REVUE DE STATISTIQUE APPLIQUÉE

J. OBADIA

L'analyse en composantes explicatives

Revue de statistique appliquée, tome 26, n° 4 (1978), p. 5-28

<http://www.numdam.org/item?id=RSA_1978__26_4_5_0>

L'ANALYSE EN COMPOSANTES EXPLICATIVES

J. OBADIA

Centre d'Enseignement Supérieur des Affaires

INTRODUCTION

De nombreuses méthodes d'analyse de données ont pour objet l'étude des relations entre un groupe de variables dites explicatives, et un autre groupe de variables dites expliquées. Ces méthodes sont pour la plupart basées sur un modèle linéaire.

La régression linéaire permet d'expliquer une variable en fonction de plusieurs variables, en établissant la combinaison linéaire :

$$z^* = a_1 x_1 + \ldots + a_p x_p$$

des vecteurs x_k, des observations des variables explicatives, présentant :

1) La corrélation la plus grande avec le vecteur z, des observations de la variable à expliquer. Cette corrélation est appelée corrélation multiple entre z et $x_1, \ldots, x_p$.

2) La variance la plus grande ; cette variance est appelée variance de z expliquée par $x_1, \ldots, x_p$.

Le rapport, variance expliquée sur variance de z, étant égal au carré du coefficient de corrélation multiple, ces deux propriétés sont équivalentes.

L'analyse canonique a pour objet d'établir m couples :

$$(f_r, g_r)$$

de combinaisons linéaires – appelées composantes canoniques – de chaque groupe de variables(*), ayant les propriétés suivantes :

1) $m \leqslant \inf(p,q)$; p étant le nombre de variables explicatives, q le nombre de variables expliquées.

2) La corrélation entre f_r et g_r – appelée corrélation canonique – est maximum, compte tenu des contraintes :

– les corrélations $COR(f_r, f_s)$ et $COR(g_r, g_s)$ sont nulles $(s < r; r = 2, \ldots, m)$;

– Les variances de f_r et g_r sont égales à 1 $(r = 1, \ldots, m)$.

(*) f_r et g_r sont, en fait, des combinaisons linéaires des vecteurs des observations associés aux variables de chaque groupe.

Mots-clés : **Analyse de données, Modèle linéaire d'analyse de données, Variables explicatives, Variables expliquées, Analyse en composantes explicatives, Modèle général d'analyse de données.**

Lorsque $q = 1$, analyse canonique et régression linéaire donnent les mêmes résultats. Le critère, à la base de la méthode, n'est qu'une généralisation de la propriété 1, présentée par la combinaison linéaire z^* calculée par la régression linéaire : il s'agit en effet d'établir la combinaison linéaire, g_r, des variables expliquées, ayant la corrélation multiple la plus grande avec les variables explicatives (compte tenu des contraintes de corrélations nulles avec g_s ; $r > 1$ et $s < r$).

La pratique de l'analyse canonique est relativement difficile : bien que conduisant à des corrélations canoniques très fortes, il est mal aisé d'en interpréter les résultats (cf.(1)). On constate souvent des corrélations canoniques artificiellement importantes, et dues essentiellement à la présence de deux seules variables – l'une expliquée, l'autre explicative – très corrélées (les autres corrélations intergroupes étant faibles). Il est donc dangereux, de se baser uniquement sur les corrélations canoniques, pour porter un jugement global sur les liaisons éventuelles entre les variables des deux groupes. En continuant le parallèle avec la régression linéaire, on constate que si, par construction, g_r a une variance expliquée par la combinaison linéaire f_r des variables explicatives, maximum, celles des variables expliquées elles-mêmes sont loin d'être optimales.

Les m premières composantes canoniques f_r ne sont pas nécessairement, les m combinaisons linéaires, les plus explicatives, des variations de l'ensemble des variables expliquées.

Nous proposons, avec l'analyse en composantes explicatives, d'aborder l'étude des liaisons entre deux groupes de variables, en déterminant les m combinaisons linéaires f_r $(r = 1, \ldots, m)$ des variables explicatives $x_1, \ldots, x_p$. non corrélées deux à deux, et expliquant au mieux, la variance du groupe de variables expliquées $z_1, \ldots, z_q$(*). La variance d'un groupe de variables est définie, comme en analyse en composantes principales, par la somme des variances des variables de ce groupe. La variance expliquée de z_j par f_r est définie, comme en régression linéaire, par la variance de z_j^*, projection orthogonale de z_j sur f_r ; celle du groupe $z_1, \ldots, z_q$, par la somme des variances des z_j^*.

Ce critère est une généralisation immédiate, de la propriété 2 de la combinaison linéaire z^* établie par la régression linéaire. Il permet d'autre part, de conserver les rôles assymétiques joués par les variables expliquées et les variables explicatives (le critère de corrélation maximum attribue, en analyse canonique, un rôle identique à chaque groupe).

Généralisations différentes de la régression linéaire, l'analyse en composantes explicatives et l'analyse canonique se présentent, comme deux méthodes complémentaires d'analyse des liaisons entre deux groupes de variables.

Après un développement théorique indiquant comment sont établies les composantes explicatives, on montrera que toutes les méthodes d'analyse des données basées sur un modèle linéaire (y compris l'analyse en composantes principales et l'analyse canonique) sont des cas particuliers de l'analyse en composantes explicatives.

Notons que l'analyse en composantes explicatives a déjà été introduite, dans un contexte fort différent par Robert et Escoufier (4) qui parlent de composantes principales des variables x_i relativement aux variables z_j, sans en détailler les

---

(*) les x_k et les z_j désignent, en fait, les vecteurs observations associés aux variables explicatives et aux variables expliquées.

propriétés. Robert et Escoufier font également une synthèse des techniques usuelles d'analyse linéaire des données, sans se servir de l'analyse en composantes explicatives, à l'aide d'un coefficient RV que l'on peut considérer comme un coefficient de liaison entre deux tableaux de données.

NOTATIONS ET DEFINITIONS

1. R^n désigne l'espace vectoriel réel à n dimensions muni de la base canonique et du produit scalaire :

$$\langle u , v \rangle = \frac{1}{n} \sum_{i=1}^{i=n} u_i v_i \quad (u \in R^n \text{ et } v \in R^n)$$

2. Soit un système de m vecteurs, $y_1, \ldots, y_m$, de R^n.

On notera :

$$E(y_1, \ldots, y_m)$$

le sous espace de R^n, engendré par ces m vecteurs.

