

REVUE DE STATISTIQUE APPLIQUÉE

S. BEN AMMOU

G. SAPORTA

Sur la normalité asymptotique des valeurs propres en ACM sous l'hypothèse d'indépendance des variables

Revue de statistique appliquée, tome 46, n° 3 (1998), p. 21-35.

http://www.numdam.org/item?id=RSA_1998__46_3_21_0

© Société française de statistique, 1998, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/legal.php>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

SUR LA NORMALITÉ ASYMPTOTIQUE DES VALEURS PROPRES EN ACM SOUS L'HYPOTHÈSE D'INDÉPENDANCE DES VARIABLES

S. Ben Ammou*, G. Saporta**

* *Département de Méthodes Quantitatives,*

Faculté de Droit et des Sciences Economiques et Politiques de Sousse (Tunisie).

** *Département de Mathématiques, Conservatoire National des Arts et Métiers, Paris.*

RÉSUMÉ

Sous l'hypothèse d'indépendance entre variables, les valeurs propres (non triviales) résultant de l'Analyse des Correspondances Multiples sont théoriquement toutes égales. Dans la pratique, à cause des fluctuations d'échantillonnage, on observe des valeurs propres différentes mais proches de la valeur théorique. Nous nous intéressons dans cet article aux lois de ces valeurs propres. Nous montrons que les valeurs propres d'une ACM sont asymptotiquement normales. La convergence étant d'autant plus lente qu'on s'approche des valeurs extrêmes. Nous proposons une procédure empirique de choix du nombre de valeurs propres en ACM.

Mots-clés : *Analyse des Correspondances Multiples, valeurs propres, distribution normale.*

ABSTRACT

Under independence hypothesis, the eigenvalues in MCA are theoretically identical. In practice we observe that the eigenvalues are different, but close to the theoretical value. This paper deals with the distributions of these eigenvalues. We show that eigenvalues in MCA are asymptotically normal. Convergence is very slow especially for the first and the last ones. we propose an empirical procedure for the selection of number of eigenvalues in MCA.

Keywords : *Multiple Correspondence Analysis, eigenvalues, normal distribution.*

1. Introduction

Soient p variables aléatoires qualitatives X_1, X_2, \dots, X_p à m_1, m_2, \dots, m_p modalités respectivement. On notera $q = \sum_{i=1}^p m_i - p$, et l'on supposera que l'on a un échantillon de taille n de X_1, X_2, \dots, X_p .

L'Analyse des Correspondances Multiples théorique de ces p variables aléatoires peut être définie comme la limite quand n tend vers l'infini de l'Analyse Factorielle des Correspondances (AFC) du tableau disjonctif complet à n lignes. Elle s'obtient en effectuant :

- soit l'AFC du tableau contenant les probabilités conjointes deux à deux des p variables ce qui correspond à l'AFC limite d'un tableau de Burt où n tend vers l'infini;
- soit en diagonalisant $1/p$ fois la matrice des probabilités conditionnelles P

$$P = \begin{pmatrix} I_{m_1} & P_{12} & \cdots & P_{1p} \\ & & \ddots & \\ & P_{ij} & & \\ P_{p1} & & & I_{m_p} \end{pmatrix}$$

P_{ij} étant la matrice des probabilités conditionnelles de \mathbb{X}_j sachant \mathbb{X}_i , ou encore la matrice des profils lignes du tableau croisant \mathbb{X}_i et \mathbb{X}_j et I_k la matrice unité d'ordre k , ce qui correspond à la solution habituelle de l'ACM.

Rappelons que les valeurs propres issue de l'AFC du tableau de Burt sont les carrés des valeurs propres issue de l'AFC du tableau disjonctif complet.

Si les \mathbb{X}_j sont deux à deux indépendantes ($\phi_{ij}^2 = 0$, $1 \leq i, j \leq p$, où ϕ_{ij}^2 est le phi-2¹ entre \mathbb{X}_i et \mathbb{X}_j), on obtient une seule valeur propre non nulle multiple, autre que les valeurs triviales 0 ou 1. Cette valeur propre, notée λ , vaut $\frac{1}{p}$ pour l'ACM (ou $\frac{1}{p^2}$ pour l'AFC du tableau de Burt) et est de multiplicité q (cf. Saporta 1990).

En réalité le résultat n'est exact que pour une taille d'échantillon infinie.

Lorsqu'on observe un échantillon, même si les variables aléatoires \mathbb{X}_i et \mathbb{X}_j sont indépendantes, les fluctuations d'échantillonnage conduisent à des tableaux tels que les phi-deux observés φ_{ij}^2 ne sont plus nuls, donc à des valeurs propres différentes de $\frac{1}{p}$.

Dans la pratique, et sous les mêmes hypothèses d'indépendance entre variables, nous n'observons plus une valeur propre multiple (de multiplicité q) égale à $\frac{1}{p}$, mais nous observons q valeurs propres notées μ_i , différentes et proches de $\frac{1}{p}$ (donc proches de la valeur propre théorique λ). Nous nous intéressons dans la suite aux lois de ces valeurs propres.

2. Lois asymptotiques des valeurs propres

Nous commençons d'abord par un rappel sur les lois des valeurs propres en AFC, ensuite nous étudions le cas de l'ACM sous l'hypothèse d'indépendance deux à deux des variables.

