REVUE DE STATISTIQUE APPLIQUÉE

PH. CASIN

L'analyse en composantes principales généralisée

Revue de statistique appliquée, tome 44, n° 3 (1996), p. 63-81

<http://www.numdam.org/item?id=RSA_1996__44_3_63_0>

Rev. Statistique Appliquée, 1996, **XLIV** (3), 63-81

L'ANALYSE EN COMPOSANTES PRINCIPALES GÉNÉRALISÉE*

Ph. Casin

Faculté de Droit et d'Economie, Ile du Saulcy
57005 Metz cedex 1

RÉSUMÉ

L'analyse en composantes principales généralisée de tableaux se propose de déterminer des variables synthétiques décrivant le mieux possible les proximités existant entre des tableaux tout en décrivant le mieux possible les tableaux de départ. Des propriétés de cette technique sont données. Une comparaison avec les techniques usuelles d'analyse de tableaux ternaires est effectuée.

Mots-clés : *Analyse canonique généralisée, technique PLS, tableaux évolutifs, analyse en composantes principales généralisée.*

ABSTRACT

The aim of generalized principal components analysis of data tables is to compute new variables both describing as well as possible proximities between data tables and being as well as possible correlated with initial variables. A theorical comparison with other methods is done.

Keywords : *Generalized canonical analysis, multivariate time series, principal components analysis.*

Lorsqu'on dispose d'un ensemble d'individus décrits par des variables, l'analyse en composantes principales normée ou non fournit une description simplifiée de la configuration des individus et des corrélations entre les variables.

Si l'on dispose de plusieurs tableaux de données définis à partir des mêmes individus, l'analyse canonique (ou ses généralisations) met en évidence les éléments communs à tout ou partie des tableaux.

L'analyse en composantes principales généralisée (ACPG) se propose de concilier ces deux approches, c'est-à-dire de comparer les tableaux tout en décrivant chacun d'entre eux.

* article reçu en mars 1995, révisé en septembre 1995.

1. Les données et leur représentation

1.1 Les données

On dispose de K tableaux X_k; chaque tableau X_k $k = 1, \ldots, K$ décrit les valeurs prises pour les mêmes n individus (les lignes du tableau) pour m_k variables (les colonnes du tableau). La variable numéro v du tableau k est notée $X_{k,v}$.

Les données décrites par le tableau X_k sont centrées pour chaque variable.

1.2 L'espace R^n

Chaque variable, pour chaque tableau, étant définie par les valeurs qu'elle prend pour les n individus peut être représentée par un vecteur de R^n.

La matrice diagonale des poids des individus D_p confère à R^n une structure d'espace vectoriel euclidien; le produit scalaire de deux vecteurs-variable X et Y est égal à leur covariance, notée $\mathrm{cov}(X, Y)$ et le cosinus de l'angle entre X et Y est égal à leur coefficient de corrélation $R(X, Y)$.

On note $ETC(X)$ l'écart-type d'une variable X et $VAR(X)$ sa variance.

Dans R^n, les m_k colonnes de chacun des tableaux X_k, $k = 1, \ldots, K$ engendrent un espace vectoriel W_k, $k = 1, \ldots, K$; soit $m = \sum_{k=1}^{K} m_k$ et $X = [X_1, \ldots, X_k, \ldots, X_K]$. P_k désigne le projecteur orthogonal sur W_k.

2. Le problème

Il s'agit de décrire simultanément un ensemble de tableaux X_k $k = 1, \ldots, K$ décrivant les mêmes individus.

Décrire simultanément, ceci signifie déterminer ce qui est commun à l'ensemble ou à une partie des tableaux, et par conséquent, ce qui différencie certains tableaux des autres, mais aussi décrire complètement chacun des tableaux, et en particulier, mettre en évidence les éléments essentiels de chaque tableau.

En termes d'analyse des données, il s'agit ici de déterminer à l'étape 1 une variable auxiliaire Z^1 telle que ses K projections orthogonales Z_k^1 sur les espaces W_k :

– soient proches les unes des autres, autrement dit que : $\sum_{k=1}^{K} R^2(Z^1, Z_k^1)$ soit élevé;

– aient une inertie élevée, c'est-à-dire que : $\sum_{v=1}^{m_k} \mathrm{cov}^2(\widetilde{Z}_k^1, X_{k,v})$ soit élevé, $\widetilde{Z}_k^1$ désignant la variable Z_k^1 normée à 1.

Aux étapes suivantes, le problème se pose dans les mêmes termes, en imposant de surcroît aux variables Z_k^j d'un même tableau d'être deux à deux orthogonales pour permettre une description de chaque tableau.

Le premier type de critère renvoie à l'analyse canonique, à la généralisation de Carroll (Carroll, 1968, Saporta, 1975)) et à sa variante l'ADT (Casin, 1995), le second à l'analyse en composantes principales (ACP) de chaque tableau X_k et la combinaison des deux, lorsqu'on dispose uniquement de deux tableaux, aux techniques PLS (Tenenhaus, Gauchi et Ménardo, 1995). Aussi, après avoir exposé le principe de la méthode proposée ici et ses propriétés, une comparaison sera faite avec ces trois types de techniques.

3. L'analyse en composantes principales généralisée

3.1 Le principe

On combine les 2 critères précédents.

Z_k^1 étant la projection de Z^1 sur W_k, si Z^1 est normé à 1, on a : $R^2(Z^1, Z_k^1) = VAR(Z_k^1)$ et donc les variables Z_k^1 sont d'autant plus liées entre elles que chaque $R^2(Z^1, Z_k^1)$ est élevé, donc que $VAR(Z_k^1)$ est élevé.

