

REVUE DE STATISTIQUE APPLIQUÉE

THIERRY FOUCART

Une nouvelle approche de la méthode STATIS

Revue de statistique appliquée, tome 31, n° 2 (1983), p. 61-75

http://www.numdam.org/item?id=RSA_1983__31_2_61_0

© Société française de statistique, 1983, tous droits réservés.

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

NUMDAM

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

UNE NOUVELLE APPROCHE DE LA METHODE STATIS

Thierry FOUCART

Laboratoire d'Informatique et Statistique Appliquées
IUT - BP 1104 - 56008 VANNES

La méthode STATIS consiste à étudier une famille d'opérateurs de covariance ou de produit scalaire (opérateurs d'Escoufier) à l'aide du produit scalaire de HILBERT-SCHMIDT entre opérateurs. Elle comporte trois étapes : l'étude de l'interstructure, chaque opérateur étant représenté par un vecteur dans un espace euclidien et un point dans les représentations graphiques, l'analyse en composantes principales du compromis, opérateur qui résume la famille en maximisant un critère et la représentation des intrastructures pour laquelle plusieurs méthodes existent déjà.

Nous présentons dans cet article une nouvelle approche de la méthode STATIS et proposons, pour représenter les intrastructures, de nouvelles procédures. Ces dernières ont été appliquées à l'analyse de trois tableaux de ratios d'entreprises pour étudier les opérateurs associés aux tableaux des données.

L'interprétation des résultats a été réalisée en collaboration avec J. GABER, assistante à l'IUT de Vannes, qui s'est également chargée de recueillir des données.

1. METHODE STATIS/INTERSTRUCTURE ET COMPROMIS [1]

Nous proposons dans ce paragraphe quelques rappels sur la méthode STATIS et précisons nos notations et le langage utilisé.

On considère une famille finie d'opérateurs symétriques positifs $\{\Gamma_i\}$ $i = 1, k$, définis sur un espace euclidien \mathbb{R}^p .

1.1. Opérateur caractéristique et interstructure

Nous allons définir un opérateur caractéristique de la famille $\{\Gamma_i\}$ $i = 1, k$, défini sur \mathbb{R}^k de façon analogue à un opérateur de covariance [2].

Soit $u \in \mathbb{R}^k$. On peut définir l'opérateur $u \cdot \Gamma$ de \mathbb{R}^p par :

$$\forall x \in \mathbb{R}^p \quad u \cdot \Gamma(x) = \sum_{i=1}^k u_i \Gamma_i(x)$$

L'opérateur $u \cdot \Gamma$ est symétrique quel que soit le vecteur u : on peut donc définir l'application $\Sigma : \mathbb{R}^k \times \mathbb{R}^k \rightarrow \mathbb{R}$

$$(u, v) \in \mathbb{R}^k \times \mathbb{R}^k \rightarrow \Sigma(u, v) = (u \cdot \Gamma) \text{hs}(v \cdot \Gamma)$$

où hs symbolise le produit scalaire de HILBERT SCHMIDT [3] munissant l'espace des opérateurs de \mathbb{R}^p d'une structure euclidienne.

L'application Σ est évidemment bilinéaire et symétrique : à tout vecteur u , on peut donc associer un vecteur unique noté $\sigma(u)$ tel que :

$$\forall v \in \mathbf{R}^k \quad \Sigma(u, v) = \sigma(u) \cdot v$$

\mathbf{R}^k étant muni du produit scalaire canonique.

Définition : on appelle opérateur caractéristique d'une famille de k opérateurs symétriques $\{\Gamma_i\}$ définis sur un espace euclidien \mathbf{R}^p l'opérateur de \mathbf{R}^k noté σ tel que :

$$\forall u \in \mathbf{R}^k \quad \forall v \in \mathbf{R}^k \quad (u \cdot \Gamma) \text{ hs}(v \cdot \Gamma) = \sigma(u) \cdot v$$

Théorème : l'opérateur caractéristique est symétrique semi-défini positif (démonstration évidente).

L'opérateur caractéristique conduit à la notion d'interstructure ; de façon analogue à la recherche des composantes principales d'un vecteur aléatoire X , on détermine les opérateurs principaux en résolvant le programme de maximisation :

$$\begin{cases} \max \|u \cdot \Gamma\|_{\text{hs}} \\ \|u\|^2 = 1 \end{cases}$$

avec itération sous contrainte d'orthogonalité.

