REVUE DE STATISTIQUE APPLIQUÉE

VIDAL COHEN
JACQUES OBADIA

Un exemple d'analyse inverse des données : cas de l'analyse en composantes principales

Revue de statistique appliquée, tome 23, n° 2 (1975), p. 47-59
<http://www.numdam.org/item?id=RSA_1975__23_2_47_0>

UN EXEMPLE D'ANALYSE INVERSE DES DONNÉES : CAS DE L'ANALYSE EN COMPOSANTES PRINCIPALES (1)

VIDAL COHEN
Université Parix IX Dauphine

Jacques OBADIA
CESA Jouy-en-Josas

L'objet de cette étude est d'aborder un problème d'analyse inverse des données dans un cas particulier : celui de l'analyse en composantes principales. Une extension possible est ensuite mentionnée, suivie d'applications numériques qui en marqueront les limites et peut-être l'intérêt.

I – PRESENTATION DES PROBLEMES

Soient $X_1, X_2, \ldots, X_p$, p variables réelles mesurées sur les n individus d'une population ($n > p$). Elles ont donné lieu à la constitution d'un tableau X de données, de n lignes et p colonnes. La colonne k donne les n observations de la kième variable. L'analyse en composantes principales conduit à extraire valeurs propres et vecteurs propres de la matrice V de covariance.

$$V = R'R \qquad \text{et} \qquad R = \frac{1}{\sqrt{n}}(X - \overline{X})$$

$\overline{X}$ est une matrice à n lignes identiques et p colonnes, la kième colonne étant constituée par la moyenne de la variable X_k n fois répétée. (Le symbole A' désignera la transposée de la matrice A).

Les facteurs $f_1, f_2, \ldots, f_p$ sont les vecteurs propres de norme 1 (ils sont rangés dans l'ordre décroissant des valeurs propres associées $\lambda_1 > \lambda_2 > \ldots > \lambda_p$); ils vérifient donc :

$$V f_j = \lambda_j f_j \qquad j = 1, 2, \ldots p \qquad (1)$$

$$f_j' \cdot f_k = \delta_{kj} \qquad k = 1, 2, \ldots p \qquad (2)$$

(δ_{kj} = symbole de Kronecker)

Notons également sous forme matricielle :

$$F = (f_1, f_2, \ldots, f_p)$$

---

(1) Article remis en Décembre 1973, révisé en Octobre 1974.

La jème colonne de F est constituée par les composantes du vecteur propre f_j (par rapport à une base donnée) ; Δ est la matrice diagonale d'ordre p ayant pour éléments diagonaux les valeurs propres λ_i, supposées toutes positives.

Les relations (1) et (2) s'écrivent alors :

$$VF = F\Delta \tag{3}$$

$$F'F = I_p \tag{4}$$

(I_m désignera dans la suite la matrice unité d'ordre m).

I.1 – Premier problème : constitution d'un tableau centré $(X - \bar{X})$ de format (n × p) conduisant à une forme factorielle donnée

Nous appellerons "forme factorielle" et nous noterons $\mathfrak{F} = (F , \Delta)$ tout couple de matrices F et Δ :

a) F matrice orthogonale d'ordre p ;

b) Δ matrice diagonale d'ordre p dont les termes diagonaux sont positifs, distincts et rangés dans l'ordre décroissant.

Nous dirons qu'une matrice X à n lignes et p colonnes admet la forme factorielle $\mathfrak{F} = (F , \Delta)$ si l'analyse en composantes principales de X conduit aux facteurs définis par les colonnes de F avec comme valeurs propres correspondantes les termes diagonaux de la matrice diagonale Δ.

(i) Existence d'une solution

Posons :

. $L = \Delta^{1/2}$.

.[a] le vecteur : colonne de $\mathbf{R}^n$ ou de $\mathbf{R}^p$, selon le cas, dont toutes les composantes sont égales à a.