Par conséquent, Z_k^1 est d'autant plus lié aux autres variables Z_k^1, et a une inertie d'autant plus élevée que $\sum_{v=1}^{m_k} \text{cov}^2(Z_k^1, X_{k,v})$ est élevé. En effet :

$$\sum_{v=1}^{m_k} \text{cov}^2(Z_k^1, X_{k,v}) = \sum_{v=1}^{m_k} R^2(Z^1, Z_k^1)\text{cov}^2(\widetilde{Z}_k^1, X_{k,v}) \qquad (1)$$

A l'étape j, du fait des contraintes d'orthogonalisation, on s'intéresse aux proximités entre les variables appartenant aux espaces W_k^j (W_k^j désignant le sous-espace de W_k^j orthogonal à $Z_k^1, \ldots, Z_k^{j-1}$).

Si Z_k^j désigne la projection de Z^j sur W_k^j, Z^j étant normé à 1, on a : $R^2(Z^j, Z_k^j) = VAR(Z_k^j)$ et les variables Z_k^j sont d'autant plus liées entre elles que chaque $R^2(Z^j, Z_k^j)$ est élevé, donc que $VAR(Z_k^j)$ est élevé.

Et, par le même raisonnement que précédemment, Z_k^j est d'autant plus lié aux autres variables Z_k^j et a une inertie d'autant plus élevée que $\sum_{v=1}^{m_k} \text{cov}^2(Z_k^j, X_{k,v})$ est élevée.

3.2 Le critère à maximiser

A l'étape 1, il s'agit donc de déterminer Z^1 et les $Z_k^1 = P_k Z^1$ tels que :

$$C^1 = \sum_{k=1}^{K} \sum_{v=1}^{m_k} \text{cov}^2(Z_k^1, X_{k,v}) \text{ soit maximal} \tag{2}$$

avec $(Z^j)' D_p Z^j = 1$.

A l'étape j, il s'agit de déterminer Z^j et les $Z_k^j = P_k^j Z_k$ (P_k^j désignant le projecteur orthogonal sur W_k^j) tels que :

$$C^j = \sum_{k=1}^{K} \sum_{v=1}^{m_k} \text{cov}^2(Z_k^j, X_{k,v}) \text{ soit maximum} \tag{3}$$

sous la condition de normalisation :

$$(Z^j)' D_p Z^j = 1$$

et les conditions d'orthogonalisation : $(Z_k^j)' D_p Z_k^r = 0$ pour $r < j$ et $k = 1, \ldots, K$.

La recherche s'arrête lorsqu'on a obtenu une base orthogonale de chacun des espaces W_k.

Remarque : il est possible de pondérer chacun des tableaux, par exemple pour tenir compte du nombre différent de variables d'un tableau à l'autre. Il suffit, en fait, de repondérer les variables et l'on se ramène au cas général.

3.3 La solution

A l'étape 1, Z_k^1 est la D_p-projection de Z^1 sur W_k, et donc $Z^1 - Z_k^1$ est orthogonal à W_k, d'où :

$$\text{Cov}\,(X_{k,v}, Z_k^1) = \text{Cov}\,(X_{k,v}, Z^1)$$

pour $v = 1, \ldots, m_k$, pour $k = 1, \ldots, K$ et donc :

$$C^1 = \sum_{k=1}^{K} \sum_{v=1}^{m_k} \text{cov}^2(Z^1, X_{k,v})$$

Par conséquent Z^1 est le vecteur propre normé de $X(X)'D_p$ associé à sa valeur propre la plus élevée. On en déduit les Z_k^1

$$Z_k^1 = P_k Z^1 \text{ pour } k = 1, \ldots, K$$

A l'étape j, Z_k^j orthogonal à Z_k^r (pour $r < j$) signifie que Z_k^j est une combinaison linéaire des variables $X_{k,v}^j$, $X_{k,v}^j$ désignant le résidu de la régression de $X_{k,v}$ par Z_k^r (pour $r = 1, \ldots, j-1$). $X_{k,v} - X_{k,v}^j$ est orthogonal à Z_k^j d'où :

$$\text{cov}(X_{k,v},\ Z_k^j) = \text{cov}\ (X_{k,v}^j,\ Z_k^j)$$

$Z^j - Z_k^j$ est orthogonal à $X_{k,v}^j$ d'où :

$$\text{cov}\ (X_{k,v}^j,\ Z_k^j) = \text{cov}\ (X_{k,v}^j,\ Z^j)$$

Et donc, le problème peut s'écrire à l'étape j : déterminer Z^j et les K variables Z_k^j tels que :

$$C^j = \sum_{k=1}^{K} \sum_{v=1}^{m_k} \text{cov}^2(Z^j,\ X_{k,v}^j) \text{ soit maximal}$$

sous la contrainte : $(Z^j)' D_p\ Z^j = 1$.

Soit $X^j = [X_1^j, \ldots, X_k^j, \ldots, X_K^j]$, X_k^j désignant le tableau dont les m_k colonnes sont les variables $X_{k,v}^j$.

Z^j est le vecteur propre normé de $X^j\ (X^j)'\ D_p$ associé à sa valeur propre la plus élevée et $Z_k^j = P_k^j Z^j$.

Le nombre total d'étapes est noté E; si aucune orthogonalité n'existe entre Z^j et un espace W_k^j (ce qui est le cas général), E est égal à la valeur la plus élevée des dimensions des espaces W_k pour $k = 1, \ldots, K$.

4. Les propriétés de l'ACPG

Propriété 1. L'ACPG généralise l'ACP dans la métrique identité.

Preuve : Lorsque $K = 1$, Z^j se confond avec Z_1^j, Z^j est déterminé en effectuant une ACP dans la métrique identité, ce qui prouve le résultat annoncé.