3. Soit une matrice A, on notera $\tilde{A}$ sa transposée.

4. Soit une variable statistique observée sur n individus. Le vecteur observation associé à cette variable, est l'élément z de R^n, dont les coordonnées – par rapport à la base canonique – sont les n observations.

Le vecteur observation est dit centré s'il est orthogonal au vecteur 1_n dont les n composantes sont égales à 1.

La dispersion de z est définie par :

$$D(z) = \|z\|^2 = \langle z , z \rangle$$

Le vecteur z est dit réduit si :

$$D(z) = 1$$

Si z est centré, D(z) représente la variance V(z) de z.

Sauf cas explicite, on supposera les vecteurs observations centrés : sous cette hypothèse, un vecteur observation quelconque z, aura donc pour ième coordonnée, $z_i - \bar{z}$, où, z_i est la ième observation, et $\bar{z} = \langle 1_n , z \rangle$, moyenne des observations z_i.

5. Soient q variables statistiques $Z_1, \ldots, Z_q$, observées sur le même ensemble de n individus.

La matrice des données relative à ces observations est la matrice à n lignes et q colonnes :

$$Z = (z_1, \ldots, z_q)$$

où la jième colonne est le vecteur observation z_j associé à la variable Z_j.

La variance totale des vecteurs observations $z_1, \ldots, z_q$, notée VT(Z), est définie par :

$$VT(Z) = \sum_{j=1}^{j=q} \|z_j\|^2 = \frac{1}{n} \text{ Trace } (\tilde{Z}Z) \tag{1}$$

6. Décomposition de la variance

Soient $f_1, \ldots, f_m$, m vecteurs de R^n orthogonaux deux à deux.

6.1. Cas d'un seul vecteur observation

Tout vecteur observation se décompose en fonction des vecteurs f_r (r = 1,..., m) suivant :

$$z = z^* + e$$

avec :

a) $$z^* = \sum_{r=1}^{r=m} < z , \frac{f_r}{\|f_r\|} > . \frac{f_r}{\|f_r\|}$$

projection de z sur le sous-espace vectoriel $E(f_1, \ldots, f_m)$.

b) $$e = z - z^*$$

vecteur résiduel orthogonal à $f_1, \ldots, f_m$.

Tenant compte de l'orthogonalité des vecteurs $f_1, \ldots, f_m$, e, la variance de z s'écrit :

$$\|z\|^2 = \sum_{r=1}^{r=m} \frac{\langle z, f_r \rangle^2}{\|f_r\|^2} + \|e\|^2 \tag{2}$$

On appellera :

– variance de z expliquée par $f_1, \ldots, f_m$ la quantité

$$VE(z ; f_1, \ldots, f_m) = \sum_{r=1}^{r=m} \frac{\langle z , f_r \rangle^2}{\|f_r\|^2} \tag{3}$$

Cette quantité représente la variance du vecteur z* ;

– variance résiduelle de z la quantité

$$VR(z ; f_1, \ldots, f_m) = \|e\|^2 \tag{4}$$

– contribution absolue de f_r à la variance de z, le rapport :

$$CAV(z ; f_r) = \frac{\langle z , f_r \rangle^2}{\|f_r\|^2} \tag{5}$$

– contribution relative de f_r à la variance de z, ou, pourcentage de la variance de z expliquée par f_r, le rapport :

$$CRV(z ; f_r) = \frac{\langle z , f_r \rangle^2}{\|z\|^2 \ \|f_r\|^2} \tag{6}$$

6.2. Cas de plusieurs vecteurs observations

Soient q vecteurs observations $z_1, \ldots, z_q$. D'après les relation (1) et (2) :

$$\|z_j\|^2 = \sum_{r=1}^{r=m} \frac{\langle z_j , f_r \rangle^2}{\|f_r\|^2} + \|e_j\|^2$$

et

$$VT(Z) = \sum_{j=1}^{j=q} \sum_{r=1}^{r=m} \frac{\langle z_j , f_r \rangle^2}{\|f_r\|^2} + \sum_{j=1}^{j=q} \|e_j\|^2$$

On appellera :

– variance de $z_1, \ldots, z_q$ expliquée par les m vecteurs $f_1, \ldots, f_m$ la quantité :

$$VE(Z ; f_1, \ldots, f_m) = \sum_{j=1}^{j=q} \sum_{r=1}^{r=m} \frac{\langle z_j , f_r \rangle^2}{\|f_r\|^2} \qquad (7)$$

– variance résiduelle la quantité :

$$VR(Z ; f_1, \ldots, f_m) = \sum_{j=1}^{j=q} \|e_j\|^2 \qquad (8)$$

Remarque

$$VE(Z ; f_1, \ldots, f_m) = \sum_{j=1}^{j=q} VE(z_j ; f_1, \ldots, f_m)$$

L'ANALYSE EN COMPOSANTES EXPLICATIVES

1. LE PROBLEME ET SA RESOLUTION

Soient deux groupes de variables notées respectivement :

$$X_1, \ldots, X_p$$

et

$$Z_1, \ldots, Z_q$$

que l'on a observées sur n individus.

Désignons par :

$$x_1, \ldots, x_p$$

et

$$z_1, \ldots, z_q$$

les vecteurs observations associés à ces variables et par :

$$X = (x_1, \ldots, x_p)$$

et

$$Z = (z_1, \ldots, z_q)$$

les matrices de données correspondantes.

1.1. Définition

On appelle composante explicative du groupe $z_1, \ldots, z_q$ relativement à $x_1, \ldots, x_p$, la combinaison linéaire f des vecteurs $x_1, \ldots, x_p$, qui conduit à une variance expliquée :

$$VE(Z\,;f) = \sum_{j=1}^{j=q} \frac{\langle z_j\,, f\rangle^2}{\|f\|^2}$$

la plus grande.

Pour déterminer cette composante explicative posons :

$$f = Xu$$

où u désigne le vecteur de R^p dont les coordonnées sont les coefficients de la combinaison linéaire f.

Nous avons alors :

$$\langle z_j\,, f\rangle^2 = \frac{1}{n^2}\,\tilde{u}\tilde{X}z_j\tilde{z}_j Xu$$

$$\|f\|^2 = \frac{1}{n}\,\tilde{u}(\tilde{X}X)u$$

$$VE(Z\,;f) = \frac{\frac{1}{n^2}\,\tilde{u}(\tilde{X}Z\tilde{Z}X)u}{\frac{1}{n}\,\tilde{u}(\tilde{X}X)u} = R(u) \qquad (9)$$

La composante explicative $f = Xu$ est donc déterminée par le vecteur u de R^p maximisant le rapport R(u) de deux formes quadratiques. L'annexe I résout ce problème.

