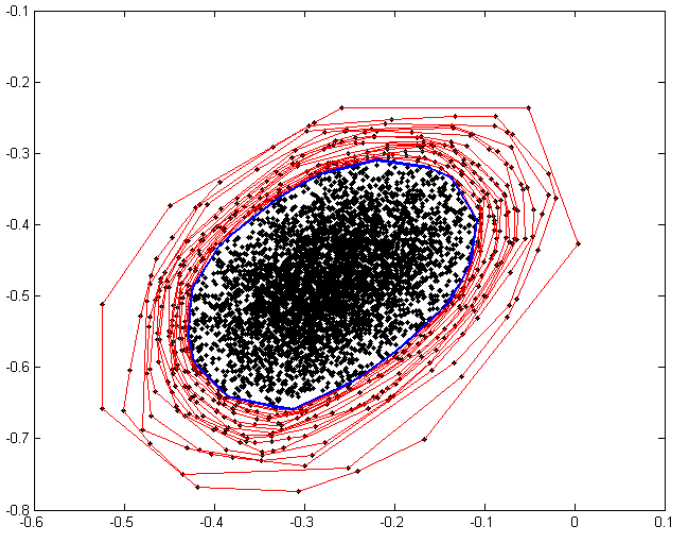

Figure 1. "Peeling" of the convex hull and working-out of a Bootstrap confidence region.

4. Application to textual data

The aggregated lexical table examined stems from the combined analysis of a closed question and an opened question, both contained in a questionnaire sent to a sample of Spanish judges in their first placements (Ayuso *et al.*, 2003). The closed question is "What is your assesment of education received in the Faculty of Law?", with five categories of response : "Muy negativa (Very negative)", "Negativa (Negative)", "Regular (Fair)", "Buena (Good)" and "Muy buena (Very good)". The open question is the complementary question "Why?". The aggregated lexical table contains 2086 occurrences distributed in 114 rows (graphical forms-

lemmas with a minimum frequency of 5) and 5 columns (categories of respondents according to their answer to the closed question).

4.1. Original inertia and bootstrap inertia

As it was indicated previously, it can be proven how the mean of bootstrap total inertias (0,4354) is sensibly bigger than the original total inertia (0,3376). In fact, the original total inertia is smaller than the minimum value of the bootstrap inertias (0,3593). This increase of the total inertia is distributed, logically among the eigenvalues, i.e. 0,1378 of the first original eigenvalue is smaller than 0,1585 of the mean of the bootstrap eigenvalues.

		Eigenvalue 1	Eigenvalue 2	Eigenvalue 3	Eigenvalue 4	Total Inertia
Mean		**0,1585**	**0,1081**	**0,0919**	**0,0770**	**0,4354**
Standard Dev.		0,0088	0,0084	0,0074	0,0072	0,0205
Skewness		0,0993	0,3396	0,1616	0,0351	0,2030
Curtosis		0,0402	0,3357	0,0384	0,0412	0,3530
Minimum		0,1285	0,0804	0,0677	0,0503	0,3593
Maximum		0,1937	0,1450	0,1196	0,1087	0,5431
Percentiles	2,5	0,1416	0,0927	0,0630	0,0781	0,3970
	5	0,1443	0,0951	0,0652	0,0801	0,4035
	95	0,1731	0,1226	0,0888	0,1044	0,4710
	97,5	0,1761	0,1259	0,0909	0,1070	0,4778
Original values		**0,1378**	**0,0797**	**0,0643**	**0,0559**	**0,3376**
Correction factor		**0,93242**	**0,85865**	**0,83647**	**0,85204**	

Table 1. Original and bootstrap total inertia and eigenvalues

Acoording to this results, for the first axe the corrected coordinate *i* for the *b*-bootstrap can be obtained from $\psi^{PBc}_{\alpha b(i)} = \psi^{+}_{\alpha b(i)} \sqrt{\frac{I_\alpha}{\bar{I}^{PB}_{\alpha}}}$ as $\psi^{PBc}_{1b(i)} = \psi^{+}_{1b(i)} \sqrt{\frac{0{,}1378}{0{,}1585}} = 0{,}93242\,\psi^{+}_{1b(i)}$

