Tests construits par la méthode du bootstrap sur les valeurs propres d'une analyse factorielle des correspondances quand elle est utilisée comme une technique de réduction des données

Daniel Pierre Loti Viaud

Pub. Inst. Stat. Univ. Paris
XXXXI, fasc. 3, 1997, 21 à 29

Tests construits par la méthode du bootstrap sur les valeurs propres d'une analyse factorielle des correspondances quand elle est utilisée comme une technique de réduction des données

Daniel PIERRE-LOTI-VIAUD
Université Paris 6 - LSTA

Résumé. L'analyse factorielle des correspondances, comme l'analyse en composantes principales, peut être utilisée comme une technique particulière de réduction des données. Après avoir rappelé le modèle probabiliste sous-jacent à cet usage de l'analyse factorielle des correspondances, ainsi que les objectifs d'une telle analyse, nous montrons qu'on peut transposer pour cette analyse la méthodologie présentée dans le cadre de l'analyse en composantes principales par Beran et Srivastava (1985). Nous obtenons ainsi des intervalles de confiance et des tests construits par la méthode du bootstrap sur les valeurs propres issues d'une analyse factorielle des correspondances. En particulier, nous proposons un test de nullité des plus petites valeurs propres et un test sur le nombre de valeurs propres à retenir pour obtenir un taux d'explication des données supérieur à une valeur fixée a priori.

Abstract. As for principal component analysis, correspondence analysis can be viewed as a particular technical tool in order to reduce dimensionnality of the data. By recalling the probabilistic model associated with such use of correspondence analysis, we show that the results of Beran and Srivastava (1985) for principal component analysis extend to correspondence analysis. Thus, we obtain confidence intervals and tests based on the bootstrap methodology for eigen values coming from correspondence analysis. In particular, we propose a test for the smallest eigen values to be null and a test for the number of significative eigen values to be considered in order to explain the data with a given fixed rate.

Adresse postale. Daniel PIERRE-LOTI-VIAUD
Université PARIS VI,
L.S.T.A., T.45-55, E.3
Boite 158
4, place Jussieu
75252 PARIS Cedex 05
FRANCE

E-mail. pilovi@ccr.jussieu.fr

Remerciements. Ce travail a été soutenu par le programme européen AIR-CT94-1111, commission européenne DG XII, 8 square de Meeus, 1024 Bruxelles. Les logiciels nécessaires à la mise en oeuvre pratique des tests ont été réalisés par la société ABC, 7, rue Chevert, 75007 Paris.

1 - Introduction

Dans ce qui suit p et n sont des entiers naturels non nuls et on observe $X_1, \ldots, X_n$ des variables aléatoires à valeurs dans $(0, \infty)^p$, les observations $X_1, \ldots, X_n$ étant supposées indépendantes et équidistribuées de loi P . On note X la matrice $p \times n$ définie par $X = [X_1 \cdots X_n]$, et $\Sigma_n(P)$ la matrice $p \times p$ qui est diagonalisée lorsqu'on effectue une analyse factorielle des correspondances (AFC) sur la matrice X .

Notre objectif dans ce document est de présenter quelques tests s'appliquant aux valeurs propres obtenues dans l'AFC de X . La méthodologie suivie, qui est la transposition de celle utilisée par Beran et Srivastava (1985) dans le cadre de l'analyse en composantes principales (ACP), est basée sur la méthode du bootstrap. On commence par valider l'utilisation de la méthode du bootstrap. Pour cela, on étudie la loi limite de la matrice $\Sigma_n(P)$, et on montre la convergence vers cette loi limite de la même matrice construite en partant d'un échantillon bootstrapé, c'est-à-dire $\Sigma_n(\widehat{P}_n)$ où $\widehat{P}_n$ désigne la loi empirique des observations. On transporte ensuite ce résultat à des fonctions suffisamment régulières de la matrice $\Sigma_n(P)$, comme les valeurs propres de cette matrice lorsque celles-ci sont toutes simples. On peut ainsi construire des ellipsoides de confiance, et par conséquent des tests, de niveau de confiance asymptotique fixé sur les valeurs propres issues de l'AFC de X . Nous renvoyons à Hall (1992), Efron et Tibshirani (1993), Shao et Tu (1995), par exemple, pour une présentation générale de la méthode du bootstrap. Indiquons juste que dans plusieurs cas il a pu être montré que la méthode du bootstrap conduit à des ellipsoides de confiance de meilleure qualité que ceux construits par la méthode de la loi limite. Dans notre cadre, l'intérêt de la méthode du bootstrap est au moins d'éviter l'estimation des paramètres de la loi limite, ceux-ci faisant intervenir des moments relativement complexes de la loi P .