Remarque : cette propriété justifie le nom donné à la technique «Analyse en composantes principales généralisée» (ACPG).

Propriété 2. Si u_k^j désigne le bloc relatif au tableau X_k du premier vecteur propre u^j de $(X^j)' D_p X^j$ et W_{kk}^j l'inverse (éventuellement généralisé) de $(X_k^j)'\ D_p\ X_k^j$, alors :

$$Z_k^j = v_j\ X_k^j\ W_{kk}^j\ u_k^j$$

v_j désignant la valeur propre associée à u^j et u^j étant normé à $(v_j)^{-0.5}$.

Preuve : Par définition de u^j :

$$\sum_{r=1}^{K}(X_k^j)'\, D_p\, X_r^j\, u_r^j = v_j\, u_k^j$$

d'où : $$\sum_{r=1}^{K} X_k^j\, W_{kk}^j\, (X_k^j)'\, D_p\, X_r^j\, u_r^j = v_j\, X_k^j\, W_{kk}^j\, u_k^j$$

soit : $$P_k^j \left(\sum_{r=1}^{K} X_r^j\, u_r^j\right) = v_j\, X_k^j\, W_{kk}^j\, u_k^j$$

or : $$Z^j = \sum_{r=1}^{K} X_r^j\, u_r^j \quad \text{et} \quad Z_k^j = P_k^j\, Z^j$$

d'où : $$Z_k^j = v_j \quad X_k^j\, W_{kk}^j\, u_k^j$$

Propriété 3. Si u_k^j désigne le bloc relatif au tableau X_k du premier vecteur propre u^j de $(X^j)'\, D_p\, X^j$:

$$Z^j = (1/v_j)\sum_{k=1}^{K}(X_k^j\,(X_k^j)')D_p\, Z_k^j$$

Preuve : On a $Z^j = \sum_{j=1}^{K} X_k^j\, u_k^j$.

D'après la propriété 2 : $v_j u_k^j = (X_k^j)'\, D_p\, Z_k^j$, d'où le résultat annoncé.

Remarque : ce résultat peut aussi s'écrire :

$$Z^j = (1/v_j)\sum_{k=1}^{K}\sum_{v=1}^{m_k} \mathrm{cov}(X_{k,v},\, Z_k^j)X_{k,v}^j$$

ce qui résulte des effets combinés des résultats de la régression et de l' ACP.

Propriété 4. Si $r > j$, alors $X_{k,v}^r$ est D_p-orthogonale à Z^j.

Preuve : Par construction, $X_{k,v}^r$ est D_p-orthogonal à Z_k^j; d'autre part, Z_k^j étant la projection de Z^j sur W_k^j, on a : $Z^j = Z_k^j + e$, avec e D_p-orthogonal à tout vecteur de W_k^j, et donc en particulier D_p-orthogonal à $X_{k,v}^r$.

Donc Z^j et $X_{k,v}^r$ sont D_p-orthogonaux.

Propriété 5. Si $r > j$, alors Z_k^r est D_p-orthogonale à Z^j.

Preuve : Z_k^r est une combinaison linéaire des $X_{k,v}^r$, donc il est orthogonal à Z^j en vertu de la propriété 4.

Propriété 6. Les variables Z^j sont deux à deux D_p-orthogonales.

Preuve : Considérons deux variables Z^j et Z^r, et supposons que $r > j$. Z^r est une combinaison linéaire des $X_{k,v}^r$:

$$Z^r = \sum_{k=1}^{K} \sum_{v=1}^{m_k} p_k^r X_{k,v}^r$$

d'après la propriété 4 :

$$\text{pour } k = 1, \ldots, K \quad \text{et} \quad v = 1, \ldots, m_k, \quad (X_{k,v}^r)' \, D_p \, Z^j = 0$$

ce qui prouve le résultat annoncé.

Propriété 7. La somme des valeurs propres v_j de l'ACPG est inférieure ou égale à $SV = \sum_{k=1}^{K} \sum_{v=1}^{m_k} VAR(X_{k,v})$.

Preuve : A l'étape j, v_j est la valeur maximale prise par C^j (équations (2) et (3)). On en déduit :

$$\sum_{j=1}^{E} v_j = \sum_{j=1}^{E} \sum_{k=1}^{K} R^2(Z^j, \, Z_k^j) \left(\sum_{v=1}^{m_k} \mathrm{Cov}^2(\widetilde{Z}_k^j, \, X_{k,v}) \right)$$

d'où :

$$\sum_{j=1}^{E} v_j \leqslant \sum_{j=1}^{E} \sum_{k=1}^{K} \sum_{v=1}^{m_k} \mathrm{Cov}^2(\widetilde{Z}_k^j, \, X_{k,v}))$$

comme les $\widetilde{Z}_k^j$ (pour $j = 1, \ldots, E$) sont (s'ils sont non nuls) orthonormés et engendrent W_k, on a :

$$\sum_{j=1}^{E} \sum_{v=1}^{m_k} \mathrm{Cov}^2(\widetilde{Z}_k^j, \, X_{k,v})) = \sum_{v=1}^{m_k} \mathrm{Var}\,(X_{k,v})$$

pour $k = 1, \ldots, K$.

Il s'ensuit que : $\sum_{j=1}^{K} v_j \leqslant SV$.

Propriété 8. L'analyse des correspondances est un cas particulier de l'ACPG.

Preuve : Si les deux variables décrivent chacune les modalités (non centrées mais réduites) d'une variable qualitative, les deux tableaux X_1 et X_2 décrivent des bases orthonormées et l'ACPG se confond avec l'analyse canonique et donc avec l'analyse des correspondances.