¹ De manière générale, on notera ϕ^2 pour des phi-deux théoriques, et φ^2 pour des phi-deux observés. Le phi-deux observé n'étant autre que le χ^2 d'écart à l'indépendance divisé par la taille n de l'échantillon.

2.1 Lois des valeurs propres en AFC

La connaissance de la distribution des valeurs propres en AFC permet de déterminer le nombre d'axes principaux à retenir pour la reconstitution des données. Compte tenu de leur complexité, la recherche de ces distributions a fait l'objet de plusieurs résultats erronés, où les valeurs propres sont supposées suivre, comme l'inertie totale, des lois du χ^2 . Lebart et al. (1995) rappellent que Lancaster a réfuté cette idée en montrant que l'espérance mathématique de la plus grande des valeurs propres est toujours supérieure à celle du χ^2 à $(m_1 - m_2 + 1)$ degrés de liberté (m_1 et m_2 étant les nombres de modalités des deux variables avec $m_1 \geq m_2$).

Lebart (1976) fut le premier à établir, que sous hypothèses d'indépendance entre les lignes et les colonnes du tableau de contingence, la distribution des valeurs propres en AFC peut être approchée par celle des valeurs propres d'une matrice dont la loi est connue ($W_{(m_1-1)(m_2-1)}(\text{Min}(m_1 - 1, m_2 - 1), I)$, matrice de Wishart à $(m_1 - 1)(m_2 - 1)$ d.d.l. en dimension $\text{Min}(m_1 - 1, m_2 - 1)$ avec la matrice unité I).

M.E. O'Neill (1978, 1980) s'est intéressé au comportement asymptotique des valeurs propres en Analyse Canonique et en AFC.

Dans le cas de l'AFC, en supposant que les valeurs propres non nulles sont simples, O'Neill a montré que la racine carrée d'une valeur propre (théoriquement non nulle) suit asymptotiquement une loi normale, et retrouve le résultat établi par Lebart (pour les valeurs propres théoriquement nulles).

A. Ghomari (1983) retrouve (par des méthodes différentes) les résultats de O'Neill, dans le cadre de l'Analyse Canonique. Il les complète par des résultats asymptotiques, valables sous des hypothèses plus larges (loi quelconque admettant uniquement un moment d'ordre 4), ce qui permet d'obtenir des tests asymptotiques en s'affranchissant des hypothèses habituelles de normalité.

Il effectue une étude asymptotique, par échantillonnage, de l'Analyse Canonique, en général, et de l'AFC comme étant un cas particulier d'Analyse Canonique non linéaire de deux variables discrètes ou bien comme Analyse Canonique linéaire de deux variables multidimensionnelles (les indicatrices des modalités des deux variables).

L'intérêt des résultats de Ghomari est de pouvoir revenir, dans un cadre inférentiel, sur des procédures statistiques de tests de diverses hypothèses, ou

d'estimation. La quantité $\sum_{i>m_2-s}^{m_2} n\lambda_i^n$ (où λ_i^n est la $i^{\text{ème}}$ valeur propre issue de

l'AFC d'un tableau de contingence $m_1 \times m_2$ avec $m_2 \leq m_1$ construit à partir de n observations) peut être considérée comme statistique de test de nullité des s dernières valeurs propres et suit asymptotiquement un $\chi_{s(m_1-m_2+s)}^2$, pour cette hypothèse.

Ghomari étudie d'abord la convergence de l'Analyse Canonique effectuée sur un échantillon de taille n vers l'Analyse Canonique théorique, ce qui lui permet de déterminer la forme particulière des éléments spectraux de l'Analyse Canonique.

En supposant que les valeurs propres étudiées ne sont pas nulles, et que

$$\frac{p_{ij}}{p_i \cdot p_j} = 1 + \sum_k \sqrt{\lambda_k} e_k^i s_k^j$$

et en adoptant les notations suivantes :

$P = (p_{ij})_{\substack{i=1,\dots,m_1 \\ j=1,\dots,m_2}}$: tableau de fréquences construit à partir de n observations, avec $m_2 \leq m_1$,

$$p_{i.} = \sum_{j=1}^{m_2} p_{ij} \text{ et } p_{.j} = \sum_{i=1}^{m_1} p_{ij} \text{ les marges du tableau } P,$$

e_k^i (respectivement s_1^j) est la $i^{\text{ème}}$ (respectivement $j^{\text{ème}}$) composante du $k^{\text{ème}}$ (respectivement $l^{\text{ème}}$) facteur dans $(\mathbb{R}^{m_1})^*$ (respectivement $(\mathbb{R}^{m_2})^*$)

λ_k^n la $k^{\text{ème}}$ valeur propre de l'analyse sur un échantillon de taille n , et λ_k la $k^{\text{ème}}$ valeur propre de l'analyse théorique.

L'application des résultats à deux variables Z_1 et Z_2 , ayant respectivement m_1 et m_2 modalités, permet d'affirmer que $\sqrt{n}(\lambda_k^n - \lambda_k)$ converge vers une gaussienne centrée et dont la variance dépend de la multiplicité de la valeur propre λ_k .