En posant :

$$W = \left(\frac{\tilde{X}Z}{n}\right)\left(\frac{\tilde{Z}X}{n}\right)$$

$$V = \left(\frac{\tilde{X}X}{n}\right)$$

il vient :

$$R(u) = \frac{\tilde{u}Wu}{\tilde{u}Vu} \qquad (10)$$

D'après l'annexe I :

Le vecteur u définissant la composante explicative f doit être vecteur propre de la matrice $S = V^{-1}W$ relatif à la plus grande valeur propre λ_1 :

$$S = \left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}Z}{n}\right)\left(\frac{\tilde{Z}X}{n}\right)$$

et

$$\lambda_1 = R(u)$$

représente la variance de $z_1, \ldots, z_q$ expliquée par f.

Remarque

Le résultat $\lambda_1 = R(u)$ ne dépend pas de la norme du vecteur u. On pourrait normaliser le vecteur u de deux façons :

a) la première conduirait à la composante explicative de variance égale à 1 :

$$\|f\|^2 = \frac{1}{n} \tilde{u}\tilde{X}Xu = 1$$

b) la seconde donnerait une variance de f égale à λ_1 :

$$\|f\|^2 = \frac{1}{n} \tilde{u}\tilde{X}Xu = \lambda_1$$

cette dernière normalisation est justifiée par la suite.

La qualité de l'explication obtenue avec la composante f peut être mesurée :

– au niveau global par l'indice :

$$I(f) = \frac{VE(Z\,;f)}{VT(Z)} = \frac{\lambda_1}{VT(Z)}$$

– au niveau de chaque variable par les contributions relatives de f à la variance de cette variable :

$$CRV(z_j\,;f) = \frac{\langle z_j\,, f\rangle^2}{\|z_j\|^2\ \|f\|^2}$$

Les variables pour lesquelles ce rapport est proche de 1 seront les mieux expliquées par la composante f. L'indice I(f) est compris entre 0 et 1 ; dans le cas d'une variance résiduelle importante, il sera nécessaire, pour améliorer l'explication, d'introduire de nouvelles composantes explicatives.

1.2. Définition

Considérons m combinaisons linéaires $f_1, \ldots, f_m$ de $x_1, \ldots, x_p$:

$$f_r = Xu_r \qquad (r = 1, \ldots, m)$$

Ces combinaisons linéaires seront appelées composantes explicatives du groupe $z_1, \ldots, z_q$ relativement à $x_1, \ldots, x_p$ si :

1) Elles sont orthogonales deux à deux :

$$\frac{1}{n} \tilde{u}_s(\tilde{X}X)u_r = 0 \qquad (r \neq s\,;r,s = 1, \ldots, m) \tag{11}$$

2) La variance expliquée par $f_1, \ldots, f_m$:

$$VE(Z\,;f_1, \ldots, f_m) = \sum_{j=1}^{j=q} \sum_{r=1}^{r=m} \frac{\langle z_j\,, f_r\rangle^2}{\|f_r\|^2}$$

est maximum.

Pour déterminer ces composantes explicatives, remarquons que :

$$VE(Z\,;f_1,\ldots,f_m) = \sum_{r=1}^{r=m} VE(Z\,;f_r)$$

et d'après (9) et (10) :

$$VE(Z\,;f_1,\ldots,f_m) = \sum_{r=1}^{r=m} \frac{\tilde{u}_r W u_r}{\tilde{u}_r V u_r}$$

Les vecteurs $u_1,\ldots,u_m$ déterminant les m composantes explicatives doivent maximiser :

$$R(u_1,\ldots,u_m) = \sum_{r=1}^{r=m} \frac{\tilde{u}_r W u_r}{\tilde{u}_r V u_r}$$

sous les contraintes (11).

D'après l'annexe I :

Les vecteurs $u_1,\ldots,u_m$, définissant les m composantes explicatives $f_1,\ldots,f_m$, sont vecteurs propres de la matrice

$$S = V^{-1}W = \left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}Z}{n}\right)\left(\frac{\tilde{Z}X}{n}\right) \qquad (12)$$

relatifs aux plus grandes valeurs propres $\lambda_1,\ldots,\lambda_m$.

De plus :

a) $$\lambda_r = \frac{\tilde{u}_r W u_r}{\tilde{u}_r V u_r}$$

est la variance expliquée par la rième composante explicative f_r

b) $$VE(Z\,;f_1,\ldots,f_m) = \sum_{r=1}^{r=m} \lambda_r$$

Remarque

Les vecteurs u_r, peuvent être normalisés de façon à avoir,

$$\text{soit} \quad V(f_r) = 1, \quad \text{soit} \quad V(f_r) = \lambda_1$$

(cf. la remarque précédente). Par la suite, on supposera les u_r normalisés suivant l'un des deux modes.

1.3. On supposera, pour que S soit définie, rang(X) = p. Le nombre maximum, t, de composantes explicatives est égal à rang $(\tilde{X}Z)$. En effet, le nombre de valeurs propres de S différentes de zéro est égal au rang de S, et, V étant inversible :

$$\text{rang}(S) = \inf(\text{rang}(W), \text{rang}(V))$$
$$\text{rang}(V) = \text{rang}(\tilde{X}X) = \text{rang}(X)$$
$$\text{rang}(W) = \text{rang}(\tilde{X}Z\tilde{Z}X) = \text{rang}(\tilde{X}Z) \leqslant \inf(\text{rang}(X), \text{rang}(Z)).$$

D'où le résultat annoncé.

2. QUALITE DE L'EXPLICATION FOURNIE PAR LES m COMPOSANTES EXPLICATIVES

2.1. Au niveau global, elle peut être mesurée par l'indice :

$$I(f_1, \ldots, f_m) = \frac{VET(Z ; f_1, \ldots, f_m)}{VT(Z)} = \frac{\Sigma \lambda_r}{VT(Z)}$$

que l'on appellera pouvoir explicatif des composantes explicatives $f_1, \ldots, f_m$. Cet indice est compris entre 0 et 1.