Concernant l'utilisation dans le contexte qui est le nôtre ici de l'AFC, les points suivants doivent être précisés. Sur un plan pratique, nous nous situons dans le cadre de la construction d'une base de données sur un type de produits présentant une grande variabilité des paramètres mesurés, les données X correspondent alors à un sondage effectué sur la population totale. Dans le but d'effectuer une classification des produits en différents groupes, une étape préliminaire de réduction des données est nécessaire afin de rechercher les paramètres mesurés ayant un caractère explicatif des différences observées. Si le sondage est effectué dans les règles de l'art, les observations $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes et équidistribuées. Lorsque les paramètres mesurés sont des teneurs en différents composants des produits étudiés, le recours à l'AFC comme technique de réduction des données peut se justifier, même si d'autres techniques peuvent être envisagées (ACP normée, ACP sur le logarithme des teneurs,...). En effet, sur un plan théorique, et de manière analogue à l'ACP, l'AFC s'apparente à une réduction du rang de la matrice X afin de condenser la taille de l'espace des observations. Ce point se déduit de la décomposition singulière de la matrice X et du théorème d'Eckart et Young (1936) qui donne la meilleure approximation de X par une matrice de rang fixé a priori. Ce contexte motive les tests sur les valeurs propres issues de l'AFC de X que nous proposons par la suite : test de nullité des plus petites valeurs propres et test sur la proportion de distance expliquée quand on remplace la matrice X par la matrice de rang ℓ la plus proche de X . Ce dernier test s'exprime comme le rapport de la somme des ℓ plus grandes valeurs propres sur la somme totale des valeurs propres (cf section 2). Signalons que nous ne sommes pas exhaustif sur les critères concernant le nombre de valeurs propres à retenir dans une AFC qui peuvent être étudiés en suivant notre démarche (cf. Greenacre (1984), et Jolliffe (1986) dans le cadre de l'ACP, pour une liste de ces critères, voir aussi Besse (1992) pour un autre critère important présenté dans

le cadre de l'ACP, mais que l'on peut transposer dans le cadre de L'AFC). Notre objectif ici ayant surtout été de montrer que la méthode du bootstrap présente une grande souplesse d'utilisation malgré sa complication apparente, nous ne détaillerons pas dans ce document d'exemple numérique de l'application des tests présentés. Indiquons cependant que ces tests ont été appliqués à de nombreuses reprises sur des données concernant des jus de fruits dans le cadre d'un contrat européen AIR, les programmes permettant d'appliquer la méthode du bootstrap à l'AFC ayant été développé par Philippe Laffez. Pour être complet, et suivant en cela une remarque d'un des rapporteurs de ce travail, nous soulignons que la comparaison des différentes méthodes qui peuvent être envisagées afin de réduire la dimension de nos données est un problème important en soit, et qu'il pourrait faire l'objet d'un travail ultérieur.

Ce document est organisé de la façon suivante. Les liaisons entre la décomposition singulière d'une matrice, le théorème d'Eckart et Young (1936) et l'AFC sont rappelées dans la section 2. La loi limite de la matrice $\Sigma_n(P)$ est étudiée dans la section 3 et celle de la matrice $\Sigma_n(\widehat{P}_n)$ dans la section 4. Les intervalles de confiance et tests sur les valeurs propres issues de l'AFC de X sont présentés dans la section 5.

2 - Décomposition singulière d'une matrice et analyse factorielle des correspondances

Pour A une matrice $K \times L$, ses coefficients sont notés $A_{k\ell}$, $1 \leq k \leq K$, $1 \leq \ell \leq L$, A' désigne sa transposée, et, si $K = L$ et A est inversible, A^{-1} désigne son inverse. La matrice identité en dimension K est notée I_K.

Soit M et N deux matrices respectivement $p \times p$ et $n \times n$ qui sont symétriques et définies positives. Les espaces $\mathbb{R}^p$ et $\mathbb{R}^n$, contenant les vecteurs colonnes et les vecteurs lignes de X, sont munis respectivement des métriques associées aux matrices M et N. Le résultat suivant se trouve par exemple dans Greenacre (1984).