Définition : on appelle opérateurs principaux les opérateurs O_h $h = 1, q$ définis sur \mathbf{R}^p par :

$$O_h = u_h \cdot \Gamma$$

où $u_h = (u_h^1, \dots, u_h^i, \dots, u_h^k)$ est le vecteur propre unitaire de l'opérateur σ associé à la valeur propre λ_h non nulle (h variant de 1 à q).

Théorème : les opérateurs unitaires $O'_h = O_h / \sqrt{\lambda_h}$ forment une base orthonormée, du sous-espace de $\mathcal{L}(\mathbf{R}^p)$ engendré par le système $\{\Gamma_i\}$ $i = 1, k$.

Conséquence : les opérateurs Γ_i sont les combinaisons linéaires des opérateurs unitaires O'_h :

$$\forall i = 1, k \quad \Gamma_i = \sum_{h=1}^q \sqrt{\lambda_h} u_h^i O'_h$$

c'est la formule de reconstitution des données.

Définition : on appelle interstructure de la famille $\{\Gamma_i\}$ $i = 1, k$ le spectre de son opérateur caractéristique.

1.2. Compromis

Le premier opérateur principal vérifie les mêmes critères que la première composante principale d'un vecteur aléatoire : il est de norme de HILBERT SCHMIDT maximale et maximise la somme des carrés des produits scalaires.

La base canonique $(e_j)_{j=1, k}$ de \mathbf{R}^k étant orthonormée, on a :

$$\sigma(u) = \sum_{j=1}^k (\sigma(u) \cdot e_j) e_j$$

$$\|\sigma(u)\|^2 = \sum_{j=1}^k (\sigma(u) \cdot e_j)^2$$

$$\Gamma_j = e_j \cdot \Gamma$$

Par définition de l'opérateur caractéristique, on peut écrire :

$$\begin{aligned} \sigma(u) \cdot e_j &= (u \cdot \Gamma) \text{hs}(e_j \cdot \Gamma) \\ &= (u \cdot \Gamma) \text{hs} \Gamma_j \end{aligned}$$

et par conséquent :

$$\|\sigma(u)\|^2 = \sum_{j=1}^k ((u \cdot \Gamma) \text{hs} \Gamma_j)^2$$

Le vecteur proposé u_1 associé à la plus grande valeur propre maximise le premier membre, donc le second ; d'autre part, les produits scalaires entre les opérateurs symétriques positifs Γ_i sont positifs ou nuls et, en vertu du théorème de FROBENIUS, les coordonnées de u_1 dans la base canonique sont toutes de même signe. On peut donc les choisir positives, et l'on en déduit :

Théorème : l'opérateur principal $O_1 = u_1 \cdot \Gamma$ est un opérateur symétrique semi-défini positif.

Cet opérateur est habituellement appelé compromis. En ce qui nous concerne, nous préférons le modifier légèrement, de façon que si par exemple :

$$\forall i = 1, k \quad \Gamma_i = \Gamma_0$$

on obtienne :

$$O_1 = \Gamma_0$$

pour cela, il suffit d'effectuer une homothétie : on pose :

$$\forall i = 1, k : \alpha_i = u_1^i / \sum_{j=1}^k u_1^j$$

Définition : on appelle compromis de la famille $\{\Gamma_i\}_{i=1, k}$ l'opérateur Γ_c défini par :

$$\Gamma_c = \alpha \cdot \Gamma$$

où α est le vecteur propre de l'opérateur caractéristique associé à la plus grande valeur propre et dont la somme des coordonnées est égale à 1.

2 OPERATEUR DE COVARIANCE [2]

Les résultats développés dans le paragraphe précédent sont généraux ; ils ne dépendent pas de la nature des opérateurs étudiés, contrairement aux étapes suivantes de la méthode STATIS.

On étudie une famille d'opérateurs de covariance lorsque l'on est en présence d'un même vecteur aléatoire défini sur des populations différentes. Nous supposons ces populations finies.

Soient $\{\Gamma_i\}$ $i = 1, k$ la famille des opérateurs de covariance, $\{I_i\}$ $i = 1, k$ la famille des populations correspondantes et $X = (X_1, \dots, X_j, \dots, X_p)$ le vecteur aléatoire étudié.