Remarque : l'analyse des correspondances multiples pouvant s'écrire comme une analyse des correspondances, elle est aussi un cas particulier de l'ACPG.

5. Les représentations graphiques de l'ACPG

Les propriétés de l'ACPG mises en évidence dans le paragraphe précédent sont utiles pour construire deux types de représentations graphiques : des représentations de chacun des tableaux et des représentations communes à l'ensemble ou une partie des tableaux.

5.1 Les représentations de chaque tableau

Pour chacun des tableaux X_k, $k = 1, \ldots, K$, les représentations graphiques sont les mêmes qu'en ACP : dans l'espace des variables, les corrélations entre les variables de départ et les variables Z_k^j sont représentées dans des cercles de corrélation, tandis que la représentation des individus est obtenue en croisant deux variables Z_k^j et $Z_k^{j'}$.

Les critères d'arrêt utilisés habituellement en analyse de données (la valeur de l'inertie de l'axe ou la valeur de l'inertie cumulée) sont ici inopérants : il faut tenir compte du pourcentage d'inertie expliqué de chaque tableau, et il se peut que pour un ou plusieurs tableaux ce pourcentage soit élevé à une étape donnée alors que la valeur propre correspondante est faible.

Et donc, en pratique, l'analyse s'arrêtera lorsque le pourcentage cumulé d'inertie expliqué de chaque tableau sera jugé suffisant.

5.2 Représentations communes à plusieurs tableaux

Lorsque les $R^2(Z^j, Z_k^j)$ sont assez proches de 1 pour l'ensemble des tableaux ou pour une partie d'entre eux, il est intéressant de représenter simultanément les variables concernées (les variables $X_{k,v}$ des tableaux pour lesquels la corrélation est élevée) par leur corrélation avec Z^j.

En ce qui concerne les individus, lorsque les $R^2(Z^j, Z_k^j)$ sont assez proches de 1, il est possible de superposer sur le même axe Z^j et Z_k^j (dont la norme égale R

(Z^j, Z_k)); l'écart, pour un individu donné, entre la valeur de Z^j et la valeur prise par Z_k^j permet de repérer éventuellement des individus «aberrants».

La variance pour un individu donné des différentes valeurs prises par Z_k^j $(k = 1, \ldots, K)$ est un indicateur de la stabilité de l'individu pour le phénomène mis en évidence par l'axe j.

6. L'évolution de 5 pays de 1964 à 1993

6.1 Les données

Les données (Economie européenne, 1994) décrivent 5 pays de l'Europe du Sud, la France (F), le Portugal (P), l'Italie (I), l'Espagne (E) et la Grèce (G), durant une période de 30 ans (de 1964 à 1993), pour les 5 variables suivantes :

- taux de chômage (CHO)
- taux de croissance du PIB (CRO)
- taux d'inflation (INF)
- taux de consommation privée (CPR)
- taux d'investissement (INV)

Chaque tableau correspond donc à un pays et comporte 5 variables (les colonnes) et 30 années (les lignes); les variables étant hétérogènes sont ici normées.

6.2 Les résultats numériques

6.2.1 Résultats généraux

Pour chaque tableau k et chaque étape j est indiqué :

– le carré de la corrélation entre la variable canonique Z^j et la variable canonique du tableau k, Z_k^j, noté Corré.

– le pourcentage d'inertie du tableau k expliqué à l'étape j par la variable Z_k^j, c'est-à-dire $\frac{1}{5}\sum_{v=1}^{5} \mathrm{cor}^2(Z_k^j, X_{k,v})$, noté Inert.

En ce qui concerne la France, l'essentiel de l'inertie se trouve sur les axes 1 et 2. L'axe 3 présente une inertie assez élevée pour les 4 autres pays et l'axe 4 pour l'Espagne et le Portugal uniquement. Le dernier axe a une faible inertie pour chacun des tableaux.

		Axe 1	Axe 2	Axe 3	Axe 4	Axe 5
France	Corré.	0.99	0.94	0.70	0.66	0.61
	Inert.	64%	24%	6%	3%	3%
Grèce	Corré.	0.91	0.87	0.33	0.39	0.02
	Inert.	46%	33%	12%	6%	3%
Espagne	Corré.	0.98	0.84	0.62	0.54	0.74
	Inert.	44%	24%	11%	14%	7%
Italie	Corré.	0.94	0.96	0.77	0.15	0.01
	Inert.	54%	25%	12%	6%	3%
Portugal	Corré.	0.87	0.91	0.75	0.84	0.32
	Inert.	30%	34%	14%	15%	7%

Ici, l'interprétation des résultats portera sur les 3 premiers axes, le but étant plus d'illustrer la méthode proposée que d'interpréter de façon exhaustive les résultats.

6.2.2 Le plan des 2 premiers axes

Les corrélations des variables Z_k^j avec les variables canoniques Z^j sont fortes pour les 2 premiers axes pour chacun des cinq pays.

Aussi est-il intéressant de représenter simultanément les 5 pays pour le plan des axes 1 et 2.

Les variances des groupes pour ces deux premiers axes sont données par le tableau suivant :

	Variance ($\times 10^4$)			Variance ($\times 10^4$)	
Année	Axe 1	Axe 2	Année	Axe 1	Axe 2
1964	12	17	1979	24	26
1965	2	46	1980	15	29
1966	23	28	1981	13	27
1967	14	30	1982	12	20
1968	20	18	1983	4	74
1969	29	37	1984	25	7
1970	8	21	1985	9	6
1971	5	14	1986	1	18
1972	37	32	1987	12	31
1973	12	44	1988	13	23
1974	48	15	1989	18	10
1975	12	25	1990	5	14
1976	40	8	1991	10	4
1977	10	81	1992	4	8
1978	81	30	1993	7	24

Sur le graphique de l'annexe 1 sont représentés les «centres» des groupes (les années de 1964 à 1993) et les groupes ayant la plus forte variance (1977, 1978 et 1983). Par souci de lisibilité, les autres groupes ne sont pas représentés sur ce graphique.