2.2. Au niveau de chaque variable z_j, elle peut être mesurée par :

- les contributions relatives à la variance de z_j

$$CRV(z_j ; f_r) = \frac{\langle z_j , f_r \rangle^2}{\|z_j\|^2 \, \|f_r\|^2}$$

- le cumul de ces contributions

$$C(z_j ; s) = \sum_{r=1}^{r=s} CRV(z_j ; f_r) \quad (s = 2, \ldots, m)$$

REMARQUE

Tous les vecteurs observations étant supposés centrés, les composantes explicatives sont également centrées, et la contribution relative d'une composante explicative f_r à la variance d'une variable z_j, est égale, au carré du coefficient de corrélation, $COR(z_j, f_r)$, entre cette variable et cette composante.

2.3. Soit $t = \text{rang}(\widetilde{XZ})$, le nombre maximum de composantes explicatives. L'annexe II montre, que les projections orthogonales z_j^* et z_j^{**} de z_j respectivement sur $E(f_1, \ldots, f_t)$ et $E(x_1, \ldots, x_p)$ sont égales :

$$z_j^* = z_j^{**}$$

Donc :

$$\frac{\|z_j^{**}\|^2}{\|z_j\|^2} = \frac{\|z_j^*\|^2}{\|z_j\|^2} = \sum_{r=1}^{r=t} \frac{\langle z_j , f_r \rangle^2}{\|z_j\|^2 \, \|f_r\|^2}$$

Les indices $C(z_j ; s)$ sont donc inférieurs ou égaux à :

$$\frac{\|z_j^{**}\|^2}{\|z_j\|^2} \quad (\text{pour tout } s \, ; s = 2, \ldots, m)$$

De plus, il est clair que, pour tout m inférieur ou égal à t :

$$I(f_1, \ldots, f_m) \leqslant \frac{VE(Z ; f_1, \ldots, f_t)}{VT(Z)} = \frac{VT(Z^{**})}{VT(Z)}$$

où Z^{**} désigne la matrice :

$$Z^{**} = (z_1^{**}, \ldots, z_q^{**})$$

3. PONDERATION DES VARIABLES

Dans ce paragraphe, nous allons examiner l'influence, sur les résultats de l'analyse en composantes explicatives, de la pondération des variables. Faire une pondération des variables, implique une transformation des vecteurs x_k et z_j du type :

$$t_k = m_k x_k \qquad (k = 1, \ldots, p)$$

$$y_j = l_j z_j \qquad (j = 1, \ldots, q)$$

où m_k et l_j sont des scalaires positifs.

Désignons par M, la matrice diagonale d'ordre p dont les termes diagonaux sont les coefficients m_k, et L la matrice diagonale d'ordre q dont les termes diagonaux sont les coefficients l_j.

Posons :

$$T = XM$$

$$Y = ZL$$

3.1. Pondération des variables explicatives.

L'analyse est faite avec le tableau T ; la composante explicative $g = Tv$ est définie par le vecteur propre de la matrice :

$$Q = \left(\frac{\tilde{T}T}{n}\right)^{-1} \left(\frac{\tilde{T}Z}{n}\right)\left(\frac{\tilde{Z}T}{n}\right)$$

En remplaçant T, par son expression XM, on obtient :

$$Qv = M^{-1} SMv = v$$

ou

$$SMv = Mv$$

Le vecteur $u = Mv$ definit la composante explicative :

$$f = Xu = XMv$$

c'est-à-dire :

$$f = Tv = g$$

Les composantes explicatives sont donc invariantes lorsque les vecteurs $x_1, \ldots, x_p$ sont pondérés. En particulier, la réduction de ces vecteurs n'affecte pas les résultats.

REMARQUE

Les résultats précédents restent encore valables, si M désigne une matrice carrée d'ordre p inversible. L'analyse en composantes explicatives est invariante pour toute transformation régulière des vecteurs $x_1, \ldots, x_p$.

3.2. Pondération des variables expliquées.

L'analyse est faite avec le tableau Y ; la composante explicative g = Yv est définie par le vecteur propre v de la matrice ;

$$P = \left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}Z}{n}\right) L^2 \left(\frac{\tilde{Z}X}{n}\right) \qquad (13)$$

Cette composante g est différente de la composante f = Xu, u étant vecteur propre de la matrice S définie par (12). Les résultats de l'analyse en composantes explicatives dépendent donc des pondérations des vecteur $z_1, \ldots, z_q$.

En particulier si ces vecteurs sont réduits, la matrice à diagonaliser est la matrice P, définie par (13), L^2, ayant alors pour termes diagonaux, les inverses des variances des vecteurs z_j.

3.3. Réductions des vecteurs observations

En réduisant les vecteurs x_k et z_j $(k = 1, \ldots, p\ ; j = 1, \ldots, q)$ les m premières composantes explicatives :

$$f_r = Xu_r \quad (r = 1, \ldots, m)$$

maximisent :

$$VE(Z\ ; f_1, \ldots, f_m) = \sum_{r=1}^{r=m} \sum_{j=1}^{j=q} \frac{\langle z_j, f_r \rangle^2}{\|z_j\|^2 \|f_r\|^2} = \sum_{r,j} COR^2(z_j . f_r)$$

D'après ce qui précède, les vecteurs u_r, sont les vecteurs propres relatifs aux m plus grandes valeurs propres de :

$$R = (R_{11})^{-1} \cdot R_{12} \cdot R_{21}$$

où :

R_{11} est la matrice des corrélations des $x_1, \ldots, x_p$;

R_{12} est la matrice des corrélations entre le groupe $(x_1, \ldots, x_p)$ et le groupe $(z_1, \ldots, z_q)$;

R_{21} est la matrice transposée de R_{12}.

La $r^{ième}$ composante explicative $f_r = Xu_r$ est donc, la combinaison linéaire des x_k, qui maximise la quantité :

$$\sum_{j=1}^{j=q} COR^2(z_j, f_r)$$

compte tenu des contraintes d'orthogonalité (11). A l'optimum, cette quantité est égale à la $r^{ième}$ valeur propre λ_r.

L'ANALYSE EN COMPOSANTES EXPLICATIVES EST UNE ANALYSE EN COMPOSANTES PRINCIPALES PARTICULIERE.

Il existe une relation étroite entre l'analyse en composantes explicatives de $z_1, \ldots, z_q$ relativement à $x_1, \ldots, x_p$, et, l'analyse en composantes principales des projections orthogonales des z_j sur $E(x_1, \ldots, x_p)$.

