N° d'ordre : 25/ 2013-M/MT

République Algérienne Démocratique et Populaire

Ministère de l'enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté des Mathématiques

Mémoire

Présenté Pour L'Obtention du Dipôlme de MAGISTER

En Mathématiques

Spécialité : Statistique Mathématique & Probabilité

Par : NEGGAZI Dalila

Thème

Etude Asymptotique Des Composantes Principales

Soutenu Publiquement le 13/06/2013, devant le jury composé de :

M^{me}K. Djaballah	Maître de conférences/A	à l'USTHB	Présidente
M M. Djedour	Professeur	à l'USTHB	Directeur de Mémoire
M A.TATACHAK	Maître de conférences/A	à l'USTHB	Examinateur

Table des matières

Table des figures

Remerciements

Tout d'abord je tiens à remercier Dieu de m'avoir donné le courage, le moral et la santé pour mener à bien ce travail.

Je tiens à remercier avec tous mes sentiments de respectueuse gratitude mon promoteur M. Mohamed Djedour Professeur à l'USTHB pour sa proposition de sujet ainsi pour son soutien, ses orientations et ses précieux conseils.

J'exprime aussi ma profonde gratitude à Mme. Khadidja Djaballah Maître de conférences à l'USTHB, pour avoir accepté de présider le jury de soutenance.

Je remercie également : M. Abdelkader Tatachak Maître de conférences à l'USTHB, pour avoir accepté d'examiner ce mémoire.

Enfin, je remercie chaleureusement toutes les personnes qui m'ont aidé, et qui ont contribué d'une manière ou d'une autre à la réalisation de ce travail.

Introduction Generale

Dans ce mémoire on s'intéresse à l'étude asymptotique des composantes principales notamment les valeurs propres et les vecteurs propres d'une population (P) décrite par p variables quantitatives de moyenne μ et de matrice de variance-covariance $\sum$.

L'étude consiste à analyser et étudier asymptotiquement les composantes principales à partir des mesures statistiques réalisées sur des échantillons.

A cet effet trois méthodes sont exposées et évaluées sur des exemples illustratifs

1-Méthode classique ou asymptotique

2-Méthode du Bootstrap paramétrique et

3-Méthode du Bootstrap non paramétrique

Dans le premier chapitre, la plupart des propriétés mathématiques et statistiques des composantes principales sont discutées, basées sur matrice de covariance (ou de corrélation) $\sum$ connue d'une population.

Le deuxième chapitre traite les propriétés des composantes principales obtenues à partir de la matrice de covariance (Ou de corrélation) d'un échantillon de taille n, les propriétés qui ne sont pertinentes que pour les composantes principales de l'échantillon seront discutées plus en détail.

On étudie par la suite les distributions de probabilité des vecteurs propres et des valeurs propres d'une matrice de covariance de l'échantillon et on montre comment ces distributions peuvent être utilisées pour faire des inférences statistiques sur les composantes principales de la population, sur la base des composantes principales de l'échantillon.

Le troisième chapitre aborde la théorie du Bootstrap.

Comme les méthodes classiques d'inférence statistique (méthode du maximum de vraisemblance utilisée dans le chapitre2) ne permettent pas d'obtenir des réponses correctes à tous les problèmes concrets que se pose l'utilisateur car elles ne sont en effet valables que sous des conditions d'application particulières. On a recours à l'utilisation de la technique du Bootstrap pour les problèmes statistiques liés à l'estimation des paramètres. Cette méthode de réechantillonage est basée sur l'utilisation intensive de l'ordinateur.

Le quatrième chapitre intitulé application est une illustration sur trois exemples (méthode classique asymptotique, méthode bootstrap paramétrique, et méthode bootstrap non paramétrique).

Les deux premières méthodes"méthode classique asymptotique"et "Méthode bootstrap paramétrique traitent le cas de la distribution connue d'une population tandis que la troisième méthode " bootstrap non paramétrique" traite le cas de la distribution inconnue d'une population.

On évalue par la suite les résultats obtenus par ces differentes méthodes, en examinant le biais, les erreurs standards et les intervalles de confiance des estimateurs.

Pour mener à terme cette illustration nous avons élaboré trois programmes informatiques sous R.

Le premier appelé "**NMéthod**" pour la méthode classique asymptotique, le second nommé "**sim.boot** " pour la méthode Bootstrap paramétrique, et le troisième nommé "**data.boot**" pour la méthode bootstrap non paramétrique.

En conclusion on peut estimer que dans l'ensemble les méthodes du bootstrap ont donné plus de satisfaction par rapport à la méthode asymptotique classique.

Nous avons evalué les estimateurs des differentes méthodes suivant les critères de comparaison :biais, intervalles de confiance et erreurs standards.

Chapitre 1

Propriétés des composantes principales d'une population

1.1 Introduction

Dans ce chapitre, la plupart des propriétés mathématiques et statistiques de composantes principales sont rappelées, basées sur la matrice de covariance (ou de corrélation) $\sum$ connue d'une population. D'autres propriétés sont incluses dans le chapitre 2, mais dans le contexte de Composantes principales de l'échantillon, plutôt que les Composantes principales de la population. En plus d'être issu d'un point de vue statistique, les Composantes principales peuvent être trouvées en utilisant des arguments purement mathématiques ; elles sont données par une transformation orthogonale linéaire d'un ensemble de variables d'optimisation d'un certain critère algébrique. En fait, les Composantes principales optimisent plusieurs différents critères algébriques et ces propriétés d'optimisation, ainsi que leurs implications statistiques, sont décrites dans la troisième section du chapitre.

En plus du calcul algébrique, les Composantes principales peuvent également être regardées d'un point de vue géométrique. Le calcul donné dans le document original sur l'ACP par Pearson (1901) est géométrique, mais il est pertinent d'échantillons, plutôt que de populations, et sera donc reportée au chapitre 2.

Cependant, d'autres propriétés des Composantes principales de la population sont également géométriques dans la nature et elles sont examinées dans la quatrième section de ce chapitre (voir page 13).

1.2 Définition des composantes principales d'une population

Nous abordons ici brièvement la définition des composantes principales d'une population. Nous introduisons en particulier leurs proprtiétés dont nous nous servirons tout au long de ce mémoire. Pour plus de détails, nous renvoyons par exemple à (réf[2]) et (réf[6]), ainsi qu'aux références qui y sont citées.

1.2.1 Recherche d'une combinaison linéaire à variance maximum

On suppose un vecteur aléatoire X centré, suit une loi normale à p composantes et de matrice de covariance $\sum$ (réf[1]).

Soit α un vecteur colonne à p composantes tel que : $\alpha'\alpha = 1$, la variance d'une combinaison linéaire $\alpha' X = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_P x_p$ est :

$$\mathbf{E}(\boldsymbol{\alpha}'\mathbf{X})^2 = \mathbf{E}(\boldsymbol{\alpha}'\mathbf{X}\mathbf{X}'\alpha) = \boldsymbol{\alpha}' \sum \boldsymbol{\alpha} \tag{1.1}$$

Pour déterminer la combinaison linéaire normée à variance maximum, nous devons trouver un vecteur α rendant maximum (1.1) ; satisfaisant à $\alpha'\alpha$=1 ; Soit :

$$\Phi = \boldsymbol{\alpha}' \sum \boldsymbol{\alpha} - \lambda(\alpha'\alpha - 1) = \sum_{i,j} \boldsymbol{\alpha}_i \sigma_{i;j} \boldsymbol{\alpha}_j - \lambda \sum_i (\boldsymbol{\alpha}_i^2 - 1) \tag{1.2}$$

Où λ est multiplicateur de Lagrange. Le vecteur de la dérivation partielle $(\partial\Phi/\partial\boldsymbol{\alpha})$ est :

$$\partial\Phi/\partial\boldsymbol{\alpha} = \mathbf{2} \sum \boldsymbol{\alpha} - \mathbf{2}\lambda\boldsymbol{\alpha} \tag{1.3}$$

Un vecteur α rendant maximum $\boldsymbol{\alpha}' \sum \boldsymbol{\alpha}$ doit satisfaire à : $\partial\Phi/\partial\boldsymbol{\alpha} = \mathbf{0}$, c'est-à-dire

$$(\sum -\lambda I)\alpha = 0 \tag{1.4}$$

Pour avoir une solution à (1.4) avec $\alpha'\alpha = 1$, $(\sum -\lambda I)$ doit être singulière, autrement dit λ doit satisfaire à :

$$\left|\sum -\lambda I\right| = 0 \tag{1.5}$$

La fonction $|\sum -\lambda I|$ est polynomiale en λ de degré p donc (1.5) de degré « p »

Supposons $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, on prémultipliant (1.4) par $\boldsymbol{\alpha}'$, on obtient :

$$\boldsymbol{\alpha}' \sum \boldsymbol{\alpha} = \lambda \alpha' \alpha = \lambda \tag{1.6}$$

Donc si α satisfait à (1.4) (avec $\alpha'\alpha$ =1), alors la variance de $\alpha' X$ donnée dans (1.1) est "λ" , pour qu'elle soit maximum on doit donc choisir λ_1. Soit $z_{_1} = \alpha_1' X$, cette combinaison linéaire, où $\boldsymbol{\alpha}_1$est le vecteur propre correspondant à λ_1.

1.2.2 **Recherche d'autres composantes**

Cherchons Maintenant une combinaison linéaire normée $\alpha' X$ à variance maximum parmi toutes les combinaisons linéaires non corrélées avec z_1 ; la corrélation nulle s'exprime par

$$0 = E(\alpha' X z_1) = E(\alpha' X X' \alpha_1) \tag{1.7}$$

$$0 = \boldsymbol{\alpha}' \sum \boldsymbol{\alpha}_1 = \lambda_1 \alpha' \alpha_1 \tag{1.8}$$

Cette dernière égalité est déjà pleine d'informations puisque la non-corrélation se traduit par une orthogonalité statistique (corrélation nulle) et une orthogonalité géométrique ($\alpha'\alpha_1 = 0$) puisque $\lambda_{_1} \neq 0$ sauf dans un cas trivial sans intérêt.

On désire donc rendre maximum

$$\Phi_2 = \boldsymbol{\alpha}' \sum \boldsymbol{\alpha} - \lambda(\alpha'\alpha - 1) - 2v_1 \boldsymbol{\alpha}' \sum \boldsymbol{\alpha}_1 \tag{1.9}$$

Où λ et v_1 sont les multiplicateurs de Lagrange, le vecteur des dérivées partielles

$$(\partial\Phi_2)/\partial\alpha = 2\sum\alpha - 2\lambda\alpha - 2v_1\sum\alpha_1 \tag{1.10}$$

la prémultiplication par α_1'de $(\partial\Phi_2)/\partial\alpha = 0$ fournit

$$0 = 2\alpha_1'\sum\boldsymbol{\alpha} - 2\lambda\alpha_1'\boldsymbol{\alpha} - 2v_1\alpha_1'\sum\alpha_1 \tag{1.11}$$

$$= -2v_1\lambda_1 \quad \text{à partir de (1.7)} \tag{1.12}$$

Donc $v_1 = 0$ et α satisfait à la même équation (1.4) que précédemment.

Anderson (1958) démontre le théorème suivant :

Théorème 1 *Si le vecteur aléatoire X est tel que* $E(X) = 0$ *et* $E(XX') = \sum$. *Il existe une transformation orthogonale* $z = A'X$ *telle que matrice de covariance de z ;* $E(zz') = \Lambda$ *(réf[2])*

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda p \end{pmatrix} \tag{1.13}$$

Où $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ *sont les racines de* $|\sum -\lambda I| = 0$. *La* $j^{ième}$ *colonnes* α_j *de A satisfait à* $(\sum -\lambda_j I)$ α_j*=0, La* $j^{ième}$ *composante de z;* $z_j = \alpha_j' X_j$ *à la variance maximum parmi toutes les combinaisons linéaires normalisées non corrélées avec* $z_1, z_2, \ldots, z_{j-1}$.

1.2.3 Autre présentation des composantes principales

En notation matricielle les composantes principales z_1, z_2,..., z_p dont le vecteur colonne sera noté Z s'écrivent : Z=A′X où A′ est la matrice transposée de A.

Comme A est orthogonale : $X = AZ$.

Supposons $\sum$ régulière, c'est-à-dire les z_j linéairement indépendant et définissons les variables y_j : $y_j = \frac{1}{\sqrt{\lambda_j}} z_j$

Elles sont non corrélées et de variance 1 ; et X s'écrit X=A $\Lambda^{1/2}$y

Où

$$\Lambda^{1/2} = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p})$$

Autrement dit : $X = BY$

$$\begin{cases} X_1 = b_{11}y_1 + b_{12}y_2 + \ldots + b_{1p}y_p \\ \vdots \\ X_p = b_{p_1}y_1 + b_{p_2}y_2 + \ldots + b_{pp}y_p \end{cases} \tag{1.14}$$

$\mathrm{Var}(X_i) = \sum_j b_{ij}^2$ pour $i = 1, \ldots, p$

$\sum_i b_{ij}^2 = \lambda_j$ pour $j = 1, \ldots, p$

On a donc décomposé les X_i selon un système y_j de variables non corrélées de variance 1 , de façon à maximiser la « portion de variance » du système X expliquée par la 1^{ere} variable y_1, puis la seconde, ect...

Si le rang de $\sum$ est inférieur à p, soit r, les $z_{(r+1)}, z_{(r+2)}, \ldots, z_p$ sont de variance nulle, donc presque surement nulles, les mêmes équations sont vraies en remplaçant par zéro les variables $y_{(r+1)}, y_{(r+2)}, \ldots, y_p$ (réf[6]).

1.2.4 Invariants statistiques dans la transformation en composantes principales

Théorème 2 *une transformation orthogonale Y=CX d'un vecteur aléatoire X conserve la variance généralisée et la somme des variances des composantes. Soit* $E(X) = 0, E(XX') = \sum$ *, alors* $E(Y) = 0$, $E(YY') = C \sum C'$ *(réf[2])*

La variance généralisée de Y est : $|C \sum C'| = |\sum| \, |CC'| = |C' \sum C| = |\sum|$

La somme des variances des composantes de Y :

$\sum_{i=1}^{p} E\{y_i^2\} = tr\{C \sum C'\} = tr\{\sum C'C\} = tr\{\sum\} = \sum_{i=1}^{p} E\{X_i^2\}$

Corollaire 3 *La variance généralisée d'un vecteur de composantes principales est la variance généralisée du vecteur origine, et la somme des variances des composantes principales est la somme des variances des variables d'origines (réf[6]).*

1.3 Propriétés algébriques des composantes principales d'une population

Il existe plusieurs propriétés statistiques et mathématiques basées sur la matrice de var-cov $\sum$ d'une population.

Soit :

$$Z = A' \, X \tag{1.15}$$

Où A : Matrice orthogonale à k colonnes, α_k est le k^{eme} vecteur propre de $\sum$ associé à la k^{eme} valeur propre. Ainsi les composantes principales sont définies par une transformation linéaire orthonormale de X.

A partir de la définition des composantes principales nous avons :

$$\sum A = A\Lambda \tag{1.16}$$

Où Λ : Matrice diagonale des valeurs propres λ_k de $\sum$

Et

$$\lambda_k = var(\alpha_k' X) = var(Z_k)$$

$$A' \sum A = \Lambda \tag{1.17}$$

Et

$$\sum = A\Lambda A' \tag{1.18}$$

Propriété $\mathbf{P}_1$

-Pour tout entier $q : 1 \leq q \leq p$; on considère la transformation linéaire :

$$y = \underset{(q\times p)}{B'} X \tag{1.19}$$

y : vecteur à q éléments, et $\sum_y = B' \sum B$:matrice Var Cov de y Alors :

$tr(\sum_y)$ est maximisée en prenant $B = A_q$ où A_q : sont les q premières colonnes de A.

Preuve

Soit β_k : le k^{eme} élément de B, comme les colonnes de A forment une base pour un sous espace à p dimensions, nous avons :

$$\beta_k = \sum_{j=1}^{p} c_{jk}\alpha_j \quad pour \quad k = 1, 2, \ldots q$$

Où c_{jk} sont définis constants à priori. Donc $B = \underset{(p\times q)}{A\ C}$

Où la matrice C d'éléments c_{jk} et :

$$B' \sum B = C'A' \sum AC = C'\Lambda C$$

utilisons (1.19)

$$= \sum_{j=1}^{p} \lambda_j c_j c_j'$$

$$tr(B' \sum B) = \sum_{j=1}^{p} \lambda_j tr(c_j c_j')$$

$$= \sum_{j=1}^{p} \lambda_j tr(c_j' c_j)$$

$$= \sum_{j=1}^{p} \lambda_j c_j' c_j$$

$$= \sum_{j=1}^{p}\sum_{k=1}^{q}\lambda_j c_{jk}^2 \tag{1.20}$$

Or $C = A'B$, d'où $C'C = B'AA'B = B'B = I_q$

Comme A est orthogonale et les colonnes de B sont orthonormées,

d'où

$$\sum_{j=1}^{p}\sum_{k=1}^{q}c_{jk}^2 = q \tag{1.21}$$

$$\sum_{k=1}^{q}c_{jk}^2 \leq 1 \tag{1.22}$$

$\sum_{k=1}^{q}c_{jk}^2 \leq 1$ est le coefficient de λ_j dans (1.20), la somme de ces coefficients est 'q' de (1.21), et aucun de ces coefficients ne peut dépasser '1 'de (1.22).

$\sum_{j=1}^{p}\sum_{k=1}^{q}c_{jk}^2$ se maximise si on peut trouver un ensemble C_{jk} dont :

$$\sum_{k=1}^{q}c_{jk}^2 = \begin{cases} 1 \quad \text{pour} \quad j = 1, \ldots, q \\ 0 \quad \text{pour } j = q+1, \ldots, p \end{cases} \tag{1.23}$$

Mais si $B' = A'_q$ alors :

$$\sum_{k=1}^{q}c_{jk} = \begin{cases} 1 \quad \text{pour} \quad j = 1, \ldots, q \\ 0 \quad \text{pour } j = q+1, \ldots, p \end{cases}$$

qui satisfait à (1.23) d'où :

$tr(\sum_y)$ Atteint sa valeur maximale quand $B' = A'_q$.

Propriété $\mathbf{P}_2$

-On considère encore la transformation orthonormale $y = \underset{(q\times p)}{B'} X$

X, B, $\sum_y$ sont définis comme précédemment, alors $tr(\sum_y)$ est minimisée en prenant $B = A_q^*$ où A_q^* sont les q dernière colonnes de A (réf[1])

Preuve

Les arguments utilisés dans P_1 peuvent êtres adaptés à P_2.

Implications statistiques

– Peut aider à détecter relations linéaires constantes entre les éléments de X.

– Peut être utilisée aussi .dans la régression.

– sélection d'un sous ensemble de variables à partir de X.

– Détection des valeurs aberrantes.

Propriété $\mathbf{P}_3$ (Décomposition spectrale de $\sum$)

$$\sum = \lambda_1\alpha_1\alpha_1' + \lambda_2\alpha_2\alpha_2' + \ldots \lambda_p\alpha_p\alpha_p' \tag{1.24}$$

Preuve

$\sum = A\Lambda A'$ D'après (1.18), et le produit matriciel donne :

$\sum = \sum_{k=1}^{p} \lambda_k\alpha_k\alpha_k'$

$Var(X_j) = \sum_{k=1}^{p} \lambda_k\alpha_{kj}^2$ les éléments de la diagonale de $\sum$.