Dans le cas d'une valeur propre simple, cette variance est donnée par $\sigma = 4\lambda_k \sigma_{kk,kk}$, avec :

$$\sigma_{kl,k'l'} = \sum_{i,i'=1}^{m_1} \sum_{j,j'=1}^{m_2} \frac{1}{4} K_{ij,i'j'} e_k^i e_k^{i'} s_l^j s_l^{j'}$$

$$K_{ijkl} = \frac{p_{ij}}{\sqrt{p_{i.} p_{k.} p_{.j} p_{.l}}} \left[-3 \frac{p_{il}}{p_{i.}} \delta_{ik} - 3 \frac{p_{kj}}{p_{.j}} \delta_{jl} + 4 \delta_{ik} \delta_{jl} + \frac{p_{kl} p_{il}}{p_{i.} p_{.l}} + \frac{p_{kl} p_{kj}}{p_{k.} p_{.j}} \right].$$

Comme application intéressante du résultat, il étudie le cas de l'indépendance et le cas où les s dernières valeurs propres λ_k sont nulles.

Pour le cas de l'indépendance ($p_{ij} = p_{i.} p_{.j}$), il retrouve que la loi limite de $(n\lambda^n)$ pour $m_2 \neq 1$ est celle des valeurs propres (rangées dans l'ordre décroissant) d'une matrice de Wishart à $(m_1 - 1)$ degrés de liberté, et d'opérateur de covariance l'identité dans \mathbb{R}^{m_2-1} (résultat déjà établi par Lebart depuis 1975).

Cette application permet de retrouver le classique test du χ^2 qui mesure l'indépendance des variables, puisque $\sum_{i=2}^{m_2} n\lambda_i^n$ converge vers un χ^2 à $(m_1 - 1)(m_2 - 1)$ degrés de liberté.

Dans le cas où les s dernières valeurs propres λ_k sont nulles, il trouve que la loi limite pour $k > (p - s)$ est identique à la loi conjointe des valeurs propres (rangées dans l'ordre décroissant) de la matrice de loi de Wishart à $(m_1 - m_2 + 1)$ degrés de liberté, et d'opérateur de covariance l'identité dans \mathbb{R}^s .

R. Siciliano (1990), trouve des résultats analogues à ceux trouvés par Lebart (1976) sur les distributions asymptotiques des valeurs propres dans le cas d'une AFC non-symétrique, utilisée lorsque les variables jouent des rôles non-symétriques.

L. Zater (1989), étudie par des techniques de simulation le comportement des valeurs propres issues de l'AFC d'un tableau de contingence. Par simulation de tableaux de contingences à marges libres, puis à marges fixes, il trouve que dans

le cas d'échantillons de taille faible, la distribution des racines carrées des valeurs propres est instable, et que dans le cas d'échantillons de taille moyenne ($n \geq 500$), on peut accepter l'hypothèse de normalité avec les tests usuels. La moyenne empirique des racines carrées des valeurs propres est biaisée positivement par rapport à la valeur théorique, la variance est stable et est proche du biais, pour des valeurs propres non nulles et une taille d'échantillon assez grande.

Il trouve que dans le cas des valeurs propres théoriquement nulles la distribution n'est pas gaussienne et ceci quelle que soit la taille de l'échantillon.

Zater observe aussi que les écarts types des valeurs propres sont rangés dans le même ordre que les valeurs propres.

Il trouve que les parties centrales (interquartiles) des valeurs propres estimées $\hat{\lambda}$ de l'AFC sont disjointes :

si F_i est la fonction de répartition de la racine carrée de $\hat{\lambda}$, $\delta_i = \inf\{\delta/F_i(\delta) \geq 0, 25\}$, le quartile inférieur, $\xi_i = \inf\{\xi/F_i(\xi) \geq 0, 75\}$, le quartile supérieur et $K_i = [\delta_i, \xi_i]$ alors on a : $K_i \cup K_j = \emptyset \forall i \neq j$.

2.2 Lois des valeurs propres en ACM sous l'hypothèse d'indépendance

Soit, X , le tableau disjonctif associé à p variables qualitatives X_i (ayant respectivement un nombre m_i de modalités) observées sur un échantillon de n individus.

Le tableau disjonctif X est formé de p blocs X_i : $X = (X_1|X_2|\dots|X_p)$.

Le rang de X est au plus $\inf\left(\sum_{i=1}^p m_i - p + 1; n\right)$, soit $\sum_{i=1}^p m_i - p + 1$ si $n > \sum_{i=1}^p m_i - p + 1$. On supposera sans nuire à la généralité que n est assez grand, donc que $\text{rg}(X) = \sum_{i=1}^p m_i - p + 1$.

L'Analyse des Correspondances Multiples est l'AFC du tableau disjonctif X . Les facteurs de l'AFC sont les vecteurs propres de $\frac{1}{p} D^{-1} B$,

où $B = X'X$ est le tableau de Burt associé à X , $D = \text{Diag}(X'X) = \text{Diag}(B)$. $D^{-1}B$ est une matrice à diagonale unité.

Sa trace est : $\text{Tr}(D^{-1}B) = \sum_{i=1}^p m_i$

donc $\frac{1}{p} \text{Tr}(D^{-1}B) = \frac{1}{p} \sum_{i=1}^p m_i$

Le nombre de valeurs propres non trivialement égales à 0 ou 1 est :

$$q = \sum_{i=1}^p m_i - p.$$

Leur somme vaut $\sum_{i=1}^q \mu_i = \frac{1}{p} \sum_{i=1}^p m_i - 1$ et leur moyenne vaut $\frac{1}{p}$.