4.2. Stability of the categories of respondents

The confidence regions of the categories of respondents, obtained from the PB coordinates (solid lines) and from the corrected PB coordinates (dotted lines), are showed in Figure 2a. The correction reduces the area of the regions for all the categories. The joint representation of rows and columns in the principal plane with the corrected regions of the columns is displayed in Figure 3a. The small size of the regions indicates that the locations of the categories of respondents are very stable. The greater stability takes place in the categories with assesment "Good" and "Fair", those of greater answering frequency, whereas the minority categories, specially "Very negative", are rather less stable. The overlapping between the regions of "Negative" and "Fair" points out that the two first factors do not allow to distinguish between the lexical profiles of these two groups of response. Though, according to the first factor, the lexical profiles of the unfavourable opinions are clearly different from those of the favourable opinions and, in accordance with the second factor, the group of respondents with the best valuation presents a lexical profile completely differentiated from the rest (Figure 3a).

4.3. Stability of the textual units

Since the number of textual units is very high, there is no point in making a simultaneous graphical representation of all of them, just as it has been made with the categories of response. That´s why, some textual units have been selected among the non-grammatical words-lemmas. However, if a simultaneous representation of several textual units is carried out, these should have some kind of semantic connection so that the compared interpretation of their stability has sense. Three sets of textual units have been selected. In each one, the units represented simultaneously belong to similar semantic networks and, in principle, they could be used as synonymous or in equivalent contexts.

Figures 2a, 2b, 2c and 2d show the confidence regions from PB and corrected PB. In general the correction produces a shrinkage of the coordinates in absolute terms, which originates translations of the confidence regions towards the origin of coordinates. In most of the cases, the areas of the regions diminish when applying the correction.

The extensive confidence regions of “Facultad” (“Faculty”) and “Universidad%” (University%) indicate the low degree of stability of these units, both with a low frequency (Figure 3b). The absence of overlapping between them means that in the principal plane their profiles of response are totally differentiated, particularly in the first factor. The lemma “Universidad%” has a profile of response rather better defined linked with the “Good” opinion.

As Figure 3c brings out, the sizes of the confidence regions are similar in the cases of “profesorado” (“teaching staff”) and “profesor%” (“teacher”) and considerably smaller than the area of “catedráticos” (“professors”). Again, the greater instability, specially in the second factor, is associated to the graphical form with smaller frequency. The partial overlapping among them in the principal plane reveals that a certain coincidence between the profiles of response of “profesorado” and “profesor%” takes place. The word “catedráticos” has a profile of response less defined but partly shared with them. In fact, the mention of “profesorado” is connected to the “Good” opinion whereas “catedráticos” is more tied to the “Very good” opinion. Anyway, the location of the three regions shows that they are dealing favourable valuations. In the set “asignatura%”-“materia%”-“contenido%” (“subject”-“matter”-“content”) (Figure 3d), the first two lemmas present locations in the plane with a similar degree of stability. By contrast, the unit with smaller frequency, “contenido%”, is less stable particularly in the first factor. A high overlapping between the confidence regions of “asignatura%” and “materia%” is observed, whereas the region of “contenido%” differs considerably from them. In this factorial plane “asignatura%” and “materia%” present similar profiles of response, associated to unfavourable opinions. These two units usually appear in plural and are cited in a similar context which refers to the different fields of Law, being used as synonymous. However, “contenido%” has a less defined profile of response, being used in a more generic sense.