Implications statistiques

– Non seulement on peut décomposer les variances des éléments de X en contribution décroissante de chaque composante, mais on peut aussi décomposer la matrice de covariance en contribution de chaque composante $\lambda_k\alpha_k\alpha_k'$ qui devient petite quand k augmente (réf[1])

Propriété $\mathbf{P}_4$

Idem aux propriétés 1 et 2, on considère la transformation :

$y = B'X$, $\det(\sum_y)$ est maximisé quand $B = A_q$

Preuve

-Pour tout entier k de 1à q, soit S_k : sous espace à p dimensions orthogonal aux vecteurs $\alpha_1, \alpha_2, \ldots, \alpha_{k-1}$ alors la $\dim(S_k) = p - k + 1$, la k^{eme} valeur propre de $\sum$ satisfait :

$$\lambda_k = \underbrace{Sup}_{\alpha \in S_k, \alpha \neq 0} \left\{ \frac{\alpha' \sum \alpha}{\alpha' \alpha} \right\}$$

-Supposons que $\mu_1 > \mu_2 > \cdots > \mu_q$ sont les valeurs propres de $B' \sum B$ et $\gamma_1, \gamma_2, \ldots, \gamma_q$ sont les vecteurs propres correspondants.

Soit T_k : sous espace à q dimensions orthogonale aux vecteurs $\gamma_{k+1}, \ldots, \gamma_q$ avec $\dim(T_k) = K$, alors pour chaque vecteurs γ différents de zéro dans T_k :

$\frac{\gamma' B' \sum B\gamma}{\gamma'\gamma} \geqslant \mu_k$

On considère le sous espace $\widetilde{S}_k$ à p dimensions de la forme $B\gamma$ pour γ dans T_k.

A partir du résultat général concernant dimension de deux espaces vectoriels on a :

$\dim(S_k \cap \widetilde{S}_k) + \dim(S_k + \widetilde{S}_k) = \dim S_k + \dim \widetilde{S}_k$

Mais $\dim(S_k + \widetilde{S}_k) \leqslant p, \dim(S_k) = p - k + 1$ et $\dim(\widetilde{S}_k) = k$

D'où $\dim(S_k + \widetilde{S}_k) \geqslant 1$.

-il existe donc un vecteur α non nul dans S_k de la forme $\alpha = B\gamma$ pour γ dans T_k. et il s'ensuit que :

$$\mu_k \leq \frac{\gamma' B' \sum B\gamma}{\gamma'\gamma} = \frac{\gamma' B' \sum B\gamma}{\gamma' B' B\gamma} = \frac{\alpha' \sum \alpha}{\alpha' \alpha} \leq \lambda_k$$

Ainsi k^{eme} valeur propre de $B' \sum B \leq$ k^{eme} valeur propre de $\sum$ pour $k = 1 \ldots q$ d'où :

$\det(\sum_y) = \prod_{k=1}^{q}$(k valeurs propres de $B' \sum B) \leq \prod_{k=1}^{q} \lambda_k$

Mais si $B = A_q$ alors les valeurs propres de $B' \sum B$ sont $\lambda_1, \ldots, \lambda_q$, d'où

$\det(\sum_y) = \prod_{k=1}^{q} \lambda_k$

Dans ce cas $\det(\sum_y)$ est maximisé quand $B = A_q$ (réf[1]).

Implications statistiques

– Pour une distribution normale multivariée X, les q premières composantes sont q fonctions linéaires de X.

1.4 Propriétés géométriques des composantes principales d'une population

Propriété G_1

-Considérer la famille des ellipsoïdes de dimensions P

$$X' \sum^{-1} X = constant \tag{1.25}$$

-Les composantes principales définissent les axes principaux de ces ellipsoïdes (réf[6]).

Preuve

Les composantes principales $Z = A'X$ et comme A est orthogonale, l'inverse de la transformation est X=AZ, on le remplace dans (1.25) :

$$(AZ)' \sum^{-1} (AZ) = constant = (Z'A') \sum^{-1} (AZ)$$

Il est bien connu que les vecteurs propres de $\sum^{-1}$ sont les mêmes que ceux de $\sum$ et les valeurs propres de $\sum^{-1}$sont les inverses de celles de $\sum$ (réf[1])

-supposons que ces valeurs sont toutes strictement positives, il s'ensuit donc à partir du résultat (1.17), que $A' \sum^{-1} A = \Lambda^{-1}$ et donc $Z'\Lambda^{-1}Z = constant$

Cette dernière équation peut s'écrire :

$$\sum_{k=1}^{P} \frac{Z_k^2}{\lambda_k} = constant \tag{1.26}$$

L'équation (1.26) implique que les demi longueurs des axes principaux sont proportionnelles à $\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_p^{1/2}$.

Ce résultat est statistiquement important si le vecteur aléatoire X à une distribution normale multivariée. Dans ce cas l'ellipsoïde donné par (1.25) définit le contour d'une probabilité constante pour la distribution de X (réf[1]).

Le premier axe principal (le plus grand) de cet ellipsoïdes définit la direction dans laquelle la variation statistique est la plus grande.

Le second axe principal maximise la variation statistique sous la condition qu'il soit orthogonal au 1^{er} axe.

Propriété G_2

-On suppose que X_1, X_2 sont deux vecteurs aléatoires, ont même distribution de probabilité q, et que X_1, X_2 ont même transformation linéaire : $y_i = B'X_i$, $i = 1, 2$

Si $\underbrace{B}_{p\times q}$ à colonnes orthonormales choisi à maximiser $E(y_1 - y_2)'(y_1 - y_2)$ Alors $B = A_q$. (on garde les mêmes notations).

Preuve

-On note que X_1, X_2 ont même moyenne "μ" et matrice de covariance $\sum$, d'où y_1, y_2 ont aussi même moyenne et matrice de covariance $B'\mu$, $B'\sum B$ respectivement.

$$E(y_1 - y_2)'(y_1 - y_2) = E\{[(y_1 - B'\mu) - (y_2 - B'\mu)]'[(y_1 - B'\mu) - (y_2 - B'\mu))]\}$$
$$= E[(y_1 - B\prime\mu)'(y_1 - B'\mu)] + E[(y_2 - B'\mu)'(y_2 - B'\mu)]$$

Le produit croisé des termes disparait à cause de l'indépendance de X_1, X_2 et par conséquent y_1, y_2.

Maintenant pour i =1, 2 , nous avons :

$E[(y_i - B'\mu)'((y_i - B'\mu)] = E[tr(\ y_i - B'\mu)'(\ y_i - B'\mu)]$

$= E[tr(\ y_i - B'\mu)(\ y_i - B'\mu)']$

$= trE[(\ y_i - B'\mu)(\ y_i - B'\mu)']$

$= tr(B'\sum B)$

Mais la $tr(B' \sum B)$ est maximisée quand $B = A_q$ (d'après P_1) et $det\{E(y_1 - y_2)(y_1 - y_2)' \}$ est maximisé quand $B = A_q$; cette propriété indique que $B = A_q$ fait la variance généralisée de $(y_1 - y_2)$ aussi large que possible (réf[2]).

La propriété peut être réservée dans le sens que si $E(y_1 - y_2)'(y_1 - y_2)$ ou $det\{E(y_1 - y_2)(y_1 - y_2)'\}$ doit être minimisé, alors cela peut être réalisé en prenant $B = A_q^*$.

1.5 **Conclusion**

Nous avons étudié dans ce chapitre les propriétés les plus importantes des composantes principales de la population, cependant dans la pratique et le plus souvent la population est inconnue et on cherche à l'estimer à partir des caractériques de l'échantillon, pour cela nous allons étudiés dans le chapitre suivant les propriétés des composantes principales de l'échantillon et nous montrons comment à partir des composantes principales de l'échantillon nous estimons celles de la poplution en se basant sur l'etude asymptotique.

Chapitre 2

Propriétés des composantes principales d'un échantillon et étude asymptotique

2.1 Introduction

Ce chapitre présente une structure similaire au chapitre 1, à l'exception qu'il traite les propriétés des Composantes principales obtenues à partir de matrice de covariance (Ou de corrélation) d'un échantillon, plutôt que de matrice de covariance (ou de corrélation) d'une population. Les deuxième et troisième sections, comme dans le chapitre 1, décrivent, respectivement, de nombreuses propriétés algébriques et géométriques des composantes principales. La plupart des propriétés décrites au chapitre 1 sont presque les mêmes pour les échantillons que pour les populations. Elles seront à nouveau mentionnées, mais seulement brièvement. Il ya, en outre, certaines propriétés qui ne sont pertinentes que pour composantes principales de l'échantillon, et celles-ci seront discutées plus en détail.

La section 2.4 traite de la décomposition en valeurs singulières, qui auraient pu être incluse dans la section 2.2 comme une propriété supplémentaire algébrique.

Cependant, le sujet est suffisamment important pour justifier sa propre section, comme elle fournit une approche alternative utile à une partie de la théorie entourant composantes principales, et donne aussi une méthode efficace pratique pour en calculer les composantes principales.

La cinquième section de ce chapitre presente une etude asymptotique, dont le premier paragraphe examine les distributions de probabilité des coefficients et les variances d'un ensemble des composantes principales d'échantillon, en d'autres termes, les distributions de probabilité des vecteurs propres et des valeurs propres d'une matrice de covariance de l'échantillon.

Le paragraphe 2.5.2 continu ensuite de montrer comment ces distributions peuvent être utilisées pour faire des inférences statistiques sur les composantes principales de la population, sur la base des Composantes principales de l'échantillon.

Et enfin comme l'inférence statistique est toujours liée aux tests statistiques, la dérnière section de ce chapitre etudie la conformité à la normale en utilisant des méthodes graphiques et tests statistiques.

2.2 **Propriétés algébriques des composantes principales d'un échantillon**

-On suppose qu'on a 'n'observations à 'p' éléments.

Soit $\widetilde{z}_{i1} = a_1' X_i, i = 1, 2, \ldots n$, et choisissons le vecteur des coefficients a_1' qui maximise la variance échantillon :

$$\frac{1}{n-1}\sum_{i=1}^{n}(\widetilde{z}_{i1} - \overline{z}_1)^2$$

Sous la contrainte de normalisation $a_1' a_1 = 1$.

Par la suite, soit $\widetilde{z}_{i2} = a_2' X_i \ i = 1, 2, \ldots n$, et choisissons a_2' qui maximise la variance échantillon de $\widetilde{z}_{i2}$ sous la contrainte de normalisation $a_2' a_2 = 1$, et aussi sous la contrainte $\widetilde{z}_{i2}$ non corrélé avec $\widetilde{z}_{i1}$ dans l'échantillon.

Continuons ce processus de cette manière, on obtient la version de la définition des composantes principales de l'échantillon, donc $a_k' X$ est défini comme k^{eme} composante principale de l'échantillon (réf[1]).

-On définit $\underbrace{\widetilde{X}}_{(n\times p)}$et $\underbrace{\widetilde{z}}_{(i,k)elements}$ tel que $z = \widetilde{X}\underbrace{A}_{p\times p}$ A matrice orthogonale à k colonnes : a_k »

-Si la moyenne de chaque élément de X est connue pour être égale à zéro, alors $S = \frac{1}{n}\widetilde{X}'\widetilde{X}$, Où S : «Matrice de variance covariance sur l'échantillon,

-Il est beaucoup plus fréquent que la moyenne de X est inconnue, et dans ce cas, le (j, k)^{eme} élément de S est : $\frac{1}{n-1}\sum_{i=1}^{n}(\widetilde{x}_{ij} - \overline{x}_j)(\widetilde{x}_{ik} - \overline{x}_k)$

Où $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_{ij}$ pour $j = 1, 2, \ldots p$

On peut écrire par la suite la matrice S comme :

$$S = \frac{1}{n-1}X'X \tag{2.1}$$

Où X est une matrice à n lignes et p colonnes , d'élément (i, j) : $(\widetilde{x}_{ij} - \overline{x}_j)$

-La matrice des scores des composantes principales

$$Z = XA \tag{2.2}$$

qui a la même variance et covariance donnée par $\widetilde{Z}$, $k = 1, 2, \ldots p$.

-Les vecteurs propres de $\frac{1}{n-1}X'X$ et $X'X$ sont identiques, et les valeurs propres de $\frac{1}{n-1}X'X$ sont simplement $\frac{1}{n-1}$ (val propre $X'X$).

On définit

$$y_i = B'X_i \quad i = 1, 2, .., n \tag{2.3}$$

où B : est matrice orthonormale.

-Les Propriétés P_1, P_2, P_4, détiennent toujours, mais en remplacement $\sum_y$ matrice de covariance de la population par S matrice de covariance de l'échantillon.

- Les preuves sont similaires à celles des populations, après avoir substituter les quantités d'échantillons à la place de quantités de population.

-La décomposition spectrale, propriété P_3, elle est de même pour les échantillons de la forme :

$$S = l_1a_1a_1' + l_2a_2a_2' + \ldots l_pa_pa_p' \tag{2.4}$$

L'implication statistique de cette expression et des autres propriétés P_1, P_2, P_4 sont les mêmes que celles de la population, sauf qu'elles doivent êtres vues dans le contexte de l'échantillon.

2.3 Propriétés géométriques des composantes principales d'un échantillon

La propriété G_1 (page 13) est aussi valide pour l'échantillon si on remplace $\sum$ par S.

Les ellipsoïdes $X'S^{-1}X = constant$ n'ont plus l'interprétation d'être contours de probabilité constante, bien qu'ils fournissent des estimations de ces contours si X_1, $X_2, \ldots, X_n$ sont tirées d'une distribution normale multivariée. Réintroduire une moyenne non nulle : les ellipsoïdes $(X - \overline{X})'S^{-1}(X - \overline{X}) = constant$ donnent le contour de la distance MAHALANOBIS égale à partir de la moyenne échantillon $\overline{X}$.

La propriété G2 reste valable pour l'échantillon comme suit :

On suppose que les observations $X_1, X_2, \ldots, X_n$ sont transformées par : $y_i = B'X_i$ pour $i = 1, 2, .., n$ où $\underbrace{B}_{p\times q}$ matrice à colonnes orthonormales. tel que : $y_1, y_2, \ldots, y_n$ sont les projections de $X_1, X_2, \ldots, X_n$ sur un sous espace à 'q' dimensions, alors $\sum_{h=1}^{n}\sum_{i=1}^{n}(y_h - y_i)'(y_h - y_i)$ est maximisée quand $B = A_q$, le même critère est minimisé quand $B = A_q^*$ (réf[1])

Cette propriété signifie que : si les n observations sont projetées sur le sous espace à q dimensions, alors la somme au carré de la distance euclidienne entre deux observations dans le sous espace est maximisée quand le sous espace est défini par les q premières composantes et minimisée quand il est défini par les q dernières composantes.

La preuve de cette propriété est similaire à celle d'une population en substituant seulement $\sum$ par S.

2.4 **Décomposition en valeurs singulières (DVS)**

-Soit X une matrice arbitraire de dimension $(n \times p)$

$$X = ULA' \tag{2.5}$$

Où $\underbrace{U}_{n\times r}$ et $\underbrace{A}_{p\times r}$: les colonnes des deux matrices sont orthonormées et r : le rang de X

$U'U = I_r$, $A'A = I_r$ et $L(r \times r)$ matrice diagonale (réf[10]).

-Pour prouver ce résultat, on considère la décomposition spectrale de $X'X$, les $(p - r)$ termes de (2.4) (page 18) et dans l'expression correspondante de $X'X$ sont nulles, d'où $X'X$ de rang r, donc :

$(n-1)S = X'X = l_1 a_1 a_1' + l_2 a_2 a_2' + \ldots + l_r a_r a_r'$

Dans cette section, il est pratique de prendre l_k les valeurs propres de $X'X$ plutôt que ceux de S ,

$\underbrace{A}_{p\times r}$: de colonnes a_k

$\underbrace{U}_{n\times r}$: de colonnes $u_k = l_k^{-1/2} X a_k$ pour $k = 1, 2, \ldots, r$

$L(r \times r)$: matrice diagonale d'élément $l_k^{1/2}$

$$X = ULA' = U \begin{pmatrix} l_1^{1/2} a_1' \\ l_2^{1/2} a_2' \\ \vdots \\ l_r^{1/2} a_r' \end{pmatrix} = \sum_{k=1}^{r} l_k^{-1/2} X a_k l_k^{1/2} a_k' = \sum_{k=1}^{p} X a_k a_k'$$

Car les a_k , $k = r+1, r+2, \ldots, p$ sont les vecteurs propres de $X'X$ correspondant aux valeurs propres nulles (réf[1])

-Le vecteur Xa_k est le vecteur de scores des k ^{eme} composantes principales $Xa_k = 0$ pour $k = r+1, r+2, \ldots, p$, donc : $ULA' = X \sum_{k=1}^{p} a_k a_k' = X$

-La DVS pour l'ACP est très importante elle fournit une méthode efficace de calcul des composantes principales ; il est clair que si on peut trouver U, L, A satisfaisant (2.5), alors A et L vont nous donner les vecteurs propres et la racine carrée des valeurs propres de $X'X$, et par conséquent le calcul standards des composantes principales de matrice de covariance S d'un échantillon.

Multiplions (2.5) à droite par A, on obtient : $XA = UL$, mais XA est la matrice des scores des composantes principales qui sont donnés par : $z_{ik} = u_{ik} l_k^{1/2}$

Pour $i = 1, 2, ..n$ et $k = 1, 2, ..r$.

Dans la matrice $Z = UL$ ou $U = ZL^{-1}$, la variance de scores pour k^{eme} composante principale est $\frac{l_k}{n-1}$ pour $k = 1, 2, ..p$ donc k^{eme} valeur propre de S est $\frac{l_k}{n-1}$.

-Notons aussi que les colonnes de U sont les vecteurs propres de XX' correspondants aux valeurs propres non nulles.

-La décomposition en valeur singulière peut aussi fournir un moyen utile de la représentation des résultats de l'ACP graphique et algébrique.

-On peut écrire les éléments de (2.5) comme suit :

$$x_{ij} = \sum_{k=1}^{r} u_{ik} l_k^{1/2} a_{jk} \tag{2.6}$$

Où u_{ik}, a_{jk} sont (i,k)^{eme}, (j, k)^{eme} éléments de U et A respectivement ; et $l_k^{1/2}$ est k^{eme} élément diagonale de L.

2.5 **Etude Asymptotique**

La cinquième section de ce chapitre présente une etude asymptotique, le premier paragraphe examine les distributions de probabilité des vecteurs propres et des valeurs propres d'une matrice de covariance de l'échantillon.

Le paragraphe 2.5.2 continu ensuite de montrer comment ces distributions peuvent être utilisées pour faire des inférences statistiques sur les composantes principales de la population, sur la base des Composantes principales de l'échantillon.

Et enfin comme l'inférence statistique est toujours liées aux tests statistiques, la dérnière section de ce chapitre etudie la conformité à la loi normale en utilisant des méthodes grahiques et tests statistiques.

2.5.1 **Distribution de probabilité des composantes principales d'un échantillon** :

Nous allons donner le théorème suivant , important dans l'étude asymptotique :

Soit un échantillon de taille n extrait d'une population (p) et S matrice de variance covariance associée.

Soient l_k, a_k pour $k = 1, 2, ..p$ les valeurs propres et vecteurs propres de S respectivement, et soient λ_k, α_k pour $k = 1, 2, ..p$ les valeurs propres et vecteurs propres de $\sum$ respectivement, aussi soient I, λ vecteurs a p éléments constitués de l_k et λ_k, respectivement. Et soient j^{eme} éléments de a_k et α_k : a_{kj} et α_{kj} respectivement.

Théorème 4

Supposons que $X \rightsquigarrow N(\mu, \sum)$ où X a une distribution normale de p variables de moyenne μ et de matrice de covariance$\sum$, quoique μ n'a pas besoin pas être donné par contre $\sum$ doit être connue alors :

$(n-1)S \rightsquigarrow W_p\ (\sum, n-1)$ (réf[6])

$(n-1)S$ Suit une distribution de wishart de paramètres $\sum$, $n-1$.