Proposition 1 (décomposition singulière d'une matrice). *Soit B une matrice $p \times n$ de rang r. Alors il existe une base orthonormée $U_1, \ldots, U_p$ de $\mathbb{R}^p$ muni de M, une base orthonormée $V_1, \ldots, V_n$ de $\mathbb{R}^n$ muni de N, et des coefficients strictement positifs $\alpha_1 \geq \cdots \geq \alpha_r$ vérifiant :*

$$B = [U_1 \cdots U_r] \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_r \end{pmatrix} [V_1 \cdots V_r]' = \sum_{k=1}^{r} \alpha_k U_k V_k' .$$

Indiquons que, pour cette décomposition de la matrice B, $U_1, \ldots, U_p$ est une base de vecteurs propres de la matrice $BNB'M$ tandis que $V_1, \ldots, V_n$ est une base de vecteurs propres de la matrice $B'MBN$, les α_k, $1 \leq k \leq r$, étant les racines carrées des valeurs propres non nulles associées (ces valeurs propres non nulles sont identiques pour $BNB'M$ et $B'MBN$).

Désignons par $\mathcal{M}(p,n)$ l'espace vectoriel des matrices $p \times n$ et par φ le produit scalaire sur $\mathcal{M}(p,n)$ défini par :

$$\varphi(A,B) = \text{trace}(ANB'M) .$$

Posons $\varphi(A) = \varphi(A,A)$ pour $A \in \mathcal{M}(p,n)$. Le théorème suivant est dû à Eckart et Young (1936) (cf. aussi Greenacre (1984)).

Théorème 1. *Soit* B *une matrice* $p \times n$ *de rang* r *et sa décomposition singulière* $\sum_{k=1}^{r} \alpha_k U_k V_k'$. *Pour* $1 \leq \ell \leq r$, $\sum_{k=1}^{\ell} \alpha_k U_k V_k'$ *est une matrice de rang inférieur ou égale à* ℓ *qui approche le mieux la matrice* B *dans* $\mathcal{M}(p,n)$ *muni de la distance associée à* φ , *c'est-à-dire, qui vérifie :*

$$\varphi\Big(B - \sum_{k=1}^{\ell} \alpha_k U_k V_k'\Big) = \inf\{\varphi(B-A)\,,\ A \in \mathcal{M}(p,n)\,,\ \mathrm{rang}(A) \leq \ell\}\ .$$

En outre, si $\ell < r$ *et* $\alpha_\ell > \alpha_{\ell+1}$, $\sum_{k=1}^{\ell} \alpha_k U_k V_k'$ *est l'unique matrice réalisant cet infimum.*

Remarquons que, pour toute décomposition singulière $B = \sum_{k=1}^{r} \alpha_k U_k V_k'$, les matrices $U_1 V_1', \ldots, U_r V_r'$ forment une famille orthonormée de $\mathcal{M}(p,n)$ muni de φ . En particulier, nous avons :

$$\varphi\Big(B - \sum_{k=1}^{\ell} \alpha_k U_k V_k'\Big) = \sum_{k=\ell+1}^{r} \alpha_k^2 \quad \text{pour} \quad 0 \leq \ell \leq r\ .$$

Désignons maintenant par P la loi de probabilité commune aux observations $X_1, \ldots, X_p$. Notons, pour $1 \leq k \leq p$ et $1 \leq \ell \leq n$, $X_{k.} = \sum_{\ell'=1}^{n} X_{k\ell'}$ et $X_{.\ell} = \sum_{k'=1}^{p} X_{k'\ell}$. Notons encore $X_{..} = \sum_{k=1}^{p} \sum_{\ell=1}^{n} X_{k\ell}$, et :

$$\mu(X) = \begin{pmatrix} X_{1.} & & 0 \\ & \ddots & \\ 0 & & X_{p.} \end{pmatrix} \quad , \qquad \nu(X) = \begin{pmatrix} X_{.1} & & 0 \\ & \ddots & \\ 0 & & X_{.n} \end{pmatrix}\ .$$

Sous l'hypothèse $P\big((0,\infty)^p\big) = 1$, les matrices $\mu(X)$ et $\nu(X)$ sont presque sûrement inversibles.