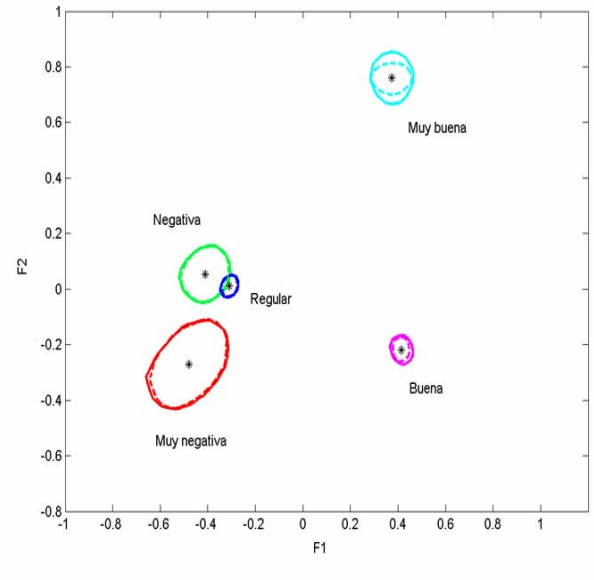

Figure 2a. Categories of respondents

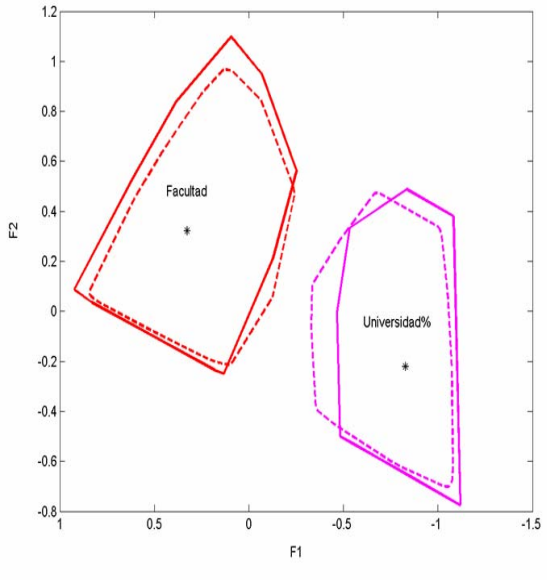

Figure 2b. Textual units : "Facultad% (Faculty)"-"Universidad% (University)"

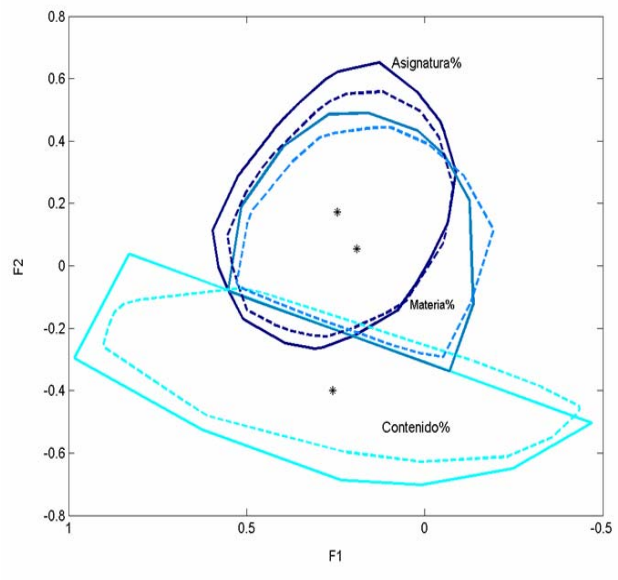

Figure 2c. Textual units : "Asignatura% (Subject)"-"Materia% (Matter)"-"Contenido% (Content)"

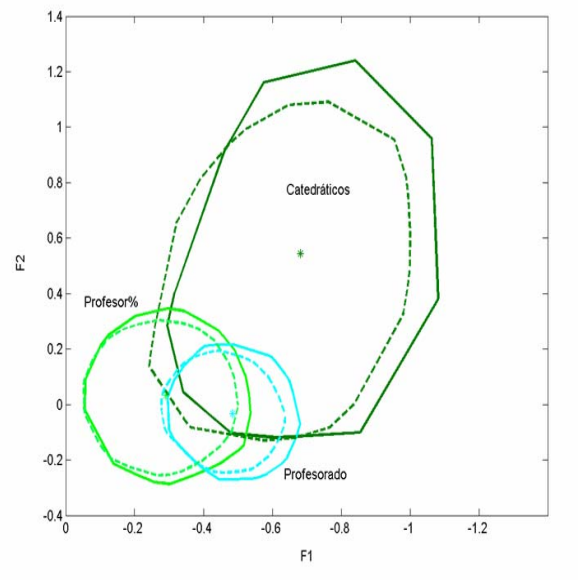

Figure 2d. Textual units : "Profesor% (Teacher)"-"Profesorado (Teaching staff)"-"Catedráticos (Professors)"

Figure 2. Bootstrap confidence regions 90%. BP (solid lines) and BP corrected (dotted lines)