La fonction densité de la matrice V qui Suit une distribution de wishart de paramètres $(\sum, n-1)$ est :

$$C|V|^{((n-p-2)/2)}e^{(-1/2tr(\sum^{-1}V))}$$

Où

$$C^{-1} = 2^{(p(n-1)/2)}\Pi^{(p(1-p)/4)}\left|\sum\right|^{(n-1)/2}\prod_{j=1}^{p}\Gamma(\frac{n-j}{2})$$

Le meilleur résultat simple et connu concernant la distribution de l_k , et a_k, suppose souvent que $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$, autrement dit toutes les valeurs propres de la population sont positives et distincts, alors les résultats suivants détiennent asymptotiquement (lorsque la taille n est grande) :(réf[1]).

- Tous les l_k sont indépendants de tous les a_k.
- I et a_k sont conjointement distribué normalement.
-

$$E(I) = \lambda, E(a_k) = \alpha_k \tag{2.7}$$

-

$$Cov(l_k, l_{k'}) = \begin{cases} \frac{2\lambda_k^2}{n-1} & k = k' \\ 0 & k \neq k' \end{cases} \tag{2.8}$$

-

$$Cov(a_{kj}, a_{k'j'}) = \frac{\lambda_k}{n-1}\sum_{l=1,l\neq k}^{p}\frac{\lambda_l\alpha_{lj}\alpha_{lj'}}{(\lambda_l - \lambda_k)^2} \tag{2.9}$$

Il convient de souligner que les résultats ci-dessus sont asymptotiques et donc approximatifs pour un échantillon fini. (réf[3])

Si une distribution autre que la normale est supposée, les résultats de répartition deviennent généralement moins maniables.

2.5.2 **Inférence basée sur composantes principales d'un échantillon**

Les résultats distributionnels précédents peuvent être utilisés pour faire inférence sur composantes principales d'une population, étant donné les composantes principales d'échantillonnage, à condition que les hypothèses nécessaires sont valides.

-L'hypothèse la plus importante est que X a une distribution normale multivariée qui est souvent non satisfaite, et valeur pratique des résultats par la suite sont limitées.

-L'ACP peut faire un outil beaucoup plus largement applicable dans l'utilisation principale dans descriptif plutôt que l'inférentiel. Elle peut fournir de précieux renseignements descriptifs pour une grande variété de données, si les variables sont continues, et distribuées normalement ou pas (réf[1])

2.5.2.1 **Estimation ponctuelle**

L'estimateur du maximum de vraisemblance(EMV) de $\sum$ n'est pas S mais $\frac{n-1}{n}$ S.

Ce résultat donne le résultat correspondant à une normale univariée.

Si λ, I, α_k, a_k et les relations correspondantes sont définis comme précédemment, alors les EMV de $\lambda, , \alpha_k$, $k = 1, 2, \ldots p$ peuvent être calculés à partir de l'EMV de $\sum$et sont égales à $\widehat{\lambda} = \frac{(n-1)}{n} I$ et $\widehat{\alpha}_k = a_k.k = 1, 2, \ldots p$, supposons que les éléments de λ sont tous positifs et distincts (réf[1])

Les EMV sont les mêmes dans ce cas, comme les estimateurs calculés par la méthode des moments. EMV de λ est biaisé mais asymptotiquement sans biais, comme EMV de $\sum$.

Comme il indiqué précédemment, $\widehat{\lambda}$ est estimateurs biaisé de λ, mais des corrections peuvent êtres faites pour réduire le biais.

Dans le cas où quelques λ_k sont égaux, EMV de leurs valeurs commune est simplement la moyenne de correspondante l_k, multiplié par $\frac{(n-1)}{n}$.

-Les EMV de α_k, correspondants à λ_k égales ne sont pas uniques, les

$(p \times q)$ matrices dont les colonnes sont EMV de α_k correspondants à λ_k égales , peuvent êtres multipliés par toute $(p \times q)$ matrice orthogonale où q est la multiplicité des valeurs propres pour obtenir un autre ensemble d'EMV.

-Le plus souvent les estimateurs de λ, α_k sont simplement donnés par I, a_k, et Ils sont rarement accompagnés par erreurs standards.

-Si la normalité multivariée ne peut pas être supposé, et s'il n'est pas évident autre hypothèse distributionnelle, alors il peut être souhaitable d'utiliser une approche robuste pour l'estimation de composantes principales (réf[6]).

2.5.2.2 **Intervalle d'estimation**

Les distributions marginales asymptotiques de l_k et a_{kj} données précédemment peuvent être utilisées pour construire intervalle de confiance approximatif de λ_k et α_{kj} respectivement. à partir de (2.7) et (2.8) la distribution de l_k est approximativement (réf[1])

$$l_k \rightsquigarrow N(\lambda_k, \frac{2\lambda_k^2}{n-1}) \tag{2.10}$$

d'où

$$\frac{(l_k - \lambda_k)}{\lambda_k[\frac{2}{n-1}]^{1/2}} \rightsquigarrow N(0,1)$$

Qui mène à un intervalle de confiance avec coefficient de confiance $1-\alpha$ pour λ_k.

$$\frac{l_k}{\left[1+\tau z_{\alpha/2}\right]} < \lambda_k < \frac{l_k}{\left[1-\tau z_{\alpha/2}\right]} \tag{2.11}$$

Où $\tau^2 = \frac{2}{n-1}$ et $z_{\alpha/2}$ est la limite supérieur $100\alpha/2$ percentile de distribution standards normale $N(0,1)$. il est supposé que "n" est assez large d'où $z_{\alpha/2} < 1$, comme le résultat distributionnel est asymptotique ceci est une supposition réaliste.

-Un intervalle de confiance alternatif approximatif est obtenu en regardant la distribution de $\ln l_k$ étant donné (2.10) il s'ensuit que :

$ln(l_k) \rightsquigarrow N(\ln(\lambda_k), \frac{2}{n-1})$ Approximativement (réf[3]).

La transformation de λ_k donne un intervalle de confiance approximatif de

$$l_k e^{-\tau z_{\alpha/2}} < \lambda_k < l_k e^{\tau z_{\alpha/2}} \tag{2.12}$$

Les l_k sont asymptotiquement indépendant, et les zones de confiance de plusieurs λ_k sont par la suite obtenues par simple combinaison intervalles à partir du (2.11) ou (2.12) ; le choix des coefficients de confiance se fait de manière à atteindre un niveau de confiance global souhaité.

-un intervalle approximatif pour chaque α_{kj} peut être obtenu à partir de distributions marginales de a_{kj} dont les moyennes et variances sont données dans (2.7) et (2.9).

Les intervalles sont construits de manière similaire à ceux de λ_k, bien que les expressions en question sont un peu plus compliquées.

Les expressions deviennent encore plus compliquées quand on regarde les régions de confiance conjointes pour plusieurs α_{kj} séparés, considérons a_k : de (2.7) et (2.9) il s'ensuit que approximativement (réf[1]) :

$$a_k \rightsquigarrow N(\alpha_k, T_k)$$

Où

$$T_k = \frac{\lambda_k}{n-1} \sum_{l=1, l\neq k}^{p} \frac{\lambda l}{(\lambda_l - \lambda_k)^2} \alpha_l \alpha_l'$$

La matrice T_k à rang (p-1) car elle a unique valeur propre nulle correspondante au vecteur propre α_k. Cela provoque d'autres complications, mais il peut être montré (Mardia et al., 1979, p. 233) que, approximativement :

$$(n-1)(a_k - \alpha_k)'(l_k S^{(-1)} + l_k^{(-1)} S - 2I_p)(a_k - \alpha_k) \rightsquigarrow \chi^2_{p-1} \tag{2.13}$$

Puisque a_k est vecteur propre de S avec valeur propre l_k, (réf[1]) il s'ensuit que :

$$l_k^{(-1)} S a_k = l_k^{(-1)} l_k a_k = a_k; l_k S^{(-1)} a_k = l_k l_k^{(-1)} a_k = a_k$$

et

$$(l_k S^{(-1)} + l_k^{(-1)} S - 2I_p) a_k = a_k + a_k - 2a_k = 0$$

D'où le résultat (2.13) se réduit à :

$$(n-1)\alpha_k'(l_k S^{(-1)} + l_k^{(-1)} S - 2I_p)\alpha_k \rightsquigarrow \chi^2_{p-1} \tag{2.14}$$

De (2.14) ; une région de confiance approximative de α_k avec coefficient de confiance $(1-\alpha)$ a la forme :

$$(n-1)\alpha_k'(l_k S^{(-1)} + l_k^{(-1)} S - 2I_p)\alpha_k \leq \chi^2_{p-1}$$

S'éloignant des hypothèses de normalité multivariée, le bootstrap non-paramétrique de Zac Efron et Tibshirani (1993), detaillé dans le chapitre3, peut être utilisé pour trouver les intervalles de confiance pour les différents paramètres, ainsi que pour estimer les erreurs-types des estimations pour $\alpha_k j$, et pour la proportion du total des variances expliquées par les premières composantes principales.

2.6 **Test de normalité**

Dans cette section, nous présenterons dans un premier temps les techniques descriptives, notamment le très populaire graphique Q-Q plot. Dans un second temps, nous détaillerons le test statistique reconnu et implémenté dans la plupart des logiciels de statistique (plus precisement le R) qui est le test de Shapiro Wilk.

On s'interesse au test de normalité de valeurs propres l_k des échantillons de taille de plus en plus grande

Pour une population Ω donnée, nous voulons étudier la conformité de la distribution d'une v.a. continue X avec la loi normale. Nous disposons pour cela de n observations x_i.

Nous allons être amenés à trier les données. Nous obtenons une série triée de manière ascendante que nous noterons $x_{(i)}$: $x_{(1)}$ correspond à la plus petite valeur observée c.-à-d. $x_{(1)} = x_{\min}$, $x_{(2)}$ est la $2^{ème}$ plus petite valeur, etc.

2.6.1 Q-Q Plot et Droite de Henry

Le Q-Q plot, quantile-quantile plot, est une technique graphique qui permet de comparer les distributions de deux ensembles de données.

Les échantillons ne sont pas forcément de même taille. Il se peut également, et c'est ce qui nous intéresse dans le cas présent, qu'un des ensembles de données soient générés à partir d'une loi de probabilité qui sert de référentiel (réf[31]) Concrètement, il s'agit :

1. de trier les données de manière croissante pour former la série $x_{(i)}$;

2. à chaque valeur $x_{(i)}$, nous associons la fonction de répartition empirique $Fi = \frac{i-0.375}{n+0.25}$ (Saporta, page361) ;

3. nous calculons les quantiles successifs $z^*(i)$ d'ordre F_i en utilisant l'inverse de la loi normale centrée et réduite ;

4. en n, les données initiales n'étant pas centrées et réduites, nous dé-normalisons les données en appliquant la transformation $x^*_{(i)} = z^*_{(i)} \times s + \bar{x}$

2.6.2 Tests statistiques

Très commodes, les approches empiriques n'ont pas la rigueur des techniques statistiques. Dans cette subsection, nous présentons un des tests de compatibilité à

la loi normale,qui est le **test de Shapiro-Wilk,**ce test represente une très bonne alternative au test de Kolmogorov-Smirnov.c'est le plus utilisé en langage R (réf[32]).

La statistique du test s'écrit :

$$w = \frac{\left[\sum_{i=1}^{n} a_i (x_{(i)})\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

où

$x_{(i)}$ (avec des parenthèses entourant l'indice i) désigne la ième statistique d'ordre , i.e., le ième plus petit nombre dans l'échantillon ;

$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ est la moyenne de l'échantillon.

$$(a_1, \cdots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

Où $m = (m_1, \cdots, m_n)^T$

et $m_1, \cdots, m_n$ sont les esperances des statistiques d'ordre d'un échantillon de variables indépendantes et identiquement distribuée suivant une loi normale, et V est la matrice de variance-covariance de ces statistiques d'ordre. L'utilisateur devra rejeter l'hypothèse nulle si W est trop petit.

Pour l'interpretation : Sachant que l'hypothèse nulle H_0 est que la population est normalement distribuée, si la p-value est inférieure au niveau alpha choisi, alors l'hypothèse nulle est rejetée (i.e. on conclut que les données ne sont pas issues d'une population normalement distribuée). Si la p-value est supérieure au niveau alpha choisi, alors on ne peut pas rejeter l'hypothèse nulle selon laquelle les données sont issues d'une population normalement distribuée. Par exemple, pour un niveau alpha de 0.05, un jeu de données avec une p-value de 0.32 n'entraîne pas le rejet de l'hypothèse nulle selon laquelle les données sont issues d'une population normalement distribuée (réf[30]).

2.7 Conclusion

Les méthodes classiques d'inférence statistique sont trés efficaces et utiles, elles fournissent de reponses trés fiables à beaucoup de questions, cependant ces méthodes ne permettent pas d'obtenir des réponses correctes à tous les problèmes concrets que se pose l'utilisateur. Elles ne sont en effet valables que sous des conditions d'application particulières. Ainsi, par exemple, le test t de Student d'égalité des moyennes suppose que les deux populations-parents sont normales, de même variance et que les deux échantillons sont aléatoires,simples et indépendants. Le calcul de l'intervalle de confiance d'une variance par l'intermédiaire des variables χ^2suppose que la population-parent est normale et que l'échantillon est aléatoire et simple. L'inférence statistique classique en régression suppose, outre les conditions d'application relatives à la population et à l'échantillon, que le modèle est ajusté au sens des moindres carrés (réf[12]).

Que peut faire l'utilisateur, en pratique, lorsque ces conditions d'application ne sont pas remplies ?

Différentes attitudes sont possibles. Dans certains cas, les méthodes classiques sont utilisées malgré le non-respect des conditions. Cette utilisation est alors justifiée par le caractère robuste des méthodes qui garantit que les résultats de L'inférence restent approximativement valables.

Le recours à des transformations de variables permet, dans certains cas, de se rapprocher des conditions d'application. Ainsi une transformation logarithmique, par exemple, peut rendre normales des distributions qui, au départ, ne le sont pas.

Une troisième attitude consiste à abandonner les méthodes paramétriques d'inférence statistique au profit de méthodes non paramétriques, pour lesquelles les conditions d'application sont bien moins restrictives.

Pour le problème de la comparaison de deux moyennes évoqué ci-dessus, le test t de Student sera par exemple remplacé par le test des rangs de Wilcoxon, aussi appelé test de Mann-Whitney (Dagnelie, 1998). Les solutions proposées ci-dessus permettent incontestablement d'élargir l'éventail des problèmes auxquels une solution peut être apportée. Elles ne permettent cependant pas de résoudre tous les problèmes.

L'accès généralisé à des moyens de calcul puissants a permis le développement de méthodes d'inférence statistique basées sur l'utilisation intensive de l'ordinateur. Le bootstrap fait partie de ces méthodes. Nous allons l'examiner dans le prochain chapitre.

Chapitre 3

Méthode du Bootstrap

3.1 Introduction

Le mot bootstrap provient de l'expression anglaise "to pull oneself up by one's bootstrap" (Efron, Tibshirani, 1993), qui signifie littéralement "se soulever en tirant sur les languettes de ses bottes". Le mot bootstrap fait penser à des traductions telles que "à la force du poignet" ou "par soi-même" ou "passe partout" (Dagnelie, 1998), mais en fait il n'est jamais traduit dans la littérature (réf[12]).

Dans ce chapitre nous allons décrire comment le Bootstrap peut être utilisé pour résoudre les problèmes d'inférence statistiques en relation avec l'estimation des paramètres, cette méthode est basée sur des simulations comme les méthodes de Monte-Carlo , nous présentons d'abord la méthode. Ensuite, nous expliquons comment évaluer les différents estimateurs par l'estimation de l'erreur-standard , l'estimation du biais et la dérmination de l'intervalle de confiance d'un paramètre.

3.2 Distributions et caractéristiques d'un échantillon

Rappelons que la théorie de l'échantillonage se propose d'étudier les propriétés d'un échantillon $(X_1, X_2\ , \ldots, X_n)$ et des caractéristiques les résumant, les statistiques, à partir de la distribution supposée connue de la variable X. Les X_i sont n

variables independantes et de même loi, les valeurs obtenues $(x_1, x_2, \ldots, x_n)$ constitues n réalisations independantes de la variable aléatoire X ou encore une réalisation unique du n couples $(X_1, X_2\ , \ldots, X_n)$.

Le problème central de l'inférence statistique est le suivant :disposant d'observations sur un échantillon de taille n $(X_1, X_2\ , \ldots, X_n)$, on désire en déduire les propriétées de la population dont il est issu.

3.2.1 **Definition statistique**

Une statistique T est une variable aléatoire fonction mesurable de X_1, X_2 , ..., X_n

$$T = f(X_1, X_2, \ldots, X_n)$$

3.2.2 **Fonction de répartition empirique d'un échantillon**

Soit F(x) la fonction de répartition de la variable aléatoire X désignons par $F_n^*(x)$ la proportion des n variables $X_1, X_2, \ldots, X_n$ qui sont inferieur à x (réf[9]).

$F_n^*(x)$ est donc une variable aléatoire pour tout x qui definit ainsi une fonction aléatoire appelée fonction de répartition empirique de léchantillon, dont les réalisations sont des fonctions en escalier de saux egaux à $\frac{1}{n}$

Si les x_i sont ordonées par valeurs croissantes

$$F_n^*(x) = \begin{cases} 0 \text{ si } x < x_1 \\ \frac{i-1}{n} \text{ si } x_{i-1} \leq x < x_i \\ 1 \text{ si } x \geq x_n \end{cases}$$

3.2.3 Théorèmes Fondamentaux

Ces trois théorèmes sont fondamentaux et ont la justification de l'usage des echantillons en statistiques (réf[9]).

Théorème 5

Pour tout x, on a

$$F_n^*(x) \underrightarrow{P.S} F(x)$$

Preuve :

A x fixé, soit Y le nombre aléatoire de variable inferieures à x, qui est une somme de variables de Bernoulli de paramètre $F(x)$, d'aprés ce qui precède $F_n^*(x)$ qui n'est pas autre que $\frac{Y}{n}$ converge presque surement vers la probabilité $F(x)$.

Théorème 6 ***(GLIVENCO-CANTELLI)***

La convergence de F_n^* vers F est presque surement uniforme, c'est à dire que

$$D_n = Sup\,|F_n^*(x) - F(x)| \to 0$$

Théorème 7 ***(KOLMOGOROV)***

$$\underrightarrow{Lim}_{n\to\infty} P(\sqrt{n}D_n < y = K(y)) = \sum_{-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2)$$

Ce théoréme signifie que la distribution asymptotique de la variable aléatoire D_n est connue et ne depend pas de la variable de départ X, et permet de calculer les limites pour les valeurs de D_n, la loi de la variable D_n a été tabulée (réf[9]).

Propriétés

$$\begin{cases} E(F_n^*(x) = F(x) \\ Var(F_n^*(x)) = \frac{F(x)(1-(F(x))}{n} \end{cases}$$

Où $F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i<x}$est un estimateur de $F(x)$ en tout point x.

Pour une réalisation donnée $(x_1, x_2, \cdots, x_n)$ de l'échantillon aléatoire. $F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{x_i<x}$est l'estimation de $F(x)$ associée à ce jeu de donnée.

3.3 Etude empirique des fluctuations d'échantillonage

Il arrive parfois que l'on soit dans l'impossibilité de calculer la distribution d'échantillonage de certaines caractéristiques utiles soit parceque la distribution de la population parente ne permet pas ces calculs, soit parcequ'elle est inconnue.

Les techniques de simulation permettent de se sortir de cette dificulté en substituant la puissance de calcul d'un ordinateur à celle d'un developpement analytique.

3.3.1 Population de distribution connue

Si on connait la loi F de la variable parente X il suffit de simuler un tres grand nombre N d'échantillons de valeurs de X. Pour chaque échantillons on calcule la statistique recherchées d'où une distribution $T_1, T_2, \ldots, T_n$. si N est grand la répartition empirique des T_i est proche de la loi de la variable T, et on aura donc ainsi des approximations de toutes les caractéristiques de T.