En annexe 2 est représenté le cercle des corrélations entre Z^1 et Z^2 et les 25 variables de départ.

L'axe 1 est fortement corrélé au chômage de la France, de l'Italie et de l'Espagne, un peu moins fortement au chômage du Portugal et de la Grèce et à l'inflation de la Grèce. Cet axe est lié négativement à la croissance (Portugal, France, Espagne) et surtout à l'investissement (tous les pays sauf le Portugal).

L'axe 2 est celui de l'inflation pour la France, l'Italie et l'Espagne et dans une moindre mesure pour le Portugal. Il est corrélé négativement avec la consommation et le chômage grecs.

La représentation des individus confirme cette interprétation : la période 1964-193 se caractérise par une augmentation continue du chômage et une baisse de la croissance et de l'investissement à partir de 1974, l'inflation culmine entre 1974 et 1984 (pour la Grèce, comme le confirme l'examen des données, la baisse de l'inflation et la montée du chômage sont plus tardives).

La forte variance des années 1977 et 1978 est dûe à la très forte inflation du Portugal et de l'Espagne ces années là (plus de 20% contre 12% en moyenne pour les 3 autres pays), tandis que la variance de 1983 s'explique par le taux d'inflation très élevé du Portugal (24.6%) et un taux relativement élevé du chômage grec.

6.2.3 L'axe 3

En ce qui concerne l'axe 3, les corrélations entre les variables canoniques de chaque tableau et la variable canonique générale étant assez différentes de 1, il est intéressant de considérer la matrice des corrélations entre les variables Z_k^j pour $k = 3$ pour les pays dont l'inertie est suffisamment élevée.

Corrélations entre les Z_k^j
Etape 3

	Grèce	Espagne	Italie	Portugal
Grèce	1.00			
Espagne	0.52	1.00		
Italie	0.39	0.57	1.00	
Portugal	0.36	0.57	0.72	1.00

Ces corrélations sont assez faibles, sauf entre l'Italie et le Portugal. Ici l'interprétation est limitée volontairement à ces pays (annexes 3 et 4), l'interprétation des résultats des 2 autres pays se faisant de la même manière qu'en ACP.

Pour l'Italie et le Portugal, l'axe 3 est lié positivement à la croissance et les configurations des individus sur l'axe 3 sont assez ressemblantes (ce qui explique le coefficient de corrélation de 0.72) puisqu'elles mettent en évidence un faible taux de croissance en 1975 qui s'oppose, pour les 2 pays, à des taux élevés en 1973 et 1979 notamment.

7. Comparaison avec d'autres techniques

7.1 Comparaison avec l'ACP

L'ACP du tableau X dans la métrique identité ou dans une autre métrique (Escofier et Pages, 1988, Lavit, 1988) se heurte aux limites mises en évidence par Pontier et Normand (Pontier et Normand, 1992) dans le cas de l'ACG de Carroll (Carroll, 1968, Saporta, 1975).

L'ACP du tableau X détermine une base de l'espace engendré par les m colonnes de X; son but est donc de décrire le tableau X, juxtaposition des tableaux X_k. Or, le but d'une technique d'analyse simultanée de tableaux, c'est de déterminer ce qui est commun à plusieurs tableaux X_k et ce qui est spécifique à chacun d'entre eux. Chaque tableau doit pouvoir être décrit à partir de ses «éléments communs» et de ses «éléments spécifiques».

En conséquence, ce qui importe est d'interpréter les Z_k^j qui doivent constituer une base de W_k; l'interprétation des variables Z^j – combinaisons linéaires de variables de tableaux différents – importe peu, le rôle de ces variables étant essentiellement de permettre des représentations graphiques simultanées des individus et des variables de tableaux différents. En ce sens, l'espace engendré par les colonnes de Z^j $j = 1, \ldots, E$ constitue un «compromis» au sens de l'ACPG entre les espaces W_k $k = 1, \ldots, K$.

L'ACP dans une métrique M du tableau X ne permet pas une interprétation aisée des Z_k^j (au nombre de m pour un tableau X_k donné comportant m_k variables, d'où forcément un nombre élevé de redondances dans l'interprétation des phénomènes particuliers du tableau k).

La différence essentielle entre l'ACP de X et l'ACPG est que l'ACP détermine une base de l'espace engendré par les colonnes de X alors que l'ACPG détermine une base de chacun des espaces W_k.

L'ACP peut d'ailleurs comme l'ACPG (ou l'ADT) être présentée, à l'étape j, comme la recherche de la première valeur propre d'une matrice suivie d'une régression. A la différence de l'ACPG cependant, la régression se fait pour toutes les variables de tous les tableaux avec la même variable explicative Z^j, et non avec une variable Z_k^j.