Rappelons (cf., par exemple, Greenacre (1984)) qu'effectuer une analyse factorielle des correspondances (AFC) de la matrice X , c'est diagonaliser les matrices $\mu(X)^{-1} X \nu(X)^{-1} X'$ et $\nu(X)^{-1} X' \mu(X)^{-1} X$, les vecteurs propres associés donnant les coordonnées des nuages des profils lignes et colonnes sur les axes principaux (coordonnées principales). Remarquons que, si $u \in \mathbb{R}^p$ et $v \in \mathbb{R}^n$ sont des vecteurs propres de $\mu(X)^{-1} X \nu(X)^{-1} X'$ et $\nu(X)^{-1} X' \mu(X)^{-1} X$ associés à la valeur propre λ , alors $\mu(X)U$ et $\nu(X)V$ sont des vecteurs propres de $X\nu(X)^{-1} X' \mu(X)^{-1}$ et $X'\mu(X)^{-1} X \nu(X)^{-1}$ toujours associés à la valeur propre λ . Ainsi, en notant $B = X$, et en munissant $\mathbb{R}^p$ de $M = M(X) = \mu(X)^{-1}$ et $\mathbb{R}^n$ de $N = N(X) = \nu(X)^{-1}$, la décomposition singulière $\sum_{k=1}^{r} \alpha_k U_k V_k'$ de la matrice X se déduit des valeurs propres et des coordonnées principales obtenues en faisant une AFC de X . Notons les observations suivantes sur la décomposition singulière associée à l'AFC de X . On a $1 = \alpha_1 > \alpha_2 \geq \cdots \geq \alpha_r > 0$, les α_k étant les racines carrées des valeurs propres non nulles issues de l'AFC. De plus, les vecteurs U_1 et V_1 associés à α_1 sont portés respectivement par les vecteurs $\mathbb{1}_p = [1 \cdots 1]' \in \mathbb{R}^p$ et $\mathbb{1}_n = [1 \cdots 1]' \in \mathbb{R}^n$.

3 - Loi limite dans une analyse factorielle des correspondances

Dans ce qui suit nous posons :

$$\Sigma_n = \Sigma_n(P) = \mu(X)^{-1} X \nu(X)^{-1} X'\ .$$

Σ_n est la matrice $p \times p$ qui est diagonalisée dans l'AFC de X. Dans cette section nous décrivons la loi limite de la suite $\Sigma_1, \Sigma_2, \ldots$. Nous utilisons les notations suivantes. Pour $n \geq 1$, soit :

$$Z_{ni} = \Big[X_{1i} \cdots X_{pi} \, \frac{X_{1i}X_{1i}}{X_{.i}} \, \frac{X_{2i}X_{1i}}{X_{.i}} \, \frac{X_{2i}X_{2i}}{X_{.i}} \cdots \frac{X_{pi}X_{pi}}{X_{.i}}\Big]' \in \mathbb{R}^{p+p(p+1)/2} \quad , \quad 1 \leq i \leq n \, .$$

Soit aussi f la fonction définie de $(0,\infty)^{p+p(p+1)/2}$ dans $(0,\infty)^{p(p+1)/2}$ par :

$$f\big([y_1 \cdots y_p \, y_{11} \, y_{21} \, y_{22} \, y_{31} \cdots y_{pp}]'\big) = \Big[\frac{y_{11}}{y_1} \, \frac{y_{21}}{y_2} \, \frac{y_{22}}{y_2} \, \frac{y_{31}}{y_3} \cdots \frac{y_{pp}}{y_p}\Big]' \, .$$

Pour une matrice A $p \times p$ symétrique, nous désignons par $\mathrm{vec}A$ le vecteur $[A_{11}A_{21}A_{22}A_{31} \cdots A_{pp}]' \in \mathbb{R}^{p(p+1)/2}$. Finalement, soit $\mathbb{V}(P)$ la matrice de variance-covariance de la loi P et $\mathbb{V}(Z_{n1})$ la matrice de variance-covariance de la variable Z_{n1} lorsque ces variances existent.

Avec ces notations nous obtenons le résultat suivant.

Théorème 2. *Si $P\big((0,\infty)^p\big) = 1$ et $\mathbb{V}(P)$ est finie, alors $\Sigma_1, \Sigma_2, \ldots$ est asymptotiquement normale. Plus précisément, nous avons :*

$$\sqrt{n}\,\mathrm{vec}\big(\Sigma_n - S\big) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}_{p(p+1)/2}(0,\Gamma) \, ,$$

avec $S = S(P) = f(\mathbb{E}(Z_{n1}))$ et $\Gamma = \Gamma(P) = \mathrm{D}f\big(S(P)\big)\,\mathbb{V}(Z_{n1})\,\mathrm{D}f\big(S(P)\big)'$ où $\mathrm{D}f(a)$ désigne la différentielle de la fonction f calculée au point a.