Soit z_j^{**} la projection de z_j sur $E(x_1, \ldots, x_p)$:

$$z_j^{**} = X\left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}z_j}{n}\right) \tag{14}$$

Soit Z^{**} la matrice à n lignes et q colonnes :

$$Z^{**} = (z_1^{**}, \ldots, z_q^{**})$$

D'après (14), Z^{**} s'exprime en fonction de Z et de X par la relation

$$Z^{**} = X\left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}Z}{n}\right) \tag{15}$$

Le vecteur v_r de R^q définissant la $r^{ième}$ composante principale $g_r = Z^{**} v_r$ du tableau Z^{**} vérifie :

$$\frac{(\tilde{Z}^{**} Z^{**})}{n} v_r = \lambda_r v_r$$

Soit en remplaçant Z^{**} par l'expression (15) :

$$\left(\frac{\tilde{Z}X}{n}\right)\left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}Z}{n}\right)v_r = \lambda_r v_r$$

Le vecteur :

$$u_r = \left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}Z}{n}\right)v_r$$

est donc vecteur propre relatif à la valeur propre λ_r, de la matrice :

$$S = \left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}Z}{n}\right)\left(\frac{\tilde{Z}X}{n}\right)$$

et la composante explicative f_r correspondante est :

$$f_r = X\left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}Z}{n}\right)v_r = Z^{**} v_r$$

Cette composante f_r, coïncide donc, avec la $r^{ième}$ composante principale de Z^{**}. Elle vérifie :

$$V(f_r) = V(g_r) = \lambda_r$$

Réciproquement, soit u_r le vecteur de R^p définissant la $r^{ième}$ composante explicative $f_r = Xu_r$. La relation

$$Su_r = \lambda_r u_r$$

implique que le vecteur v_r de R^q

$$v_r = \frac{\widetilde{Z}X}{n} u_r$$

est vecteur propre de $\widetilde{Z}^{**} Z^{**}/n$ relatif à λ_r.

La composante principale g_r correspondante est :

$$g_r = Z^{**} \frac{v_r}{(\widetilde{v}_r v_r)^{1/2}}$$

or

$$\widetilde{v}_r v_r = \lambda_r \widetilde{u}_r \frac{(\widetilde{X}X)}{n} u_r$$

a) Si l'on prend :

$$V(f_r) = \widetilde{u}_r \frac{(\widetilde{X}X)}{n} u_r = \lambda_r$$

alors

$$(\widetilde{v}_r v_r)^{1/2} = \lambda_r$$

et

$$g_r = \frac{1}{\lambda_r} (Z^{**} v_r) = \frac{1}{\lambda_r} X\left(\frac{\widetilde{X}X}{n}\right)^{-1}\left(\frac{\widetilde{X}Z}{n}\right)\left(\frac{\widetilde{Z}X}{n}\right)u_r = Xu_r$$

c'est-à-dire

$$g_r = f_r$$

b) Si l'on prend :

$$V(f_r) = 1$$

alors

$$(\widetilde{v}_r\ v_r)^{1/2} = (\lambda_r)^{1/2}$$

et la composante principale

$$g_r = Z^{**} \frac{v_r}{(\lambda_r)^{1/2}}$$

est de variance égale à λ_r et identique à $f_r(\lambda_r)^{1/2}$.

La démontration ci-dessus montre que, l'analyse en composantes explicatives de $z_1, \ldots, z_q$ relativement à $x_1, \ldots, x_p$, coïncide avec l'analyse en composantes principales des projections de $z_1, \ldots, z_q$ sur $E(x_1, \ldots, x_p)$.

Ce résultat permet de déterminer les composantes explicatives en diagonalisant la matrice symétrique :

$$\left(\frac{\tilde{Z}X}{n}\right)\left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}Z}{n}\right)$$

à la place de la matrice non symétrique S définie par (11). Ce qui d'un point de vue programmation informatique est appréciable.

L'ANALYSE EN COMPOSANTES EXPLICATIVES : UN MODELE GENERAL D'ANALYSE DE DONNEES

L'analyse en composantes explicatives peut être considérée comme un modèle général d'analyse de données. En effet, toutes les méthodes linéaires en sont des cas particuliers.

1. LA REGRESSION

Soient $Z_1, X_1, \ldots, X_p$ des variables quantitatives observées sur n individus. L'analyse en composantes explicatives de z_1 relativement à $x_1, \ldots, x_p$ coïncide avec la régression linéaire de z_1 en $x_1, \ldots, x_p$. Dans ce cas particulier, nous avons en effet une seule composante explicative :

$$f_1 = Xu_1$$

et il suffit de montrer (cf. (3)) que :

$$u_1 = \left(\frac{\tilde{X}X}{n}\right)^{-1}\left(\frac{\tilde{X}z_1}{n}\right)$$

est vecteur propre de la matrice :

$$S = \left(\frac{\tilde{X}X}{n}\right)^{-1}\frac{\tilde{X}z_1}{n}\frac{\tilde{z}_1X}{n}$$

Cette propriété est immédiate si l'on prend comme valeur propre :

$$\lambda_1 = \left(\frac{\tilde{z}_1X}{n}\right)\left(\frac{\tilde{X}X}{n}\right)^{-1}\frac{\tilde{X}z_1}{n}$$

Cette valeur propre représente le carré du coefficient de corrélation multiple entre z_1 et $x_1, \ldots, x_p$.

2. L'ANALYSE EN COMPOSANTES PRINCIPALES

Soient $X_1, \ldots, X_p$, p variables quantitatives observées sur n individus. L'analyse en composantes explicatives des vecteurs $x_1, \ldots, x_p$ relativement à

$x_1, \ldots, x_p$ coïncide avec l'analyse en composantes principales du tableau :

$$X = (x_1, \ldots, x_p)$$

En effet, les vecteurs u_r, définissant les composantes explicatives sont dans ce cas, les vecteurs propres de la matrice :

$$V = \left(\frac{\tilde{X}X}{n}\right)$$

matrice des covariances des x_k $(k = 1, \ldots, p)$.

3. L'ANALYSE DISCRIMINANTE

Soient une variable qualitative Q ayant q modalités, et :

$$Z_1, \ldots, Z_q$$

les variables indicatrices associées à ces q modalités ;

Z_j (i) = 1 si i présente la modalité j

Z_j (i) = 0 dans le cas contraire

Supposons ces variables observées sur n individus. Soient d'autre part, $X_1, \ldots, X_p$, p variables quantitatives observées sur ces mêmes individus. Déterminons les composantes explicatives :

$$f_r = Xu_r$$

de $z_1, \ldots, z_q$ relativement à $x_1, \ldots, x_p$.

