Étude de la stabilité des valeurs propres de l'AFC d'un tableau lexical au moyen de procédures de rééchantillonage

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

[1]Universidad de León. León – Spain – dderae@unileon.es
[2]Universidad Politécnica de Cataluña. Barcelona – Spain – monica.becue@upc.es
[3]Universidad de Burgos. Burgos – Spain – oval@ubu.es

Abstract

This paper studies the external stability of eigenvalues issued from Correspondence Analysis of a lexical table, by means of resampling. The problems of lack of stability are more likely to happen in this particular kind of contingency tables, the lexical tables, due to the sparsity of the matrix and other characteristics.
The non-parametric Bootstrap applied in combination with factorial methods which use a Singular Value Decomposition (SVD) algorithm, as Correspondence Analysis does, results in "extra variability" of the statistics involved, including the eigenvalues. Reflections, permutations and rotations of axes may take place in the Bootstrapped samples, that make comparisons between each simulated configuration and the original one quite difficult to understand and senseless in some way. But the whole variability observed doesn't measure the real external stability degree. To study in depth the latter, subsequent corrections must be applied to the results issued from the Total Bootstrap. This may be achieved by a suitable Orthogonal Procustes Rotation.

Résumé

Cet article étudie la stabilité externe des valeurs propres fournies par l'Analyse de Correspondances d'un tableau lexical au moyen de rééchantillonnage. Les problèmes de stabilité des résultats sont plus fréquents dans ce type de tableau de contingence quasi-vide ou creux.
Le Bootstrap non paramétrique est appliqué en combinaison avec les méthodes factorielles qui emploient un algorithme de décomposition en valeurs singulières (SVD), comme l'Analyse de Correspondances, ce qui entraîne une variabilité supplémentaire des statistiques impliquées, y compris les valeurs propres. On peut avoir à faire face à des réflexions, permutations et rotations des axes dans les analyses effectuées sur les échantillons bootstrap, ce qui rend difficile les comparaisons entre chaque configuration simulée et l'originale. Mais la variabilité totale observée ne mesure pas le vrai degré de stabilité externe. Pour vraiment étudier la stabilité, il faut appliquer quelques corrections aux résultats, en particulier en applicant une rotation orthogonale de type Procustes.

Keywords: eigenvalues, correspondence analysis, bootstrap, lexical tables, Procustes analysis

1. Introduction

Le travail présenté ici[1] considère que le caractère exploratoire d'une technique n'implique pas que ses résultats ne soient pas soumis à un certain type de « contrôle de qualité ». En ce sens, la stabilité des résultats fournis par les analyses factorielles exploratoires a été étudiée par divers auteurs. Lebart (1998) étudie la portée et la validité des résultats des Analyses en Composantes Principales (ACP), de l'Analyse Factorielle des Correspondances (AFC) et de

[1] Travail développé grâce au Projet « Analyse Statistique des Enquêtes aux Jeunes Juges » (SEC2001-2581-C02-02, Ministère espagnol pour la Science et la Technologie).

l'Analyse en Correspondances Multiples (ACM) en introduisant certaines perturbations dans la matrice initiale et en suggérant des outils probatoires de diverses natures.

Notre intérêt est centré sur la stabilité externe, ce qui implique la nécessité de vérifier la stabilité face aux fluctuations de l'échantillonnage. En suivant les suggestions de Lebart et d'autres auteurs mentionnés dans la suite de cet article, nous considérons le rééchantillonnage comme une option adéquate pour analyser le comportement de statistiques complexes, comme sont les valeurs propres fournies par une AFC. L'AFC appliquée à des tableaux de contingence lexicaux est un cas particulier qui présente des caractéristiques propres, puisque ces tableaux sont plus propices à l'instabilité.

La procédure de rééchantillonnage choisie est le Bootstrap non paramétrique (Efron, 1993), à cause de son caractère de rééchantillonnage basé sur les données sans hypothèses formelles de travail. Son application pratique requiert cependant une attention spéciale quand il est combiné avec des techniques factorielles, comme l'AFC, qui font usage d'un algorithme de Décomposition en Valeurs Singulières. Cet algorithme, comme l'indiquent Milan et Whittaker (1995), produit une variabilité « supplémentaire » non imputable aux fluctuations du rééchantillonnage, variabilité qui peut affecter la perception initiale du degré de stabilité externe des résultats mais ne doit pas être pris en considération dans son évaluation finale.