2.1. Etude du compromis

Le compromis $\Gamma_c = \alpha \cdot \Gamma$ est symétrique semi-défini positif : c'est l'opérateur de covariance du vecteur aléatoire X défini sur un espace probabilisé particulier ; on pose :

$$I = \bigcup_{i=1}^k I_i$$

$$\forall i = 1, k \quad \forall \omega_i \in I_i \quad P(\omega_i) = \alpha_i P_i(\omega_i)$$

où $P_i(\omega_i)$ est la probabilité (ou le poids) de l'individu ω_i dans la population I_i .

On peut définir sur l'ensemble $\{I_i\}$ $i = 1, k$ la probabilité P_I :

$$\forall i = 1, k \quad P_I(I_i) = \alpha_i$$

Les probabilités P_I , P et P_i sont liées par une relation de probabilité conditionnelle.

On sait que l'opérateur de covariance total Γ_T est égal à la somme de l'opérateur de covariance interclasse Γ_G et de l'opérateur de covariance interclasse Γ_I :

$$\Gamma_T = \Gamma_G + \Gamma_I$$

plus précisément on a :

$$\Gamma_I = \sum_{i=1}^k \alpha_i \Gamma_i$$

Théorème : le compromis Γ_c est égal à l'opérateur de covariance intraclasse du vecteur aléatoire X défini sur l'espace probabilisé $[I, \mathcal{R}(I), P]$

On supposera désormais que les variables X_j sont centrées sur chaque population I_i : l'opérateur de covariance interclasse est alors nul et l'opérateur compromis est égal à l'opérateur de covariance totale. Dans ce cas, la diagonalisation de l'opérateur compromis revient à l'ACP du vecteur aléatoire X défini sur $[I, \mathcal{R}(I), P]$.

2.2. Les infrastructures [4]

Nous allons chercher à représenter les variables X_j $j = 1, p$ telles qu'elles sont vues par chaque opérateur Γ_i $i = 1, k$: on représente chaque variable par

$k + 1$ points, le premier correspondant à l'opérateur compromis et les suivants aux opérateurs Γ_i $i = 1, k$. La simple superposition des représentations graphiques obtenues par les analyses en composantes principales des opérateurs Γ_i est difficilement interprétable, puisque les axes ne sont pas les mêmes.

La base orthonormée du sous-espace engendré par les variables X_j que nous considérons est constituée par les composantes principales réduites du compromis ; on note :

C_h $h = 1, q$: les composantes principales du compromis
 u_h : les vecteurs principaux
 λ_h : les valeurs propres.

on a donc :

$$\forall h = 1, q \quad C_h = u_h \cdot X$$

Les composantes principales réduites sont :

$$C'_h = \frac{1}{\sqrt{\lambda_h}} u_h \cdot X$$

dans l'analyse en composantes principales du compromis, les coordonnées des variables sont les covariances avec les composantes principales réduites calculées sur la population I :

$$\forall j = 1, p \quad X_j = \sum_{h=1}^q \text{cov}(X_j, C'_h) C'_h$$

On peut estimer les covariances entre les variables X_j et les composantes principales C'_h en considérant leurs restrictions aux populations I_i ; on pose :

$$\forall i = 1, k \quad \forall j = 1, p \quad X_j^i = \sum_{h=1}^q \text{cov}_{I_i}(X_j, C'_h) C'_h$$

sur la population I_i , on a :

$$\text{cov}_{I_i}(X_j, C'_h) = \frac{1}{\sqrt{\lambda_h}} \Gamma_i(e_j) \cdot u_h$$

et par conséquent :

$$\forall i = 1, k \quad \forall j = 1, p \quad X_j^i = \sum_{h=1}^q \frac{1}{\sqrt{\lambda_h}} (\Gamma_i(e_j) \cdot u_h) C'_h$$

Dans le cas où Γ_i est égal à l'opérateur Γ_c et I_i à I, on retrouve la formule classique de reconstitution des données. A partir de la h^e composante principale du compromis, on représentera donc la variable X_j par les $k + 1$ points de coordonnées :

$$\text{cov}(X_j, C'_h), \{ \text{cov}_{I_i}(X_j, C'_h), i = 1, k \}$$

Les "déformations" de X_j dans I_i sont dues à plusieurs facteurs :

- le changement des corrélations entre les variables.
- la "déformation" des composantes principales du compromis par I_i .