3.3.2 Population de distribution inconnue, la méthode de réechantillonage Bootstrap

Les méthodes de réechantillonage Bootstrap peuvent etre utilisées sur un échantiilon issu d'une populaion mère de distribution F connue ou inconnue.

Si n la taille de l'échantillon, est grande F_n^* est proche de F,on aura donc une bonne approximation de la loi de T en utilisant F_n^* à la place F.

On est donc amené à tirer des échantillons de n valeurs dans la loi F_n^* ce qui revient à réechantilloner dans l'échantillon $x_1, x_2, \ldots, x_n$, c'est à dire à effectuer des tirage aléatoire et simple de n valeurs parmi les n valeurs observées : les valeurs observées $x_1, x_2, \ldots, x_n$ sont donc répetées selon les réalisations d'un vecteur multi-dimentionnel K_1, K_2, ..., K_n d'effectif n et de probabilité $p_i = \frac{1}{n}$

Lorsque n n'est pas trés elevés on peut enumérés tous les échantillons possibles equiprobables, il y'en a n^n,sinon on se contente d'en tirer un nombre suffisement grand à l'aide d'une technique de tirage dans une population finie.

3.4 Les principes de base du bootstrap non paramétrique

3.4.1 Bootstrap des individus

On considère un échantillon de n observations :

$x_1, x_2, \ldots, x_i, \ldots, x_n$, prélevé de manière aléatoire et simple dans une population. Ces observations peuvent concerner une seule variable, ou, au contraire, être relatives à plusieurs variables. Dans ce cas, les x_i représentent des vecteurs de dimension p, p étant le nombre de variables. Afin de ne pas alourdir les notations, nous ne distinguerons pas ces deux situations et, de manière plus condensée, nous désignerons l'échantillon initial par le symbole x, qu'il s'agisse d'un vecteur ou d'une matrice. Le principe de la méthode du bootstrap est de prélever une série d'échantillons aléatoires et simples avec remise de n observations dans l'échantillon initial, considéré comme une population. Ces échantillons successifs seront notés :

$\{x_{11}^*, x_{12}^*, \ldots, x_{1n}^*\}, \{x_{21}^*, x_{22}^*, \ldots, x_{2n}^*\} \ldots, \{x_{B1}^*, x_{B2}^*, \ldots, x_{Bn}^*\}$, B étant le nombre de rééchantillonnages effectués (réf[12]).

L'échantillon sera utilisé pour des inférences sur n^{eme}caractéristique (paramètre) de la population noté θ,en utilisant une statistique T dont la valeur sur l'échantillon est t. Nous supposons que T a été choisi et qu'il est un estimateur de θ,que nous supposons être un scalaire. Notre attention sera portée sur la distribution de probabilité de T, et ensuite étudier le biais, l'erreur standard, la variation du pourcentage d'inertie, et les intervalles de confiance. Mais comment calculer les intervalles de confiance pour θ en utilisant T ?

3.4.2 Distribution empirique $\hat{F}$ et fonction statistique t(·)

Dans le cas non paramétrique la distribution empirique qu'affecte une probabilité de $\frac{1}{n}$ sur chaque valeur x_n de l'échantillon joue un rôle important.

L'estimateur correspendant de F est la distribution empirique $\hat{F}$ definie par :

$$\hat{F}(x) = \frac{\neq \{x_j \leq x\}}{n} \tag{3.1}$$

Où $\neq \{A\}$ signifiant le nombre de fois l'evenement A apparait.

plus formellement

$$\hat{F}(x) = \frac{1}{n}\sum_{j=1}^{n} H(x - x_j) \tag{3.2}$$

avec H(u) la fonction en escalier avec un saut de 0 à 1 pour u=0.

Notons que les valeurs de $\hat{F}$ sont prises aux points $(0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n})$, pour la statistique T, l'estimateur obtenu pour les valeurs observées $x_1, x_2, \ldots, x_n$ permet d'écrire $t = t(\hat{F})$.

Ceci correspond à la relation

$$\theta = t(F)$$

entre la caractéristique d'interêt et la distribution F sous djacente .

Par exemple on peut avoir $t(F) = \int x dF(x)$, $t(F) = \int x^2 dF(x) - \left\{\int x dF(x)\right\}^2$

La fonction $t(\cdot)$ definit le paramètre et son estimation ; on utilisera $t(\cdot)$ pour representer la fonction et t pour l'estimation de θ obtenue par les valeurs observées $x_1, x_2, \ldots, x_n$

Remarque

On note aussi $t = t(\hat{F}) = \hat{\theta}$ estimation de θ.

Exemple 8 *La moyenne $\overline{x}$ de l'échantillon est une estimation de la moyenne de la population, en effet :*

$\mu = \int x dF(x)$ pour montrer que $\overline{x} = t(\hat{F})$, on remplace $\hat{F}$ par sa valeur dans (3.2).

$$\begin{aligned} t(\hat{F}) = \int x d\hat{F}(x) &= \int x d(\tfrac{1}{n}\sum_{j=1}^{n} H(x - x_j)) \\ &= \tfrac{1}{n}\sum_{\substack{j=1 \\ j=1}}^{n} \int x dH(x - x_j) \\ &= \tfrac{1}{n}\sum_{j=1}^{n} x_j = \overline{x} \end{aligned}$$

car $\int a(j) dH(g - x) = a(x)$ pour fonction continue.

3.4.3 Erreur standard et biais d'un paramètre

3.4.3.1 Estimation de l'Erreur standard

Les Méthodes de bootstrap dépendent de la notion d'échantillon bootstrap, soit $\widehat{F}$ la distribution empirique, mettant probabilité $\frac{1}{n}$ pour chaque valeur observée x_i ,$i = 1, 2, ..n$, un échantillon bootstrap est défini comme étant un échantillon aléatoire de taille n tiré de $\hat{F}$, soit $X^* = x_1^*, x_2^*, \ldots, x_n^*$ (réf[13]).

$$\hat{F} \rightarrow x_1^*, x_2^*, \ldots, x_n^* \tag{3.3}$$

Il ya une autre façon de (3.3) : les données bootstrap des points $x_1^*, x_2^*, \ldots, x_n^*$ sont un échantillon aléatoire de taille n avec replacement de la population de n individus $x_1, x_1, \ldots, x_n$, ainsi nous pourrions avoir $x_1^* = x_7, x_1^* = x_3, x_1^* = x_2, \ldots, x_1^* = x_7$

L'ensemble des données bootstrap x_1^*, x_2^*,..., x_n^* est constitué des données initiales $x_1, x_1, \ldots, x_n$, apparaissant parfois zéro, certains apparaissant une fois, certains apparaissant deux fois, etc. Correspondant à un ensemble de données bootstrap X^* est une réplication bootstrap de $\widehat{\theta}$.

$$\widehat{\theta}^* = S(X^*) \tag{3.4}$$

la quantité $S(X^*)$ est le résultat de l'application de la même fonction $S(.)$ de X^*qui avait été appliqué à X.

Par exemple si $S(X)$ est la moyenne $\bar{x}$ d'échantillon alors $S(X^*)$ est la moyenne de l'ensemble de données bootstrap.

$$\bar{x}^* = \sum_{i=1}^{n} \frac{x_i^*}{n}$$

L'estimation bootstrap de $Se_{F(\widehat{\theta})}$, l'erreur standard de la statistique $\widehat{\theta}$, est un plug-in qui utilise l'estimation de la fonction de répartition empirique $\widehat{F}$ à la place de la distribution inconnue F. Spécifiquement, l'estimation bootstrap de $Se_{F(\widehat{\theta})}$ est définie par :

$$Se_{\widehat{F}}(\widehat{\theta}^*) \tag{3.5}$$

en d'autres termes, l'estimation bootstrap de $Se_{F(\widehat{\theta})}$ est l'erreur standard de $\widehat{\theta}$ pour les ensembles de données de taille n échantillonnés aléatoirement de $\widehat{F}$.

Formule (3.5) est appelée l'estimation bootstrap idéale de l'erreur standard de $\widehat{\theta}$, pour pratiquement n'importe quel estimation $\widehat{\theta}$ autre que la moyenne, il n'y a pas de formule exacte qui nous permet de calculer la valeur numérique de l'estimation exacte.

l'algorithme de bootstrap, décrit ci-dessous, est une façon d'obtenir une bonne approximation de la valeur numérique de $Se_{\widehat{F}}(\widehat{\theta}^*)$.

L'échantillon bootstrap est constitué des membres correspondants de X,

$$x_1^* = x_{i1}, x_2^* = x_{i2}, \ldots, x_n^* = x_{in} \tag{3.6}$$

L'algorithme de bootstrap fonctionne en créant de nombreux échantillons bootstrap indépendants, évaluation des réplications bootstrap correspondants, et l'estimation de l'erreur standard de $\widehat{\theta}$ par l'écart-type empirique du résultat réplications. le résultat s'appelle l'estimation bootstrap de l'erreur-standard, noté $\widehat{Se}_B$, où B est le nombre d'échantillons bootstrap utilisés.

L'algorithme suivant est une description plus explicite de la procédure de bootstrap pour estimer l'erreur standard de $\widehat{\theta} = S(X)$ de données observées X.

Algorithme 9 ***pour estimer les erreurs standards***

- *Sélectionner B échantillon bootstrap indépendants $X^{*1}, X^{*2}, \ldots, X^{*B}$, composés chacun de n valeurs de données tirées de replacement de X.*
- *évaluer la réplication bootstrap correspondant à chaque échantillon bootstrap,*

$$\widehat{\theta}^*(b) = S(X^{*b}) \; pour \; b = 1, 2, \ldots, B \tag{3.7}$$

- *Estimer L'erreur standard $Se_F(\widehat{\theta})$ par l'écart type des B réplications .*

$$\widehat{Se}_B = \left\{ \frac{\sum_{b=1}^{B} [\widehat{\theta}^*(b) - \widehat{\theta}^*(.)]^2}{(B-1)} \right\}^{1/2} \tag{3.8}$$

Où

$$\widehat{\theta}^*(.) = \frac{\sum_{b=1}^{B} \widehat{\theta}^*(b)}{B}$$

$$\lim_{B \to \infty} \widehat{Se}_B = Se_{\widehat{F}}(\widehat{\theta}^*)$$

Le fait que $\widehat{Se}_B$ se rapproche de $Se_{\hat{F}}$ quand B tend vers l'infini, l'écart-type empirique se rapproche de l'écart type de la population. Dans ce cas $\widehat{\theta}^* = S(X^*)$, où

$$\hat{F} \rightarrow x^{*1}, x^{*2}, \ldots, x^{*n} = X^*$$

Chaque échantillon bootstrap est un échantillon aléatoire indépendant de taille n de $\hat{F}$. Le nombre de réplications B bootstrap pour estimer l'erreur-standard est généralement compris entre 25 et 200, quand $B \rightarrow \infty$, $\widehat{Se}_B$ se rapproche de $Se_F(\widehat{\theta})$ (réf[13]).

Remarque : Bootstrap paramétrique

Dans le cas ou la population est connue, nous allons utiliser aussi le bootstrap paramétrique et le comparer à la méthodes classique .

Comme dans le cas non paramétrique, Après la génération des échantillons bootstrap, nous évaluons notre statistique sur chaque échantillon bootstrap, puis calculer l'écart type des réplications bootstrap B.

3.4.3.2 **Estimation du biais**

Nous nous sommes concentrés sur l'erreur standard comme une mesure de la précision d'un estimateur $\widehat{\theta}$, il existe d'autres mesures utiles de la précision statistique (ou erreur statistique), en mesurant différents aspects de $\widehat{\theta}$ les préoccupations biais, la différence entre l'estimation $\widehat{\theta}$ et la quantité estimée θ.

L'algorithme de bootstrap est facilement adaptable à donner des estimations de biais, ainsi que de l'erreur standard (réf[13]).

Nous supposons que nous sommes dans le cas non paramétrique d'un échantillon.

La probabilité de répartition F inconnue a donné les données $X = x_1, x_1, \ldots, x_n$ par échantillonnage aléatoire, $F \rightarrow X$. Nous voulons estimer un paramètre à valeurs réelles $\theta = t(F)$.

Pour l'instant, nous allons prendre l'estimateur d'être une statistique $\widehat{\theta} = S(X)$, par la suite nous serons particulièrement intéressés par le plug-in estimation $\widehat{\theta} = t(\hat{F})$.

Le biais de $\widehat{\theta} = S(X)$, comme une estimation de θ est définie comme étant la différence entre l'espérance mathématique de la valeur $\widehat{\theta}$ et du paramètre θ(réf[15]).

$$biais_F = bias_F(\widehat{\theta}, \theta) = E_F[S(X)] - t(F) \tag{3.9}$$

Un biais important est généralement un aspect indésirable de performance.

Des estimateurs non biaisés, ceux pour lesquels $E_F(\widehat{\theta}) = \theta$, jouent un rôle important dans la théorie statistique et pratique. les estimateurs plug in $\widehat{\theta} = t(\hat{F})$ ne sont pas nécessairement non biaisés, mais ils ont tendance à avoir des biais petits par rapport à la magnitude de leurs quelques erreurs standard est l'une des bonnes caractéristiques de la prise en principale. nous pouvons utiliser le bootstrap pour évaluer le biais d'un estimateur $\widehat{\theta} = S(X)$, l'estimation bootsrap du biais est définie comme étant l'estimation $biais_{\widehat{F}}$, nous obtenons $biais_{\widehat{F}}$ en remplaçant $\hat{F}$ par F dans (3.9).

$$biais_{\widehat{F}} = E_{\widehat{F}}[S(X^*)] - t(\hat{F}) \tag{3.10}$$

ici $t(\hat{F})$, l'estimation plug-in de θ, peut différer de $\widehat{\theta} = S(X)$. En d'autres termes $biais_{\widehat{F}}$ est l'estimation plug -in des $biais_F$ (réf[19]).

Algorithme 10 ***Pour estimer le biais***

- Nous générons B échantillons bootstrap indépendants $X^{*1}, X^{*2}, \ldots, X^{*B}$,
- évaluer les réplications bootstrap $\widehat{\theta}^*(b) = S(X^{*b})$,
- et approximer l'espérance bootstrap $E_{\hat{F}}[S(X^*)]$ par la moyenne :

$$\widehat{\theta}^*(.) = \frac{\sum_{b=1}^{B} \widehat{\theta}^*(b)}{B} = \frac{\sum_{b=1}^{B} S(X^{*b})}{B} \tag{3.11}$$

L'estimation bootstrap du biais basée sur B réplications $\widehat{Biais}_B$, est :

$$\widehat{Biais}_B = \widehat{\theta}^*(.) - t(\hat{F}) \tag{3.12}$$

avec $\widehat{\theta}^*(.)$ est substituée à $E_{\widehat{F}}[S(X^*)]$.

3.4.4 **Intervalle de confiance Bootstrap**

Il y a plusieurs types d'intervalles de confiance utilisant la simulation bootstrap, deux d'entre eux, sont présentés ici : Méthode de l'erreur standard et intervalle de confiance bootstrap empirique (intervalle de confiance bootstrap bilatéral basé sur les percentiles) (réf[19]).

3.4.4.1 **Méthode de l'Erreur standard**

Une première solution consiste à définir l'intervalle de confiance par la méthode de l'erreur- standard (standard bootstrap confidence interval) :

$$\widehat{\theta} \pm u_{1-\alpha} \widehat{Se}_B$$

Où

$$\phi(u_{1-\alpha}) = 1 - \alpha$$

Pour que cette approche soit satisfaisante, il faut que la distribution d'échantillonnage du paramètre étudié soit approximativement normale, que l'estimateur soit non biaisé, et que $\widehat{Se}_B$ soit une bonne estimation de l'erreur-standard de la distribution du paramètre.

Le fait que ces conditions soient remplies ou non dépend des circonstances. La condition de normalité peut être vérifiée à partir de la distribution des $\widehat{\theta}^*(b)$ et il peut être utile éventuellement d'effectuer une transformation de manière à rendre la distribution plus proche de la normale. La qualité de l'estimation de l'erreur-standard est liée au nombre de répétitions B considéré.

3.4.4.2 **Méthode des percentiles simples**

Dans la méthode des percentiles simples (simple percentile confidence interval), les limites de confiance sont données par les percentiles α et $1-\alpha$ de la distribution d'échantillonnage empirique, c'est-à-dire de la distribution des $\widehat{\theta}^*(b)$. Nous les notons $\widehat{\theta}^*_{\alpha}$; et $\widehat{\theta}^*_{1-\alpha}$.

Contrairement à la méthode de l'erreur-standard, la distribution d'échantillonnage du paramètre étudié ne doit pas être normale pour que la méthode des percentiles soit satisfaisante. Par contre, le nombre de rééchantillonnages B doit être plus élevé que dans le cas de la méthode de l'erreur-standard, car il faut un plus grand nombre d'observations pour estimer, avec une précision suffisante, un percentile que pour estimer un écart-type. B sera par exemple de l'ordre de 1000.

L'intervalle $\left[\widehat{\theta}^*_{\alpha}; \widehat{\theta}^*_{1-\alpha}\right]$ entre les α-ème et 1–$\alpha^{ème}$ percentile de la distribution bootstrap d'une statistique est un intervalle de confiance au niveau $1-\alpha$ pour le paramètre correspondant.

Si $\widehat{\theta}_{\text{inf}} = \widehat{\theta}_{(\alpha)}$ est la borne inferieure et $\widehat{\theta}_{Sup} = \widehat{\theta}_{(1-\alpha)}$est la borne supérieure d'un tel intervalle de confiance bootstrap, on peut obtenir la probabilité que la vraie valeur du paramètre soit hors de l'intervalle trouvé comme suit :

$$P(\widehat{\theta}_{\text{inf}} < \theta) = \frac{\alpha}{2} + \bigcirc(n^{-1/2}) \tag{3.13}$$

et

$$P(\widehat{\theta}_{\sup} > \theta) = \frac{\alpha}{2} + \bigcirc(n^{-1/2}) \tag{3.14}$$

ce qui nous donne un intervalle de confiance précis de premier ordre.

Les avantages de l'intervalle de confiance empirique (l'intervalle bootstrap de confiance bilatéral basé sur les percentiles) sont :

- c'est un intervalle calculé automatiquement, intuitif et utilisable avec chaque transformation vers la normalité ;
- on peut l'employer même si la distribution de $\theta^*(b)$ est asymétrique ;
- il conserve le domaine de θ (une fois fourni) ;
- les transformations sont respectées (intervalle de confiance cohérent par exemple pour θ ou $\log(\theta)$.

Etape de Construction d'Intervalle de Confiance

$$"Prob(\theta \in [\widehat{\theta}_{\inf}, \widehat{\theta}_{\sup}]) = 1 - 2\alpha"$$

- Classement des B valeurs de $S(X^{*b})$ par ordre croissant
- Intervalle de confiance $[\widehat{\theta}_{\inf}, \widehat{\theta}_{\sup}]$ couvrant $1 - 2\alpha$,

Intervalle contenant $100(1 - 2\alpha)\%$ des valeurs Avec :

$\widehat{\theta}_{\inf} = 100.\alpha^{ième}$ percentile des $S(X^{*b})$ calculés, i.e., B. $\alpha^{ième}$ valeur de la liste classée par ordre croissant

$\widehat{\theta}_{\sup} = 100.(1 - \alpha)^{ième}$ Percentile des $S(X^{*b})$ calculés, i.e., $B.(1 - \alpha)^{ième}$ valeur de la liste classée par ordre croissant (réf[18]).