La section 2 précise l'objectif de l'étude ; la section 3 présente la méthodologie conçue pour aborder le problème de la stabilité des valeurs propres dans l'AFC d'un tableau lexical ; la section 4 décrit l'exemple choisi (tableau lexical) tandis que la section 5 présente les résultats obtenus ; finalement, la section 6 offre quelques conclusions.

2. Objectif du travail

Le but de ce travail est l'étude de la stabilité externe des valeurs propres de l'AFC d'un tableau lexical au sein d'une étude globale qui inclut aussi l'analyse des coordonnées, des contributions absolues et relatives et des vecteurs propres.

Les tableaux lexicaux sont un type spécial de tableaux de contingence plus enclins à présenter des problèmes d'instabilité dûs à leur nature. En effet, il s'agit de matrices creuses, avec une grande différence entre le nombre de lignes (unités lexicales) et le nombre de colonnes (réponses regroupées par catégories, réponse individuelle ou document textuel classique), dans lesquelles abondent les fréquences marginales très faibles (ce qui concerne tout particulièrement les unités lexicales). D'autre part, l'inertie totale est généralement distribuée presque uniformément entre les axes produits par l'AFC, ce qui entraîne une grande proximité entre les inerties principales ou les valeurs propres. Tout cela est généralement source de problèmes, tant dans la génération des simulations bootstrap elles-mêmes comme dans l'interprétation des facteurs et la représentation graphique des plans factoriels obtenus. Ces difficultés peuvent conditionner la validité des résultats. Dans ce travail, nous centrons notre attention sur la stabilité des valeurs propres.

Nous utilisons une procédure de rééchantillonnage en trois phases :

a. La génération d'un nombre B élevé d'échantillons simulés (B=5000), par Bootstrap non paramétrique.

b. L'application de l'AFC à chacun des échantillons Bootstrap construits (Bootstrap Total).

c. La comparaison des B ensembles de statistiques, associées à chacune des AFC effectuées, avec les statistiques obtenues par l'AFC appliquée à la matrice originale, en particulier comparaison des valeurs propres.

Une comparaison directe des statistiques fournies par le Bootstrap Total d'une AFC avec les statistiques originales peut conduire à une évaluation erronée du degré de stabilité externe, puisque les configurations des échantillons Bootstrap peuvent différer pour des causes « apparentes » et/ou pour des causes « réelles » :

– Nous appelons « apparentes » les différences provoquées par la combinaison du rééchantillonnage Bootstrap avec les méthodes qui, comme les méthodes factorielles, utilisent un algorithme de décomposition en valeurs singulières (SVD) (réflexions de certains facteurs, permutations arbitraires entre des facteurs de rang voisin, changements d'orientation des sous-espaces ou simplement changements d'échelle). Ces modifications des sous-espaces ne supposent pas une réelle modification des distances relatives entre les points et, pour cela, on considère qu'il n'y a pas de véritables différences entre les configurations. Ce pourquoi cette variabilité « supplémentaire » ne doit pas être considérée comme un signe d'instabilité.

– On considère « réelles » les différences causées par les fluctuations de rééchantillonnage. Ces différences doivent servir à évaluer la stabilité externe des résultats. On utilise l'Analyse Procustes Orthogonal pour effectuer une rotation qui permette de superposer chaque configuration bootstrap à la configuration correspondant à l'échantillon original de manière optimale, c'est-à-dire, en obtenant la réduction maximale de la variabilité apparente, et en permettant une comparaison plus adéquate des configurations, a partir de laquelle il soit possible d'effectuer l'analyse de stabilité externe souhaitée.

3. Méthodologie

La méthodologie que nous proposons effectue la comparaison des configurations et, par conséquent, des valeurs propres de trois manières différentes :

1. Par comparaison directe des résultats fournis par le Bootstrap Total. Cette analyse présente des limitations dues à l'influence de la variabilité « supplémentaire » des statistiques, en particulier des valeurs propres, qui vient du fait que les statistiques ne correspondent pas nécessairement aux axes auxquelles elles sont assignées.