Observons que :

$$(S(P))_{k\ell} = \frac{1}{\mathbb{E}(X_{k1})}\,\mathbb{E}\Big(\frac{X_{k1}X_{\ell 1}}{X_{.1}}\Big) \qquad \text{pour} \quad 1 \leq k \leq p \quad \text{et} \quad 1 \leq \ell \leq p \, .$$

Preuve. Pour $n \geq 1$, remarquons que les Z_{ni}, $1 \leq i \leq n$, sont des variables aléatoires indépendantes et équidistribuées qui sont de variance finie puisque nous avons :

$$0 < X_{ki}X_{\ell i}/X_{.i} \leq X_{ki} \qquad \text{pour} \qquad 1 \leq k \leq p \ \text{ et } \ 1 \leq \ell \leq p \, .$$

Ainsi, posant, pour $n \geq 1$, $Y_n = 1/n \sum_{i=1}^n Z_{ni}$, la suite $Y_1, Y_2, \cdots$ est asymptotiquement normale par le théorème limite central. La preuve du théorème 2 est alors complétée en remarquant que, pour $n \geq 1$, on a $\mathrm{vec}\,\Sigma_n = f(Y_n)$, et en utilisant le lemme classique suivant, la fonction f étant différentiable sur $(0,\infty)^{p+p(p+1)/2}$. Pour une démonstration de ce lemme nous renvoyons à Mardia, Kent et Bibby (1994).

Lemme 1. *Soit $Y_1, Y_2, \ldots$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^K$ qui est asymptotiquement normale, c'est-à-dire qui vérifie :*

$$\sqrt{n}\,\big(Y_n - a\big) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}_K(0,V) \, ,$$

pour $a \in \mathbb{R}^K$ et V une matrice $K \times K$ symétrique positive. Soit aussi g une application définie de $\mathbb{R}^K$ dans $\mathbb{R}^L$ qui est différentiable au point a. Alors $g(Y_1), g(Y_2), \ldots$ est asymptotiquement normale. Plus précisément :

$$\sqrt{n}\,\big(g(Y_n) - g(a)\big) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}_L\big(0, \mathrm{D}g(a)V(\mathrm{D}g(a))'\big)\ ,$$

où $\mathrm{D}g(a)$ désigne la différentielle de g en a.

4 - Bootstrap et Analyse Factorielle des Correspondances

Pour $n \geq 1$, soit $\mathcal{L}_n(P)$ la loi de $\sqrt{n}\,\mathrm{vec}\big(\Sigma_n(P) - S(P)\big)$ et $\widehat{P}_n$ la loi empirique des observations, qui est définie pour tout borélien $\mathcal{A}$ de $\mathbb{R}^p$ par :

$$\widehat{P}_n(\mathcal{A}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \in \mathcal{A})\ ,$$

où $\mathbb{I}(\mathcal{B})$ désigne la fonction indicatrice de l'événement $\mathcal{B}$. Pour estimer $\mathcal{L}_n(P)$ par la méthode du bootstrap, on approche cette loi par la loi $\mathcal{L}_n(\widehat{P}_n)$ qui est elle-même estimée par une méthode de Monte-Carlo, c'est-à-dire, à partir de la simulation d'un grand nombre de n-échantillons bootstrapés $X_1^{\star i}, \ldots, X_n^{\star i}$ de loi $\widehat{P}_n$, pour $1 \leq i \leq I$.

Le résultat présenté maintenant est similaire au théorème 1 de Beran et Srivastava (1985). Il valide l'utilisation de la méthode du bootstrap dans le cadre qui est le nôtre ici.

Théorème 3. *Sous les hypothèses du théorème 2 on a :*

$$\mathbb{P}\Big(\mathcal{L}_n(\widehat{P}_n) \xrightarrow[n\to\infty]{\text{étroitement}} \mathcal{N}_{p(p+1)/2}(0,\Gamma)\Big) = 1\ .$$

Preuve. Notons, pour $P_1, P_2, \ldots$ une suite de lois de probabilité sur $(0,\infty)^p$:

$$P_n \underset{n\to\infty}{\Longrightarrow} P\ ,$$

si $P_1, P_2, \ldots$ converge étroitement vers P et si, pour tout $1 \leq k_1, k_2, k_3, k_4 \leq p$, tout $0 \leq i_1, i_2, i_3, i_4 \leq 1$ et tout $0 \leq i_5 \leq 2$ vérifiant $1 \leq i_1 + i_2 + i_3 + i_4 - i_5 \leq 2$:

$$\lim_{n\to\infty} \int \frac{x_{k_1}^{i_1} x_{k_2}^{i_2} x_{k_3}^{i_3} x_{k_4}^{i_4}}{\big(\sum_{j=1}^{p} x_j\big)^{i_5}}\, \mathrm{d}P_n(x_1, \ldots, x_p) = \int \frac{x_{k_1}^{i_1} x_{k_2}^{i_2} x_{k_3}^{i_3} x_{k_4}^{i_4}}{\big(\sum_{j=1}^{p} x_j\big)^{i_5}}\, \mathrm{d}P(x_1, \ldots, x_p)\ .$$