La somme des carrés des valeurs propres non triviales (cf. Saporta 1990) est donnée par :

$$\sum_{i=1}^q (\mu_i)^2 = \frac{1}{p^2} \sum_{i=1}^p (m_i - 1) + \frac{1}{p^2} \sum_{i \neq j} \varphi_{ij}^2$$

où φ_{ij}^2 est le phi-deux observé de K . Pearson du croisement de \mathbb{X}_i avec \mathbb{X}_j .

Si tous ces φ_{ij}^2 étaient nuls on aurait

$$\sum_{i=1}^q (\mu_i)^2 = \frac{1}{p^2} \sum_{i=1}^p (m_i - 1) = \frac{1}{p^2} q$$

et donc

$$\frac{1}{q} \sum_i (\mu_i)^2 = \frac{1}{p^2} = \left[\frac{1}{q} \sum_i (\mu_i) \right]^2$$

Comme la moyenne des carrés ne peut être égale au carré de la moyenne que si toutes les valeurs propres sont égales, on en déduit qu'alors $\mu_i = \frac{1}{p} \forall i$.

On en déduit que l'Analyse des Correspondances Multiples théorique (i.e. échantillon infini), sous l'hypothèse d'indépendance entre les variables \mathbb{X}_i conduit donc à une valeur propre non nulle notée λ égale à $\frac{1}{p}$ de multiplicité q , une valeur propre égale à 1, et $\sum_{i=1}^p m_i - q - 1$ valeurs propres nulles.

Si l'on fait l'hypothèse que les données observées sont un échantillon i.i.d. issu d'une population où les variables aléatoires $\mathbb{X}_1, \mathbb{X}_2, \dots, \mathbb{X}_p$ sont deux à deux indépendantes, donc avec des ϕ^2 (théoriques) nuls, en raison des fluctuations d'échantillonnage les φ^2 (observés) sont non nuls et on observe des valeurs propres différentes de $\frac{1}{p}$.

Si on connaît les probabilités marginales des données, les p blocs du tableau disjonctif sont des variables aléatoires suivant des lois multinomiales d'effectif n et de paramètres $(p_{i1}, p_{i2}, \dots, p_{im_i}) \forall i = 1, \dots, p$. (p_{ik} : proportion d'individus ayant la $k^{\text{ème}}$ modalité de la variable \mathbb{X}_i).

La variable aléatoire \mathbb{X}_i s'écrit alors sous la forme $\mathbb{X}_i = (\mathbb{X}_{i1}, \mathbb{X}_{i2}, \dots, \mathbb{X}_{im_i})$, avec $p_{ik} = p_i^k$ la probabilité pour qu'un individu soit dans la $k^{\text{ème}}$ catégorie du $i^{\text{ème}}$ caractère et p_{kl}^{ij} la probabilité pour qu'un individu soit dans la $k^{\text{ème}}$ catégorie du $i^{\text{ème}}$ caractère et dans la $l^{\text{ème}}$ catégorie du $j^{\text{ème}}$ caractère.

Le ϕ^2 théorique est donné par :

$$\phi^2 = \sum_{k=1}^{m_i} \sum_{l=1}^{m_j} \frac{(p_{kl}^{ij} - p_k^i p_l^j)^2}{p_k^i p_l^j}$$

En raison de la convergence de la loi multinomiale vers la loi normale et donc de toute fonction régulière de ses composantes, chaque valeur propre de $\frac{1}{p} D^{-1} B$ est asymptotiquement gaussienne, donc les μ_i suivent asymptotiquement des lois normales mais celle-ci sont différentes en raison du fait que les valeurs propres sont ordonnées, (par exemple $E(\mu_1) > \frac{1}{p}$ mais $E(\mu_1) \xrightarrow[n \rightarrow \infty]{} \frac{1}{p}$).

L'étude de la convergence des valeurs propres vers des distributions normales et en particulier le problème de la vitesse de convergence, sera effectuée par simulation d'échantillons de tailles croissantes ($n = 100, \dots, n = 10000$).

3. Exemple

Pour illustrer le problème de vitesse de convergence des valeurs propres nous avons effectué des simulations extensives de variables aléatoires multinomiales indépendantes (12 variables dont 4 ayant 2 modalités, 3 ayant 3 modalités, 4 ayant 4 modalités, et une ayant 5 modalités) et pour des tableaux de tailles croissantes ($n = 100, n = 200, n = 500, n = 1000, n = 5000$ et $n = 10000$).

3.1 Processus de simulation

Nous commençons par construire une variable \mathbb{X}_i de la manière suivante : dans une population partagée en m_i catégories, en proportions respectives $(p_{i1}, p_{i2}, \dots, p_{im_i})$, on tire avec remise n individus. L'échantillon ainsi construit comporte des individus appartenant aux différentes catégories. On compte ensuite les effectifs d'individus de cet échantillon appartenant aux diverses catégories; chacun des individus a une probabilité p_{ik} de tomber dans la catégorie k . La variable aléatoire \mathbb{X}_i suit alors, une loi multinomiale d'effectif n et de paramètres $(p_{i1}, p_{i2}, \dots, p_{im_i})$. Concrètement on utilise n fois la fonction 'rantbl' de SAS pour construire les variables multinomiales.