2. Par comparaison des résultats fournis par le Bootstrap Total, mais après une rotation Procustes orthogonale destinée à mettre en rapport les axes, et en reconstruisant les valeurs propres (appelées dans ce qui suit pseudo-valeurs propres) à partir des coordonnées

3. Par l'obtention des pseudo-valeurs propres à partir d'un Bootstrap Partiel

Dans ce qui suit, on détaille la méthodologie correspondant au point 2, qui permet de résoudre certains des inconvénients mentionnés dans la section 2, et on précise le rôle et le fonctionnement de l'Analyse Procustes dans ce contexte.

Le principal problème est de comparer chacune des B configurations obtenues C_1, C_2, ..., C_B, avec la configuration associée à l'échantillon original, C_o. L'Analyse Procustes Orthogonale permet d'effectuer un ajustement de chaque configuration bootstrap C_b à la configuration originale C_o, en appliquant une série d'opérations (translation rigide, rotation rigide et dilatation uniforme) de sorte que soit minimale la somme des carrés des distances entre chaque paire de points correspondant à chacune des deux configurations (Krzanowski, 2001).

– La translation est utilisée pour modifier l'origine de l'espace, habituellement pour le faire coïncider avec l'origine de coordonnées et obtenir ainsi une origine commune pour les configurations comparées. Pour cela, il est usuel de travailler avec des matrices centrées par colonne, mais dans notre cas, la translation est faite en multipliant les coordonnées, tant originales C_o comme Bootstrap C_b, par les fréquences marginales respectives F_o et F_b : (C_{op}= F_o C_o

et $C_{bp}= F_b\ C_b$). F_b est la matrice diagonale des fréquences marginales de la simulation b. Ainsi, la somme des coordonnées pondérées C_{bp} est nulle.

– La rotation implique le calcul d'une matrice de rotation R_b pour chacune des matrices bootstrap pondérées C_{bp}, à fin d'obtenir les distances minimales avec la matrice originale pondérée C_{op}.

– La dilatation suppose l'obtention d'une constante scalaire c_b pour adapter la configuration C_{bp} à la taille de la configuration originale pondérée C_{op}.

De cette manière, on obtient de nouvelles matrices avec les configurations pondérées adaptées $C_{bpR} = c_b\ C_{bp}\ Rot_b$, à partir desquelles, en divisant par les fréquences correspondantes F_b, on définit finalement les matrices de coordonnées après rotation C_{bR}. Ces coordonnées représentent des configurations comparables puisque les opérations précédentes ont réduit le plus possible les effets des éventuels réflexions, permutations et changements d'orientation des axes de chaque sous-espace. Les différences entre les configurations demeurant après ces opérations peuvent être considérées comme des différences réelles.

Une fois déterminées les coordonnées après rotation, il est possible de reconstruire les valeurs propres correspondant à chaque axe, à partir des coordonnées sur ces axes ; on obtient ainsi les « pseudo-valeurs propres », ou valeurs propres reconstruites, mentionnées plus haut. Elles correspondent aux inerties principales des nuages de points adaptés de manière beaucoup plus réelle qu'avant la rotation.

Le calcul des pseudo-valeurs propres peut être effectué ou à partir des coordonnées des lignes après rotation ou bien à partir des coordonnées des colonnes après rotation :

$$\lambda^*_{\alpha b} = \sum_{j=1}^{p} f_{.jb} \varphi^{*\ 2}_{j\alpha b} \quad , \text{ ou bien } \lambda^*_{\alpha b} = \sum_{i=1}^{n} f_{i.b} \psi^{*\ 2}_{i\alpha b}$$

La pseudo-inertie totale d'un échantillon bootstrap est égale à la somme des pseudo-valeurs propres.

4. Tableau lexical

Le questionnaire remis aux juges formés à l'École Supérieure de la Magistrature de Barcelone comprenait une question fermée : "*Comment évaluez-vous la formation obtenue à la Faculté de Droit ?* ", avec cinq modalités de réponse : *très négativement, négativement, moyennement, positivement et très positivement.* La question ouverte : "*Pourquoi ?* " était ensuite posée.
On dispose seulement de 268 réponses, ce qui est peu pour procéder à l'analyse de questions ouvertes et pourrait, en principe, affecter négativement la stabilité des résultats. Le tableau lexical comprend 2086 occurrences, réparties en 114 unités lexicales-ligne (formes-lemmes prononcées avec une fréquence supérieure à 5) et 5 colonnes correspondant aux 5 modalités de la question fermée.