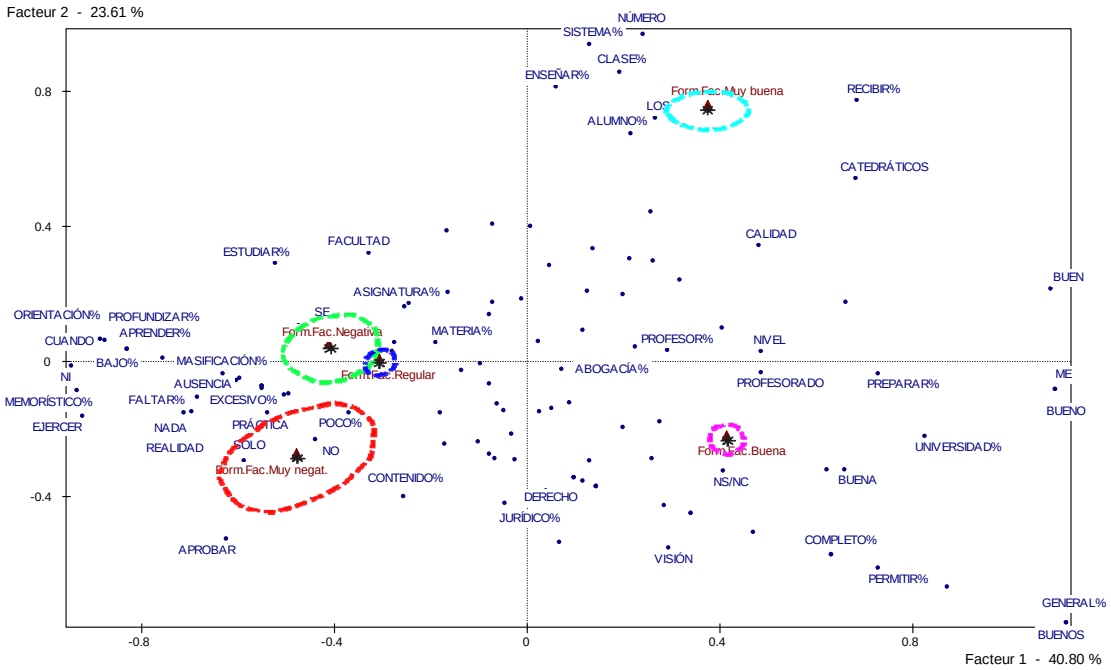

Figure 3a. Principal factorial plane and Bootstrap confidence regions (corrected PB). Categories of respondents

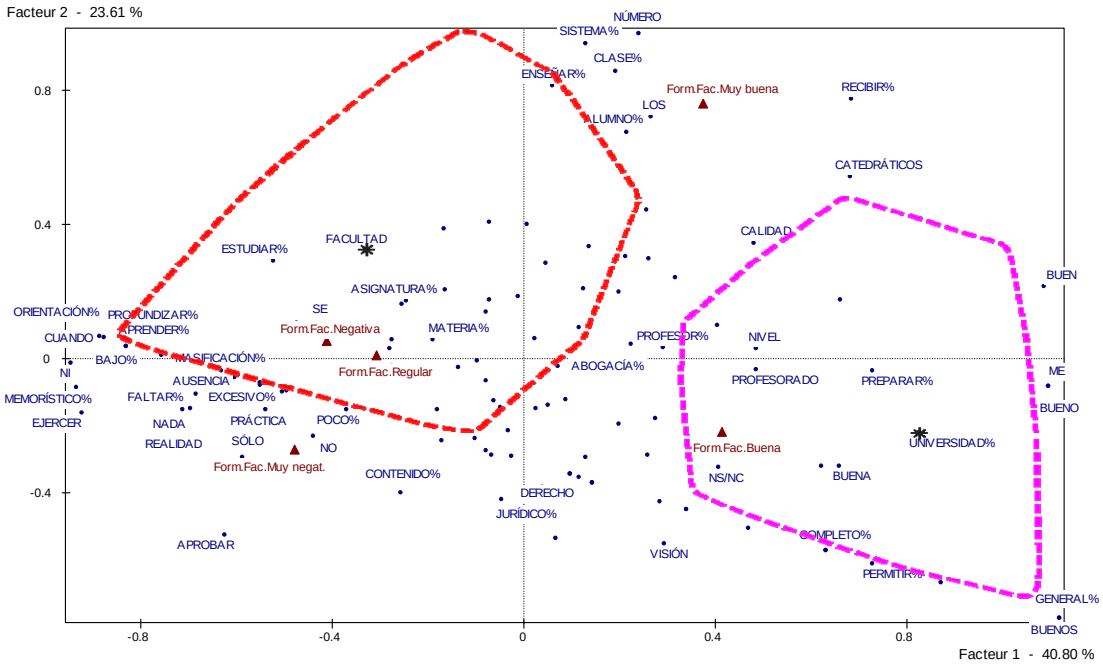

Figure 3b. Principal factorial plane and Bootstrap confidence regions (corrected PB). Textual units : "Facultad% (Faculty)"-"Universidad% (University)"