Nous construisons, donc, un tableau de 12 variables aléatoires $\mathbb{X}_1, \mathbb{X}_2, \dots, \mathbb{X}_{12}$ suivant des lois multinomiales indépendantes. Pour cette étude les probabilités marginales ont été prises dans le tableau (page suivante).

L'ACM conduit donc à 26 valeurs propres.

Chaque tableau disjonctif d'effectif n est alors simulé A fois ($A = 150$ ou 600).

3.2 Résultats

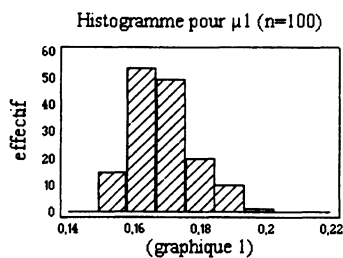
Nous remarquons que la moyenne des réalisations de μ_i diffère significativement de $\frac{1}{p} = \frac{1}{12} = 0.083333$ pour des petites tailles d'échantillons. Le biais est d'autant plus faible que la taille de l'échantillon est grande. De plus la convergence des lois des valeurs propres vers des lois normales se fait de manière très lente. Cette convergence n'est pas très visible pour les valeurs propres extrêmes (μ_1 et μ_q) même pour des échantillons de très grandes tailles.

X_1	$m_1 = 2$	$p_{11} = 0.3$	$p_{12} = 0.7$			
X_2	$m_2 = 2$	$p_{21} = 0.2$	$p_{22} = 0.8$			
X_3	$m_3 = 2$	$p_{31} = 0.6$	$p_{32} = 0.4$			
X_4	$m_4 = 2$	$p_{41} = 0.9$	$p_{42} = 0.1$			
X_5	$m_5 = 3$	$p_{51} = 0.1$	$p_{52} = 0.4$	$p_{53} = 0.5$		
X_6	$m_6 = 3$	$p_{61} = 0.3$	$p_{62} = 0.3$	$p_{63} = 0.4$		
X_7	$m_7 = 3$	$p_{71} = 0.5$	$p_{72} = 0.2$	$p_{73} = 0.3$		
X_8	$m_8 = 4$	$p_{81} = 0.3$	$p_{82} = 0.3$	$p_{83} = 0.3$	$p_{84} = 0.1$	
X_9	$m_9 = 4$	$p_{91} = 0.1$	$p_{92} = 0.1$	$p_{93} = 0.2$	$p_{94} = 0.6$	
X_{10}	$m_{10} = 4$	$p_{101} = 0.5$	$p_{102} = 0.1$	$p_{103} = 0.2$	$p_{104} = 0.2$	
X_{11}	$m_{11} = 4$	$p_{111} = 0.2$	$p_{112} = 0.6$	$p_{113} = 0.1$	$p_{114} = 0.1$	
X_{12}	$m_{12} = 5$	$p_{121} = 0.1$	$p_{122} = 0.1$	$p_{123} = 0.3$	$p_{124} = 0.4$	$p_{125} = 0.1$

Nous présentons ici les histogrammes et les droites de Henry associés à quelques unes des valeurs propres.

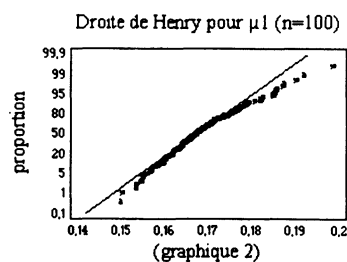
Ces graphiques sont réalisés pour un échantillon de petite taille ($n = 100$) et un échantillon de grande taille ($n = 10\,000$), avec respectivement 150 et 600 réalisations pour chacune des 26 valeurs propres.

La convergence de la loi de la plus grande valeur propre (μ_1) vers une loi normale n'est pas très visible, pour l'échantillon de petite taille ($n = 100$) comme le montrent les graphiques 1 et 2.



GRAPHIQUE 1

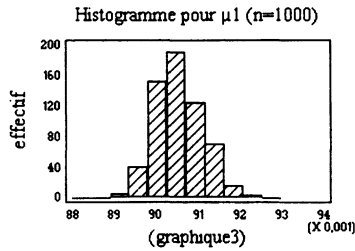
Histogramme pour μ_1 ($n = 100$).



GRAPHIQUE 2

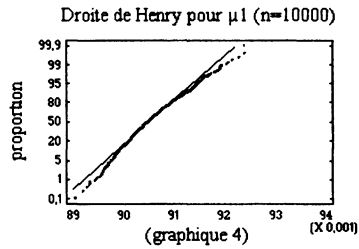
Droite de Henry pour μ_1 ($n = 100$).

L'augmentation de la taille de l'échantillon (ici $n = 10\,000$) ne résout pas le problème, comme le montrent les graphiques 3 et 4 :



GRAPHIQUE 3

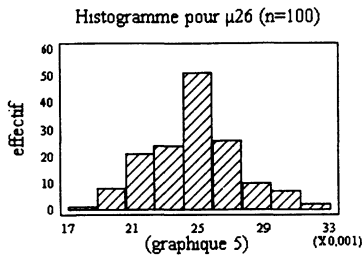
Histogramme pour μ_1 ($n = 1\ 000$).