Observons que, sous les hypothèses du théorème 3, la loi des grands nombres implique :

$$\mathbb{P}\Big(\widehat{P}_n \underset{n\to\infty}{\Longrightarrow} P\Big) = 1\ .$$

Pour compléter la preuve du théorème 3, nous montrons maintenant que, pour toute suite $P_1, P_2, \ldots$ vérifiant $P_n \underset{n\to\infty}{\Longrightarrow} P$, on a $\mathcal{L}_n(P_n) \xrightarrow[n\to\infty]{\text{étroitement}} \mathcal{N}_{p(p+1)/2}(0,\Gamma)$.

Considérons donc, pour $n \geq 1$, $X^\dagger = [X_1^\dagger \cdots X_n^\dagger]$, où $X_1^\dagger, \ldots, X_n^\dagger$ sont des vecteurs aléatoires indépendants et équidistribués de loi P_n . Posons, pour $n \geq 1$:

$$Z_{ni}^\dagger = \Big[X_{1i}^\dagger \cdots X_{pi}^\dagger \; \frac{X_{1i}^\dagger X_{1i}^\dagger}{X_{.i}^\dagger} \; \frac{X_{2i}^\dagger X_{1i}^\dagger}{X_{.i}^\dagger} \; \frac{X_{2i}^\dagger X_{2i}^\dagger}{X_{.i}^\dagger} \cdots \frac{X_{pi}^\dagger X_{pi}^\dagger}{X_{.i}^\dagger}\Big] \in \mathbb{R}^{p+p(p+1)/2} \quad , \quad 1 \leq i \leq n \; ,$$

$M_n = \mathbb{E}(Z_{n1}^\dagger)$, V_n , la variance de $Z_{n1}^\dagger$, et, $Y_n^\dagger = 1/n \sum_{i=1}^n Z_{ni}^\dagger$. Les vecteurs aléatoires $Z_{n1}^\dagger, \ldots, Z_{nn}^\dagger$ sont indépendants et équidistribués de variance finie, de plus, $\lim_{n\to\infty} V_n = V$, où V est la variance commune aux Z_{ni} , $1 \leq i \leq n$, $n \geq 1$, définis dans la section précédente. Ainsi, en notant $\Phi_Z(t) = \mathbb{E}(\exp(it'Z))$, $t \in \mathbb{R}^{p+p(p+1)/2}$, la fonction caractéristique d'un vecteur aléatoire Z à valeurs dans $\mathbb{R}^{p+p(p+1)/2}$, nous avons, pour $t \in \mathbb{R}^{p+p(p+1)/2}$:

$$\Phi_{\sqrt{n}(Y_n^\dagger - M_n)}(t) = \Big(\Phi_{(Z_{n1}^\dagger - M_n)}\big(\frac{t}{\sqrt{n}}\big)\Big)^n = \Big(1 - \frac{1}{n} t'V_n t + o\big(\frac{t't}{n}\big)\Big)^n \xrightarrow[n\to\infty]{} exp\big(-\frac{t'Vt}{2}\big) \; ,$$

la convergence étant un résultat d'analyse classique. Pour la fonction f définie dans la section 3, il en résulte par le lemme 2 suivant qui est une extension facile du lemme 1 que :

$$\sqrt{n}\,\big(f(Y_n^\dagger) - f(M_n)\big) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}_{p(p+1)/2}(0, \Gamma) \; ,$$

c'est-à-dire, puisque $f(M_n) = f(\mathbb{E}(Z_{n1}^\dagger)) = S(P_n)$:

$$\mathcal{L}_n(P_n) \xrightarrow[n\to\infty]{\text{étroitement}} \mathcal{N}_{p(p+1)/2}(0, \Gamma) \; .$$

Ce qui complète la preuve du théorème 3.

Lemme 2 (extension du lemme 1). *Soit $Y_1, Y_2, \ldots$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^K$ qui est asymptotiquement normale dans le sens étendu suivant :*

$$\sqrt{n}\,\big(Y_n - a_n\big) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}_K(0, V) \; ,$$

pour $a_1, a_2 \ldots$ une suite d'éléments de $\mathbb{R}^K$ qui converge vers $a \in \mathbb{R}^K$ et V une matrice $K \times K$ symétrique positive. Soit aussi g une application définie de $\mathbb{R}^K$ dans $\mathbb{R}^L$ qui est uniformément différentiable et de différentielle continue dans un voisinage du point a . Alors $g(Y_1), g(Y_2), \ldots$ est asymptotiquement normale. Plus précisément :

$$\sqrt{n}\,\big(g(Y_n) - g(a_n)\big) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}_L\big(0, \mathrm{D}g(a) V (\mathrm{D}g(a))'\big) \; ,$$

où $\mathrm{D}g(a)$ désigne la différentielle de g en a .