Le premier facteur est celui que nous voulons mettre en évidence. Le second vient du fait que les composantes principales C'_h ne sont ni réduites, ni non cor-

réelles sur chaque population I_i : leur emploi est justifié par les propriétés de l'opérateur compromis citées plus haut et par le théorème suivant :

Théorème : la variable X_j est la moyenne des variables X_j^i $i = 1, k$ $j = 1, p$ pondérées par les coefficients α_i $i = 1, k$. Ce théorème résulte de la définition de l'opérateur compromis.

Représentation des composantes principales

La démarche précédente peut être utilisée pour représenter toute combinaison linéaire des variables X_j $j = 1, p$: on peut donc représenter les composantes principales du compromis "déformées" par la population I_i : en les notant C_m^i on a les formules ci-dessous :

$$\forall i = 1, k \quad \forall m = 1, q \quad C_m^i = \sum_{h=1}^q \text{cov}_{I_i}(C_m, C_h') C_h'$$

$$C_m^i = \sum_{h=1}^q \frac{1}{\sqrt{\lambda_h}} \Gamma_i(u_m) \cdot u_h C_h'$$

Définition : on appelle infrastructures d'une famille d'opérateurs de covariance $\{\Gamma_i\}$ $i = 1, k$ l'ensemble des variables et des composantes principales du compromis reconstruites par chaque opérateur Γ_i .

On peut également s'intéresser aux composantes principales de chaque population, mais l'interprétation des résultats se révèle dans la pratique assez délicate.

3. OPERATEURS DE PRODUIT SCALAIRE

On étudie une famille d'opérateurs de produit scalaire (plus connus sous le nom d'opérateurs d'Escoufier) lorsque l'on est en présence de plusieurs vecteurs aléatoires $Z_1, \dots, Z_j, \dots, Z_k$ à valeurs dans les espaces euclidiens $\mathbb{R}^{p_1}, \dots, \mathbb{R}^{p_j}, \dots, \mathbb{R}^{p_k}$, et définis sur une même population I que l'on supposera finie et munie d'une probabilité P . Chaque vecteur Z_j définit un opérateur Ω_j de l'espace $\mathcal{L}^2(I)$ et l'on note Π_j le produit scalaire défini sur \mathbb{R}^{p_j} .

3.1. Interprétation du compromis [5]

Soit $\Omega_c = \alpha \cdot \Omega$ le compromis de la famille $\{\Omega_j\}$ $j = 1, k$: cet opérateur est symétrique, semi-défini et positif : il possède toutes les propriétés d'un opérateur et nous savons qu'en fait, c'est l'opérateur d'un vecteur aléatoire Z [5].

On pose :

$$P = \sum_{j=1}^k p_j$$

$$Z = (Z_1^1, Z_1^2, \dots, Z_1^{p_1}, \dots, Z_j^1, Z_j^2, \dots, Z_j^{p_j}, \dots, Z_k^1, Z_k^2, \dots, Z_k^{p_k})$$

$$= (Z_1, \dots, Z_j, \dots, Z_k)$$

où Z^{h_j} ($1 \leq h_j \leq p_j$) est la h_j^e coordonnée du vecteur aléatoire Z_j .

Le vecteur aléatoire Z est à valeurs dans l'espace vectoriel \mathbf{R}^p , sur lequel on définit la métrique Π :

$$\forall \omega = (\omega_1, \dots, \omega_j, \dots, \omega_k); \omega_j \in \mathbf{R}^{p_j}$$

$$\forall \omega' = (\omega'_1, \dots, \omega'_j, \dots, \omega'_k); \omega'_j \in \mathbf{R}^{p_j} \quad \Pi(\omega, \omega') = \sum_{j=1}^k \alpha_j \Pi_j(\omega_j, \omega'_j)$$

On note : $\Pi = \oplus \alpha_j \Pi_j$

Théorème : l'opérateur compromis Ω_c est l'opérateur de produit scalaire du vecteur aléatoire Z défini sur I et à valeurs dans l'espace euclidien \mathbf{R}^p muni de la métrique Π .

L'analogie avec le problème précédent est évidente :

– à la juxtaposition des individus repérés par les mêmes variables accompagnée d'une définition appropriée de la probabilité sur l'ensemble I obtenu correspond l'opérateur de covariance.

$$\Gamma_c = \sum_{i=1}^k \alpha_i \Gamma_i$$

– à la juxtaposition des variables accompagnée d'une définition appropriée d'une métrique sur \mathbf{R}^p correspond l'opérateur de produit scalaire $\Omega_c = \sum_{j=1}^k \alpha_j \Omega_j$

3.2. Intrastructure [6]

Comme dans le cas d'une famille d'opérateurs de covariance, nous allons chercher à représenter chaque individu ω par $k + 1$ points, le premier correspondant au compromis et les suivants aux opérateurs Ω_j .