5. Résultats

5.1. *Bootstrap Total du tableau lexical sans rotation*

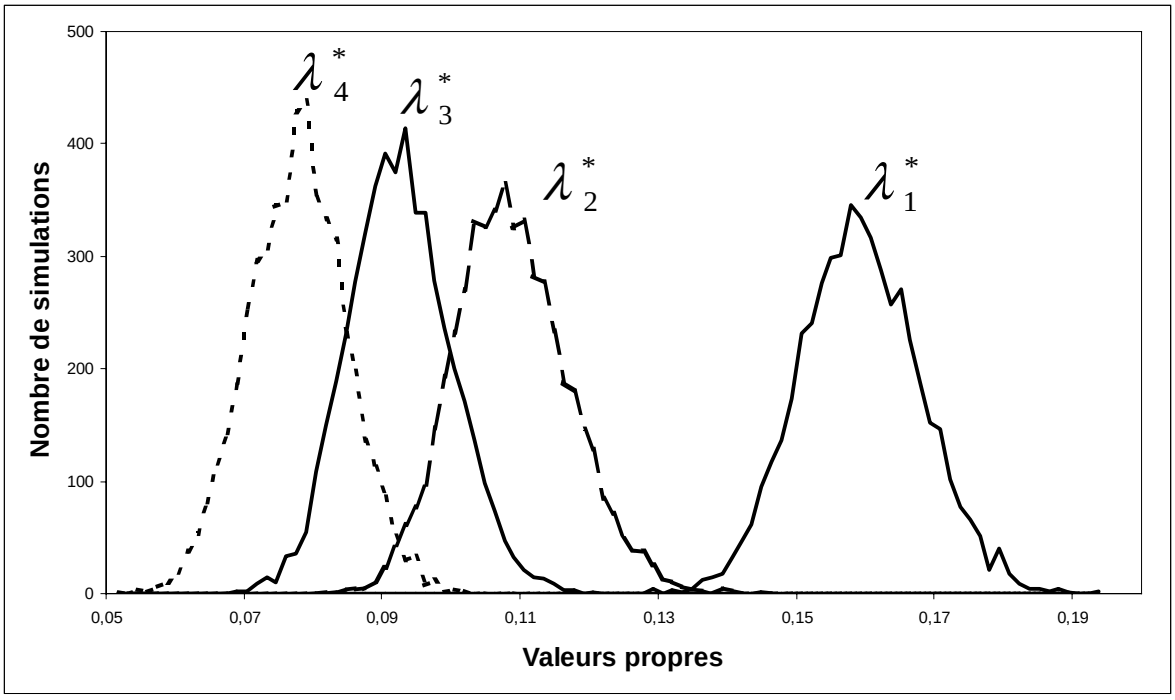
La distribution de l'inertie totale des échantillons bootstrap présente une valeur moyenne (0.4354) très supérieure à l'inertie totale originale (0.3377). Les intervalles percentiles calculés n'incluent pas l'inertie totale d'échantillonnage, ni même en considérant 99 % des simulations. Ceci confirme que le Bootstrap produit une notable augmentation de l'inertie totale et

provoque une perturbation excessive des données initiales. On constate aussi une absence de normalité des inerties totales bootstrap, ce qui indique que la construction d'intervalles de confiance ne peut utiliser le présupposé de la normalité et impose de travailler avec des intervalles percentiles.