Preuve. Nous écrivons :

$$\sqrt{n}\,\big(g(Y_n) - g(a_n)\big) =$$

$$\sqrt{n}\ \mathrm{D}g(a)(Y_n - a_n) + \sqrt{n}\ \big(\mathrm{D}g(a_n) - \mathrm{D}g(a)\big)(Y_n - a_n) + \sqrt{n}\ ||Y_n - a_n||\ \varepsilon_{a_n}(Y_n - a_n)\ ,$$

où ε_{a_n} est une suite de fonctions qui est uniformément continue en 0 pour a_n assez proche de a et $||.||$ est la norme euclidienne usuelle sur $\mathbb{R}^K$. On applique alors les hypothèses et les théorèmes de Slutsky pour conclure la preuve du lemme 2.

5 - Intervalles de confiance et tests sur les valeurs propres issues d'une AFC

Faisons pour commencer l'hypothèse que les valeurs propres de la matrice $S(P)$ sont toutes distinctes. Sous cette hypothèse, la fonction associant ses valeurs propres à une matrice $p \times p$ et symétrique est différentiable au point $S(P)$. Grâce au lemme 1, il en résulte (cf. Beran et Srivastava (1985) pour plus de détails) le corollaire suivant du théorème 3. Pour une matrice A $p \times p$ et symétrique, nous notons $\lambda_1(A) \geq \cdots \geq \lambda_p(A)$ ses valeurs propres, et, pour $X_1^\star, \ldots, X_n^\star$ des vecteurs aléatoires indépendants et équidistribués de loi $\widehat{P}_n$, nous notons $X^\star = [X_1^\star \cdots X_n^\star]$ et $\Sigma_n^\star = \Sigma_n(\widehat{P}_n) = \mu(X^\star)^{-1} X^\star \nu(X^\star)^{-1} (X^\star)'$.

Corollaire 1 (intervalles de confiance simultanés pour les valeurs propres). *Sous les hypothèses du théorème 3, et si les valeurs propres de $S(P)$ sont toutes distinctes, alors, pour $0 < \beta < 1$:*

$$\lim_{n\to\infty} \mathbb{P}\Big(\lambda_k(\Sigma_n) - a_n^\star \leq \lambda_k(S(P)) \leq \lambda_k(\Sigma_n) + a_n^\star \quad \textit{simultanément pour } 1 \leq k \leq p\Big) = 1 - \beta\ ,$$

lorsque $a_n^\star$ vérifie :

$$\mathbb{P}\Big(\max_{1\leq k\leq p} |\lambda_k(\Sigma_n^\star) - \lambda_k(S(\widehat{P}_n))| \leq a_n^\star\Big) = 1 - \beta\ .$$

Rappelons que pour déterminer approximativement $a_n^\star$ on utilise une méthode de Monte-Carlo en simulant un grand nombre d'échantillons bootstrapés (de loi $\widehat{P}_n$).

Le premier test que nous proposons est lui aussi construit dans le cadre de l'hypothèse les valeurs propres de la matrice $S(P)$ sont toutes distinctes. Rappelons que, pour $1 \leq \ell \leq p$, la statistique $\sum_{k=1}^{\ell} \alpha_k^2 / \sum_{k=1}^{p} \alpha_k^2 = \sum_{k=1}^{\ell} \lambda_k(\Sigma_n) / \sum_{k=1}^{p} \lambda_k(\Sigma_n)$ que nous utilisons mesure la proportion de distance expliquée quand on retient l'approximation de rang ℓ de X construite à partir de L'AFC (cf. section 2). Le résultat suivant permet d'utiliser la méthode du bootstrap pour construire un test de l'hypothèse la proportion de distance expliquée est plus grande qu'un seuil fixé a priori. Il se prouve comme le corollaire 1.