GRAPHIQUE 4

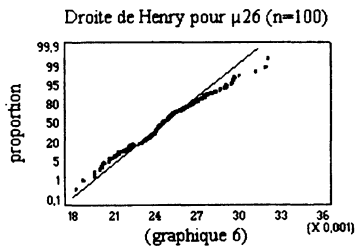
Droite de Henry pour μ_1 ($n = 10\ 000$).

Le même problème se pose pour la plus petite valeur propre (μ_{26}) comme le montrent les graphiques 5 et 6, pour un échantillon de taille $n = 100$.



GRAPHIQUE 5

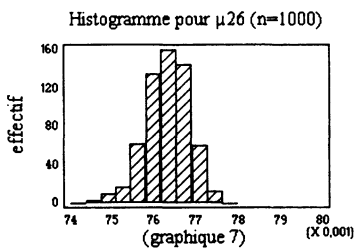
Histogramme pour μ_{26} ($n = 100$).



GRAPHIQUE 6

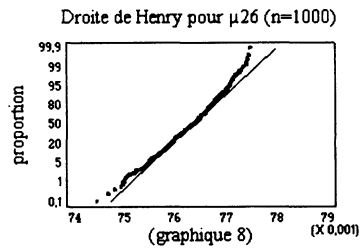
Droite de Henry pour μ_{26} ($n = 100$).

et pour la plus petite valeur propre (μ_{26}) pour un échantillon de taille $n = 1\ 000$ (cf. graphiques 7 et 8).



GRAPHIQUE 7

Histogramme pour μ_{26} ($n = 1\ 000$).

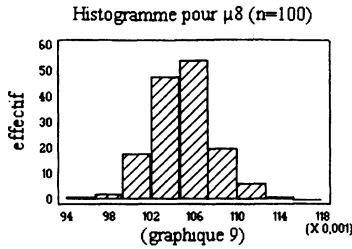


GRAPHIQUE 8

Droite de Henry pour μ_{26} ($n = 1\ 000$).

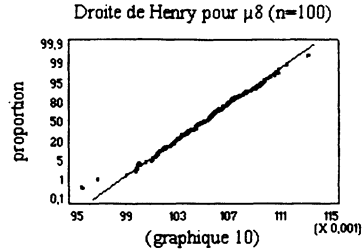
Nous remarquons, par contre, que les graphiques correspondant aux valeurs propres intermédiaires (autres que μ_1 et μ_{26}) montrent bien la convergence des lois de ces valeurs propres vers des lois normales et ceci même pour les échantillons de petite taille, l'augmentation de la taille de l'échantillon permet d'améliorer cette convergence.

Nous présentons ici, pour un échantillon de petite taille ($n = 100$), les graphiques 9 et 10 correspondant à l'une des valeurs propres intermédiaires (ici μ_8).



GRAPHIQUE 9

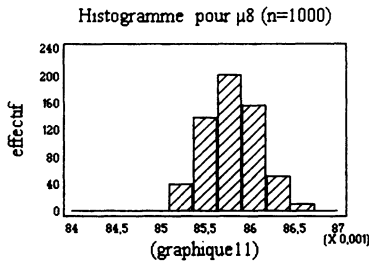
Histogramme pour μ_8 ($n = 100$).



GRAPHIQUE 10

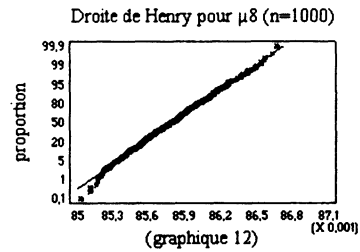
Droite de Henry pour μ_8 ($n = 100$).

et pour un échantillon de grande taille ($n = 1000$), les graphiques 11 et 12 correspondant à cette même valeur propre (μ_8).



GRAPHIQUE 11

Histogramme pour μ_8 ($n = 1000$).



GRAPHIQUE 12

Droite de Henry pour μ_8 ($n = 1000$).

Nous avons refait plusieurs fois l'expérience en changeant les probabilités marginales, le résultat sur la convergence des valeurs propres reste inchangé.

4. Propriétés du spectre et intervalle de variation des valeurs propres

Nous nous intéressons maintenant à la forme du spectre des valeurs propres dans le cas de l'indépendance, puisque dans la pratique et en raison des fluctuations d'échantillonnage, on observe des valeurs propres différentes de $\frac{1}{p}$, $\mu_i = \frac{1}{p} + \varepsilon_i$.

La trace de $\frac{1}{p} D^{-1} B$ étant constante, le spectre moyen $\bar{\mu}$ reste inchangé.

L'étude de la dispersion des valeurs propres, à l'intérieur du spectre correspondant à une réalisation donnée du tableau de Burt, nous permettra de fixer empiriquement un intervalle de variation des valeurs propres sous hypothèse d'indépendance.