Le comportement des inerties principales est spécifié dans le tableau suivant :

	VPROPRE 1	VPROPRE 2	VPROPRE 3	VPROPRE 4
Valeurs propres originaux	**,1378**	**,0797**	**,0643**	**,0559**
Moyenne v.p. Bootstrap	,1584	,1081	,0918	,0770
Écart-type	,0088	,0084	,0074	,0072
Minimum	,1285	,0804	,0678	,0503
Maximum	,1937	,1450	,1196	,1087
Percentiles 0,5	,1364	,0889	,0731	,0590
2,5	,1416	,0927	,0781	,0631
5	,1442	,0951	,0801	,0652
95	,1731	,1226	,1044	,0888
97,5	,1761	,1259	,1070	,0909
99,5	,1810	,1325	,1128	,0960
Test normalité K-S (Sig)	,451	,001	,053	,443
Test norm. Lilliefors (Sig)	,089	,000	,000	,085

Tableau 1. Statistiques des valeurs propres sans rotation



Graphique 1. Distribution des valeurs propres sans rotation

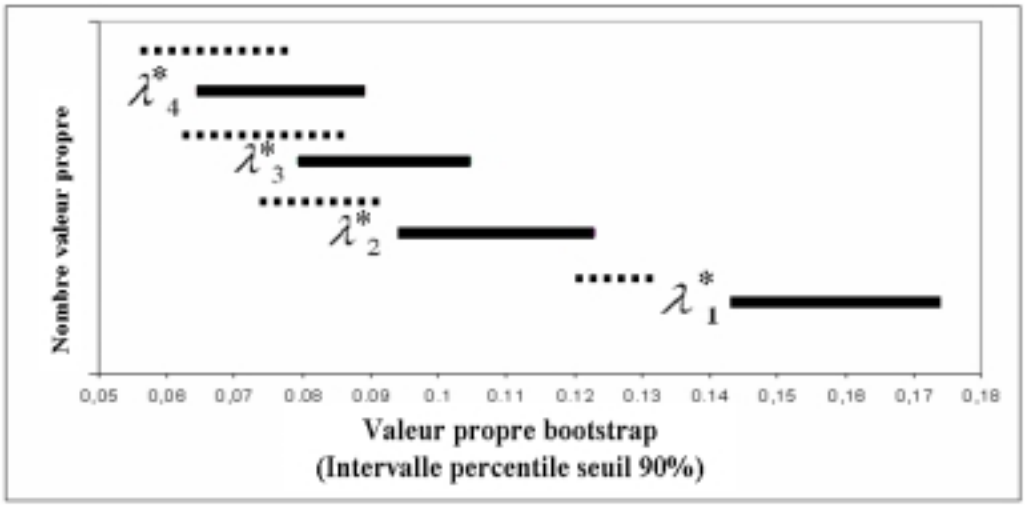
L'interprétation des résultats du Tableau 1 et du Graphique 1 permet d'effectuer quelques observations préliminaires :

– La moyenne des valeurs propres simulées est, dans tous les axes, supérieure à la valeur propre de l'échantillon original : $\overline{\lambda}^*_\alpha > \lambda_\alpha \ \forall\alpha$

– Les inerties principales d'échantillonnage ne sont pas comprises dans les intervalles percentiles au niveau 95 %, ni même au niveau 99 %, sauf l'inertie principale du premier axe. Ceci suggère que la perturbation produite par le bootstrap non seulement donne lieu à une augmentation de l'inertie au niveau global mais à un accroissement de celle-ci dans toutes les dimensions de l'espace factoriel, surestimant systématiquement les véritables inerties principales.

– Au niveau de 95 %, on rejette l'hypothèse de normalité de la distribution des deuxième et troisième valeurs propres, alors que la normalité des première et quatrième valeurs propres ne peut pas être écartée (cette normalité serait cependant rejetée au niveau 90 %).

– Le Graphique 2 représente les intervalles de confiance percentiles au niveau 90 % comme lignes continues. Seul l'intervalle correspondant à la première valeur propre ne recouvre aucun des intervalles correspondant aux autres valeurs propres.



Graphique 2. Intervalles de confiance des percentiles pour les valeurs propres au seuil 90 %

L'examen des résultats directs du Bootstrap Total suggère la stabilité du premier facteur et la possible existence d'instabilité pour le reste des facteurs étant donnée la proximité entre leurs inerties principales. Le Bootstrap Total donne lieu à des estimations non paramétriques des valeurs propres qui se caractérisent tant par leur non-normalité comme par leur biais positif conduisant à surestimer les valeurs-propres.

Ceci nous conduit :

– à rejeter l'utilisation d'intervalles de confiance basés sur la normalité, comme l'intervalle standard avec estimation bootstrap de l'écart type (σ*), et à recourir à des estimations non paramétriques à partir des intervalles percentiles.