.Q une matrice à p lignes et n colonnes uniquement soumise aux contraintes :

$$Q . Q' = I_p \tag{5}$$

$$Q[1] = [0] \tag{6}$$

$$. R' = FLQ \tag{7}$$

Dans ces conditions, la matrice $V = R'R$ vérifiera :

$$V = F\Delta F'$$

donc :

$$V f_j = \lambda_j f_j \, (\forall_j)$$

Ce qui établit que $\sqrt{n}\, R = \sqrt{n}\, Q'LF'$ convient comme tableau $(X - \bar{X})$ cherché, la contrainte (6) sur Q assurant le centrage.

(ii) *Unicité de Q à R fixée* : les λ_j étant supposés tous positifs (ce qui n'est pas, dans ce cadre, une limitation importante), L est inversible. Tout tableau centré $\sqrt{n}\,R$ admettant la forme factorielle $\mathcal{F} = (F, \Delta)$ peut s'écrire sous la forme $\sqrt{n}\,Q'LF'$ avec $Q = L^{-1}F'R'$ qui vérifie les contraintes (5) et (6). En résumé, notre problème a en général une infinité de solutions que nous atteignons toutes.

Remarque : Le cas de valeurs propres égales ne présente pas de difficultés particulières, Q étant alors définie à une pré-multiplication près par une matrice orthogonale d'ordre p.

I.2 – Deuxième problème

Un tableau centré $(Y - \bar{Y})$ à n lignes et p colonnes étant donné, comment constituer un tableau centré $(X - \bar{X})$ à n lignes et p colonnes, "proche" du précédent et conduisant à la forme factorielle $\mathcal{F} = (F, \Delta)$ envisagée plus haut ?

D'après le premier problème, si l'on pose $S = \frac{1}{\sqrt{n}}(Y - \bar{Y})$, ce problème se ramène à celui de la recherche d'une matrice Q telle que FLQ soit proche, au sens d'une certaine métrique, de S′ sous les contraintes (5) et (6) :

$$QQ' = I_p$$

$$Q[1] = [0].$$

Adoptant une distance quadratique entre tableaux, nous chercherons à minimiser la somme des carrés des éléments de $S' - FLQ$, et aurons à résoudre le problème d'optimisation (M1) suivant :

Chercher Q minimisant :

$$\mathcal{C} = \text{trace } (S'-FLQ)\,(S-Q'LF')$$

sous les contraintes :

$$QQ' = I_p$$

$$Q[1] = [0]$$

Or, $\mathcal{C} = \text{trace } (S'S) + \Sigma\,\lambda_j - 2 \text{ trace } (FLQS)$ et si nous posons $B = SFL$, le problème (M1) se ramène au problème :

Etant donné une matrice B, quel est le maximum de trace (QB) sous les contraintes $QQ' = I_p$ et $Q[1] = [0]$.

Interprétation géométrique : dans $\mathbf{R}^n$ ou plus précisément dans son plan orthogonal au vecteur [1], on considère p vecteurs : $b_1, b_2, \ldots, b_p$ (vecteurs colonnes de B). Il s'agit de construire dans ce plan p vecteurs orthonormés $q_1, \ldots, q_p$ tels que la somme des produits scalaires

$$K = \sum_{j=1}^{j=p} q_j \,.\, b_j \qquad \text{soit maximum.}$$

Solution algébrique : elle découle de la remarque que si les vecteurs b_j sont orthogonaux deux à deux, K est maximum si chaque q_j est colinéaire à b_j et de même sens. Or, pour toute matrice T orthogonale d'ordre p :

$$\text{Trace } (QB) = \text{Trace } (T'QBT)$$

$$\text{et} \quad \left\{ \begin{array}{l} QQ' = I_p \\ Q[1] = [0] \end{array} \right\} \Longleftrightarrow \left\{ \begin{array}{l} (T'Q)\,(Q'T) = I_p \\ T'Q[1] \quad = [0] \end{array} \right\}$$

On cherchera donc une matrice orthogonale T d'ordre p telle que BT soit formée de p vecteurs-colonnes orthogonaux deux à deux, soit :

$$T'B'BT = A$$

où A désigne une matrice diagonale d'ordre p de terme diagonal général a_j ou encore

$$B'BT = TA\,.$$

Il en résulte que T devra avoir pour colonnes les vecteurs propres orthonormés de B'B, les a_j en étant les valeurs propres associées (supposées dans la suite strictement positives). D'après la remarque précédente, si M est matrice d'ordre p, diagonale de terme général $\sqrt{a_j}$, notre problème a pour solution :

$$Q = TM^{-1}T'B' = TM^{-1}T'LF'S'$$

la condition de centrage (6) étant du même coup vérifiée.