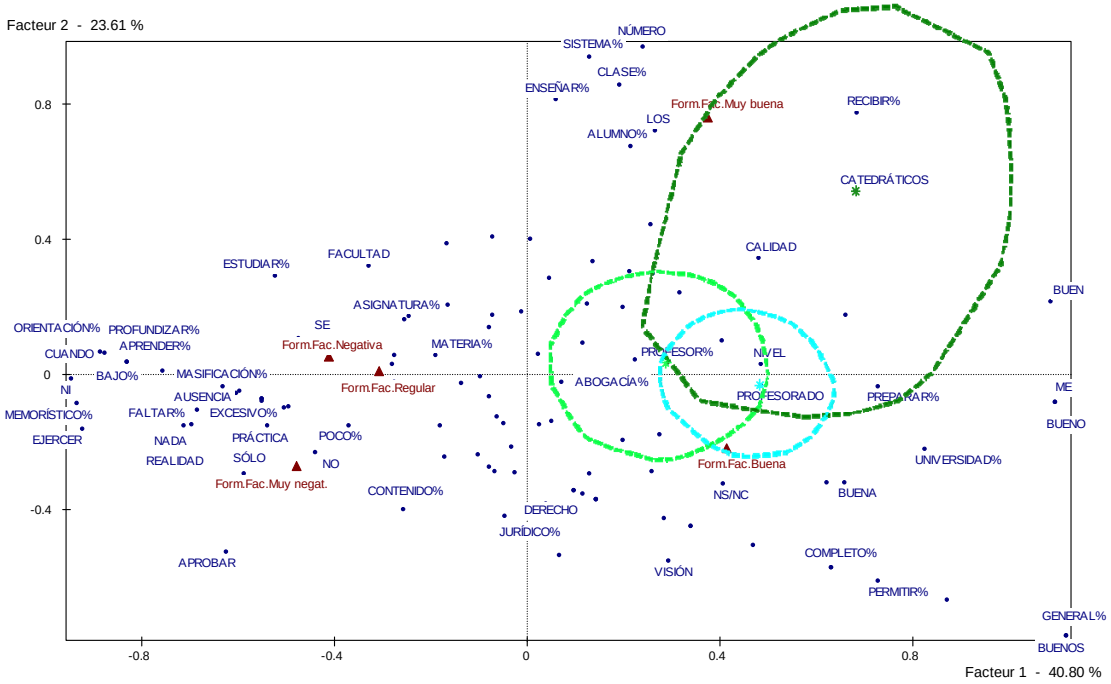

Figure 3c. Principal factorial plane and Bootstrap confidence regions (corrected PB).
Textual units : "Profesor% (Teacher)"-"Profesorado (Teaching staff)"-"Catedráticos (Professors)"