7.2 Comparaison avec l'analyse discriminante de tableaux

L'analyse discriminante de tableaux (ADT, Casin, 1995), généralisation de l'analyse canonique à plus de deux tableaux se présente ainsi à l'étape j :

Déterminer Z^j et les Z_k^j tels que :

$$C^j = \sum_{k=1}^{K} R^2(Z^j,\ Z_k^j)$$

soit maximal sous les contraintes de normalisation :

$$(Z^j)'\ D_p\ Z^j = 1$$

et les conditions d'orthogonalisation :

$$(Z_k^j)'\ D_p\ Z_k^r = 0 \text{ pour } r < j \text{ et } k = 1, \ldots, K$$

L'ADT peut être considérée comme un cas particulier de l'ACPG : il s'agit d'une ACPG des espaces W_k décrits chacun par une base orthonormée. En effet, si pour un tableau k donné les variables $X_{k,v}$ sont normées, deux à deux orthogonales, et forment un système libre, alors quelque soit $\tilde{Z}_k^j$: $\sum_{v=1}^{m_k} \text{cov}^2(\tilde{Z}_k^j,\ X_{k,v}) = 1$ et l'équation (2) s'écrit alors :

$$\sum_{k=1}^{K} \sum_{v=1}^{m_k} \text{cov}^2(Z_k^j,\ X_{k,v}) = \sum_{k=1}^{K} R^2(Z^j,\ Z_k^j)$$

L'ADT recherche des proximités entre des tableaux, mais ne s'intéresse pas à l'inertie des variables Z_k^j, et il se peut que ses résultats dans l'espace des variables soient difficilement interprétables; dans le cas de l'ACPG, l'interprétation dans l'espace des variables est plus aisée, surtout lorsque les variables de départ sont normées : on obtient alors une représentation simultanée optimale des matrices de corrélation des variables des différents tableaux.

Les résultats dans l'espace des individus sont interprétés en termes d'analyse discriminante (c'est-à-dire de représentation simultanée optimale des différents nuages de points-individus) dans le cas de l'ADT. L'interprétation des résultats de l'ACPG dans l'espace des individus se fait tableau par tableau, en termes d'analyse en composantes principales; la comparaison des différents nuages de points-individus ne possède pas l'optimalité de l'ADT.

Dans le cas où l'on dispose de variables nombreuses et d'un nombre restreint d'individus (ce qui est souvent le cas en Economie où un petit nombre de pays comparables sont décrits, pour une période donnée, par des variables macroéconomiques nombreuses), les résultats de l'ADT sont indéterminés alors que l'ACPG reste praticable.

Enfin, on notera que certaines propriétés sont communes à l'ACPG et à l'ADT (en particulier l'orthogonalité des variables Z^j et $Z^{j'}$, pour j différent de j'), que l'ADT possède des propriétés que ne possède pas l'ACPG (Z^j n'est ni composante

principale, ni combinaison linéaire des Z_j^k en ACPG), mais que l'ACPG est une méthode plus générale que l'ADT (l'ACPG admet comme cas particulier l'ACP dans la métrique identité).

7.3 Comparaison avec les techniques PLS

Les techniques PLS (Tenenhaus, Gauchi et Ménardo, 1995) s'appliquent dans le cas de 2 tableaux; on cherche dans chacun des deux tableaux une variable ayant une inertie élevée et bien liée linéairement à la variable de l'autre tableau.

Dans cette optique, il est possible de proposer la généralisation suivante à plus de deux tableaux des techniques PLS : la liaison entre une variable auxiliaire Z^j de R^n et une variable synthétique $Z_k^j = X_k^j v_k^j$ représentative du tableau k étant définie par $L_k = \text{cov}^2(Z^j, Z_k^j)$, on cherche Z^j maximisant $\sum_{k=1}^{K} L_k$ les conditions d'orthogonalisation étant celles de l'ACPG, les conditions de normalisation portant sur v_k^j (on impose aux v_k^j d'être normés à 1).

Les solutions de ce problème ont certaines propriétés que ne possède pas l'ACPG :

– Z^j est première composante principale des Z_k^j, donc combinaison linéaire des Z_k^j;

– pour Z^j donné, Z_k^j est obtenu à la première étape de la régression PLS où Z^j est la variable «à expliquer», c'est-à-dire maximise cov (Z^j, Z_k^j) sous la contrainte de normalisation de v_k^j.

Mais comme Z_k^j n'est pas la projection de Z^j sur W_k^j, on n'obtient plus comme en ACPG la propriété de non-corrélation des Z^j entre eux, ce qui limite les représentations graphiques.

L'ACPG respecte l'esprit des techniques PLS – déterminer des variables dans chaque espace bien liées entre elles et ayant une inertie élevée –, tout en procurant des résultats graphiques complets.

D'un point de vue algorithmique, les techniques PLS et l'ACPG sont proches, faisant à chaque étape succéder des régressions sur les tableaux de départ à la recherche du premier vecteur propre d'une matrice.

8. Une généralisation

L'ACPG peut être étendue à d'autres métriques que la métrique identité et à l'étape j :

– on effectue d'abord la première étape d'une ACP du tableau X^j dans la métrique M_j

– on projette ensuite Z^j sur W_k^j pour déterminer Z_k^j, puis W_k^{j+1}.

Lorsqu'on dispose d'un seul tableau, et que la métrique est la même à chaque étape j, c'est-à-dire lorsque $M_j = M$, il s'agit de l'ACP du tableau X dans la métrique M.

Si l'on dispose de plusieurs tableaux et que M_j est la métrique identité, on retrouve l'ACPG; si M_j est la pseudo-métrique donnée par la matrice diagonale par blocs des inverses généralisés des matrices de corrélation de X_k^j sur chaque tableau, on obtient l'ADT. Mais on peut aussi utiliser d'autres métriques (de type «AFM» ou «STATIS») : on conserve alors la propriété d'orthogonalité des variables Z_k^j de chaque tableau et la propriété d'orthogonalité de Z^j et de $Z^{j'}$, et donc la possibilité d'obtenir des représentations graphiques pour un seul tableau ou pour plusieurs tableaux simultanément.

On peut ainsi distinguer 2 familles de méthodes : celles dont le but essentiel est la description d'un seul tableau X, et celles dont le but essentiel est la description simultanée de K tableaux X_k.