Conséquences

1/ Le tableau cherché est ici :

$(X - \bar{X}) = \sqrt{n}SFLTM^{-1}T'LF'$, donc de la forme $(Y - \bar{Y})Z$, de sorte que les p variables constituant le tableau $(X - \bar{X})$ sont des combinaisons linéaires de celles de $(Y - \bar{Y})$ sur la population envisagée.

2/ A l'optimum :

$$\mathfrak{G} = \text{Trace } (S'S) + \sum_{j=1}^{p} \lambda_j - 2 \sum_{j=1}^{p} \sqrt{a_j}$$

où les a_j sont les valeurs propres de LF'S'SFL ; la somme des carrés des écarts entre tableau donné $(Y - \bar{Y})$ et tableau calculé $(X - \bar{X})$ est égale à $n\,\mathfrak{G}$.

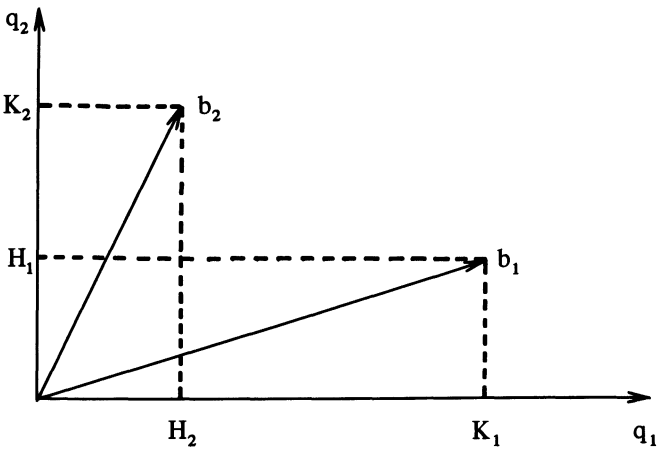
Dans le cas particulier où la forme $\mathfrak{F}$ correspondrait précisément au tableau $(Y - \bar{Y})$, il est clair que :

$$\Sigma\sqrt{a_j} = \Sigma\lambda_j$$

et, par suite, à l'optimum, $\mathfrak{G} = 0$ (ce qui était prévisible !)

3/ Comme, à l'optimum, la matrice QB est symétrique, on peut dire, reprenant l'interprétation géométrique : si les vecteurs q_j et b_j sont choisis

de manière à maximiser $\Sigma b_j \cdot q_j$, la projection de b_i sur q_k est égale à la projection de b_k sur q_i pour tout couple d'indices i , k distincts.



Cas p = 2 : le maximum de $OK_1 + OK_2$ est realisé quand $OH_1 = OH_2$.

Quelques remarques de nature géométrique (par exemple construction des q_j à partir des b_j) ainsi qu'une autre voie (de type JACOBI) de détermination du tableau optimal sont mentionnées par ailleurs[1].

II – GENERALISATIONS ET EXTENSIONS

II.1. – L'étude précédente peut paraître artificielle dans la mesure où, dans une analyse en composantes principales, on se limite généralement aux premiers facteurs correspondant aux plus fortes valeurs propres de la matrice de covariance. Les deux problèmes envisagés peuvent alors être repris dans cette perspective. En particulier, le second conduirait à la construction d'un "tableau optimal" encore plus proche du tableau donné si seuls étaient imposés les r ($r < p$) premiers facteurs et les valeurs propres correspondantes[2].