Exemple

$B = 2000$ et $\alpha = 5\%$

$\widehat{\theta}_{\inf} = 100^{\ ème}$ valeur de la liste classée

$\widehat{\theta}_{\sup} = 1900^{\ ème}$ valeur de la liste classée

Remarque sur le nombre de réplications Bootstrap B

Combien de Bootstrap ? (réf[14]).

– B=25 à B=50 : pour obtenir un début d'information

– B=200 : pour estimer l'erreur standard

– B=500 : pour l'évaluation d'intervalles de confiance.

Mais les moyens de calculs puissants et modernes permettent d'aller bien au delà de 500 pour donner plus de précision.

En Résumé "on donne les formules du Bootstrap "

Estimateur $\hat{\theta} = f(x_1, x_2, \cdots, x_n)$

Réplique Bootstrap, pour les b répliques$(b = 1, \cdots B)$

$\widehat{\theta}^*(b) = f(x_1^*, x_2^*, \cdots, x_n^*)$

Les propriétés statistiques de l'estimation seront calculées sur la distribution des répliques (dite distribution Bootstrap). En particulier :

Estimation Bootstrap du paramètre

$$\widehat{\theta}^*(.) = \frac{\sum_{b=1}^{B} \widehat{\theta}^*(b)}{B}$$

Variance Bootstrap de la distribution du paramètre

$$\frac{\sum_{b=1}^{B} [\widehat{\theta}^*(b) - \widehat{\theta}^*(.)]^2}{(B-1)}$$

Définition du biais d'un estimateur $Biais(\widehat{\theta}) = E(\widehat{\theta}) - \theta$

Estimation Bootstrap du biais $Biais_{Boot}(\widehat{\theta}) = \widehat{\theta}_{Boot} - \widehat{\theta}$

Etape de construction d'intervalle de confiance

$$"Prob(\theta \in [\widehat{\theta}_{\inf}, \widehat{\theta}_{\sup}]) = 1 - 2\alpha"$$

– Classement des B valeurs de $S(X^{*b})$ par ordre croissant

– Intervalle de confiance $[\widehat{\theta}_{\text{inf}}, \widehat{\theta}_{\text{sup}}]$ couvrant $1-2\alpha$,

Intervalle contenant $100(1-2\alpha)\%$ des valeurs Avec :

$\widehat{\theta}_{\text{inf}}$= $100.\alpha^{ième}$ percentile des $S(X^{*b})$calculés, i.e., B. $\alpha^{ième}$ valeur de la liste classée par ordre croissant

$\widehat{\theta}_{\text{sup}} = 100.(1-\alpha)^{ième}$ percentile des $S(X^{*b})$ calculés, i.e., $B.(1-\alpha)^{ième}$ valeur de la liste classée par ordre croissant .

3.5 **Conclusion**

Le bootstrap est une méthode d'inférence statistique basée sur l'utilisation de l'ordinateur qui peut répondre sans formule à beaucoup de questions statistiques réelles. Il est incontestable que l'utilisation des techniques de rééchantillonnage a été rendue possible grâce à la généralisation des moyens de calculs performants. Ces techniques reposent, au départ, sur des idées simples. Toutefois, il faut bien admettre que les développements apportés aux méthodes de base leur ont fait perdre une partie de cette simplicité.

Dans ce chapitre, nous nous sommes interessés à l'estimation du biais,de l'erreur-standard d'un paramètre, et à la détermination des limites de confiance d'un paramètre. Il ne s'agit cependant pas là des seules applications des méthodes de rééchantillonnage. Celles-ci peuvent, en effet, aussi être utilisées pour la réalisation de différents tests d'hypothèses, pour le choix des variables et l'estimation de l'erreur de prédiction en régression, pour l'estimation du taux d'erreur en analyse discriminante, notamment. Bien qu'elles puissent être utilisées dans des situations très variées, leur mise en œuvre ne présente guère d'intérêt lorsque l'inférence statistique peut être réalisée par des méthodes analytiques classiques, pour lesquelles les conditions d'application sont remplies.

Elles ne sont donc pas destinées à remplacer les méthodes d'inférence statistique classiques lorsque celles-ci sont applicables mais plutôt à fournir des réponses à des questions pour lesquelles les méthodes classiques sont inapplicables ou non disponibles.

Le caractère relativement général des problèmes qui peuvent être résolus par bootstrap ne doit pas faire perdre de vue que la qualité de l'inférence dépend de la nature de la question posée et de la disponibilité des données. Comme le suggère Manly (1997), le bootstrap doit être utilisé avec prudence dans les situations où il n'a pas encore été testé de manière approfondie.

Chapitre 4

Application

Nous commençeons d'abord par donner la liste des abréviations utilisées dans ce chapitre.

MCA :méthode classique asymptotique

MBP :méthode bootstrap paramétrique

MPNP :méthode bootstrap non paramétrique

(p) : la population

n"taille de l'échantillon"

$\widehat{\mu}_{MCA}$: estimation de la moyenne par la méthode classique asymptotique

$\widehat{\mu}_{MBP}$: estimation de la moyenne par la méthode Bootstrap paramétrique

$\widehat{\mu}_{MBNP}$: estimation de la moyenne par la méthode Bootstrap non paramétrique

echp(n) : échantillon simulé de taille n.

$\sum$: matrice de variance covariance de la population

S_n :matrice de variance covariance associé à l'échantillon simulé de taille n

λ_i :i^{eme} valeur propre de la population

B_{MCA} : biais estimé par la méthode classique

l_i : i^{eme} valeur propre de l'échantillon

α : matrice des vecteurs propres sur la population et α_k :le k^{ième} vecteur propre de cette matrice

a : matrice des vecteurs propres sur l'échantillon et a_k :le k^{ième} vecteur propre de cette matrice

$\widehat{Se}_{MCA}$: estimation de l'erreur standard par la méthode classique

echp : échantillon initial bootstrap

B : le nombre de bootstrap

X_i i^{eme} échantillon prélevé

X_i^* : i^{eme} échantillon bootstrap prélevé

$\widehat{Se}_{MBP}$: estimation de l'erreur standard par la méthode bootstrap paramétrique

$\widehat{Se}_{MBNP}$: estimation de l'erreur standard par la méthode bootstrap non paramétrique

IDC : intervalle de confiance.

4.1 **Présentation des populations**

Pour illustrer les performances des méthodes présentées Asymptotique (classique) et Bootstrap paramétrique, nous considérons d'abord une population (P) connue caractérisée par quatre variables de distribution multinormale de moyenne $\mu = (0, 0, 0, 0)$ et de matrice de variance-covariance $\sum$ ci-dessous.

Ensuite pour la troisième méthode présentée bootstrap non paramétrique nous considérons un échantillons de 88 étudiants dont chacun d'entre eux a subit 04 tests(Mardia,Kent and Bibby (1979) .

Matrice des Variances-covariances de (P)

$$\sum = \begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & 11.39 & 9.92 & 2.66 & 4.82 \\ [,2] & 9.92 & 8.94 & 4.12 & 5.48 \\ [,3] & 2.66 & 4.12 & 12.06 & 9.29 \\ [,4] & 4.82 & 5.48 & 9.29 & 7.91 \end{pmatrix}$$

Nous pouvons donc calculer les valeurs et les vecteurs propres pour cette population.

a)**Les valeurs propres** de la Matrice des Variances Covariances $\sum$ sont :

[1] 28.231344528 12.029303657 0.031195129 0.008156685

Et les vecteurs propres associés

$$\alpha = \begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & -0.5153370 & 0.5685961 & 0.06379894 & 0.6380094 \\ [,2] & -0.5076565 & 0.3711511 & -0.31941014 & -0.7088786 \\ [,3] & -0.4922408 & -0.6581561 & -0.51637719 & 0.2405912 \\ [,4] & -0.4841608 & -0.3252320 & 0.79199839 & -0.1804190 \end{pmatrix}$$

Ensuite à partir des échantillons aléatoires et simples de cette loi normale multivariée, on calcule les vecteurs des moyennes, les matrices des variances covariances, les vecteurs propres, et les valeurs propres, et on estime celles de la population, en utilisant les méthodes :

-classique (convergence asymptotique)

-Et Bootstrap paramétrique

On compare les résultats obtenus par ces deux méthodes, en examinant le biais,les erreurs standards, et les intervalles de confiance des estimateurs.

Pour mener à terme cette illustration nous avons élaboré deux programmes informatiques sous R.

Le premier appelé «**NMéthod** » nous permet de générer des échantillons aléatoires de tailles quelconques tirés de la population (p)connue, et le second nommé « **sim.boot** » pour la méthode Bootstrap paramétrique.

4.2 Méthode classique ou Asymptotique"MCA"

Nous allons présenter dans ce qui suit, les résultats obtenus à partir de la simulation de plusieurs échantillons aléatoires de tailles différentes tirés de la population (P) connue en utilisant le programme élaboré sous R « NMéthod ». Les résultats obtenus sont exposés dans différents tableaux et différentes figures pour la clarté de l'exposé.

4.2.1 Estimation de la moyenne

	n	$\widehat{\mu}_{MCA}$
(P)	∞	(0 0 0 0)
Echp(n)	n=100	(-0.487477 -0.4970356 -0.4657812 -0.4436922)
Echp(n)	n=200	(-0.03362451 0.02224984 0.29536147 0.19363434)
...	...	...
Echp(n)	n=10 000	(0.006746307 0.005134577 0.004360062 0.004712367)

Table 4.1 : Moyennes estimées"MCA"

-On note que lorsque n augmente le vecteur des moyennes se rapproche de la moyenne de la population qui est nulle.

4.2.2 **Estimation des Matrices variances-Covariances**

$$\sum = \begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 11.39 & 9.92 & 2.66 & 4.82 \\ X_2 & 9.92 & 8.94 & 4.12 & 5.48 \\ X_3 & 2.66 & 4.12 & 12.06 & 9.29 \\ X_4 & 4.82 & 5.48 & 9.29 & 7.91 \end{pmatrix} \quad S_{100} = \begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 10.85 & 9.81 & 4.42 & 5.93 \\ X_2 & 9.81 & 9.19 & 5.90 & 6.70 \\ X_3 & 4.42 & 5.90 & 13.72 & 10.88 \\ X_4 & 5.93 & 6.70 & 10.88 & 9.29 \end{pmatrix}$$

$$S_{200} = \begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 11.25 & 9.75 & 2.32 & 4.52 \\ X_2 & 9.75 & 8.76 & 3.88 & 5.25 \\ X_3 & 2.32 & 3.88 & 11.97 & 9.16 \\ X_4 & 4.52 & 5.25 & 9.16 & 7.74 \end{pmatrix} \quad \cdots$$

$$S_{9900} = \begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 11.58 & 10.06 & 2.67 & 4.87 \\ X_2 & 10.06 & 9.06 & 4.13 & 5.52 \\ X_3 & 2.67 & 4.13 & 12.05 & 9.29 \\ X_4 & 4.87 & 5.52 & 9.29 & 7.92 \end{pmatrix} \quad S_{10000} = \begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 11.49 & 10.01 & 2.74 & 4.90 \\ X_2 & 10.01 & 9.03 & 4.20 & 5.55 \\ X_3 & 2.74 & 4.20 & 12.10 & 9.35 \\ X_4 & 4.90 & 5.55 & 9.35 & 7.97 \end{pmatrix}$$

Table 4.2 : Matrices variances covariances estimées"MCA"

-Nous constatons que les variances et les covariances obtenues se rapprochent des valeurs initiales de la population (P) quand n augmente.

4.2.3 **Test de normalité de Shapiro-Wilk sur les valeurs propres**

a)L'estimation des valeurs propres (ponctuelle ou par intervalle de confiance) necessite la normalité des valeurs propres des échantillons(théorie chapitre1),pour cela nous allons d'abord etudier leur distributions. Nous proposons d'utiliser dans

cet etude le test de normalité de **Shapiro-Wilk** qui est une très bonne alternative au test de Kolmogorov-Smirnov (réf[32]).

La théorie de ce test est présentée par Legendre (1998) p.181. Le test consiste à mesurer la conformité de la distribution observée avec une distribution normale théorique, sur une représentation permettant de visualiser la distribution de fréquence cumulée normale comme une droite .

Plus W est grand, plus la distribution est proche de la normale et plus la probabilité p du test s'approche de 1. Lorsque p est égal ou inférieur au seuil a préétablité (p.ex. 0.05), l'hypothèse H_0 de normalité est rejetée (réf[30]).

Ce test Shapiro-Wilk de normalité est le plus utilisé en langage R. La fonction est : shapiro.test(donnees)

```
        Shapiro-Wilk normality test

data:  testnormsh$Valpro1
W = 0.9948, p-value = 0.969

data:  testnormsh$Valpro2
W = 0.9853, p-value = 0.3355

data:  testnormsh$Valpro3
W = 0.9903, p-value = 0.6912

data:  testnormsh$Valpro4
W = 0.9943, p-value = 0.9512
```

Sachant que l'hypothèse nulle est que la population est normalement distribuée, nous remarquons que la p-value est supérieure au niveau alpha choisi (0.05) pour les quatres valeurs propres, alors on ne peut pas rejeter l'hypothèse nulle selon laquelle les données sont issues d'une population normalement distribuée.

b)Les techniques graphiques (présentées aux chapitre1) permettent aussi de voir la normalité mais d'une façon plus simple et directe."Le quantile-quantile (QQ) plot

normal"est un excellent moyen de voir si les données s'écartent de la normale (ici nous nous intéressons uniquement à la distribution normale).

Si les points sur ce dernier sont proches d'une droite cela signifie que les observations ont une distribution normale (réf[31]) (voir la figure 4.1)

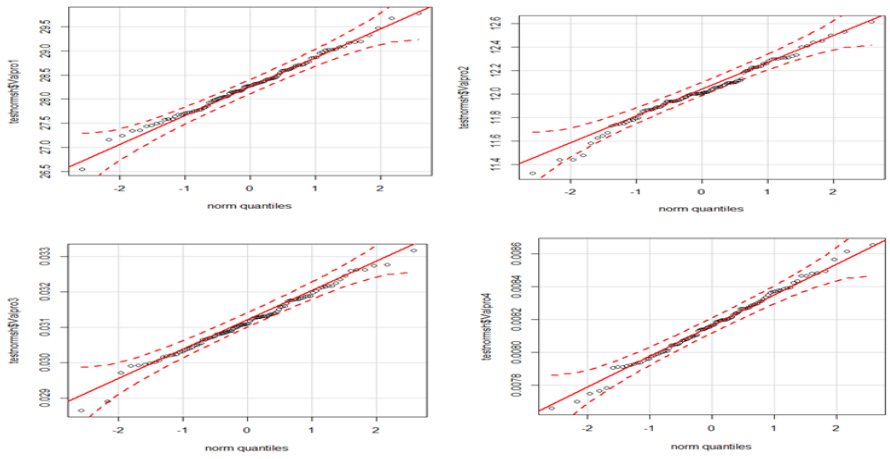

FIG. 4.1 – QQ plots pour les quatres valeurs propres"MCA"

Nous constatons que les points sont relativement alignés. Nous n'observons pas un écartement signicatif, et aucun point ne semble non plus se démarquer des autres, donc nous concluons que les valeurs propres ont une distribution normale. Passons maintenant à l'estimation et à l'evaluation de ces valeurs.

4.2.4 **Estimation des valeurs propres de la population, et du Biais**

		$1^{ere}valpro$	$2^{ere}valpro$	$3^{ere}valpro$	$4^{ere}valpro$
(p)	$\boldsymbol{\lambda}_i$	**28.231**	**12.029**	**0.031**	**0.008**
echp(100)	l_i	32.677	10.329	0.033	0.007
	$\mathbf{B}_{MCA}$	**4.446**	**-1.700**	**1.969e-03**	**-1.057e-03**
echp(200)	l_i	27.377	12.308	0.025	0.007
	$\mathbf{B}_{MCA}$	**-0.853**	**0.278**	**-5.293e-03**	**-6.9684e-04**
...	...	...	...	...	...
echp(9900)	l_i	28.443	12.115	0.031	0.008
	$\mathbf{B}_{MCA}$	**0.212**	**0.085**	**5.464e-04**	**5.816e-05**
echp (10000)	l_i	28.540	12.014	0.031	0.008
	$\mathbf{B}_{MCA}$	**0.308**	**-0.014**	**-1.035e-04**	**4.044e-05**

Table 4.3 : Valeurs propres estimés et biais"MCA"

λ_k : Vraie valeur (valeur propre de la population) pour $k = 1, \ldots, 4$

l_k : Valeur propre estimée

$Biais_k = l_k - \lambda_k$.

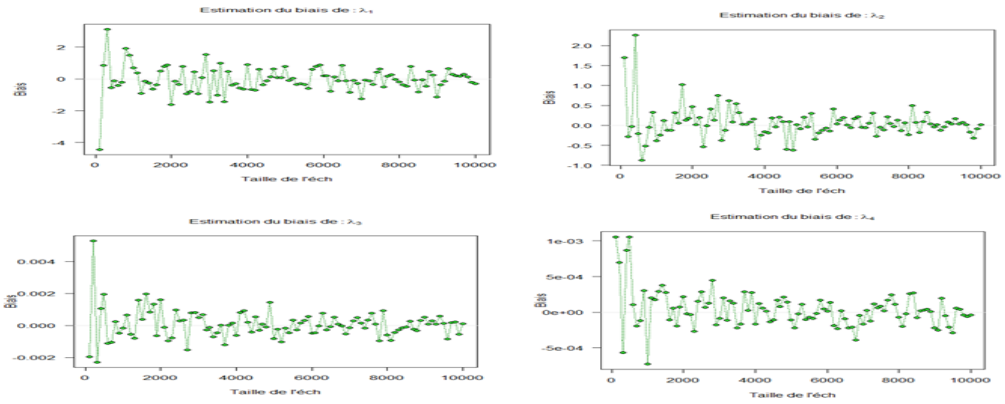

FIG. 4.2 – Biais des valeurs propres"MCA"

La lecture des graphiques ci dessus indique que les biais des valeurs propres $Biais_k$ lorsque la taille de l'échantillon est grande convergent vers zéro, on peut conclure que les estimateurs des valeurs propres sont asymptotiquement non biaisés, ce qui confirme la théorie du chapitre1.

4.2.5 **Determination des intervalles de confiance des valeurs propres**

plutôt que de determiner une valeur approchée d'un estimateur λ obtenu à l'aide d'un estimateur l, pour chaque valeur propre on va rechercher un intervalle de confiance dans lequel on sait avec une probabilité satisfaisante que la valeur λ s'y trouve. l'intervalle de confiance approximatif (démonstration au chapitre1) de λ_k (dans notre exemple $k = 1, \ldots, 4$) avec coefficient de confiance $1 - \alpha$ est : $l_k e^{-\tau z_{\alpha/2}} < \lambda_k < l_k e^{\tau z_{\alpha/2}}$ ceci signifie, pour $\alpha = 0,05$, que si l'on calculait les intervalles de confiance pour le paramètre à partir de N échantillons différents, 95 % des intervalles contiendraient la vraie valeur et pour 5 % d'entre eux la vraie valeur serait à l'extérieur.

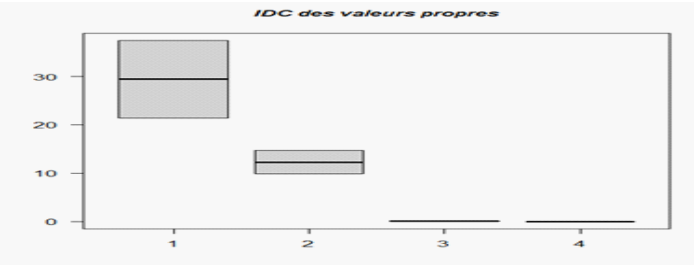

FIG. 4.3 – Intervalle de confiance des valeurs propres "MCA"

La figure determine les limites des intervalles de confiance pour les quatre valeurs propres :

- La première valeur propre se situe dans l'intervalle [21.39854, 37.35538] avec un niveau de confiance de 0.95.