– à proposer une correction du biais mentionné, en faisant usage de la même technique Bootstrap pour l'estimer :

$$biais_{\alpha}^{*} = \overline{\lambda_{\alpha}^{*}} - \lambda_{\alpha\,(échantillon)}$$

Le Tableau 2 montre les nouveaux intervalles pour les valeurs propres corrigées, qui comprennent maintenant les valeurs propres originales, construits à partir des intervalles précédents mais en éliminant le biais.

	$\lambda_{éch}$	$\bar{\lambda}^*$	Biais*	$P^*_{5(corr)}$	$P^*_{95(corr)}$	$P^*_{2,5(corr)}$	$P^*_{97,5(corr)}$	$P^*_{0,5(corr)}$	$P^*_{99,5(corr)}$
1	0,1378	0,1584	0,0206	0,1236	0,1525	0,1210	0,1555	0,1158	0,1604
2	0,0797	0,1081	0,0284	0,0667	0,0942	0,0643	0,0975	0,0605	0,1041
3	0,0643	0,0918	0,0275	0,0526	0,0769	0,0506	0,0795	0,0456	0,0853
4	0,0559	0,0770	0,0211	0,0441	0,0677	0,0420	0,0698	0,0379	0,0749

Tableau 2. Intervalles des valeurs propres avec biais corrigés. Percentile au 90 %, 95 % et 99 %.

5.2. *Bootstrap Total du tableau lexical avec rotation*

On considère le tableau de coordonnées C^Y de dimensions ($n+p$, q), juxtaposant la matrice de coordonnées-ligne (n, q) et la matrice de coordonnées-colonne (p, q). On pondère le tableau simulé bootstrap ainsi obtenu ($C^Y_{bp}= F^Y_b\ C^Y_b$) et on le compare au tableau juxtaposant les coordonnées originales (lignes et colonnes) en pondérant chacun des éléments par les fréquences marginales ($C^Y_{op}= F^Y_o\ C^Y_o$).

La rotation Procustes effectuée diminue l'inertie des nuages simulés et offre une pseudo-valeur propre proche de la valeur propre du nuage original. Cela signifie que le caractère excessivement perturbateur du Bootstrap dans l'étude de l'AFC d'un tableau lexical, n'est pas aussi grand que l'on pouvait initialement le penser. Il répond à la nature intrinsèque de la technique Bootstrap, fondamentalement à la variabilité « supplémentaire » produite par sa combinaison avec l'algorithme SVD utilisé par les méthodes factorielles.

En éliminant cette variabilité, l'intervalle percentile au niveau 90 % inclut la valeur originale de l'inertie totale, très proche de la moyenne de celle correspondant aux échantillons bootstrap.

La distribution des pseudo-valeurs propres calculées à partir des coordonnées après rotation, (Tableau 3) permet d'observer que les moyennes des pseudo-valeurs propres sont beaucoup plus proches des valeurs propres originales. Les intervalles percentiles au niveau 90 % incluent les valeurs propres originales λ_2 et λ_3 ; le niveau doit être augmenté jusqu'à 95 % pour λ_4 et jusqu'à 99 % pour λ_1.

La distribution des seconde troisième et quatrième pseudo-valeurs propres ne vérifie pas la normalité, avec un risque de 5 %

Les superpositions entre les intervalles percentiles au niveau 90 % des pseudo-valeurs propres se donnent pour les deuxième, troisième et quatrième axes, tandis que la première valeur propre est encore parfaitement éloignée et différenciée du reste. On ne peut être interpréter, en aucune façon, ces résultats comme un symptôme d'instabilité puisque les trois facteurs mentionnés se sont montrés hautement stables comme l'indiquent les petits intervalles percentiles des valeurs propres correspondantes. Le graphique 2 montre (lignes discontinues) les intervalles de confiance pour les pseudo-valeurs propres.