II.2. – L'analyse en composantes principales fournit une représentation particulière d'un ensemble de données possédant certaines propriétés. Le point de vue "analyse inverse" concerne l'étude de la compatibilité d'ensembles de données avec des "formes" à travers certains modes de représentation. A cet égard, les liens avec les problèmes de stabilité et de test de structure apparaissent clairement. Des études concernant, dans cette perspective, d'autres modes de représentation et les métriques associées feront l'objet de publications ultérieures[3].

---

(1) Publication du laboratoire de Statistique de l'Université PARIS IX Dauphine,

(2) Vidal COHEN-Jacques OBADIA – Inverse Data Analysis. Proceedings du Congrès international de Statistique de Vienne (COMPSTAT). Physica Verlag (1974).

(3) Ces études sont menées dans le cadre d'une recherche D.G.R.S.T. sur le thème : – "Informatique et Sciences Humaines".

III – APPLICATIONS

L'étude précédente n'a pas pour objet principal de montrer comment l'on "doit" choisir ses données pour obtenir une forme de représentation fournie d'avance. Elle s'est imposée à nous dans la mesure où l'on dispose parfois d'une "forme de représentation" dont on désire "tester" la compatibilité avec un ensemble de données recueillies sur des individus. Cette forme de représentation peut être aussi bien issue de l'étude d'un autre ensemble d'individus que d'une analyse du même ensemble d'individus mais par une autre voie (exemples : représentation d'un ensemble de malades selon l'évolution clinique et selon les valeurs de certaines autres variables mesurées sur eux ; analyse de données recueillies à l'instant t et test de la compatibilité d'une structure sous-jacente ainsi dégagée avec un ensemble de données analogues recueillies à un autre instant t'). L'écart à l'optimum n $\mathfrak{C}$ pourrait relever de tests du type χ^2 moyennant certaines hypothèses sur l'incertitude des mesures.

Notons que plusieurs études –dont certaines récentes– concernent plutôt l'approximation des représentations que celle des données. Nous avons choisi cette dernière car il est souvent plus difficile d'interpréter pratiquement un écart entre facteurs, par exemple, qu'un écart entre données.

Signalons aussi que c'est à titre expérimental que nous faisons usage dans la suite de tableaux simulés de variables indépendantes, ce qui n'est guère, bien sûr, un cas "normal" d'application de l'analyse en composantes principales : l'importance des lois de répartition des variables par rapport à ce type d'analyse apparaîtra ainsi peut-être plus clairement.

III.1. – Enrichissement d'une population

Soit X_0 un tableau de données, résultat de la mesure de p variables sur les n individus d'une population $\mathcal{R}$. Ce tableau X_0 a conduit à la forme factorielle $\mathcal{F} = (F, \Delta)$. Nous nous intéressons au problème suivant : cette structure sous-jacente reste-t-elle "acceptable" lorsque nous ajoutons k individus supplémentaires à la population $\mathcal{R}$ ou, en d'autres termes, la forme factorielle est-elle compatible avec le tableau de données X_k obtenu en ajoutant, au tableau X_0, k lignes, profils des k individus supplémentaires ?

L'analyse inverse des données nous permet d'avoir un premier élément de réponse à ce problème. Adoptons, pour cela, le modèle suivant :

$$X_k = \sqrt{n+k}\ Q'LF' + E$$

(les notations sont celles des sections précédentes) ;

E désigne la matrice des écarts entre X_k et $R = \sqrt{n+k}\ Q'LF'$; la matrice R admet la forme factorielle $\mathcal{F} = (F, \Delta)$; en choisissant Q' telle que la somme des carrés des écarts $(n + k)\mathfrak{C}_k$ soit minimum, nous disposons d'un indice permettant de mesurer l'écart entre la population initiale et la population enrichie, relativement à l'analyse adoptée.

Nous avons considéré le cas d'une population de n = 50 individus décrits par 10 variables dont les observations ont été simulées et donc constituées par des nombres pseudo-aléatoires.

Dans un premier cas, ces nombres pseudo-aléatoires ont été "tirés" dans une loi normale $\mathcal{N}$ (0,1) (résultats donnés dans le tableau I), dans un second cas, ils ont été "tirés" dans une loi uniforme sur l'intervalle [0,1], (résultats donnés dans le tableau II).