Plusieurs procédures existent déjà ou sont en train d'être mises au point [6]. Celles que nous proposons consistent à reconstruire les composantes principales C_h $h = 1, q$ du compromis par chaque vecteur Z_j $j = 1, k$: la représentation graphique à laquelle on aboutit met en évidence les pertes d'informations sur chaque individu lorsque l'on se restreint à un vecteur Z_j .

i) On note $C'_{m,j}$ $j = 1, k$ $m = 1, q_j$ les composantes principales réduites du vecteur aléatoire Z_j . On reconstruit les composantes principales C_h par la formule de reconstitutions des données :

$$C_h^j = \sum_{m=1}^{q_j} \text{cov}(C_h, C'_{m,j}) C'_{m,j}$$

Il n'existe aucune raison pour que la reconstruction soit exacte, même lorsque toutes les composantes principales $C'_{m,j}$ de Z_j sont prises en compte, puisque la variable aléatoire C_h est combinaison linéaire de toutes les variables $X_j^{h_j}$ $j = 1, k$ $h_j = 1, p_j$.

A un individu ω sont associés $k + 1$ vecteurs de \mathbf{R}^q (q étant le nombre de composantes principales du compromis) :

– le premier est obtenu directement de Ω_c :

$$(C_1(\omega), \dots, C_h(\omega), \dots, C_q(\omega))$$

– les suivants correspondent aux opérateurs Ω_j :

$$(C_1^j(\omega), \dots, C_h^j(\omega), \dots, C_q^j(\omega))$$

On peut envisager de représenter des individus supplémentaires, dont les valeurs ne sont pas toutes connues par exemple, ou qui caractérisent les composantes principales de chaque opérateur ; j et m étant fixés on considère l'individu ϵ défini par :

$$C'_{m,j}(\epsilon) = 1 \quad \text{et} \quad C'_{m',j}(\epsilon) = 0 \quad \forall m' \neq m$$

Cet individu engendre le m^e axe principal du tableau Z_j .

ii) On peut aussi procéder de la façon suivante : on sait que la composante principale C_h est vecteur propre de l'opérateur compromis associé à la valeur propre λ_h :

$$\forall h = 1, q \quad C_h = \frac{1}{\lambda_h} \Omega_c(C_h)$$

De la même façon que dans le cas des opérateurs de covariance, on remplace Ω_c par Ω_j , $j = 1, k$ pour trouver :

$$C_h^j = \frac{1}{\lambda_h} \Omega_j(C_h)$$

Lorsque l'on utilise cette procédure, on aboutit à la propriété suivante :

Théorème : la composante principale C_h est la moyenne des composantes principales C_h^j pondérées par les coefficients α_j , $j = 1, k$.

On peut mettre en évidence la différence entre les deux méthodes ; l'opérateur Ω_j est égal à la somme des projecteurs $P_{m,j}$ sur les espaces propres engendrés par les composantes principales $C_{m,j}$ de Z_j et pondérés par les valeurs propres associées $\lambda_{m,j}$:

$$\Omega_j = \sum_{m=1}^{q_j} \lambda_{m,j} P_{m,j}$$

on en déduit :

$$\Omega_j(C_h) = \sum_{m=1}^{q_j} \lambda_{m,j} \text{cov}(C_h, C'_{m,j}) C'_{m,j}$$

et enfin :

$$C_h^j = \frac{1}{\lambda_h} \sum_{m=1}^{q_j} \lambda_{m,j} \text{cov}(C_h, C'_{m,j}) C'_{m,j}$$

cette dernière formule fait intervenir de façon explicite les valeurs propres de l'opérateur Ω_j .

4. APPLICATION

L'application que nous présentons ci-dessous concerne trois tableaux de données définis par les mêmes individus et les mêmes variables et échelonnés sur trois ans. Après avoir précisé ces données, nous interprétons les résultats des analyses des opérateurs de covariance et des opérateurs de produit scalaire.