Les z_j seront pondérés, dans l'analyse, par les poids l_j dont les carrés sont égaux aux coefficients :

$$\frac{1}{p_j}$$

avec :

$$p_j = \frac{n_j}{n}$$

où n_j désigne le nombre d'éléments de l'ensemble M_j, des individus présentant la modalité j de la variable Q.

D'après (13), les vecteurs u_r sont vecteurs propres de la matrice :

$$P = \left(\frac{\tilde{X}X}{n}\right)^{-1} \left(\frac{\tilde{X}Z}{n}\right) L^2 \left(\frac{\tilde{Z}X}{n}\right)$$

avec L, matrice diagonale dont les termes diagonaux sont les l_j définis ci-dessus.

a) La matrice $\widetilde{XZ}/n$ a pour terme général :

$$\sum_{i=1}^{i=n} (x_{ik} - \bar{x}_k)(z_{ij} - \bar{z}_j) = \frac{n_j}{n} (\bar{x}_{jk} - \bar{x}_k)$$

avec :

$$\bar{x}_{jk} = \frac{1}{n_j} \sum_{i \in M_j} x_{ik}$$

moyenne de x_k, calculée sur les éléments appartenant au groupe M_j.

b) D'après a) et la définition de L^2, le terme général de la matrice :

$$\left(\frac{\widetilde{XZ}}{n}\right) L^2 \left(\frac{\widetilde{ZX}}{n}\right)$$

est :

$$w_{kl} = \sum_{j=1}^{j=q} \frac{n_j}{n} (\bar{x}_{jk} - \bar{x}_k)(\bar{x}_{jl} - \bar{x}_l) \qquad (16)$$

On reconnait là, le terme général de la matrice W des covariances intergroupes des $x_1, \ldots, x_p$ (cf. (3)).

Si on impose aux composantes explicatives f_r d'avoir une variance égale à 1, les vecteurs propres u_r de $P = V^{-1}W$ définissent les axes discriminants : fonctions discriminantes et composantes explicatives sont donc égales.

Remarque

Nous avons supposé, dans la démonstration ci-dessus, les z_j centrés. Lorsqu'on abandonne cette hypothèse, les coefficients p_j représentent la dispersion des z_j.

L'analyse en composantes explicatives faite avec les z_j non centrés et réduits, donne encore les mêmes résultats. En effet, les x_k étant centrés, le terme général de la matrice $\widetilde{XZ}/n$ ne change pas et donc la relation (16) reste encore valable.

4. L'ANALYSE DES CORRESPONDANCES

Soient Q_1 et Q_2 deux variables qualitatives, ayant respectivement p et q modalités. Désignons par :

$$X_1, \ldots, X_p$$
$$Z_1, \ldots, Z_q$$

les variables indicatrices correspondantes :

$X_k(i) = 1$ si i présente la modalité k de Q_1
$X_k(i) = 0$ dans le cas contraire

$Z_j(i) = 1$ si i présente la modalité j de Q_2
$Z_j(i) = 0$ dans le cas contraire.

On suppose ces deux variables observées sur un ensemble de n individus. Déterminons les composantes explicatives $f_r = Xu_r$ de $z_1, \ldots, z_q$ relativement à $x_1, \ldots, x_p$.

On fera l'analyse avec :

– les x_k non centrés ;

– les z_j non centrés et réduits.

Si on note :

$p(k, j)$ la proportion d'individus présentant les modalités k de Q_1 et j de Q_2

$p_K(k)$ la proportion d'individus présentant la modalité k de Q_1

$p_J(j)$ la proportion d'individus présentant la modalité j de Q_2

on constate que la matrice :

$$P = \left(\frac{\widetilde{XX}}{n}\right)^{-1} \left(\frac{\widetilde{XZ}}{n}\right) L^2 \left(\frac{\widetilde{ZX}}{n}\right)$$

à diagonaliser est telle que :

a) $\widetilde{XX}/n$ est diagonale ; le kième terme de la diagonale étant $p_K(k)$ $(k = 1, \ldots, p)$;

b) les termes diagonaux de L^2 sont égaux à $1/p_J(j)$ $(j = 1, \ldots, q)$;

c) le terme général de la matrice $\widetilde{XZ}/n$ est $p(k, j)$.

La matrice P a donc, pour terme général :

$$r_{kl} = \sum_{j=1}^{j=q} \frac{p(k, j)}{p_K(k)} \frac{p(l, j)}{p_J(j)} \tag{17}$$

et ses vecteurs propres u_r doivent vérifier

$$\sum_{l=1}^{l=p} \left(\sum_{j=1}^{j=q} \frac{p(k, j)}{p_K(k)} \frac{p(l, j)}{p_J(j)} \right) u_{lr} = \lambda_r u_{kr} \tag{18}$$

Si de plus, on impose aux composantes explicatives une norme égale à 1, il vient :

$$\sum_{k=1}^{k=p} p_K(k) \cdot (u_{kr})^2 = 1 \tag{19}$$

Désignons par :

– O_k l'ensemble des éléments vérifiant la modalité k de Q_1 ;

– M_j l'ensemble des éléments vérifiant la modalité j de Q_2 ;

– $N_{kj} = O_k \cap M_j$.

Nous avons, pour i appartenant à O_k :

$$f_r(i) = u_{kr} \quad \text{pour tout} \quad r \ (r = 1, \ldots, m)$$

et la moyenne de f_r calculée sur M_j s'écrit :

$$f_{jr} = \frac{1}{n_j} \sum_{i \in M_j} f_r(i) = \frac{1}{n_j} \sum_{k=1}^{k=p} \sum_{i \in N_{kj}} f_r(i)$$

soit

$$f_{jr} = \sum_{k=1}^{k=p} \frac{p(k\,,j)}{p_J(j)}\, u_{kr} \qquad (20)$$

Les relations (17), (18), (19), (20) sont celles de l'analyse des correspondances du tableau T, dont le terme général est p(k , j) (cf. (2)).

Ce tableau définit une correspondance entre l'ensemble K des modalités de Q_1 et l'ensemble J des modalités de Q_2. Le nuage, N(J), des q points t_j de R^q dont les kièmes composantes sont :

$$t_{kj} = \frac{p(k\,,j)}{p_J(j)} \quad (k = 1, \ldots, p)$$

admet pour facteurs, les vecteurs u_r, définissant les composantes explicatives f_r ; les composantes principales de ce nuage sont définies par (20).