Les k "individus" supplémentaires enrichissant la population sont chaque fois décrits suivant le même mode que ceux de la population initiale (c.a.d. tirage des observations dans la même loi).

Tableau I — Cas de loi normale $\mathcal{N}$ (0,1)

n + k	trace (S'S)	$\Sigma\lambda_j$	2 trace (QB)	$\mathfrak{C}$	(n + k)$\mathfrak{C}$
50	10.7262	10.7262	21.4523	0.0000	0.0000
55	10.5388	10.7262	21.2229	0.0420	2.2098
60	10.3601	10.7262	21.0333	0.0529	3.1767
65	10.4394	10.7262	21.0836	0.0819	5.3218
70	10.4025	10.7262	21.0355	0.0931	6.5185
75	10.4895	10.7262	21.0776	0.1380	10.3493
80	10.2994	10.7262	20.8735	0.1521	12.1667
85	10.2826	10.7262	20.8329	0.1758	14.9449
90	10.2401	10.7262	20.7711	0.1952	17.5651
95	10.2075	10.7262	20.7179	0.2157	20.4958
100	10.2040	10.7262	20.6740	0.2561	25.6138
105	10.2041	10.7262	20.6947	0.2356	24.7360
110	10.2068	10.7262	20.7076	0.2253	24.7848
115	10.0911	10.7262	20.5805	0.2367	27.2241
120	10.2307	10.7262	20.7008	0.2560	30.7179
125	10.1288	10.7262	20.5778	0.2772	34.6495
130	10.1156	10.7262	20.5568	0.2849	37.0416
135	10.1362	10.7262	20.5710	0.2913	39.3298
140	10.1088	10.7262	20.5547	0.2862	39.2235
145	10.0567	10.7262	20.4905	0.2924	42.3945

III.2. — Compatibilité d'une structure avec un tableau de données

Donnons-nous une forme factorielle $\mathcal{F}_1 = (F_1, \Delta_1)$ issue d'un tableau X_1 de format (n x p) et un tableau de données X_2 de même format. Ces deux tableaux sont supposés décrire la même population $\mathcal{P}$ constituée de n individus.

La somme des carrés des écarts $n\mathfrak{C}$ entre le tableau X_2 et le tableau X_2^* le plus proche de X_2 (au sens des moindres carrés) et admettant la forme factorielle $\mathcal{F}_1 = (F_1, \Delta_1)$ permettra de "tester" la compatibilité de cette forme avec le tableau X_2 ("tester" n'étant pas pris ici au sens strict, puisque la loi de ces écarts n'est pas connue).

Nous avons traité le cas où X_2 de format (20 , 10) était constitué de nombres pseudo-aléatoires tirés dans une loi normale $\mathcal{N}(0,1)$; X_2 a été ensuite centré (tableau III). La forme factorielle est donnée par le tableau V. L'analyse inverse des données conduit à la matrice X_2^* de format (20 , 10) donnée par le tableau IV. Les deux ensembles de données ont été constitués sans introduire de dépendance statistique entre les variables. Cependant, les lois de répartition des variables sont différentes : loi uniforme dans le cas de X_1, loi normale dans celui de X_2. L'écart $n\mathfrak{C}$ qui en résulte est de 53,5968. Mais on pourrait comparer aussi les tableaux III et IV par lignes ou par colonnes.

Remarque : Lorsque nous échangeons les rôles des tableaux X_1 et X_2 – c'est-à-dire que nous cherchons le tableau X_1^* le plus proche de X_1 et admettant la forme factorielle $\mathcal{F}_2 = (F_2 , \Delta_2)$ issue de X_2, nous constatons le même écart $n\mathfrak{C}$. Cette propriété est démontrée en annexe.