9. Conclusion

L'ACPG est une technique d'analyse de tableaux à 3 dimensions à mi-chemin entre l'ADT et les techniques PLS; ses graphiques s'interprètent comme ceux d'une analyse en composantes principales, ou comme la superposition de graphiques d'analyses en composantes principales. La simplicité de cette méthode la rend accessible à un grand nombre d'utilisateurs potentiels.

Bibliographie

Carroll J.D. «A generalisation of canonical correlation analysis to three or more sets of variables», *Proc. 76th Conv. Amer. Psych. Ass.*, 1968.

Casin Ph. «L'analyse discriminante de tableaux évolutifs», *Revue de Statistique Appliquée*, volume XLIII, n°3, pp 73-91, 1995.

Economie européenne, n°58, 1994.

Escofier B. et Pages J. «Analyses factorielles simples et multiples», Dunod, 1988.

Lavit Ch. «Analyse conjointe de tableaux quantitatifs», Masson, 1988.

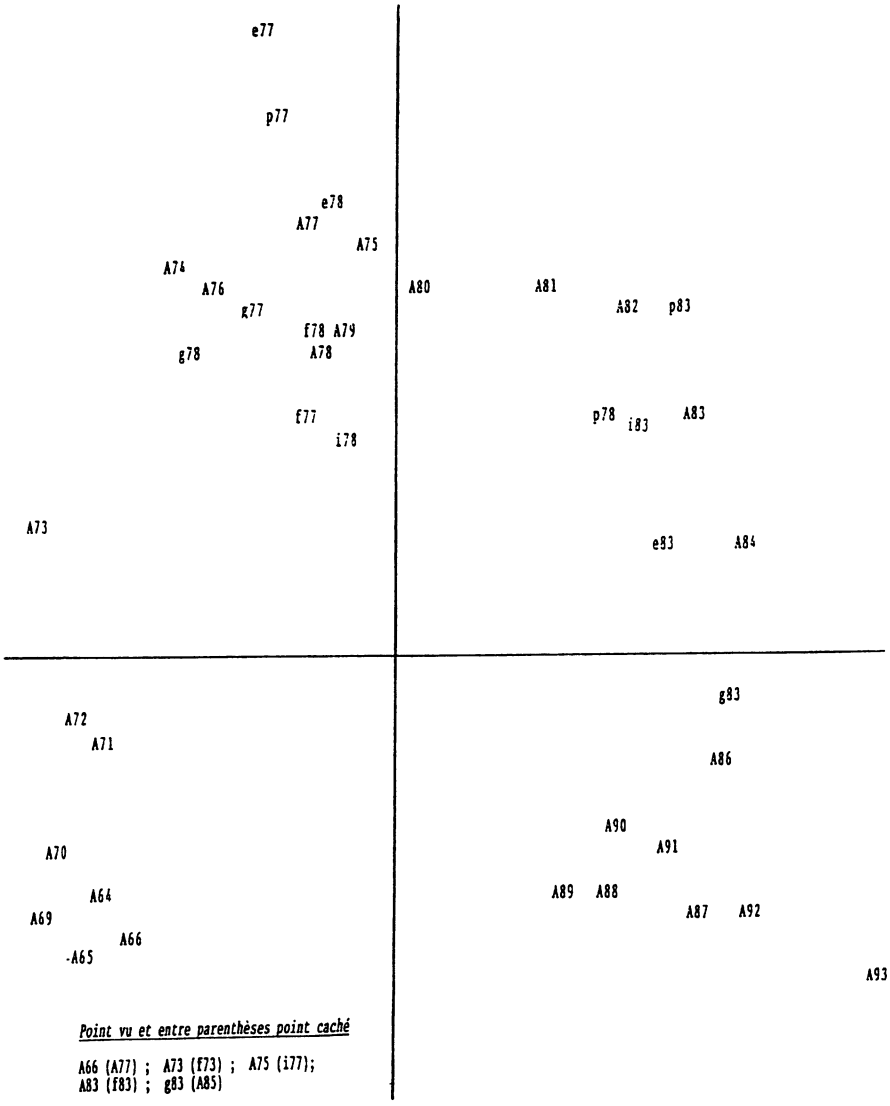
Pontier J. et Normand M. «A propos de généralisation de l'analyse canonique», *Revue de Statistique Appliquée*, volume XL, n°1, pp 57-75, 1992.

Saporta G. «Liaisons entre plusieurs ensembles de variables et codage de données qualitatives», Thèse de 3ème cycle, Université de Paris VI, 1975.

Tenenhaus M., Gauchi J.P. et Ménardo C. «Régression PLS et applications», *Revue de Statistique Appliquée*, volume XLIII, n°1, pp 7-63, 1995.