5. L'ANALYSE CANONIQUE

Les couples canoniques (u_r, v_r) définissant les composantes canoniques :

$$f_r = Xu_r \quad \text{et} \quad g_r = Zv_r$$

vérifient :

a) $V(f_r) = V(g_r) = 1 \qquad (r = 1, \ldots, m)$

b) $COR(f_r\,, f_s) = COR(g_r\,, g_s) = 0 \quad (r \neq s\,; r, s = 1, \ldots, m)$

c)

$$\left(\frac{\tilde{X}X}{n}\right)^{-1} \left(\frac{\tilde{X}Z}{n}\right) \left(\frac{\tilde{Z}Z}{n}\right)^{-1} \left(\frac{\tilde{Z}X}{n}\right) u_r = \lambda_r u_r \qquad (21)$$

d)

$$v_r = \frac{1}{(\lambda_r)^{1/2}} \left(\frac{\tilde{Z}Z}{n}\right)^{-1} \left(\frac{\tilde{Z}X}{n}\right) u_r$$

La matrice $\tilde{Z}Z/n$, étant symétrique définie positive d'ordre q, il existe une matrice T carrée d'ordre q, inversible, telle que :

$$\frac{\tilde{Z}Z}{n} = \tilde{T}T$$

En posant :

$$Y = ZT^{-1}$$

et :

$$Y = (y_1, \ldots, y_q)$$

La relation (21) s'écrit :

$$\left(\frac{\widetilde{XX}}{n}\right)^{-1}\left(\frac{\widetilde{XY}}{n}\right)\left(\frac{\widetilde{YX}}{n}\right)u_r = \lambda_r u_r \qquad (22)$$

Cette relation montre que le vecteur u_r, définit la rième composante explicative de $y_1, \ldots, y_q$ relativement à $x_1, \ldots, x_p$; de plus, la composante canonique f_r, explique, de façon optimale, non pas la variance du groupe $z_1, \ldots, z_q$, mais celle du groupe $y_1, \ldots, y_q$.

Remarque

La matrice T n'est pas unique ; elle peut-être calculée :

– par factorisation de $\widetilde{ZZ}/n$ suivant la méthode de Choleski : T est alors une matrice triangulaire haute. Dans ce cas :

$$Y = ZT^{-1}$$

est la matrice obtenue – au facteur n prés – par orthogonalisation des vecteurs $z_1, \ldots, z_q$, suivant la méthode de Gram-Schmidt.

– par diagonalisation de $\widetilde{ZZ}/n$:

$$\widetilde{ZZ}/n = PD\widetilde{P}$$

P est la matrice des vecteurs propres, de norme 1, de $\widetilde{ZZ}/n$, et D, la matrice diagonale ayant pour termes diagonaux les valeurs propres de $\widetilde{ZZ}/n$; T est alors définie par :

$$T = D^{1/2}\widetilde{P}$$

$$T^{-1} = PD^{-1/2}$$

et

$$Y = ZPD^{-1/2}$$

est la matrice des composantes principales – de variance 1 – du tableau Z.

ANNEXE I

Nous rappelons, dans un premier temps, en 1., certains résultats concernant les vecteurs propres d'une matrice de la forme $V^{-1}W$, V et W étant deux matrices symétriques ; V définie positive ; en 2. et 3. sont établis deux résultats, utilisés en 4. Ce dernier paragraphe est consacré à la résolution du problème de l'optimisation d'une somme de rapports de deux formes quadratiques définies par W et V.

1. Rappels

Soient :

– W et V, deux matrices symétriques d'ordre p définies positives ;

– $u_1, \ldots, u_p$, les vecteurs propres de la matrice :

$$S = V^{-1}W$$

relatifs aux valeurs propres $\lambda_1, \ldots, \lambda_p$ (que l'on suppose toutes différentes et rangées par ordre décroissant) tels que

$$\tilde{u}_r V u_r = 1 \qquad (r = 1, \ldots, p) \tag{1}$$

– U la matrice :

$$U = (u_1, \ldots, u_p)$$

la rième colonne étant formée par la coordonnées du vecteur u_r (R^p est muni de la base canonique).

– D la matrice diagonale dont le rième terme diagonal est la vapeur propre λ_r.

1.1. Les vecteurs propres u_r de $V^{-1}W$ vérifient :

$$\tilde{u}_r V u_s = 0 \qquad (r \neq s\ ; r, s = 1, \ldots, p) \tag{2}$$

1.2. Les relations (1) et (2) impliquent :

$$\tilde{U} V U = I_p \tag{3}$$

où I_p désigne la matrice unité d'ordre p.

1.3. D'après (3) et la relation

$$V^{-1} W U = UD$$

il vient :

$$W = VUD\tilde{U}V$$

ou

$$W = \sum_{k=1}^{k=p} \lambda_k V u_k \tilde{u}_k V \tag{4}$$

2. Soient $v_1, \ldots, v_p$, p vecteurs de R^p vérifiant :

$$\tilde{v}_r V v_r = 1 \qquad (r = 1, \ldots, p) \tag{5}$$

$$\tilde{v}_r V v_s = 0 \qquad (r \neq s\ ; r, s = 1, \ldots, p) \tag{6}$$

et z un vecteur de R^p tel que :

$$\tilde{z} V z = 1 \tag{7}$$

Nous avons alors pour tout $m \leqslant p$:

$$0 \leqslant \sum_{r=1}^{r=m} (\tilde{v}_r V z)^2 \leqslant 1$$

En effet, le vecteur z se décompose suivant :

$$z = \sum_{r=1}^{r=p} (\tilde{v}_r V z)\, v_r$$

et donc d'après (5), (6), (7) :

$$\tilde{z}Vz = 1 = \sum_{r=1}^{r=p} (\tilde{v}_r Vz)^2$$

d'où la relation (8) pour tout m inférieur à p ; l'égalité étant obtenue bien sûr pour m = p.

3. Soient m un entier inférieur à p, et, p scalaires $a_1, \ldots, a_p$ vérifiant :

$$\sum_{k=1}^{k=p} a_k = m$$

$$0 \leqslant a_k \leqslant 1$$

Nous avons alors :

$$\sum_{k=1}^{k=p} \lambda_k a_k \leqslant \sum_{r=1}^{r=m} \lambda_r \qquad (9)$$

$\lambda_1, \ldots, \lambda_p$ désignant toujours, les valeurs propres de $V^{-1}W$ rangées dans l'ordre décroissant.