4.1. Les données étudiées

Les individus sont définis par un échantillon de 45 entreprises bretonnes qui ont déposé leur bilan entre les années 1971 et 1980. Parmi elles, une moitié environ ont été mises en liquidation judiciaire (disparition de l'entreprise) alors que les autres ont bénéficié d'un règlement judiciaire (sauvetage de l'entreprise). Cette information est connue pour chaque entreprise, mais n'a pas été prise en compte dans les analyses (c'est l'objet d'un travail ultérieur). Les entreprises sont codées de 01 à 53 la lettre E (échec : disparition de l'entreprise) ou R (réussite : sauvetage) complétant l'identificateur. Certains codes correspondent à des entreprises qui ont été écartées des tableaux à cause de leurs résultats aberrants ; il ne figurent pas dans la liste.

A partir des bilans annuels, un certain nombre de ratios ont été calculés. Parmi eux nous avons retenu les huit suivants :

Ratios d'équilibre de bilan ou de structure

- SOL : capitaux propres/dettes totales ;
- LQG : capitaux circulants/dettes à court terme ;
- TRL : valeurs réalisables et disponibles/dettes à court terme ;
- RFR : fonds de roulement/passif.

Ratios de gestion

- ATF : cash-flow/dettes totales ;
- FIN : frais financiers/chiffres d'affaires ;
- RVA : frais de personnel/valeur ajoutée.

Ratio de rentabilité

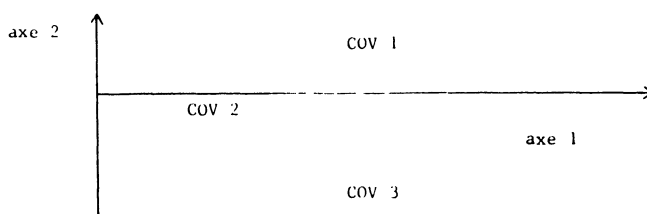
REN : résultat d'exploitation/chiffre d'affaires hors taxes.

La valeur ajoutée a été calculée suivant la définition donnée dans l'ordonnance du 17 août 1967 sur la participation des travailleurs aux fruits de l'entreprise.

4.2. Analyse des opérateurs de covariance : interstructure.

L'analyse des opérateurs de covariance comporte trois étapes. En fait, après avoir étudié rapidement l'interstructure dont l'intérêt est limité par le petit nombre de tableaux, nous interpréterons les représentations graphiques simultanées au compromis et des intrastructures.

Le schéma ci-dessous représente l'interstructure de la famille définie par les trois opérateurs de covariance :



Interstructure des opérateurs de covariance.

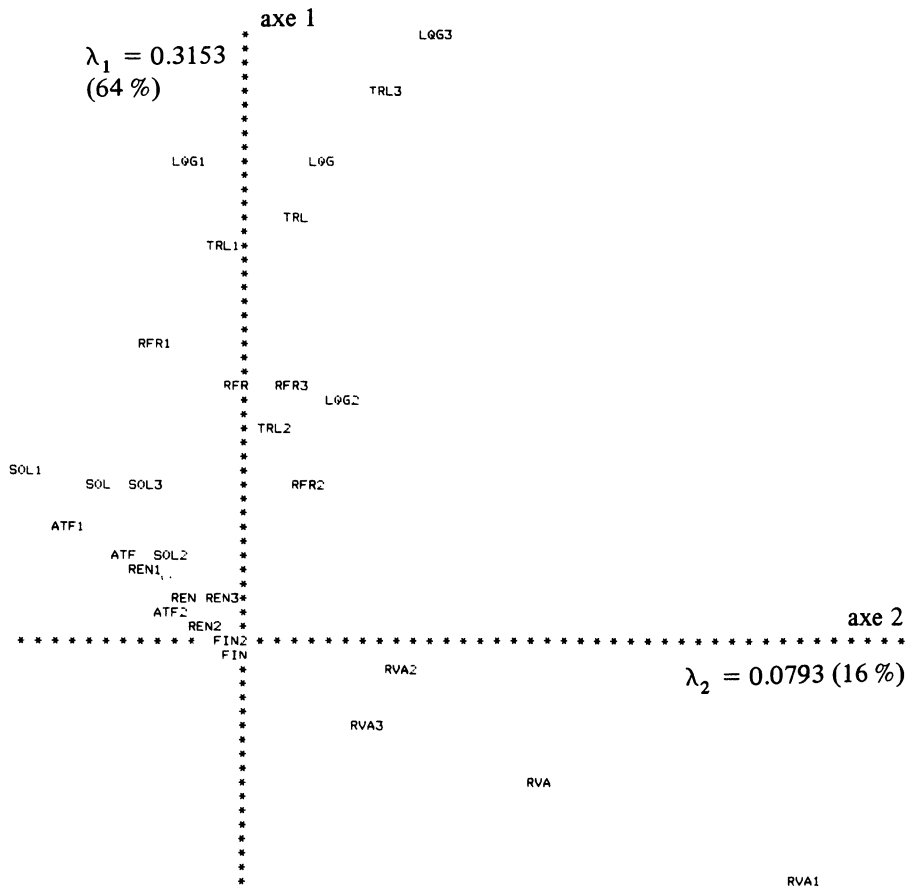
La première propriété qui apparaît est un échelonnement chronologique le long de l'axe 2. Nous nous garderons bien de l'interpréter immédiatement puisque, par un simple tirage au hasard, la probabilité d'un tel échelonnement est de 1/3. Par contre, le décrochement du point COV 2, qui représente l'opérateur de covariance 2 années avant le dépôt de bilan est significatif : la norme de cet opérateur est nettement plus faible que celle des deux autres. Reprenons un peu cette notion de norme. La métrique sur R^p étant canonique l'opérateur de covariance Γ a pour matrice la matrice de covariance V :