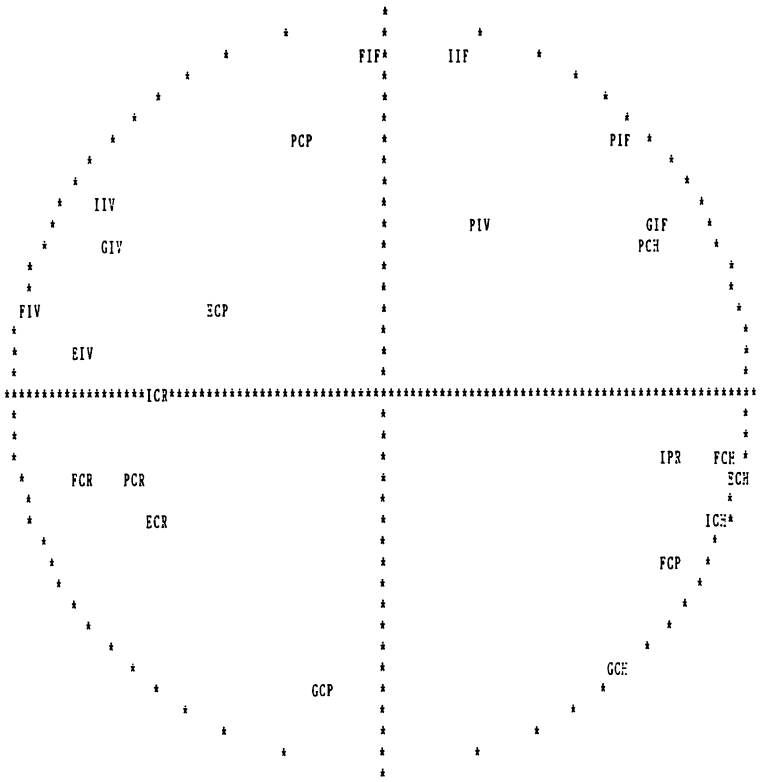
Annexe 1

Représentation simultanée
Axes 1 (horizontal) et 2 (vertical)



Annexe 2

Corrélation
Axes 1 (horizontal) et 2 (vertical)



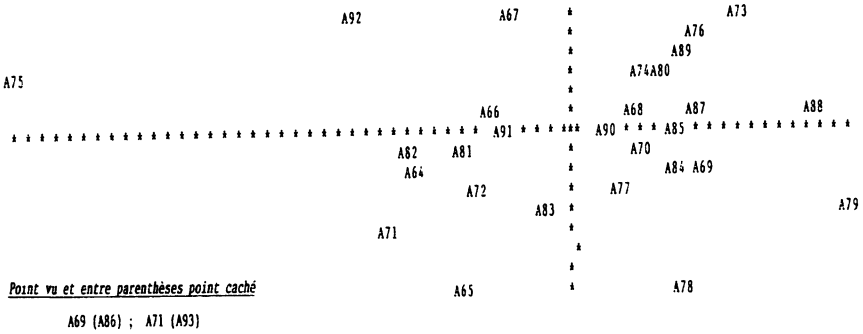
Point vu et entre parenthèses point caché

FCR (GCR) ; FIF (EIF)

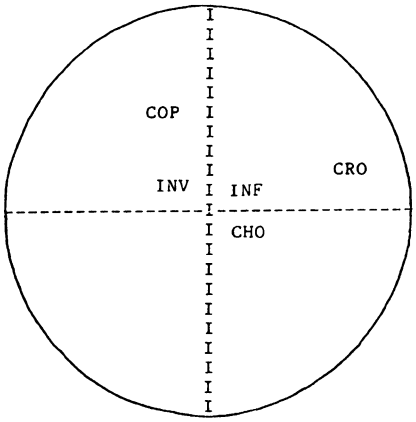
La première lettre est l'initiale du pays. Les 2 suivantes indiquent la variable: Consommation privée (CP), Chômage (CH), Investissement (IV), Inflation (IF) et Croissance (CR).

Annexe 3

Italie
Axes 3 (horizontal) et 4 (vertical)



Représentation des individus

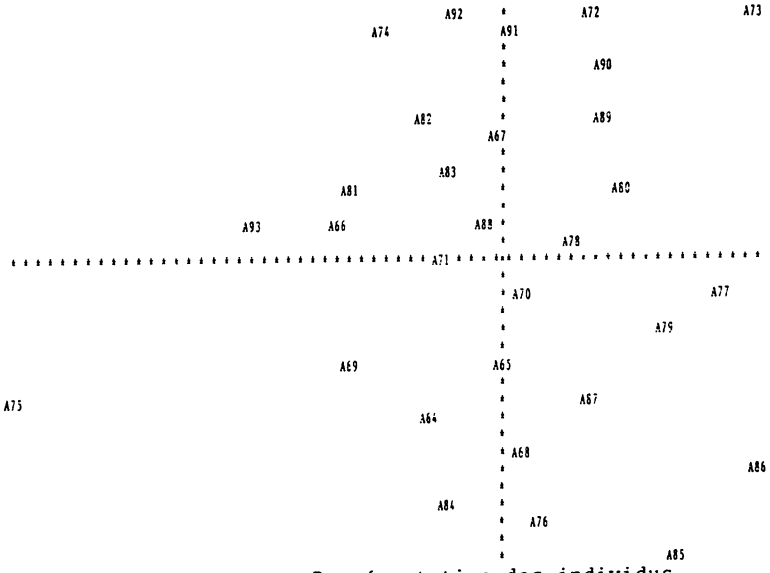


CHO: Chômage
COP: Consommation privée
CRO: Croissance
INV: Investissement
INF: Inflation

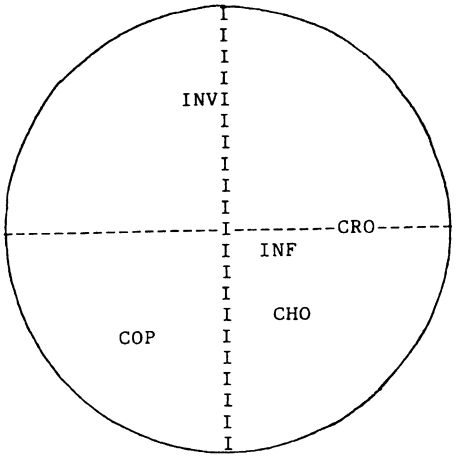
Cercle des corrélations

Annexe 4

Corrélation
Axes 3 (horizontal) et 4 (vertical)



Représentation des individus



Cercle des corrélations

CHO: Chômage
COP: Consommation privée
CRO: Croissance
INV: Investissement
INF: Inflation