Tableau II – Cas de loi uniforme sur l'intervalle [0,1]

n + k	trace (S'S)	$\Sigma\lambda_j$	2 trace (QB)	$\mathfrak{C}$	(n + k) $\mathfrak{C}$
50	0.8700	0.8700	1.7400	0.0000	0.0000
55	0.8613	0.8700	1.7287	0.0026	0.1456
60	0.8615	0.8700	1.7260	0.0056	0.3358
65	0.8619	0.8700	1.7227	0.0092	0.9942
70	0.8516	0.8700	1.7074	0.0142	0.9942
75	0.8609	0.8700	1.7128	0.0181	1.3590
80	0.8590	0.8700	1.7106	0.0184	1.4692
85	0.8514	0.8700	1.7015	0.0199	1.6923
90	0.8502	0.8700	1.7006	0.0196	1.7636
95	0.8497	0.8700	1.6983	0.0213	2.0258
100	0.8497	0.8700	1.6961	0.0236	2.3645
105	0.8475	0.8700	1.6925	0.0250	2.6275
110	0.8424	0.8700	1.6884	0.0240	2.6428
115	0.8457	0.8700	1.6896	0.0261	3.0068
120	0.8449	0.8700	1.6882	0.0267	3.2052
125	0.8457	0.8700	1.6865	0.0292	3.6486
130	0.8512	0.8700	1.6930	0.0282	3.6638
135	0.8556	0.8700	1.6975	0.0281	3.7936
140	0.8475	0.8700	1.6891	0.0283	3.9687
145	0.8460	0.8700	1.6887	0.0273	3.9610

Remarque : Pour rendre les résultats du tableau II comparables à ceux du tableau I, il suffit de les multiplier par 12, inverse de la variance d'une loi uniforme sur [0,1]. (Les deux lois considérées ont alors même variance 1).

Tableau III : données *simulées* selon la loi normale centrée et réduite

.84261	−.98900	.52190	.16371	.00361	−1.04014	−1.89650	1.64863	.07199	1.05650
−1.01665	.17366	.46619	−.30131	1.21658	−.86737	−.38390	.04113	.04833	.87926
.41796	1.69666	.08166	.18250	.72741	−1.24365	−.40454	−.63249	.35689	.04547
−.23319	.40619	−.70622	−.98349	.40789	−.44947	.61868	−.95647	−2.34196	−.62098
.08618	−.88475	−.89189	.61278	.81064	.38034	.07183	−.11652	1.06127	−.58686
.26053	−1.61186	−.63447	.16193	−.55548	−1.23487	.84701	.98005	.18406	.27648
−.16351	−.11478	−.52146	−.07884	−.72239	1.64530	−.13261	−.63826	1.67045	.22264
−.99763	−.23004	.57059	−.86717	1.85247	−.95376	.26513	.28663	.39352	1.98952
.58788	1.11977	−.34963	1.76060	−2.66504	−.10817	−.44661	−.25802	−.18532	−1.09463
−.95788	.04090	1.18287	1.90512	−.35158	.78591	1.54917	−.31108	.12240	−.29972
1.92646	−1.71259	−.88496	−.91698	.70932	−1.27341	−1.50653	.36583	1.17712	.88301
−.51701	.78128	−.31230	−.95309	.35180	.36136	−.28414	.06235	1.37198	−.91769
.78218	−.02993	−.47441	.24959	−1.36055	−1.50952	.26657	.60648	−.37392	1.94048
−.49490	1.64884	−.42963	−1.25297	.20096	1.01071	−.26842	.99095	2.25117	.57013
.59291	1.52579	−.73266	.36070	−1.40987	1.91638	.80413	−.19060	−.27039	−.42197
.80936	−.89938	−.16050	−.57148	1.95709	−1.25426	1.38483	−1.86895	−1.52390	−1.18012
.19997	−.05312	1.37163	.61862	−.17208	2.26845	.90777	−.85139	−1.83531	−1.16150
−.43815	−.50188	−.61908	.26778	.11166	1.12404	−.79450	1.09014	−.18247	−.27988
−.48997	−1.12671	2.36409	−1.35390	1.26631	−.59662	−.92793	−.89250	−1.38766	.20207
−1.19712	.76094	.15831	99591	−2.37876	1.03876	.33052	.64409	−.60826	−1.50222

Tableau IV : données *calculées* (à partir de la forme factorielle décrite au tableau V) de manière à approcher "au mieux" celles du tableau III