Corollaire 2. *Sous les hypothèses du théorème 3, et si les valeurs propres de $S(P)$ sont toutes distinctes, alors, pour $0 < \beta < 1$:*

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{\sum_{k=1}^{\ell} \lambda_k(S(P))}{\sum_{k=1}^{p} \lambda_k(S(P))} \geq \frac{\sum_{k=1}^{\ell} \lambda_k(\Sigma_n)}{\sum_{k=1}^{p} \lambda_k(\Sigma_n)} - b_n^\star \quad \textit{simultanément pour } 1 \leq \ell \leq p\Big) = 1 - \beta\ ,$$

lorsque $b_n^\star$ vérifie :

$$\mathbb{P}\Big(\max_{1\leq \ell\leq p} \Big[\frac{\sum_{k=1}^{\ell} \lambda_k(S(\widehat{P}_n))}{\sum_{k=1}^{p} \lambda_k(S(\widehat{P}_n))} - \frac{\sum_{k=1}^{\ell} \lambda_k(\Sigma_n^\star)}{\sum_{k=1}^{p} \lambda_k(\Sigma_n^\star)}\Big] \geq -b_n^\star\Big) = 1 - \beta\ .$$

Nous proposons maintenant un test de l'hypothèse les $p-\ell$ plus petites valeurs propres de $S(P)$ sont nulles. Dans ce cadre on ne peut plus supposer que les valeurs propres de $S(P)$ sont toutes distinctes et c'est la fonction associant les polynômes symétriques de ses valeurs propres à une matrice $p \times p$ symétrique qui est différentiable au point $S(P)$ (cf. Beran et Srivastava (1985)). Pour une matrice A $p \times p$ symétrique, soit $\Lambda_1(A), \ldots, \Lambda_p(A)$ les p polynômes symétriques élémentaires des valeurs propres, définis par :

$$\Lambda_1(A) = \sum_{k=1}^{p} \lambda_k(A) \ , \ \Lambda_2(A) = \sum_{1 \leq k < \ell \leq p} \lambda_k(A)\lambda_\ell(A) \ , \ \ldots \ , \ \lambda_p(A) = \prod_{k=1}^{p} \lambda_k(A) \ .$$

Remarquons que les $p-\ell$ dernières valeurs propres de $S(P)$ sont nulles si et seulement si les $p-\ell$ derniers $\Lambda_k(S(P))$ sont nuls. Ainsi, on peut tester que les $p-\ell$ dernières valeurs propres de $S(P)$ sont nulles à partir d'un test de nullité des $p-\ell$ derniers $\Lambda_k(S(P))$. En procédant comme pour les deux corollaires précédents, on obtient le résultat suivant.

Corollaire 3 (intervalles de confiance simultanés pour les polynômes symétriques des valeurs propres). *Sous les hypothèses du théorème 3, on a , pour* $0 < \beta < 1$:

$$\lim_{n \to \infty} \mathbb{P}\Big(\Lambda_k(\Sigma_n) - c_n^\star \leq \Lambda_k(S(P)) \leq \Lambda_k(\Sigma_n) + c_n^\star \quad \textit{simultanément pour } 1 \leq k \leq p\Big) = 1 - \beta \ ,$$

lorsque $c_n^\star$ *vérifie :*

$$\mathbb{P}\Big(\max_{1 \leq k \leq p} |\Lambda_k(\Sigma_n^\star) - \Lambda_k(S(\widehat{P}_n))| \leq c_n^\star\Big) = 1 - \beta \ .$$

Pour tout $1 \leq \ell \leq p$, on en déduit un test de niveau de confiance asymptotique inférieur à $1-\beta$ de l'hypothèse les $p-\ell$ dernières valeurs propres de $S(P)$ sont nulles en prenant comme région de rejet de ce test la région :

$$max_{\ell < k \leq p}[\Lambda_k(\Sigma_n) - c_n^\star] > 0 \ .$$

Bibiographie

- Beran R. et Srivastava M.S., Annals of Statistics 13 (1985) 95-115 : "Bootstrap tests and confidence regions for functions of a covariance matrix".
- Besse P., Statistics and Probability Letters 13 (1992) 405-410 : "PCA stability and choice of dimensionality".
- Efron B. et Tibshirani R.J., An Introduction to the Bootstrap, Chapman and Hall (1993).
- Eckart C. et Young G., Psychometrika 1 (1936) 211-218 : "The approximation of one matrix by another of lower rank".
- Greenacre M.J., Theory and Application of Correspondence Analysis, Academic Press (1984).
- Hall P., The Bootstrap and Edgeworth expansion, Springer-Verlag (1992).
- Jolliffe I.T., Principal Component Analysis, Springer-Verlag (1986).
- Mardia K.V., Kent J.T. et Bibby J.M., Multivariate Analysis, Academic Press (1994).
- Shao J. et Tu D., The Jackknife and Bootstrap, Springer-Verlag (1995).