- La deuxième valeur propre se situe dans l'intervalle [9.840644, 14.577767] avec un niveau de confiance de 0.95.

-La troisième valeur propre se situe dans l'intervalle [0.02652056, 0.03654396] avec un niveau de confiance de 0.95.

-La quatrième valeur propre se situe dans l'intervalle [0.007067152, 0.009327673] avec un niveau de confiance 0.95.

4.2.6 **Estimation des vecteurs propres**

α

	[1,]	[,2]	[,3]	[,4]
[1,]	−0.52	0.57	0.06	0.64
[,2]	−0.51	0.37	−0.32	−0.71
[,3]	−0.49	−0.66	−0.52	0.24
[,4]	−0.48	−0.33	0.79	−0.18

a_{100}

	[1,]	[,2]	[,3]	[,4]
[1,]	0.46	0.60	0.11	0.63
[,2]	0.47	0.41	−0.37	−0.67
[,3]	0.54	−0.61	−0.49	0.28
[,4]	0.50	−0.28	0.77	−0.24

a_{200}

	[1,]	[,2]	[,3]	[,4]
[1,]	−0.50	0.54	0.02	0.64
[,2]	−0.49	0.34	−0.26	−0.72
[,3]	−0.49	−0.65	−0.53	0.20
[,4]	−0.48	−0.32	0.80	−0.12

$\cdots$

a_{9900}

	[1,]	[,2]	[,3]	[,4]
[1,]	−0.52	0.56	0.06	0.63
[,2]	−0.51	−0.366	0.318	−0.70
[,3]	−0.48	−0.66	−0.51	0.24
[,4]	−0.48	−0.32	0.79	−0.17

a_{10000}

	[1,]	[,2]	[,3]	[,4]
[1,]	0.51	0.56	0.06	0.63
[,2]	0.50	0.37	−0.31	−0.70
[,3]	0.49	−0.65	−0.51	0.24
[,4]	0.48	−0.32	0.791	−0.18

Table 4.4 : vecteurs propres estimés"MCA"

-Les vecteurs propres calculés à partir des échantillons de taille petite (par exemple n=100) ne reflettent pas les vraies vecteurs, tandis que lorsqu'on augmente la taille, les vecteurs propres estiment avec plus de précision ceux de la population.

4.2.7 **Biais des vecteurs propres estimés**

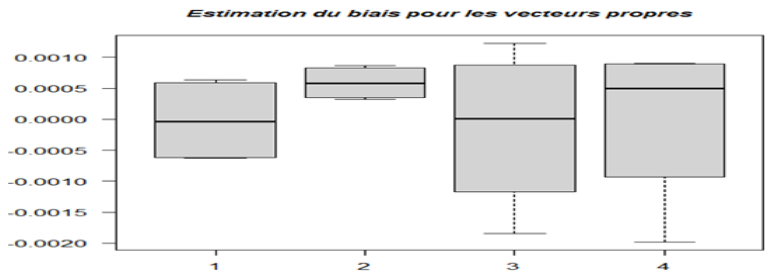

FIG. 4.4 – Biais des composantes du premier vecteur propre"MCA"

-La boite à moustache indique que le biais des composantes du 1^{er} vecteur propre lorsque la taille de l'échantillon est grande est presque égale à zéro, donc on peut en déduire que le premier vecteur propre qui correspond à la plus grande valeur propre est asymptotiquement sans biais. (le temps de calcul n'est pas excessif).

4.2.8 **Erreurs standards du premier vecteur propre estimé**

Erreur standard=ecart type=$\sqrt{Variance}$

	MCA
$\widehat{Se}_{MCA}(a_1)$	$\begin{pmatrix} 0.463 \\ 0.455 \\ 0.435 \\ 0.430 \end{pmatrix}$

Table 4.5 : erreur standard du 1er vecteur propre"MCA"

-L'Erreur standard est la mesure la plus importante pour évaluer la précision d'un estimateur, le tableau montre que le premier vecteur propre est marqué par de faibles erreurs standards.

4.2.9 Pourcentage d'inertie expliqué par la première composante principale en fonction de la taille de l'échantillon

-La première composante principale représente la plus grande part de la variance expliquée($\simeq 70\%$)

$$\widehat{\theta} = \frac{l_1}{\sum_{i=1}^{4} l_i}$$

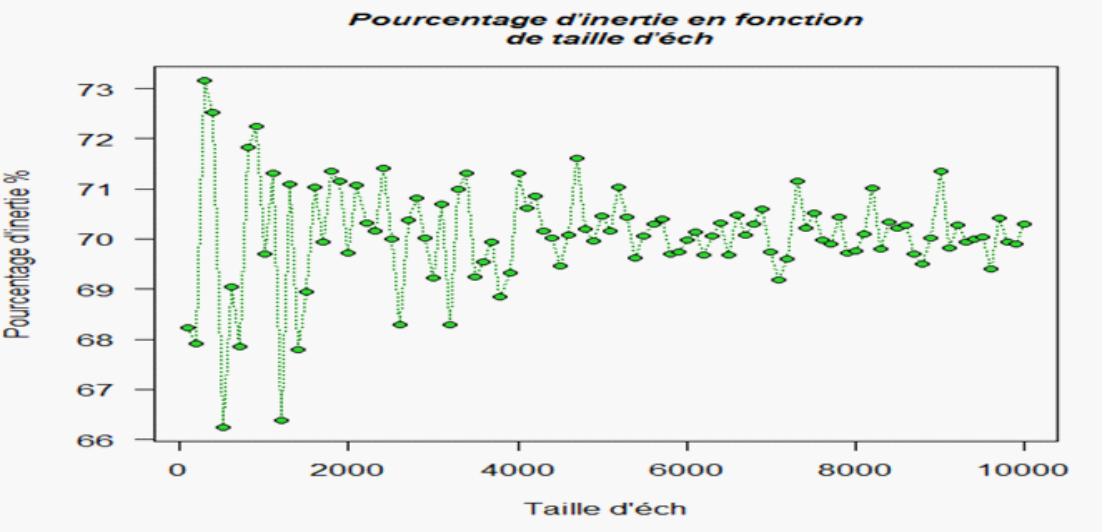

FIG. 4.5 – Pourcentages d'inertie en fonction de la taille de l'échantillon"MCA"

-Le graphe montre que le pourcentage d'inertie expliqué par la première composante principale quand n est grand se stabilise à enviren 70% qui est proche du pourcentage d'inertie de la population. Donc on peut déduire que les informations dans les échantillons de tailles grandes refletent celles de la population avec une précision trés elevée.

-On remarque qu'on peut avoir un nombre de points infini dénombrable de mesure nulle et la somme de mesure nulle est toujours nulle.

Conclusion 11

L'application précédente "MCA"sur une population de distribution connue a donné de bonnes approximations des moyennes, valeurs propres, et vecteurs propres,donc cette méthode(asymptotique classique) a bien confirmé l'etude asymptotique des composantes principale de la thérorie du chapitre1, dans le cas général où l'on ne connait pas la distribution de la population, on ne peut pas appliquer la méthode asymptotique mais la méthode bootstrap que nous propsons nous permet de surmonter cette handicape.

4.3 Méthodes Bootstrap

Nous présentons dans cette section deux méthodes de Bootstrap :

- le cas où la distribution de la population est connue (Bootstrap paramétrique)

-Et le cas où la distribution de la population est inconnue (Bootstrap non paramétrique)

4.3.1 Méthode du Bootstrap paramétrique"MBP"

On considère un échantillon de 100 observations noté "echp", prélevé de manière aléatoire et simple dans la même population (P) connue donnée précedement, de distribution multinormale, de moyenne $\mu = (0, 0, 0, 0)$ et de matrice de variance-covariance $\sum$. Nous allons prélever de cet échantillon "echp" une série de B échantillons aléatoires et simples avec remise .

Ces échantillons successifs seront notés : X_1^*, X_2^*,..., X_B^* , B étant le nombre de réplications (de bootstrap) effectuées .

Dans notre exemple on prendra $B = 5000$. (pour les deux cas etudiés).

-Le deuxième programme élaboré avec "R" (**sim.boot**) nous permet de rééchantillonner B fois avec remise l'échantillon initial .

On calcule pour chaque échantillon la moyenne , la matrice de variance covariance, les valeurs propres, les vecteurs propres, et le pourcentage d'inertie.

On obtient ainsi une série de B valeurs pour les moyennes, de même une série de B matrices de variance-covariance , une série de B valeurs propres , et de B vecteurs propres.

On s'interesse principalement aux erreurs standards des valeurs propres, des vecteurs propres , et des matrices de variance-covariance , à la constrcution des intervalles de confiance des valeurs propres, et aussi aux biais des valeurs propres et des vecteurs propres Bootstrap, ceci en utilisant toujours le programme sim.boot.

Et pour plus de clareté, ce programme donne aussi differents graphiques pour l'ensemble des estimateurs tel que : les histogrammes et les boites à moustache des valeurs propres, graphiques des intervalles de confiance et des pourcentages d'inertie expliqués par la première composante principale pour différentes valeurs de B, ainsi que les graphiques associés aux biais des valeurs propres et des vecteurs propres.

4.3.1.1 Présentation de quelques échantillons Bootstrap paramétrique

a) **Echantillon initial (X)**

> echp

$Data

	X1	X2	X3	X4
1	0.18324754	-0.147552071	-0.65848295	-0.58507585
2	-2.44012699	-2.374972954	-2.31094672	-2.06024351
3	4.85326200	4.405585189	1.17487919	1.57797570
4	...	...	...	...

5	...	...	...	...
⋮	...	...	...	...
⋮	...	...	...	...
99	3.55608379	3.391806125	2.74856250	3.09593862
100	2.27848277	1.025686710	-5.24731061	-3.28769305

b) **Premier échantillon Bootstrap X_1^* (B=1)**

> boot$Data[[1]]

	X1	X2	X3	X4
1	-3.51828132	-3.1333723	-1.64195890	-1.84994186
2	-4.44692353	-4.2898635	-3.07371607	-3.06221330
3	1.49419367	1.2193700	-1.14148392	-0.68639151
4	...	...	...	...
5	...	...	...	...
⋮	...	...	...	...
⋮	...	...	...	...
99	-1.14594575	-0.4742232	1.80377157	0.75373100
100	2.01936598	1.9054862	1.65208843	2.11947580

c) X_{5000}^* **Echantillon Bootstrap pour B=5000**

> boot$Data[[5000]]

	X1	X2	X3	X4
1	0.07349553	0.210852292	1.48254057	1.35221142
2	3.56352671	2.625931285	-1.22677080	0.50596455
3	-2.19834629	-1.722877990	1.27765710	0.24999138
4	...	...	...	...
5	...	...	...	...
⋮	...	...	...	...

⋮	...	...	...	...
99	1.35051436	1.653780871	4.19760956	3.56267262
100	-4.30855143	-3.475461664	0.15276601	-1.28057054

L'algorithme montre que l'ensemble des données bootstrap X_1^*, X_2^*,..., X_n^* est constitué des données initiales $X_1, X_2, \ldots, X_n$, certains n'apparaissant aucune fois, certains apparaissant une fois et certains deux fois, etc. Les Statistiques d'intérêts appliquées sur les 5000 échantillons Bootstrap sont exposées dans les tableaux ci-dessous.

Nous allons d'abord commencer par etudier la distribution asymptotique des valeurs propres.

4.3.1.2 Distribution asymptotique des valeurs propres

Pour B=200 les histogrammes donnent :

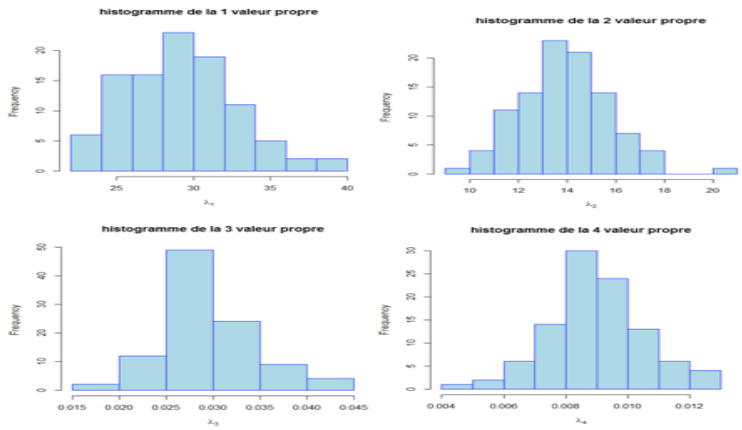

FIG. 4.6 – Histogramme des valeurs propres "MBP"

Nous constatons que pour B=200, les valeurs propres ne sont pas distribuées selon une loi normale. Tandis que pour B=5000 les histogrammes donnent :

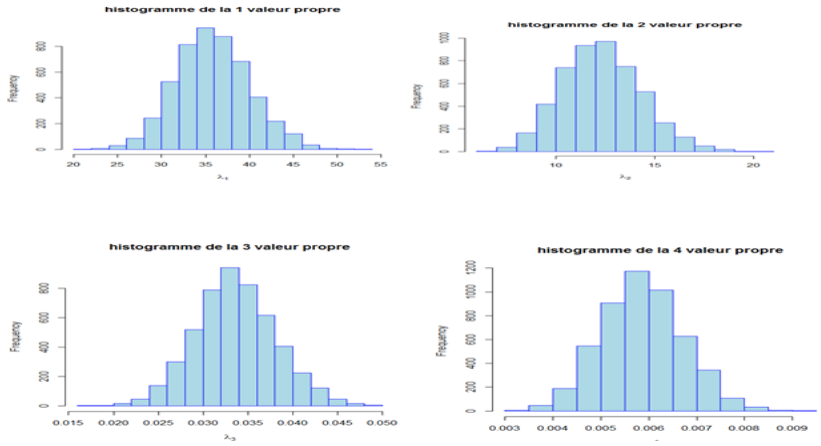

FIG. 4.7 – Histogramme des valeurs propres pour B=5000("MBP")

Les distributions se rapprochent d'une normale lorsqu'on augmente le nombre de bootstrap B, donc on peut penser que les valeurs propres ont asymptotiquement une distribution normale.

4.3.1.3 Estimation de la moyenne

	n	$\widehat{\mu}_{MBp}$
echp	n=100	(–0.3019954 -0.2034684 0.2883195 0.1167068)
B=1	n=100	(0.27162635 0.21541230 -0.13166108 -0.05102669)
B=2	n=100	(-0.4528529 -0.3121646 0.4720458 0.2239080)
...	...	...
B=4999	n=100	(0.1735502 0.2207014 0.3450916 0.2722289)
B=5000	n=100	(-0.15616872 -0.04897612 0.62277005 0.40797177)

Table 4.6 : Moyennes estimées"MBP"

L'estimation des moyennes des échantillons bootstrap est proche de zéro.

4.3.1.4 **Estimation des Matrices Variances Covariances**

echp	$\begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 12.364 & 10.8633 & 3.693 & 5.873 \\ X_2 & 10.901 & 9.805 & 4.214 & 5.868 \\ X_3 & 3.693 & 5.310 & 13.438 & 10.465 \\ X_4 & 5.873 & 6.612 & 10.465 & 8.988 \end{pmatrix}$
B=1	$\begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 12.605 & 10.901 & 2.341 & 4.995 \\ X_2 & 10.901 & 9.805 & 4.214 & 5.868 \\ X_3 & 2.341 & 4.214 & 13.704 & 10.340 \\ X_4 & 4.995 & 5.868 & 10.340 & 8.701 \end{pmatrix}$
...	...
B=4999	$\begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 12.713 & 11.440 & 5.231 & 7.026 \\ X_2 & 11.440 & 10.652 & 6.737 & 7.748 \\ X_3 & 5.231 & 6.737 & 14.302 & 11.461 \\ X_4 & 7.026 & 7.748 & 11.461 & 9.987 \end{pmatrix}$
B=5000	$\begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 13.109 & 11.555 & 3.986 & 6.229 \\ X_2 & 11.555 & 10.560 & 5.595 & 6.958 \\ X_3 & 3.986 & 5.595 & 13.500 & 10.586 \\ X_4 & 6.229 & 6.958 & 10.586 & 9.150 \end{pmatrix}$

Table 4.7 : Matrices variances covariances estimées "MBP"

Les matrices de variance-covariance pour chaque échantillons bootstrap sont estimées en utilsant toujours le programme élaboré"sim.boot".

4.3.1.5 Estimation ponctuelle des Valeurs propres ainsi par intervalles de confiance

	l
echp	(32.589694437 12.067974543 0.034468718 0.009636206)
B=1	(30.54172127 14.23423308 0.03201978 0.00857553)
B=2	(33.49414127 12.76579535 0.03419367 0.01127649)
...	...
B=4999	(36.748267418 10.855716249 0.042929326 0.009790794)
B=5000	(34.06109489 12.21739497 0.03181525 0.01107366)
IDC	$\begin{cases} IDC(\lambda_1) = [26.71070; 38.48951] \\ IDC(\lambda_2) = [8.690323; 15.153350] \\ IDC(\lambda_3) = [0.02726142; 0.04044146] \\ IDC(\lambda_4) = [0.00717364; 0.01137233] \end{cases}$

Table 4.8 : estimation ponctuelle des valeurs propres ainsi par intervalle de confiance"MBP"

La construction des intervalles Bootstrap est différente de la méthode asymptotique classique (chapitre03), nous avons : $B = 5000$ et $\alpha = 5\%$.

$\widehat{\theta}_{\text{inf}} = 250^{\,ème}$ valeur de la liste classée (borne inférieure calculée pour chaque valeur propre), le calcul est fait pour chaque valeur propre.

$\widehat{\theta}_{\text{sup}} = 4750^{\,ème}$ valeur de la liste classée (borne supérieure calculée pour chaque valeur propres), l'estimation est faite pour chaque valeur propre.

4.3.1.6 Estimation des Biais des Valeurs Propres

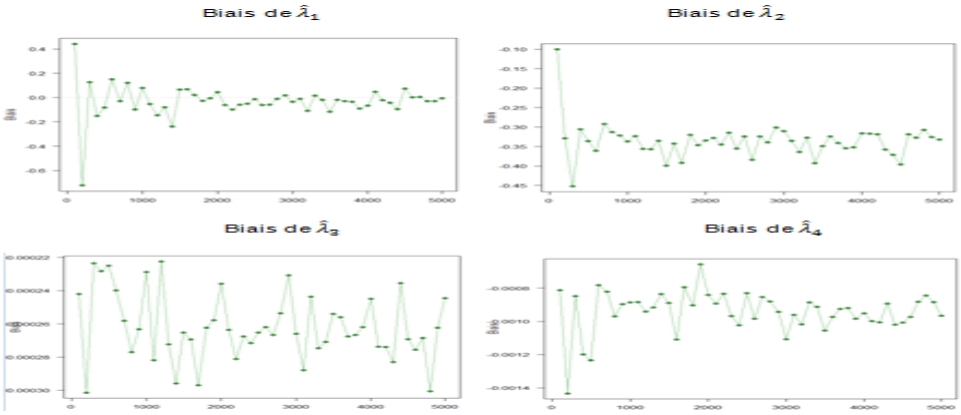

FIG. 4.8 – Biais des valeurs propres"MBP"

-les biais des valeurs propres lorsque B augmente convergent vers zéro, on peut conclure que les estimateurs des valeurs propres sont asymptotiquement sans biais.

4.3.1.7 Estimation des vecteurs propres, de l'erreur standard et du biais

Rappelons que le tableau des valeurs propres de la population est donné dans la section(4.1).