.55093	−.94078	.67789	.32222	−.59706	−.84963	−1.84856	1.68731	−.22106	.06625
−1.67513	.35161	−.29713	−.01048	1.02430	−.61935	−.73917	−.09572	−.30611	.35851
.12667	1.79370	−.28901	.17143	.44760	−1.38700	−.72990	−1.01261	.21728	−.38754
−.81264	1.32355	−1.44284	−.70901	.39423	−.37990	.25822	−.49115	−2.39837	.46333
−.18179	−.72526	−.46475	.79593	.95341	.50101	.25259	.05259	1.21726	−.53539
.60305	−1.41365	−.03604	.00136	−.63087	−.95279	1.42493	1.54460	.32618	−.01478
.20343	−.90736	.03510	−.23335	−.68903	1.40727	−.26426	−2.08500	1.72067	.67749
−1.30370	−.19824	−.17039	−.59757	1.47508	−.24193	.57589	.06589	−.06270	1.26563
74206	.86847	−.24031	1.49516	−2.25886	−.91426	−1.38357	−.70923	−.27490	−.58775
−.42134	−.28070	.71650	1.67589	41822	.95919	1.65850	.40770	.36210	−.16227
1.89257	−1.57038	.21101	−1.04169	−.71777	−.93042	−1.07012	−.61827	1.58224	.43481
−.58689	.47899	.34043	−1.01617	.60263	−.21789	−.30241	−.02261	1.49822	−1.61417
1.15453	−.15087	−.72718	.18221	−1.73981	−1.17873	.39488	−.17572	−1.05876	2.06083
−.22308	.70818	.16980	−1.10708	.79156	.52747	.24959	.43780	1.44950	−.58236
1.23102	1.06888	−.61663	.39639	−.52051	1.55808	1.11449	−.29313	−.79166	.47443
.74859	27243	−.22231	−.84441	1.13581	−.65302	1.76651	−.98854	−.21076	−.33488
.81685	−.00639	1.04917	.40809	.25539	2.41064	1.21297	.18178	−1.13146	−.03929
−1.08166	−.33987	−.69363	.89356	.77043	1.04375	−.86847	1.69230	−.85012	−.28055
−.68293	−.79790	1.88194	−1.74115	−.03123	−.25361	−1.37943	−.97164	−.18907	.05840
−1.10051	.46561	.11839	.95867	−1.08353	.17119	−.32326	1.39366	−.87849	−1.32060

Tableau V – "forme factorielle", obtenue par traitement d'un tableau X_1 simulé selon la loi uniforme centrée et réduite.

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
1.8155	1.5953	1.3943	1.2917	.8611	.6816	.4686	.2773	.1884	.1268
f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
−.34280	−.22905	.54334	−.01341	−.03822	−.21639	−.31168	−.43405	.44629	.04020
.28401	−.16332	−.20416	−.40385	−.51768	−.20967	−.18344	−.12768	.02384	−.57045
−.14777	.21451	.02853	.25304	−.02319	.29590	−.49502	.54460	.32742	−.36108
.24659	−.35259	.11091	.38795	−.29542	−.15677	.50719	.30656	.43420	.02151
.34693	.55881	−.19109	−.18878	.21320	−.07733	.09454	−.20398	.62046	.09090
.38980	.24251	.41407	.24482	−.34431	.54493	.01267	−.34121	−.15943	−.01907
.42781	.14147	.57766	−.17023	.22474	−.42615	−.12931	.36399	−.22992	−.01056
.27345	−.14601	−.20049	.60686	.41418	−.23287	−.18577	−.33016	−.10016	−.34076
−.42776	.55297	.11191	.21871	−.21087	−.36435	.35365	−.06672	−.15589	−.34619
−.07818	−.17984	.23580	−.28553	.46066	.35436	.42605	−.02739	.08329	−.54537

ANNEXE

Soient X_1 et X_2 deux tableaux de données centrées de formats respectifs (n, p) et (m, p) ; $\mathcal{F}_1 = (F_1, \Delta_1)$ et $\mathcal{F}_2 = (F_2, \Delta_2)$ les formes factorielles associées à ces deux tableaux. Désignons par X_1^* et X_2^* les tableaux proches de X_1 et X_2 et admettant respectivement les formes factorielles $\mathcal{F}_2 = (F_2, \Delta_2)$ et $\mathcal{F}_1 = (F_1, \Delta_1)$.