4.1 Etude de la dispersion des valeurs propres

Soit S_μ^2 la statistique qui mesure la dispersion des μ_i autour de $\frac{1}{p}$ donnée par :

$$\begin{aligned} S_\mu^2 &= \frac{1}{q} \sum_{i=1}^q \left(\mu_i - \frac{1}{p} \right)^2 = \frac{1}{q} \sum_{i=1}^q \mu_i^2 - \frac{1}{p^2} \\ &\Rightarrow \sum_{i=1}^q \mu_i^2 = q \left(S_\mu^2 + \frac{1}{p^2} \right) \end{aligned}$$

D'après le résultat rappelé en 2.2 on a :

$$\sum_{i=1}^q \mu_i^2 = \frac{1}{p^2} \sum_{i=1}^p (m_i - 1) + \frac{1}{p^2} \sum_{i \neq j} \sum \varphi_{ij}^2$$

or

$$q = \sum_{i=1}^p (m_i - 1)$$

donc

$$\sum_{i=1}^q \mu_i^2 = \frac{q}{p^2} + \frac{1}{p^2} \sum_{i \neq j} \sum \varphi_{ij}^2 = \frac{q}{p^2} + \frac{1}{np^2} \sum_{i \neq j} \sum \chi_{ij}^2$$

En supposant les variables deux à deux indépendantes, les χ_{ij}^2 sont des réalisations de variables suivant des lois $\chi_{(m_i-1)(m_j-1)}^2$ donc d'espérance $(m_i - 1)(m_j - 1)$, ce qui permet de calculer facilement $E\left(\sum_{i=1}^q \mu_i^2\right)$.

En effet nous avons :

$$E\left(\sum_{i=1}^q \mu_i^2\right) = E\left(\frac{q}{p^2} + \frac{1}{p^2} \sum_{i \neq j} \sum \frac{\chi_{ij}^2}{n}\right) = \frac{q}{p^2} + \frac{1}{p^2} \frac{1}{n} \sum_{i \neq j} \sum (m_i - 1)(m_j - 1)$$

Par suite l'espérance de S_μ^2 est :

$$E(S_\mu^2) = \frac{1}{q} E\left(\sum_{i=1}^q \mu_i^2\right) - \frac{1}{p^2} = \frac{1}{q} \left[\frac{q}{p^2} + \frac{1}{p^2} \frac{1}{n} \sum_{i \neq j} \sum (m_i - 1)(m_j - 1) \right] - \frac{1}{p^2}$$

d'où l'on déduit :

$$E(S_\mu^2) = \frac{1}{p^2 n q} \sum_{i \neq j} (m_i - 1)(m_j - 1) \quad (1)$$

Bien que l'échantillon non ordonné des μ_i ne puisse être considéré comme un ensemble de variables aléatoires identiquement distribuées, $E(S_\mu^2)$ représente l'espérance de leur variance expérimentale, nous la noterons par σ^2 .

Par analogie avec une pratique courante, l'intervalle $\frac{1}{p} \pm 2\sigma$ devrait contenir environ 95 % du spectre si celui-ci était constitué de variables normales i.i.d. Cette dernière hypothèse étant fautive, la probabilité de couverture de cet intervalle n'est plus 95 %.

En réalité, cette probabilité est encore plus élevée car la répartition du spectre, si elle est à peu près symétrique par rapport à $\frac{1}{p}$ est bien plus concentrée que celle d'une loi normale (dans le sens où le coefficient d'aplatissement ou de kurtosis est inférieur à celui d'une loi normale).

On peut donc déduire la procédure empirique suivante pour détecter le nombre d'axes utiles en ACM lorsque l'hypothèse d'indépendance ne convient pas (ce qui est le cas le plus usuel ...) : on considérera comme significatives les valeurs propres supérieures à $\frac{1}{p} + 2\sigma$.

Signalons ici que la procédure de «test» à 2 écarts-type est une procédure empirique pragmatique, le facteur 2 (utilisé par analogie à une loi normale) ayant été confirmé par des simulations.

4.2 application à un cas de variables indépendantes

Reprenons les données du § 3, avec des variables indépendantes.

Le spectre moyen est de manière évidente égal à $\frac{1}{12} = 0.083333$.

La première simulation ($n = 100$) nous donne le spectre suivant, avec un écart type, calculé à partir de (1), interne au spectre de 0.0403.

L'écart type estimé est de 0.04198, et l'intervalle [0.0027; 0.1639] contient toutes les valeurs propres, sauf la première et la dernière.

Une autre simulation ($n = 10\,000$) nous donne le spectre suivant, avec un écart type, calculé à partir de (1), interne au spectre de 0.00403.

L'écart type estimé est de 0.004199, et l'intervalle [0.07527; 0.09139] contient toutes les valeurs propres (sauf la dernière).

La concentration du spectre des valeurs propres augmente avec la taille de l'échantillon. En effet, en répétant plusieurs fois les simulations, nous remarquons que pour $n=100$, dans 9 cas sur 10 la première valeur propre n'appartient pas à l'intervalle, soit 3.85 % des valeurs propres sont à l'extérieur de l'intervalle, et la fourchette de 5 % n'est pas dépassée.

0.16902	*****	0.08987	*****
0.15126	*****	0.08910	*****
0.14483	*****	0.08899	*****
0.12939	*****	0.08863	*****
0.12245	*****	0.08677	*****
0.11691	*****	0.08665	*****
0.11256	*****	0.08602	*****
0.11021	*****	0.08575	*****
0.09771	*****	0.08547	*****
0.09411	*****	0.08494	*****
0.08849	*****	0.08484	*****
0.08450	*****	0.08431	*****
0.07451	*****	0.08349	*****
0.06908	*****	0.08283	*****
0.06630	*****	0.08234	*****
0.06114	*****	0.08192	*****
0.05762	*****	0.08140	*****
0.05535	*****	0.08092	*****
0.05187	*****	0.08057	*****
0.04927	*****	0.07971	*****
0.04284	*****	0.07916	*****
0.04211	*****	0.07868	*****
0.03724	****	0.07810	*****
0.02799	***	0.07774	*****
0.02659	***	0.07512	*****

Première simulation

Deuxième simulation

Pour $n = 500$, dans 2 cas sur 10 la première valeur propre n'appartient pas à l'intervalle. Pour $n = 1\ 000$, et $n = 10\ 000$ toutes les valeurs propres appartiennent à l'intervalle (pour 10 simulations).