$$V = (\text{cov}(X_i, X_j)) \quad i = 1, p, j = 1, p.$$

Le carré de la norme est défini par :

$$\begin{aligned} \|\Gamma\|^2 &= \sum_{i,j} \text{cov}(X_i, X_j)^2 \\ &= \sum_{i,j} \rho_{i,j}^2 V(X_i) V(X_j) \end{aligned}$$

$\rho_{i,j}$ étant le coefficient de corrélation entre les variables X_i et X_j .



Compromis et infrastructures des covariances

La petitesse de la norme de l'opérateur COV 2 s'explique donc par une diminution des covariances entre les différentes variables ou, ce qui est équivalent sous l'hypothèse de l'invariance des corrélations, par une diminution des variances. Cette dernière interprétation est la plus vraisemblable, étant donnée la proximité de l'opérateur COV 2 et de l'opérateur compromis qui est –rappelons-le– défini par le 1^{er} axe dans l'interstructure (nous verrons d'ailleurs qu'elle est confirmée par l'étude des infrastructures). Cette diminution des variances signifie que, dans le tableau des ratios 2 années avant le dépôt de bilan, les entreprises sont moins dispersées que dans les deux autres ; à ce niveau, la poursuite de l'interprétation est délicate, compte-tenu des données : ce resserrement des entreprises est-il dû aux variances interclasses – les 2 classes étant définies par les entreprises disparues et sauvées – ou aux variances intra-classes ? On ne peut répondre à cette question en se bornant aux résultats précédents.

4.3. Infrastructures et compromis

Le schéma ci-dessus donne les représentations des variables sur le plan engendré par les deux premières composantes principales du compromis. Sur la composante principale C'_h ($h = 1$ ou 2), la variable X_j dans le groupe i ($1 \leq i \leq 3$) est, rappelons le, représentée par $\text{cov}_{i_1}(X_j, C'_h)$ (cf. § 2.2). Les notations sont faciles : à l'identificateur du ratio on ajoute 1, 2 ou 3 suivant qu'il s'agit de l'opérateur COV 1, COV 2 ou COV 3.

Les deux premières composantes principales expliquent 80 % de la variance totale il n'est donc pas étonnant que les variables soient bien représentées sur le plan sauf l'une d'entre elles : FIN.

Les résultats concernant ce ratio sont d'ailleurs surprenants : il n'est corrélé à aucune des quatre premières composantes principales et varie donc de façon à peu près indépendante de l'ensemble des autres. Or, sur le plan gestion, il est donné comme indicateur de l'état de santé d'une entreprise, qui, lorsqu'elle marche mal, doit supporter des frais financiers élevés. Par ailleurs, on constate en examinant les matrices des opérateurs que les covariances de ce ratio avec toutes les autres variables sont toujours très faibles (il n'en est pas de même des corrélations) : sa variance étant en particulier très petite, son influence dans l'analyse en est d'autant réduite, et cela suggère d'étudier les corrélations plutôt que les covariances dans la prochaine étude.