Pour construire les tableaux X_1^* et X_2^*, nous sommes amenés à chercher les valeurs et vecteurs propres des matrices

$$U_1 = \frac{1}{n} (L_2 F_2' X_1' X_1 F_2 L_2)$$

$$U_2 = \frac{1}{m} (L_1 F_1' X_2' X_2 F_1 L_1)$$

I – RELATION ENTRE VALEURS ET VECTEURS PROPRES DE U_1 ET U_2

Soient : a_j et $u_j (j = 1, 2, \ldots, p)$ les valeurs propres et vecteurs propres correspondants de U_2 ;

b_j et $v_j (j = 1, 2, \ldots, p)$ les valeurs propres et vecteurs propres correspondants de U_1.

Nous avons :

$$\frac{1}{m} L_1 F_1' X_2' X_2 F_1 L_1 u_j = a_j u_j \tag{1}$$

En pré-multipliant par $F_1 L_1$ et tenant compte de :

$$\frac{1}{m} X_2' X_2 = F_2 L_2 L_2 F_2'$$

(1) s'écrit

$$F_1 L_1 L_1 F_1' F_2 L_2 L_2 F_2' F_1 L_1 u_j = a_j F_1 L_1 u_j \tag{2}$$

La relation $\frac{1}{n} (X_1' X_1) = F_1 L_1 L_1 F_1'$ et la pré-multiplication des deux membres de la relation (2) par $L_2 F_2'$ donnent :

$$\frac{1}{n} L_2 F_2' \, (X_1' X_1) \, F_2 L_2 \, (L_2 F_2' F_1 L_1 u_j) = a_j L_2 F_2' F_1 L_1 u_j \qquad (3)$$

soit : $$U_1 (L_2 F_2' F_1 L_1 u_j) = a_j L_2 F_2' F_1 L_1 u_j$$

Les matrices U_1 et U_2 ont mêmes valeurs propres et les vecteurs propres de ces matrices vérifient :

$$v_j = L_2 F_2' F_1 L_1 u_j$$

II – RELATION ENTRE $n\mathfrak{C}_1$ et $m\mathfrak{C}_2$

Désignons par $n\mathfrak{C}_1$ la somme des carrés des écarts entre X_1 et X_1^* et par $m\mathfrak{C}_2$ celle entre X_2 et X_2^* .

$$n\mathfrak{C}_1 = \text{trace}\,(X_1' X_1) + n \,\text{trace}\,(\Delta_2) - 2n \Sigma \sqrt{b_j}$$

$$m\mathfrak{C}_2 = \text{trace}\,(X_2' X_2) + m \,\text{trace}\,(\Delta_1) - 2m \Sigma \sqrt{a_j}$$

tenant compte de :

$$a_j = b_j \qquad (\forall j)$$

$$\text{trace}\,(X_1' X_1) = n \,\text{trace}\,(\Delta_1)$$

$$\text{trace}\,(X_2' X_2) = m \,\text{trace}\,(\Delta_2)$$

$$n\mathfrak{C}_1 = n \,\text{trace}\,(\Delta_1) + n \,\text{trace}\,(\Delta_2) - 2n\Sigma\sqrt{a_j}$$

$$m\mathfrak{C}_2 = m \,\text{trace}\,(\Delta_2) + m \,\text{trace}\,(\Delta_1) - 2m\Sigma\sqrt{b_j}$$

et $$n\mathfrak{C}_1 - m\mathfrak{C}_2 = (n - m)\,(\text{trace}\,(\Delta_1) + \text{trace}\,(\Delta_2) - 2\Sigma\sqrt{a_j})$$

Dans le cas particulier où $n = m$, la somme des carrés des écarts entre X_1 et X_1^* et celle entre X_2 et X_2^* est la même.