-Le tableau ci-dessous donne les estimations des vecteurs propres pour les différents échantillons Bootstrap ainsi que l'évaluation de ces estimateurs "erreur standard et biais", on s'interesse en particulier à l'erreur standard du premier vecteur propre.

	a
echp	$\begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & -0.503 & 0.581 & 0.088 & 0.632 \\ [,2] & -0.502 & 0.373 & -0.356 & -0.693 \\ [,3] & -0.504 & -0.653 & -0.497 & 0.268 \\ [,4] & -0.489 & -0.309 & 0.786 & 0.215 \end{pmatrix}$
B=1	$\begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & 0.508 & 0.574 & 0.063 & 0.637 \\ [,2] & -0.506 & 0.371 & -0.327 & -0.705 \\ [,3] & 0.497 & -0.657 & -0.509 & 0.247 \\ [,4] & -0.487 & -0.315 & 0.792 & -0.184 \end{pmatrix}$
...	...
B=4999	$\begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & 0.493 & 0.589 & 0.038 & 0.638 \\ [,2] & 0.496 & 0.383 & -0.298 & -0.719 \\ [,3] & 0.516 & -0.643 & -0.517 & 0.227 \\ [,4] & 0.494 & -0.301 & 0.801 & -0.151 \end{pmatrix}$
B=5000	$\begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & -0.518 & 0.568 & 0.108 & 0.629 \\ [,2] & -0.513 & 0.360 & -0.374 & -0.683 \\ [,3] & -0.487 & -0.665 & -0.489 & 0.283 \\ [,4] & -0.480 & -0.323 & 0.779 & -0.237 \end{pmatrix}$

Table 4.9 : Vecteurs propres estimés"MBP"

Et les erreurs standards estimées associées sont :

	MBP
$\widehat{Se}_{MBP}(a_1)$	$\begin{pmatrix} 0.409 \\ 0.403 \\ 0.399 \\ 0.387 \end{pmatrix}$

Table 4.10 : erreur standard du 1er vecteur propre"MBP"

$\widehat{Se}_{5000}$ est l'estimation habituelle d'erreur standard bootstrap basée sur $B = 5000$ réplications bootstrap, les valeurs de $\widehat{Se}_{5000}$ du 1^{er} vecteur propre noté a_1 sont caractérisées par des erreurs standards faibles.

4.3.1.8 Bais des vecteurs propres

$$\begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & 0.2195571 & -0.0038625382 & -0.0001207815 & -0.025109765 \\ [,2] & 0.2164554 & -0.0029558482 & 0.0010020102 & 0.027487210 \\ [,3] & 0.2060969 & 0.0024398373 & 0.0019210980 & -0.010471631 \\ [,4] & 0.2038363 & 0.0006083576 & -0.0030053708 & 0.008308971 \end{pmatrix}$$

-Les estimateurs des vecteurs propres sont asymptotiquement non biaisés

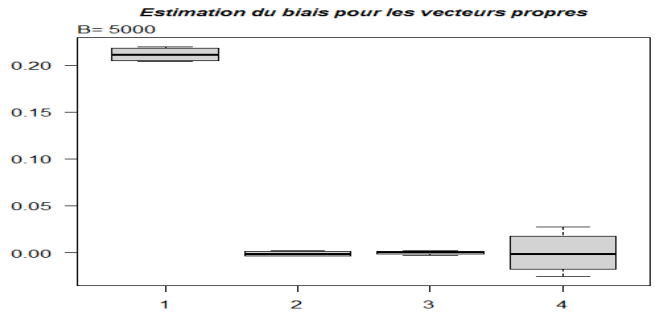

FIG. 4.9 – Biais des composantes du 1er vecteur propre"MBP"

4.3.1.9 Pourcentage d'inertie expliqué par la première composante principale en fonction du nombre de bootstrap

La valeur de $\widehat{\theta}$ mesure le pourcentage de la variance expliquée par la première composante principale.

Le degré de précision est $\widehat{\theta}$? C'est le genre de question que le bootstrap permet d'hapréhender. nous pouvons calculer $\widehat{\theta}$ pour tout ensemble de données bootstrap.

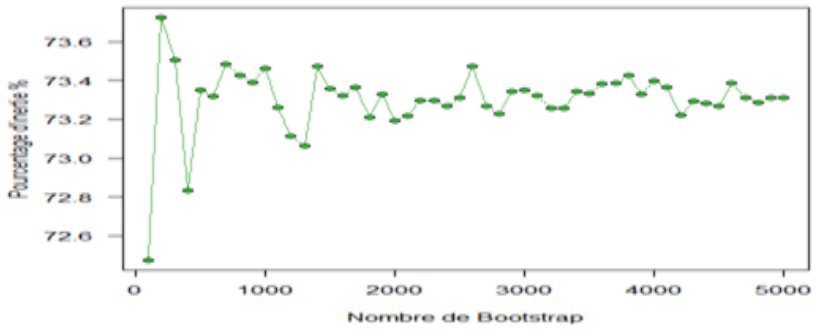

FIG. 4.10 – Pourcentage d'inertie en fonction du nombre de bootstrap"MBP"

-Le graphe montre que le pourcentage d'inertie expliqué par la première composante principale se stabilise à ($\backsimeq$73 %) qui est proche du pourcentage d'inertie de l'échantillon initial ($\backsimeq$72.9%), mais il est un peu plus élevé par rapport à celui de la population($\backsimeq$70 %), c'est pour cela nous allons faire un test d'hypothèse dans la section (4.4)"étude comparative".

-Identiquement à la méthode asymptotique, on remarque qu'on peut avoir un nombre de points infini dénombrable de mesures nulles et la somme de mesures nulles est toujours nulle.

En conclusion La méthode du Bootstrap paramétrique a donné de bonnes approximations des moyennes, des valeurs propres, et des vecteurs propres, cependant cette méthode a concerné le cas de la population connue, et dans la pratique, cette situation est rare car le plus souvent il est impossible de travailler sur toute la population . Nous avons proposé la méthode du bootstrap non paramétrique.

4.3.2 **Méthode du Bootstrap non paramétrique"MBNP"**

Contrairement aux méthodes exposées ci-dessus, le bootstrap non paramétrique n'exige aucune hypothèse à priori ni sur la loi ni sur ses paramètres, mais le principe de rééchantillonage et l'évaluation des estimateurs sont les mêmes dans les deux cas de bootstrap : paramétrique ou non paramétrique.

Nous traitons un échantillon de 88 étudiants dont chacun d'entre eux a subit 04 tests. comme le cas paramétrique on prend B=5000 et on suit les mêmes étapes que precedement mais en utilisant un autre programme que nous avons réalisé nommé "**data.boot**" qui est différent dans sa conception au "sim.boot".

a)**Echantillon initial (X)**

	test1	**test2**	**test3**	**test4**

i=1	82	67	67	81
i=2	78	80	70	81
i=3	73	71	66	81
.	.	.	.	.
.	.	.	.	.
i=86	30	32	35	21
i=87	26	15	20	20
i=88	40	21	9	14

-Le troisième programme élaboré sur le logiciel R (data.boot) nous permet de ré échantillonner B fois avec remise l'échantillon initial, on calcule pour chaque échantillon la moyenne, la matrice de variance covariance, les valeurs propres, les vecteurs propres, le pourcentage d'inertie ainsi que les composantes principales.

On évalue par la suite les estimateurs obtenus selon plusieurs critères : les erreurs standards , les intervalles de confiance , et les biais .

Et pour une meilleure lecture des résultats, le programme affiche aussi des graphiques représentatifs en boites à moustache, histogrammes et en courbes.

4.3.2.1 **Estimation des Matrices Variances Covariances**

echp	$\begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 172.842 & 85.157 & 94.672 & 99.012 \\ X_2 & 85.157 & 112.885 & 112.113 & 121.870 \\ X_3 & 94.672 & 112.113 & 220.380 & 155.535 \\ X_4 & 99.012 & 121.870 & 155.535 & 297.755 \end{pmatrix}$
B=1	$\begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 162.298 & 68.277 & 76.991 & 74.136 \\ X_2 & 68.277 & 111.805 & 107.974 & 106.246 \\ X_3 & 76.991 & 107.974 & 210.781 & 133.587 \\ X_4 & 74.136 & 106.246 & 133.587 & 263.784 \end{pmatrix}$
...	...
B=4999	$\begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 137.778 & 67.824 & 72.446 & 76.821 \\ X_2 & 67.824 & 94.651 & 108.837 & 97.315 \\ X_3 & 72.446 & 108.837 & 247.711 & 126.674 \\ X_4 & 76.821 & 97.315 & 126.674 & 241.702 \end{pmatrix}$
B=5000	$\begin{pmatrix} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 182.388 & 75.670 & 91.322 & 113.774 \\ X_2 & 75.670 & 102.838 & 116.688 & 137.125 \\ X_3 & 91.322 & 116.688 & 243.420 & 188.226 \\ X_4 & 113.774 & 137.125 & 188.226 & 328.618 \end{pmatrix}$

Table 4.11 : Matrices variances covariances estimées"MBNP"

-Pour toutes les réplications bootstrap ,les variances et les covariances sont estimées par le programme réalisé"data.boot".

4.3.2.2 **Estimation ponctuelle des Valeurs propres, ainsi par intervalle de confiance**

L'estimation par intervalle de confiance dans ce cas "bootstrap non paramétrique" est similaire au cas du "bootstrap paramétrique"

	l
echp	(557.506, 121.74443, 91.51626, 33.09661)
B=1	(490.199, 125.026, 96.540, 36.904)
B=2	(623.180, 123.182, 69.073, 28.677)
...	...
B=4999	(479.064, 118.999, 98.105,25.67436)
B=5000	(613.393, 122.694, 93.425,27.752)
IDC	$\begin{cases} IDC(\lambda_1) = [547.250, 558.346] \\ IDC(\lambda_2) = [122.982, 125.021] \\ IDC(\lambda_3) = [83.642, 85.307] \\ IDC(\lambda_4) = [30.668, 31.195] \end{cases}$

Table 4.12 : estimation ponctuelle des valeurs propres
ainsi par intervalle de confiance"MBNP"

-Le tableau indique que les intervalles de confiance Bootstrap estiment avec une très grande précision les quatre valeurs propres.

4.3.2.3 Histogrammes des Valeurs Propres

Pour B=200 les histogrammes donnent :

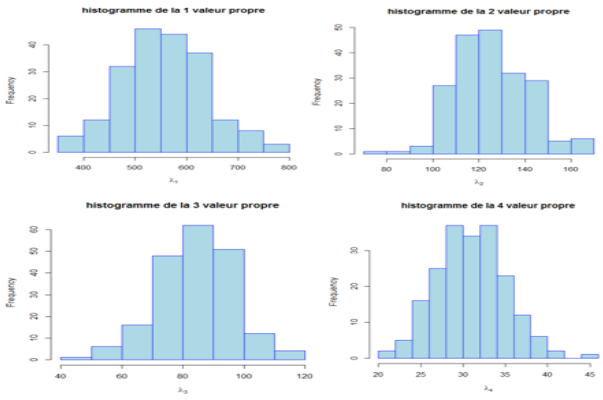

FIG. 4.11 – Histogramme des valeurs propres"MBNP"

Nous constatons que pour B=200, les valeurs propres ne sont pas distribuées selon une loi normale.Tandis que pour B=5000 les hidtogrammes donnent :

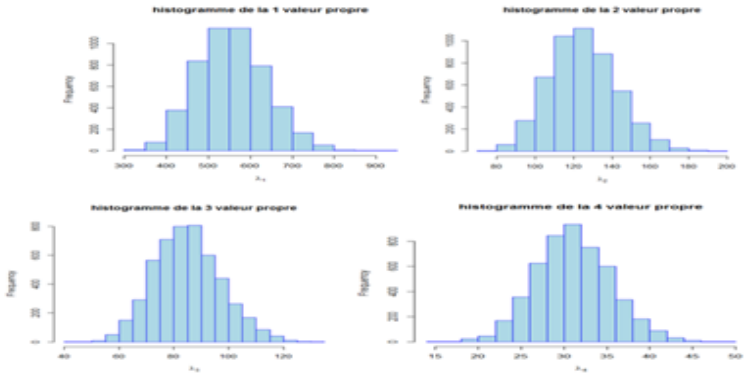

FIG. 4.12 – Histogramme des valeurs propres"MBNP"

On peut déduire d'aprés les histogrammes que les valeurs propres ont asymptotiquement une distribution normale.

4.3.2.4 Estimation des Biais des Valeurs Propres

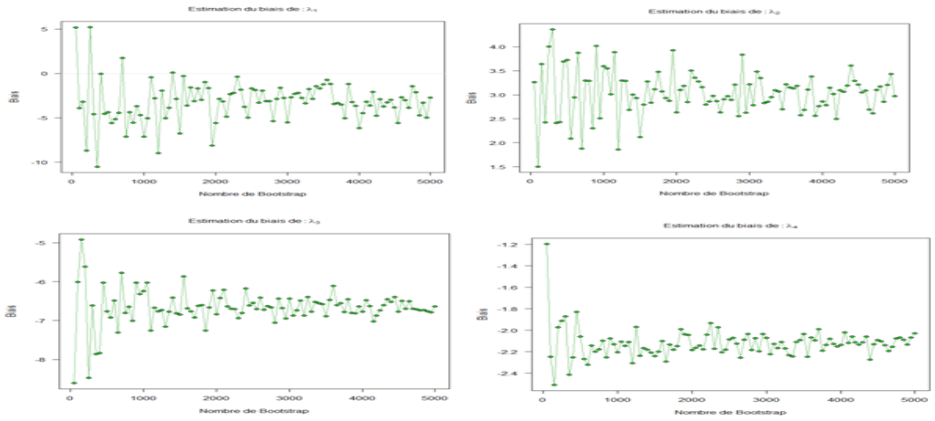

FIG. 4.13 – Estimation des biais des valeurs propres"MBNP"

On constate que les biais des valeurs propres lorsque B augmente se stabilisent donc ils sont convergents, mais dans cette méthode les estimateurs des valeurs propres sont fortement biaisés par rapports aux méthodes bootstrap paramétrique et asymptotique classique, et comme on a déja mentionné le biais reste une caractéristique d'un estimateur mais pas un critère de performance.

4.3.2.5 Estimation des vecteurs propres et de l'erreur standard

Les quatres vecteurs propres sont calculés pour chaque echantillon bootstrap et par la suite l'erreur standard bootstrap est estimée pour B=5000 en utilisant le programme élaboré "data.boot"

Commençons d'abord par estimer les vecteurs propres

	a
echp	$\begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & -0.383 & 0.736 & 0.499 & -0.247 \\ [,2] & -0.385 & 0.138 & -0.048 & 0.910 \\ [,3] & -0.534 & 0.152 & -0.779 & -0.290 \\ [,4] & -0.647 & -0.644 & 0.376 & -0.156 \end{pmatrix}$
B=1	$\begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & -0.357 & 0.780 & 0.481 & -0.177 \\ [,2] & -0.401 & 0.101 & -0.130 & 0.900 \\ [,3] & -0.556 & 0.115 & -0.736 & -0.367 \\ [,4] & -0.633 & -0.606 & 0.457 & -0.147 \end{pmatrix}$
...	...
B=4999	$\begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & -0.341 & 0.156 & 0.886 & -0.269 \\ [,2] & -0.385 & -0.048 & 0.137 & 0.911 \\ [,3] & -0.615 & -0.715 & -0.192 & -0.269 \\ [,4] & -0.596 & 0.679 & -0.397 & -0.156 \end{pmatrix}$
B=5000	$\begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [,1] & 0.355 & 0.922 & 0.025 & -0.151 \\ [,2] & 0.357 & 0.017 & -0.071 & 0.931 \\ [,3] & 0.542 & -0.230 & -0.763 & -0.262 \\ [,4] & 0.672 & -0.310 & 0.640 & -0.202 \end{pmatrix}$

Table 4.13 : Vecteurs propres estimés"MBNP"

On s'interesse au 1er vecteur propre, l'estimation de son erreur standard donne :

	MBNP
$\widehat{se}_{MBNP}(a_1)$	$\begin{pmatrix} 0.219 \\ 0.213 \\ 0.296 \\ 0.355 \end{pmatrix}$

Table 4.14 : erreur standard du 1er vecteur propre"MBNP"

- Le tableau montre que les erreurs standard du 1^{er} vecteur propre sont faibles, ce qui nous permet de deduire que les estimateurs des vecteurs propres donnés par cette méthode sont plus precis que ceux donnés par les méthodes précédentes bootstrap paramétrique et asymptotique classique.

4.3.2.6 Bais des vecteurs propres

$$\begin{pmatrix} & [,1] & [,2] & [,3] & [,4] \\ [1,] & 0.06831418 & -0.20421241 & -0.05014972 & 0.04592525 \\ [2,] & 0.06792428 & -0.05975435 & 0.03553980 & -0.19767860 \\ [3,] & 0.09096541 & -0.18365837 & 0.16388398 & 0.06339185 \\ [4,] & 0.11155150 & 0.31087179 & -0.12402331 & 0.03831265 \end{pmatrix}$$

-les estimateurs des biais des vecteurs propres composante par composante sont légèrement biaisés pour B=5000 mais ils sont asymptotiquement non biaisés, ce qui est vérifiable lorsqu'on augmente le nombre de réplications de plus en plus. et pour plus de clarté on donne le graphe ci dessous.

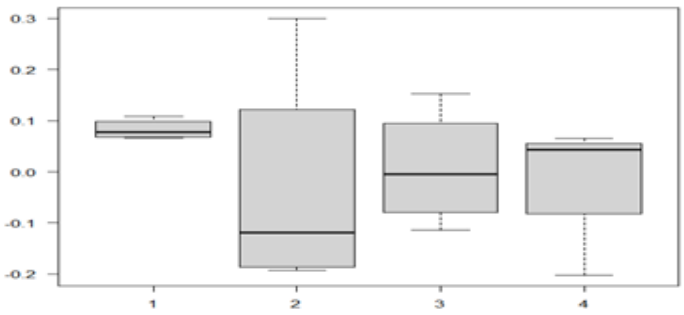

FIG. 4.14 – Biais des composantes du premier vecteur propre"MBNP"

4.3.2.7 Pourcentage d'inertie expliqué par la première composante principale en fonction du nombre de Bootstrap

Nous avons calculé le pourcentage d'inertie $\hat{\theta}$ pour tout ensemble de données bootstrap.

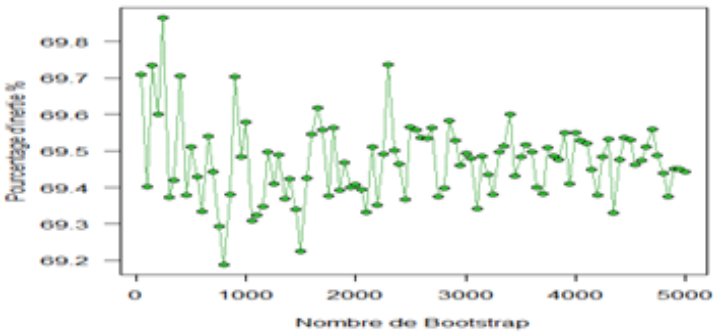

FIG. 4.15 – Pourcentage d'inertie en fonction du nombre de Bootstrap "MBNP"

-Le graphe montre que le pourcentage d'inertie expliqué par la première composante principale est convergent et il se stabilise à ($\backsimeq 69.5\%$).

-Identiquement aux méthodes asymptotique et bootstrap paramétrique, on remarque qu'on peut avoir un nombre de points infini dénombrable de mesures nulles et la somme de mesures nulles est toujours nulle.

4.4 Etude comparative" MCA et MBP"

Nous allons comparer dans cette section les deux Méthodes etudiées asymptotique classique et bootstrap paramétrique , rappelons que dans ce cas la population est connue.

L' évaluation des performances de chacune des méthodes, est basée sur les résultats obtenus des Biais, des valeurs propres, et des vecteurs propres, les erreurs standards des vecteurs propres ainsi que les intervalles de confiance des valeurs propres.