En effet :

$$\sum_{k=1}^{k=p} \lambda_k a_k = \sum_{r=1}^{r=m} \lambda_r + \sum_{r=1}^{r=m} \lambda_r (a_r - 1) + \sum_{r=m+1}^{r=p} \lambda_r a_r \qquad (10)$$

Or

$$\sum_{r=m+1}^{r=p} \lambda_r a_r \leqslant \lambda_{m+1} \sum_{r=m+1}^{r=p} a_r = \lambda_{m+1} \left(m - \sum_{r=1}^{r=m} a_r \right)$$

$$\leqslant \sum_{r=1}^{r=m} \lambda_{m+1} (1 - a_r) \leqslant \sum_{r=1}^{r=m} \lambda_r (1 - a_r)$$

Ce qui implique

$$\sum_{r=m+1}^{r=p} \lambda_r a_r + \sum_{r=1}^{r=m} \lambda_r (1 - a_r) \leqslant 0$$

et donc, d'après (10), l'inégalité (9).

4. Le maximum de l'expression :

$$R(y_1, \ldots, y_m) = \sum_{r=1}^{r=m} \frac{\tilde{y}_r W y_r}{\tilde{y}_r V y_r}$$

sous les contraintes

$$\tilde{y}_r V y_s = 0 \quad (r = s \,; r = 1, \ldots, m) \tag{11}$$

est atteint pour l'ensemble des vecteurs propres $u_1, \ldots, u_m$ de $V^{-1}W$; ce maximum a pour valeur :

$$R(u_1, \ldots, u_m) = \sum_{r=1}^{r=m} \lambda_r$$

Démonstration

a) Pour tout système de scalaires $c_1, \ldots, c_m$ différents de zéro :

$$R(y_1, \ldots, y_m) = R(c_1 y_1, \ldots, c_m y_m)$$

On peut donc, sans perdre de généralité, supposer vérifiées les relations :

$$\tilde{y}_r V y_r = 1 \quad (r = 1, \ldots, m) \tag{12}$$

et le problème revient à chercher le maximum de :

$$H(y_1, \ldots, y_m) = \sum_{r=1}^{r=m} \tilde{y}_r W y_r \tag{13}$$

sous les contraintes (11) et (12).

b) D'après (4), l'expression (13) devient :

$$H(y_1, \ldots, y_m) = \sum_{r=1}^{r=m} \sum_{k=1}^{k=p} \lambda_k \tilde{y}_r V u_k \tilde{u}_k V y_r$$

$$= \sum_{k=1}^{k=p} \lambda_k a_k$$

avec

$$a_k = \sum_{r=1}^{r=m} \tilde{y}_r V u_k \tilde{u}_k V y_r = \sum_{r=1}^{r=m} (\tilde{y}_r V u_k)^2$$

D'après les résultats du paragraphe 2 appliqués aux vecteurs $u_1, \ldots, u_p$ et au vecteur y_r on obtient :

b1) $$\sum_{k=1}^{k=p} a_k = \sum_{k=1}^{k=p} \sum_{r=1}^{r=m} (\tilde{y}_r V u_k)^2 = m$$

b2) $$0 \leqslant a_k \leqslant 1$$

c) D'après b) les hypothèses du paragraphe 3 sont vérifiées et nous avons :

$$H(y_1, \ldots, y_m) \leqslant \sum_{r=1}^{r=m} \lambda_r \tag{15}$$

Or

$$H(u_1, \ldots, u_m) = \sum_{r=1}^{r=m} \sum_{k=1}^{k=p} \lambda_k (\tilde{u}_r V u_k)^2$$

d'après (2) et prenant des vecteurs propres u_r tels que

$$\tilde{u}_r V u_r = 1 \tag{16}$$

il vient

$$H(u_1, \ldots, u_m) = \sum_{r=1}^{r=m} \lambda_r \tag{17}$$

Le maximum de $H(y_1, \ldots, y_m)$ sous les contraintes (11) et (12) est donc, d'après (15) et (17) atteint pour les vecteurs propres $u_1, \ldots, u_m$ vérifiant (16).

d) D'après a) tout système $v_1, \ldots, v_m$ de vecteur vérifiant

$$v_r = c_r u_r$$

c_r étant un scalaire différent de zéro, est encore solution du problème d'optimisation de l'expression $R(y_1, \ldots, y_m)$ sous les contraintes (11).

ANNEXE II

Soit $t = \text{rang}(\tilde{X}Y)$ le nombre maximum de composantes explicatives. Il s'agit d'établir le résultat suivant : les projections z_j^* et z_j^{**} de z_j, respectivement sur $E(f_1, \ldots, f_t)$ et $E(x_1, \ldots, x_p)$ sont égales.

1) Supposons $p = \text{rang}(\tilde{X}Z)$; (ceci est vérifié, par exemple, lorsque $\text{rang}(\tilde{X}Z) = \inf(p, q)$ et $p < q$). Dans ce cas, $t = p$, et, les composantes explicatives f_r $(r = 1, \ldots, p)$, orthogonales deux à deux, forment une base de $E(x_1, \ldots, x_p)$:

$$E(x_1, \ldots, x_p) = E(f_1, \ldots, f_p)$$

D'où le résultat annoncé.

2) Supposons $p > \text{rang}(\tilde{X}Z)$.

Dans ce cas $t = \text{rang}(\tilde{X}Z) = \text{rang}(Z^{**})$ et les t composantes explicatives f_r, relatives aux t valeurs propres différents de zéro, forment une base de $E(z_1^{**}, \ldots, z_q^{**})$, et donc

$$E(f_1, \ldots, f_t) = E(z_1^{**}, \ldots, z_q^{**})$$

D'où le résultat annoncé.

REFERENCES

[1] BERTIER P., BOUROCHE J.M. – Analyse multidimensionnelle des données (Puf).

[2] BENZECRI J.P. – L'analyse des données (Dunod).

[3] CAILLEZ F., PAGES J.P. – Introduction à l'analyse des données (Smash).

[4] ROBERT P. et ESCOFIER Y. – A unifying tool for linear multivariate statistical methods. The R.V. coefficient. *Applied Statistics*, Vol 25, n° 3 (1976), p. 257/65.