Dans ces simulations, les variables étaient deux à deux indépendantes.

Le risque α de rejeter à tort l'hypothèse d'indépendance deux à deux des variables quand il existe des valeurs propres à l'extérieur de l'intervalle, est alors d'environ 5 % pour n petit, mais diminue lorsque n augmente.

4.3 Application à un cas de variables non indépendantes

Pour cet exemple nous avons repris des données médicales publiées dans Andersen (1991), concernant l'espérance de vie de malades du cancer de l'ovaire, ayant subi différents traitements postopératoires.

Les données sont sous forme d'un tableau de contingence croisant quatre variables dichotomiques observées sur 299 individus.

A : traitement aux rayons X

B : survie après 10 ans

C : opération radicale ou limitée

D : stade avancé ou non du cancer lors de l'opération

D stade	C opération	B survie	A rayons-X	
			non	oui
non avancé	radicale	non	10	17
		oui	41	64
	limitée	non	1	3
		oui	13	9
avancé	radicale	non	38	64
		oui	6	11
	limitée	non	3	13
		oui	1	5

Nous savons (cf. Andersen 1991) qu'il existe une interaction entre les deux variables B et D.

Notons par X_1, X_2, X_3 et X_4 les quatre variables A, B, C et D.

le nombre d'individus : $n = 299$

le nombre de variables : $p = 4$

le nombre de modalités des variables : $m_1 = m_2 = m_3 = m_4 = 2$

le rang de la matrice X : $q = \sum_{i=1}^p (m_i - 1) = 4$

Nous calculons ensuite la moyenne m et la variance estimée σ :

$$m = \frac{1}{p} = 0.25 \quad \sigma = \frac{1}{p} \sqrt{\frac{1}{q} \frac{1}{n} \sum_{i \neq j} (m_i - 1)(m_j - 1)} = 0.025042$$

L'intervalle à deux écarts-type associé aux valeurs propres est donc : [0.19992, 0.30008]

Nous obtenons les valeurs propres suivantes :

```
0.4145 *****
0.2512 *****
0.2449 *****
0.0894 *****
```

La première et la dernière valeur propre ne sont pas dans l'intervalle, soit 50 % des valeurs propres, le risque α dépasse de loin les 5 %, ce qui confirme la non indépendance entre les variables aléatoires étudiées.

5. Conclusion

L'examen de la distribution des valeurs propres en ACM peut permettre de déterminer le nombre d'axes principaux à retenir pour la reconstitution des données. La complexité des distributions des valeurs propres dans le cas général nous a amené à étudier le cas simple de l'indépendance dans le but de préciser le seuil d'élimination des valeurs propres. Nous proposons ainsi une procédure empirique qui fixe ce seuil en fonction de la variance interne au spectre des valeurs propres. Nous justifions que seules les valeurs propres supérieures à ce seuil sont significatives.

Références bibliographiques

- [1] ANDERSEN E.B. (1991). *Statistical Analysis of Categorical Data*. (Second edition), Springer-Verlag.
- [2] BEN AMMOU S. (1996). *Comportement des valeurs propres en Analyse des Correspondances Multiples sous certaines hypothèses de modèles*. Thèse de l'Université Paris IX Dauphine.
- [3] GHOMARI A. (1983). *Analyse Canonique d'une table de contingence. Etude asymptotique par échantillonnage*. Thèse de l'Université de Pau et des Pays de l'Adour
- [4] LEBART L. (1976). The significance of Eigenvalues issued from Correspondence Analysis. *COMPSTAT*, Physica Verlag, Vienne p. 38-45.
- [5] LEBART L., MORINEAU A., PIRON M. (1995). *Statistique exploratoire multidimensionnelle*. Dunod.
- [6] O'NEILL M.E. (1978). Asymptotic distributions of the canonical correlations from contingency tables. *Austral. J. Statist.* 20 (1) p. 75-82.
- [7] O'NEILL M.E. (1978). Distributional expansion for canonical correlations from contingency tables. *J.R. Statist. Soc. B.* 40, n° 3 p. 301-312.
- [8] O'NEILL M.E. (1980). A note on the canonical correlations from contingency tables. *Austral. J. Statist.* 22 (1) p. 58-66.
- [9] Norme Française NF X 06-050 (Décembre 1991). *Application de la Statistique : Etude de la normalité d'une distribution*.
- [10] SAPORTA G. (1990). *Probabilités, Analyse des Données et Statistique*. Editions Technip.
- [11] SICILIANO R. (1990). Asymptotic distribution of eigenvalues and statistical tests in non symmetric correspondence analysis. *Statistica Applicata*, vol. 2 n° 3, p. 259-276.
- [12] ZATER L. (1989). *Contribution à l'étude de la variabilité des valeurs propres et du choix de la dimension en analyse factorielle des correspondances*. Thèse de l'Université Paris IX Dauphine.