		VPROPRE 1	VPROPRE 2	VPROPRE 3	VPROPRE 4
Valeurs propres originaux		**,1378**	**,0797**	**,0643**	**,0559**
Moyenne v.p. Bootstrap		,1263	,0830	,0740	,0675
Écart-type		,0048	,0063	,0080	,0070
Minimum		,1086	,0617	,0460	,0437
Maximum		,1456	,1059	,1102	,0935
	Percentiles 0,5	,1143	,0682	,0557	,0506
	2,5	,1170	,0710	,0594	,0543
	5	,1184	,0728	,0615	,0563
	95	,1343	,0935	,0877	,0795
	97,5	,1360	,0954	,0904	,0818
	99,5	,1392	,1003	,0961	,0872
Test normalité K-S (Sig)		,458	,309	,016	,156
Test norm. Lilliefors (Sig)		,092	,033	,000	,005

Tableau 3. Statistiques des pseudo-valeurs propres. Bootstrap Total. Rotation

5.3. *Bootstrap Partiel du tableau lexical*

Le nombre différent de lignes (114) et de colonnes (5) fait que les pseudo-valeurs propres sont différentes si elles sont reconstruites ou bien à partir des lignes ou bien à partir des colonnes.

L'inertie totale des échantillons bootstrap obtenue à partir des pseudo-valeurs propres des colonnes (0.3422) est légèrement supérieure à l'originale (0.3377). Quand on obtient les pseudo-valeurs à partir des lignes la différence est beaucoup plus grande (0.4362). Il y a aussi absence de normalité pour les trois derniers facteurs. Ceci confirme que le Bootstrap produit une augmentation notable de l'inertie totale (surtout pour les lignes dans un tableau lexical), en provoquant une perturbation excessive des données initiales ; on doit donc effectuer la correction de biais proposée au paragraphe 5.1 ou bien une rotation de la configuration bootstrap obtenue avec le Bootstrap Partiel.

Dans le Tableau 4 on observe que les moyennes des pseudo-valeurs propres reconstruites à partir des coordonnées des colonnes sont proches des valeurs propres originales et que l'on rejette l'hypothèse de normalité pour la troisième et la quatrième.

		VPROPRE 1	VPROPRE 2	VPROPRE 3	VPROPRE 4
Valeurs propres originaux		**,1378**	**,0797**	**,0643**	**,0559**
Moyenne v.p. Bootstrap		,1390	,0808	,0656	,0568
Écart-type		,0099	,0085	,0101	,0077
Minimum		,1053	,0529	,0309	,0308
Maximum		,1748	,1123	,1121	,0916
	Percentiles 0,5	,1139	,0602	,0410	,0382
	2,5	,1198	,0645	,0467	,0425
	5	,1228	,0671	,0495	,0445
	95	,1555	,0952	,0826	,0702
	97,5	,1587	,0979	,0858	,0727
	99,5	,1661	,1031	,0931	,0778
Test normalité K-S (Sig)		,492	,683	,111	,033
Test norm. Lilliefors (Sig)		,108	,200*	,002	,000

* Borne inférieure de la signification vraie

Tableau 4. Statistiques des pseudo-valeurs propres colonnes. Bootstrap Partiel. Sans rotation

Mais dans le Tableau 5, les moyennes des pseudo-valeurs propres obtenues à partir des lignes sont assez différentes des valeurs propres originales. Par exemple, pour le facteur 2 la valeur 0.797 est hors des intervalles de confiance. L'absence de normalité dans ce cas affecte les trois derniers facteurs.

	VPROPRE 1	VPROPRE 2	VPROPRE 3	VPROPRE 4
Valeurs propres originaux	**,1378**	**,0797**	**,0643**	**,0559**
Moyenne v.p. Bootstrap	,1566	,1032	,0921	,0843
Écart-type	,0087	,0098	,0119	,0097
Minimum	,1257	,0722	,0514	,0536
Maximum	,1909	,1476	,1470	,1282
Percentiles 0,5	,1355	,0809	,0648	,0623
2,5	,1399	,0850	,0707	,0667
5	,1425	,0878	,0735	,0690
95	,1710	,1199	,1123	,1012
97,5	,1741	,1230	,1166	,1048
99,5	,1790	,1304	,1257	,1123
Test normalité K-S (Sig)	,604	,007	,015	,000
Test norm. Lilliefors (Sig)	,200*	,000	,000	,000

* Borne inférieure de la signification vraie

Tableau 5. Statistiques des pseudo-valeurs propres lignes. Bootstrap Partiel. Sans rotation

Le Tableau 6 montre maintenant les nouveaux intervalles pour les valeurs propres lignes corrigées, intervalles qui comprennent les valeurs propres d'échantillonnage.