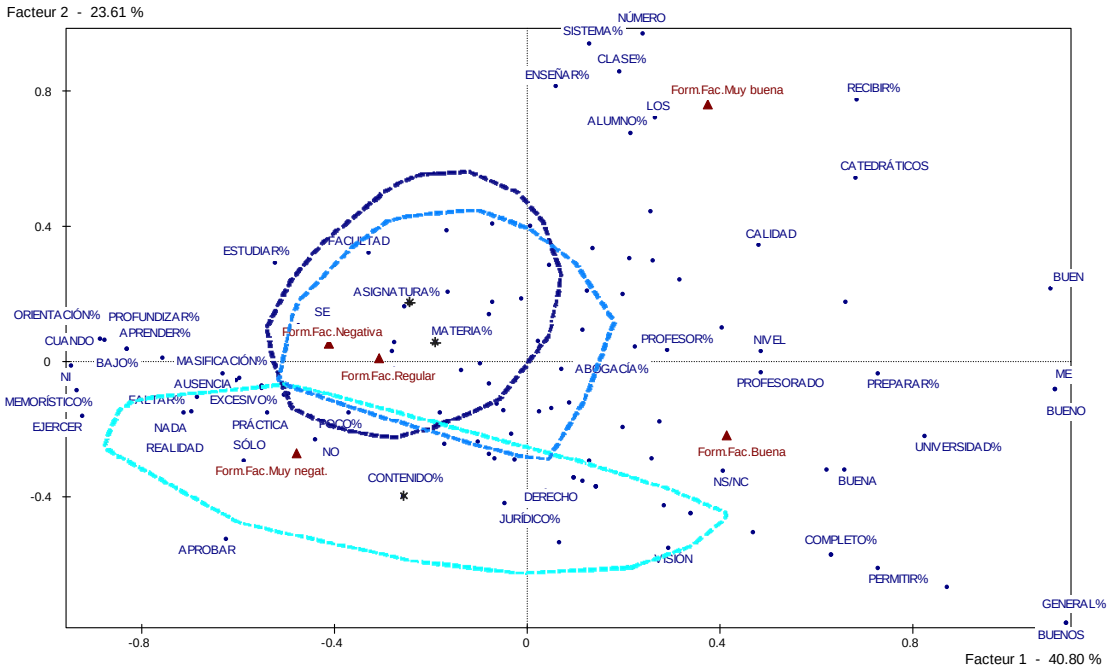

Figure 3d. Main plane and regions of confidence Bootstrap (BP corrected).
Textual units : "Asignatura% (Subject)"-"Materia% (Matter)"-"Contenido% (Content)"

5. Conclusions

Confidence regions allow to evaluate the stability on a factorial plane of both categories of respondents and textual units. As far as categories of respondents are concerned, their lexical profiles can be compared. With regard to textual units, their profiles of response are

compared. The absence of overlapping of the confidence regions indicate different profiles, partial overlapping (respectively, high overlapping) show that the corresponding elements present shared profiles (respectiveley, similar profiles). The confidence regions resulting from the correction of the PB coordinates are more suitable. Moreover, the most fitting confidence regions are those elaborated by means of a "peeling" of the corresponding convex hulls, since they are based on the concrete shape of each scatterplot of Bootstrap points, without privileging any type of possible contour.

References

Álvarez, R., Bécue, M., Valencia, O. (2004). Étude de la stabilité des valeurs propres de l´AFC d´un tableau lexical au moyen de procédures de reéchantillonage. *Actes des JADT 2004. 7es Journées internationales d´Analyse statistique des Données Textuelles*. Louvain, Belgique : 42-51.

Ayuso *et al*. (2003). *Jueces en su primer destino 2002. Análisis Estadístico de las Encuestas realizadas a Jueces en sus primeros destinos (Promociones 48/49 y 50) y Análisis Comparativo con Jueces de mayor experiencia*. Informe de resultados nº 2, elaborado para la Escuela Judicial de Barcelona.

Tan *et al*. (2004). Correspondence analysis of microarray time-course data in case-control design. *Journal of Biomedical Informatics*, 37 : 358-365.

Gifi A. (1990). *Nonlinear multivariate analysis*. John Wiley & Sons Ltd.

Greenacre, M. (1984). *Theory and Aplications of Correspondence Analysis*. Academic Press, London.

Lebart, L. (2004). Validité des visualisations de données textuelles. *Actes des JADT 2004. 7es Journées internationales d´Analyse statistique des Données Textuelles*. Louvain, Belgique : 708-715.

Lebart, L., Morineau, A, Piron, M. (2000). *Statistique exploratoire multidimensionnelle*. Dunod, París.

Lebart, L., Salem, A., Bécue, M. (1999). *Análisis estadístico de textos*. Milenio, Lleida.

Markus, M. (1994). *Bootstrap Confidence Regions in Nonlinear Multivariate Analysis*. Leiden, DSWO Press.

Meulman, J. J. (1984). Correspondence Analysis and Stability. *Research Report 84-01, Dept of Data Theory, Leiden University.*

Reiczigel, J. (1996). Bootstrap tests in Correspondence Analysis. *Applied Stochastic Models and Data Analysis*, 12 : 107-117.

Ringrose, T. J. (1992). Bootstrapping and Correspondence Analysis in Archaeology. *Journal of Archaeological Science*, 19 : 615-629.