4.4.1 Comparaison des Biais de la première valeur propre

La comparaison des biais montre que le biais Bootstrap de la première valeur propre est inférieur au biais de la première valeur propre de la méthode asymptotique, donc on peut conclure que la première valeur propre qui a la plus grande variance est mieux estimée par la méthode du Bootstrap que l'asymtotique classique. Voir les figures ci dessous :

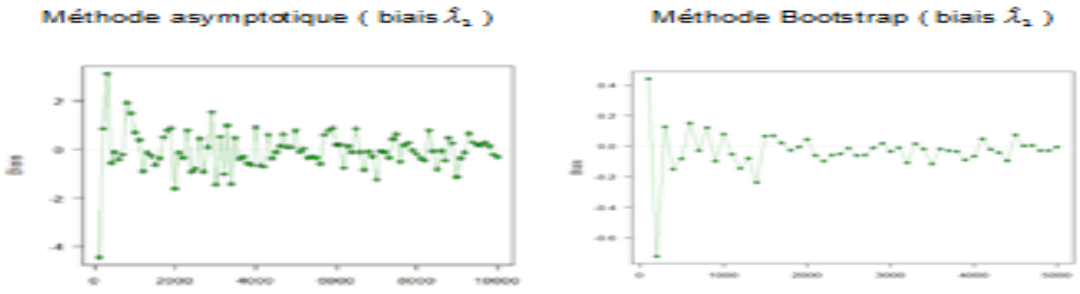

FIG. 4.16 – Comparaison des biais de la première valeur propre

4.4.2 Comparaison des intervalles de confiance des valeurs propres

Idéalement tout problème d'estimation devrait être productif d'un intervalle de confiance. ne donner qu'une estimation ponctuelle masque l'incertitude qui accompagne tout résultat.

	MCA	**MBP**
$IDC(\lambda_1)$	[21.39854, 37.35538]	[26.71070, 38.48951]

Table 4.15 : Comparaison des intervalles de confiance des valeurs propres

-les estimations par intervalles de confiance à 95% présentent plus de précision dans la méthode Bootstrap que dans la méthode asymptotique où l'intervalle de confiance bootstrap est plus etroit que l'intervalle de confiance classique.

-Les probabilités de couverture des valeurs propres dans la méthode Bootstrap sont plus fortes que dans la méthode asymptotique.

4.4.3 Comparaison des pourcentages d'inerties expliqué par la première composante principale

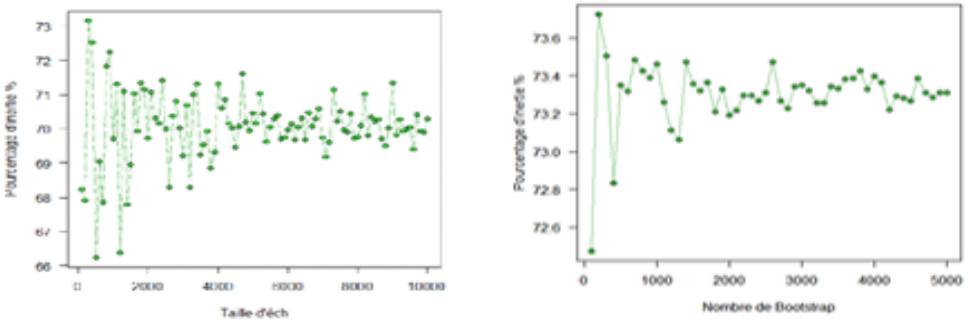

FIG. 4.17 – comparaison des pourcentages d'inerties

-Lorsqu'on augmente la taille de l'échantillon et le nombre de bootstrap, les pourcentages d'inertie se stabilisent dans les deux méthodes.

-on constate que le pourcentage d'inertie estimé par la méthode classique est trés proche du pourcentage d'inertie de la population quand la taille de l'échantillon augmente, tandis que celui du bootstrap est un peu elevé, nous proposons le test suivant pour situer ce pourcentage.

avant de commencer le test etudions d'abord graphiquement la distribution des pourcentages d'inertie bootstrap :

4.4.4 **Distribution asymptotique des pourcentages d'inerties en fonction du nombre de bootstrap**

Pour B=200 et pour B=5000 le porgramme sim.boot affiche

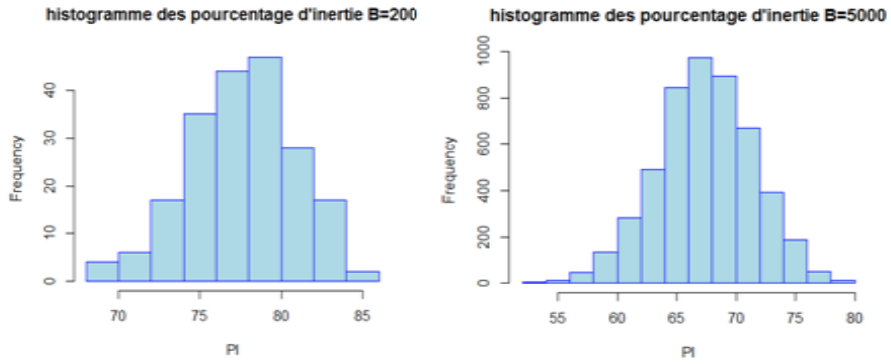

FIG. 4.18 – Histogrammes des pourcentages d'inerties pour differents nombre de bootstrap

il est remarquable que la distribution des pourcentages d'inertie expliqués par la première composante principale se rapproche de la normale de plus en plus.

effectuons maintenant le test

4.4.5 **Test d'hypothèse**

La population connue que nous avons consideré avait les valeurs propres $\lambda = (28.231344528 \;\; 12.029303657 \;\; 0.031195129 \;\; 0.008156685)^t$ où le pourcentage d'inertie expliqué par la première valeur propre represente $\backsimeq 70\%$.

Par les différentes méthodes nous avons obtenu deux estimateurs du pourcentage d'inertie, le premier $\backsimeq 70\%$ estimé par la méthode asymptotique et le deuxième $\backsimeq$ 73% estimé par la méthode bootstrap.

On teste l'hypothèse H_0 que la proportion de la variance expliquée par la première composante principale est égale à $\Psi = 0.70$ (voir[16]).

contre H_1 la proportion de la variance expliquée par la première composante principale est supérieur à $\Psi = 0.70$

-la variance expliquée par les q premières composantes principales est :

$$\Psi_q = \frac{\lambda_1 + \lambda_2 + \ldots \lambda q}{\sum_{j=1}^{p} \lambda_j}$$

-Dans la pratique elle est estimée par :

$$\widehat{\Psi}_q = \frac{l_1 + l_2 + \ldots lq}{\sum_{j=1}^{p} l_j}$$

Qui suit asymptotiquement une distribution normale (réf[16]), on obtient :

$$\sqrt{n-1}(\widehat{\Psi}_q - \Psi_q) \underrightarrow{L} N(0, D^T V D)$$

Où

$$V = 2\Lambda^2$$

et

$$D = (d_1, \ldots, d_p)^T$$

Il s'ensuit que :

$$\sqrt{n-1}(\widehat{\Psi}_q - \Psi_q) \underrightarrow{L} N(0, \omega^2)$$

Où

$$\omega^2 = \frac{2tr(\sum^2)}{\{tr(\sum)\}^2}(\Psi^2 - 2\beta\Psi_q + \beta)$$

et

$$\beta = \frac{\lambda_1^2 + \ldots + \lambda_q^2}{\lambda_1^2 + \ldots + \lambda_p^2} = \frac{\lambda_1^2 + \ldots + \lambda_q^2}{tr(\sum^2)}$$

a)Test appliqué sur la méthode asymptotique

$$l = (28.5400628 \quad 12.0146365 \quad 0.03109154 \quad 0.08197131)^t$$

et

IL s'ensuit que pour $q = 1$, on obtient $\widehat{\beta} = 0,84945247$

Et $\hat{w} = 0,41662387.$

Sous l'hypothèse nulle $H_0 : \Psi_1 = 0.70$,le test statistique $\frac{\sqrt{n-1}(\widehat{\Psi}_q - 0.7)}{\omega}$ a une distribution normale standard.la valeur du test statistique est $t = 0,4286431$ plus petite que la valeur critique de la distribution normale $\emptyset^{(-1)}(0.975) = 1.96$ et on accepte l'hypothèse nulle H_0.

D'où à un intervalle de confiance $\alpha = 0.95$, nous avons prouvé que la proportion de la variance expliquée par la première composante principale dans le cas de la méthode asymptotique est égale à 70%.

b)Test appliqué sur la méthode Bootstrap

$$l = (31.53373529 \quad 11.3600536 \quad 0.036389762 \quad 0.007374383 \quad)^t$$

et

$$\Psi = (0.734409231 \; 0.998980749 \; 0.999828253 \quad 1)^t$$

IL s'ensuit que pour $q = 1$, on obtient $\hat{B} = 0.885126487$

Et $\hat{w} = 0.389357833$

Sous l'hypothèse nulle $H_0 : \Psi_1 = 0.70$, le test statistique $\frac{\sqrt{n-1}(\widehat{\Psi}_q - 0.7)}{\omega}$ a une distribution normale standard.la valeur du test statistique est $t = 6.248382592$ plus large que la valeur critique de la distribution normale $\emptyset^{(-1)}(0.975) = 1.96$ et on rejette l'hypothèse nulle H_0.

D'où à un intervalle de confiance $\alpha = 0.95$, nous avons prouvé que la proportion de la variance expliquée par la première composante principale dans le cas du Bootstrap est plus large que 70%.

On constate que la méthode du Bootstrap dans ce cas ne donne pas les résultats réels.

4.4.6 Comparaison des Biais du 1er vecteur propre

-Les estimateurs des premiers vecteurs propres calculés par les deux méthodes sont légèrement biaisés, mais ils ont tendance d'être asymptotiquement sans biais. Cependant on doit mentionner q'un estimateur sans biais n'est pas toujours meilleur qu'un estimateur biaisé, et notre fort critère de comparaison est l'erreur standard.

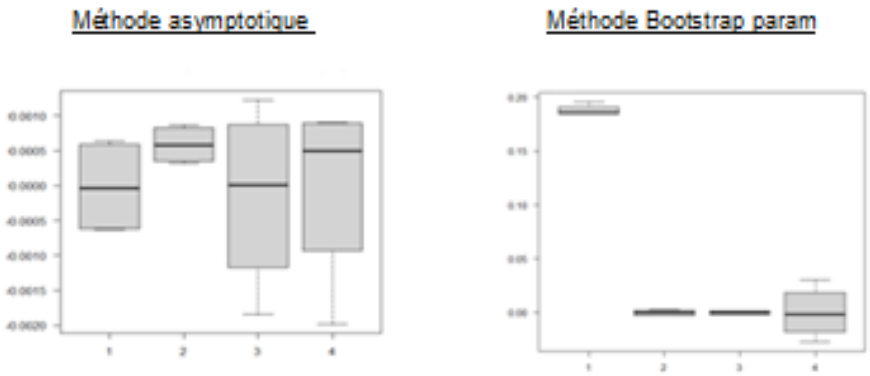

FIG. 4.19 – Comparaison des biais des vecteurs propres

4.4.7 **Comparaison des erreurs standards des vecteurs propres**

L'absence des biais n'est pas suffisante pour s'assurer de l'efficacité d'un estimateur car le paramère peut avoir plusieurs estimateurs sans biais. Dans ce cas c'est la variance des estimateurs qui permet de les comparer.Si cette variance est elevée, l'estimateur peut prendre des valeurs trés eloignées de la valeur effective du paramère.

	MCA	**MBP**
$\widehat{\mathbf{Se}}(a_1)$	$\begin{pmatrix} 0.463 \\ 0.455 \\ 0.435 \\ 0.430 \end{pmatrix}$	$\begin{pmatrix} 0.409 \\ 0.403 \\ 0.399 \\ 0.387 \end{pmatrix}$

Table 4.16 : comparaison des erreurs standards des vecteurs prores

-En effectuant 5000 réplications bootstrap d'une taille de 100 seulement et en se basant sur un seul échantillon, on obtient des erreurs standards du 1^{er} vecteur

propre (qui correspond à la plus grande variance) plus faibles que celles données par la méthode classique assymptotique, donc selon notre critère de comparaison (erreurs standards) qui est le meilleur parmi tout les autres critères on peut classer les méthodes selon leurs performances comme suit :

1-Méthode Bootstrap

2-Méthode asymptotique classique

Mais on doit tenir compte des avantages et des inconveinients de chacune et aussi de la nature du problème posé.

4.4.8 **Comparaison des temps d'exécutions**

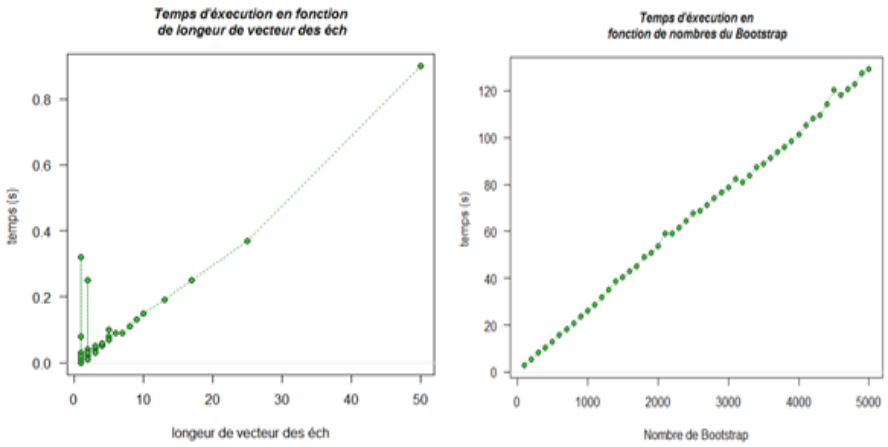

FIG. 4.20 – Comparaison des temps d'executions

L'exécution du programme Bootstrap dure (123) secondes en tenant compte des caractéristiques du micro utilisé (TOSHIBA- I5 ,4,00 GO et 64bits) (sans compter l'exécution de l'exemple traité qui dure plus de deux jours), par contre la simulation de plusieurs échantillons de tailles différentes de 100 à 10 000 ne prend que 0.82 (S), donc on peut conclure que le bootstrap donne de meilleurs résultats par rapport à

la méthode inférentielle asymptotique mais il est plus couteux, il met plus de temps et il demande de matériel informatiques puissants pour faciliter son exécution.

Conclusion Generale

L'étude sur les propriétés asymptotiques des composantes principales d'un échantillon issu d'une population a été abordé de deux approches :

-L'approche classique asymptotique et l'approche de réechantillonage Bootstrap.

Pour mener à terme cette etude nous avons dû réaliser trois programmes sous le logiciel "R"nommés "N method", "sim.boot", et "data.boot".

Nous avons dû etudier les estimateurs des valeurs propres et des vecteurs propres et de leur précisions grace à l'etude du biais, de l'intervalle de confiance et de l'erreur standard.

Les résultats obtenus ont été comparé par les deux méthodes "classique asymptotique" et "bootstrap paramétrique", ces résultats nous ont permis de conclure que le cas de bootstrap paramétrique donne de meilleurs resultats par rapport à la méthode classique asymptotique,donne l'erreur standard la plus faible et l'intervalle de confiance le plus precis.

Par contre en ce qui concerne le pourcentage d'inertie expliqué par la première composante principale la méthode du bootstrap ne donne pas le même résultat que la méthode asymptotique.

Pour la première méthode la première composante principale explique 70% de l'inertie qui est conforme à celle de la population, alors que la méthode Bootstrap donne plus de 70%.

Fort de ces resultats nous avons abordé l'etude asymptotique dans le cas non paramétrique où l'échantillon initial est issu d'une population de distribution inconnue,là encore nous avons obtenu des erreurs standards des vecteurs propres composante par composante faibles et des intervalles de confiance pour les valeurs propres d'amplitude petite donc précis.

On peut conclure que la méthode Bootstrap donne de bonnes approximations des paramètres,cependant on peut avoir des echecs pour certains types complexe de paramètres par exemple le pourcentage d'inertie. On doit noter aussi que le temps d'exécution des programmes Bootstrap est trés important.

Cette méthode de Bootstrap couplée avec des outils informatiques de plus en plus puissants ouvre de nouveaux horizons de recherches pour les populations de distributions complexes différentes de celles classiques utilisées jusqu'a présent.

Bibliography

[1] IT.Jolliffee(second edition2002) : In introduction to the multivariate statistical

[2] A.Anderson (1958) : In introduction to the multivariate statistical analysis

[3] A.Anderson(1963) : Asymptotic theory for principal component analysis Ann.Math.Stat

[4] Jean Jacque Droesbeke, Bernard Fichet et Philippe Tassi : Modeles pour l'analyse des données multidimensionnelles

[5] Lebart L.et Fenelon J.P(1971) : statistique et information appliquées Duno,Paris

[6] R.Tomassone (2008) : L' Analyses en composantes principales

[7] Flury, B.§Riedwyl, H(1988) : Multivariate statistics, A pratical Approach

[8] D. Chessel, A.B. Dufour & J. Thioulouse (2003) : Analyses en composantes principales

[9] G.saporta(1990) : Probabilités, Analyse des données et statistiques

[10] Marie Chavent,Vanessa Kuentz,Jérome Saracco (2007) : Analyse en facteurs"presentation et comparaison des logiciels : SAS, SPAD et SPSS.

[11] Lary Wasserman (2003) : All of Statistics Aconcise course in statistical inference

[12] Utilisation du Bootstrap pour les problèmes statistiques lies a l'estimation des parametres (Biotechnol Agron 2002)

[13] Bredely Efron and Robert j.Tishirani : An introduction to the Bootstrap

[14] Bernard Rappchi (Decembre1994) : Une introduction au Bootstrap

[15] Fisher, R.A. : Statistical Methods and Scientific Inference. , 1993, Macmillan, New York, third edition., Reprinted in Fisher (1990)

[16] Wolfgang Hardle, LEOPOLD SIMAR (April 2003) : Applied multivariate statistical analysis

[17] Jean Bouyer : Methodes statistiques

[18] Nataliya Dragieva(fevrier2008) : Construction d'un intervalle de confiance par la methode de bootstrap et test de permuation presente

[19] Efron.B(1985) : Bootstrap confidence intervals for a class of parametric problems, Biometrica,

[20] C.Urber(septembre2006) : Une methode de reechantillonage :le bootstrap

[21] Irene Buvat (Septembre 2000) : Introduction a l'approche Bootstrap

[22] François Husson, sébastien Lé et Jérome Pagés (2009) : Analyse de données avec R

[23] André Bouchier(2010) : L'analyse des données multivariées à l'aide du logiciel R

[24] Pnger Loic(2008) : Réechantillonage sous R : Bootstrap et Jackknife

[25] D. Chessel & J. Thioulouse(2000) : Fiche d'utilisation du logiciel R– Statistique non paramétrique

[26] AC Davison and Diego Kuonen : An introduction to the Bootstrap with application in R

[27] Nils Pénard, UCB BIOSCIENCES GmbH, Monheim am Rhein, Germany(2012) : Bootstrap Analysis Double-Independent Programming : Issues and Solutions

[28] Christian Jost(2010/11) : Analyses statistiques de base avec R et Rcmdr comme interface graphique

[29] Hubert RAYMONDAUD LEGTA de Carpentras (2012) : Simulations, Algorithmes en Probabilité -Applications avec R

[30] Scherrer (2007, vol, 1) ; Legende(1988) ; Socal & Rohlf(1981) : Tests de normalité

[31] Anthony Davision (2007) : cours de probabilité et statistique

[32] Ricco Rakotomalala(2011) : tests de normalité, techniques empiriques et test statistiques