	$\lambda_{éch}$	$\bar{\lambda}^*$	Biais*	$P^*_{5(corr)}$	$P^*_{95(corr)}$	$P^*_{2,5(corr)}$	$P^*_{97,5(corr)}$	$P^*_{0,5(corr)}$	$P^*_{99,5(corr)}$
1	0,1378	,1566	0,0188	0,1237	0,1522	0,1211	0,1553	0,1167	0,1602
2	0,0797	,1032	0,0235	0,0643	0,0964	0,0615	0,0995	0,0574	0,1069
3	0,0643	,0921	0,0278	0,0457	0,0845	0,0429	0,0888	0,0370	0,0979
4	0,0559	,0843	0,0284	0,0406	0,0728	0,0383	0,0764	0,0339	0,0839

Tableau 6. Intervalles des pseudo-valeurs propres lignes avec correction du biais. Percentile au 90 %, 95 % et 99 %. Bootstrap Partiel. Sans rotation

6. Conclusions

L'application du Bootstrap non paramétrique est fréquente dans les études de stabilité des résultats des méthodes factorielles exploratoires, en particulier pour l'analyse des coordonnées, ce qui peut être effectué avec un Bootstrap Partiel.

L'inertie totale des tableaux lexicaux obtenus par rééchantillonnage est plus grande que l'originale. Même si on utilise un Bootstrap Partiel, les coordonnées (surtout des lignes-formes) sont surestimées, comme on a pu le vérifier avec le calcul des pseudo-valeurs à partir des coordonnées lignes ou des colonnes.

Le fait que l'inertie des analyses répliquées augmente suggère que les coordonnées bootstrap « vraies » sont plus petites que celles obtenues avec la simulation. Cela conduit à effectuer une correction des pseudo-valeurs propres et à corriger postérieurement les coordonnées, ou à chercher un autre type de simulation moins perturbateur qu'un bootstrap pur, ou à effectuer les rotations nécessaires.

Si on prétend approfondir dans l'étude du comportement d'autres statistiques il est indispensable d'utiliser un Bootstrap Total, ce qui implique la comparaison de configurations qui correspondent à des espaces factoriels différents. Pour cette raison, la simple considération des résultats bruts du Bootstrap Total peut conduire à des conclusions erronées. Dans notre cas, initialement seule la première dimension de l'analyse paraissait stable, tandis que les autres trois semblaient montrer une certaine instabilité. Après la rotation Procustes, qui réduit dans la mesure du possible les effets des réflexions, échanges entre facteurs et changements d'orientation, on montre que tous les facteurs sont stables. Ce travail ne prétendait pas étudier les autres statistiques.

Références

Álvarez R., Bécue M., Lanero J.J. et Valencia O. (2002). Results stability in textual analysis: Its application to the study of the Spanish investiture speeches (1979-2000). In *Actes des JADT 2002 :*1-12.

Chateau F. et Lebart L. (1996). Assessing sample variability and stability in the visualization techniques related to principal component analysis: Bootstrap and alternative simulation methods. In *Proceedings Computational Statistics. COMPSTAT* : 205-210.

Efron B. et Tibshirani R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

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

Krzanowski W.J. (2001). *Principles of Multivariate Analysis. A User´s Perspective*. Oxford Statistical Science Series.

Lebart L. (1976). The significance of eigenvalues issued from correspondence analysis. In *Proceedings in Computational Statistics*. *COMPSTAT*. Physica Verlag : 38-45.

Lebart L., Morineau A. et Piron M. (1998). *Statistique exploratoire multidimensionnelle*. Dunod.

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

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

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

Michailidis G. et De Leeuw J. (1998). The Gify System of descriptive multivariate analysis. *Statistical Science*, vol. (13) : 307-336.

Milan L. et Whittaker J. (1995). Application of the parametric Bootstrap to models that incorporate a singular value decomposition. *Appl. Statist*, vol. (44/1) : 31-49.

O´Neill M. E. (1978). Aymptotic distributions of the canonical correlations form contingency tables. *Australian Journal of Statistics*, vol. (20/1) : 75-82.

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