On peut interpréter en deux temps les représentations graphiques des autres variables. Tout d'abord, l'analyse en composantes principales du compromis met bien en évidence les propriétés des ratios que nous avons définis précédemment. Les ratios ATF, REN et SOL varient en sens inverse du ratio RVA de valeur ajoutée, ce qui s'explique par le compte "résultat d'exploitation" : lorsque ce dernier diminue, le ratio REN diminue et le ratio RVA augmente ; il suffit pour cela de considérer leurs définitions :

REN = résultat d'exploitation/chiffre d'affaires

RVA = frais de personnel/valeur ajoutée.

REN est donc une fonction croissante alors que RVA est une fonction décroissante du résultat d'exploitation qui est inclus dans la valeur ajoutée. L'opposition du ratio ATF au ratio RVA s'explique de la même façon.

L'opposition entre les ratios de solvabilité LQG, TRL et RFR et le ratio RVA est beaucoup moins nette ; les corrélations sont toutefois négatives, alors qu'elles sont positives avec les ratios de rentabilité.

On peut conclure cette interprétation typique de l'analyse en composantes principales en donnant un sens aux axes : le premier (axe vertical du schéma) représente en fait l'équilibre du bilan des entreprises ; c'est suivant ce caractère qu'elles se distinguent le plus les unes des autres. Le second met en cause des ratios concernant la gestion des entreprises.

La deuxième étape de notre interprétation consiste à étudier les intra-structures : elle est caractéristique de la méthode STATIS.

La première constatation est évidente : les points représentatifs d'une même variable associés aux opérateurs COV 1, COV 2 et COV 3 restent bien groupés autour du point représentatif associé au compromis (qui est le point moyen).

On en déduit que les ratios définis par chaque opérateur présentent les mêmes propriétés générales. On constate également que les positions des variables LQG 1, LQG 2, LQG 3 autour de LQG par exemple correspondent exactement aux positions de SOL 1, SOL 2 et SOL 3 autour de SOL. C'est même presque toujours le cas : une rotation-homothétie unique permet de superposer les variables (sauf le ratio RVA) de l'opérateur COV 3 avec les variables correspondantes de l'opérateur COV 2 ; il en est de même pour COV 2 et COV 1, l'homothétie étant semble-t-il de rapport inverse. Cette propriété est très intéressante car elle permet d'affirmer que l'ordre chronologique obtenu sur l'interstructure n'est pas dû à un facteur aléatoire, puisqu'on le retrouve au niveau de chaque variable, et que les angles entre les ratios d'un même tableau restent à peu près constants, donc leurs corrélations : nous justifions ici l'hypothèse faite dans l'étude de l'interstructure et d'ailleurs, l'étude des intrastructures montre bien que la variance des ratios est beaucoup plus faible deux années avant le dépôt de bilan : cette variance est en effet donnée par le carré de la distance du point représentatif à l'origine des axes.

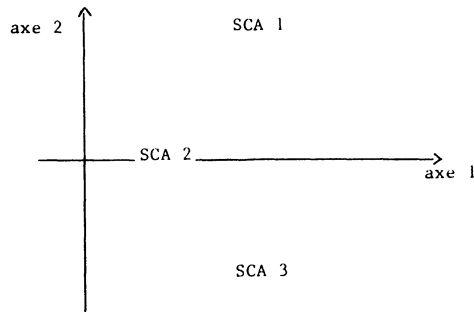
Nous ne pouvons aller plus loin dans l'interprétation avec les résultats établis : l'analyse ne prend pas en compte en effet les deux groupes d'entreprises, mais il semblerait, d'après l'étude des moyennes, que la variation des variances soit due à la variation des variances interclasses.

4.4. Analyse des opérateurs de produit scalaire

Les entreprises de chaque tableau étant les mêmes, nous avons effectué l'analyse des opérateurs de produit scalaire. Les deux analyses ne sont pas équivalentes puisque une permutation des individus dans un tableau ne modifie pas les covariances.

En ce qui concerne l'interstructure, l'interprétation est la même que la précédente : l'ordre chronologique n'est significatif que si un certain nombre d'entreprises le respecte, ce qui sera confirmé. La diminution de la norme du deuxième opérateur caractérise la diminution des produits scalaires entre individus, ou de leur norme : on retrouve exactement la même propriété que sur les covariances et variances.

Le graphique ci-joint représente le plan principal des deux premières composantes principales du compromis, sur lequel on a été placés les points représentatifs des individus, suivant la deuxième méthode exposée dans le paragraphe 3.2. Comme un individu est représenté par 4 points, certains sont absents de la représentation graphique (en cas de point double, seul l'identificateur du premier individu rencontré dans la liste indiqué sur le schéma).



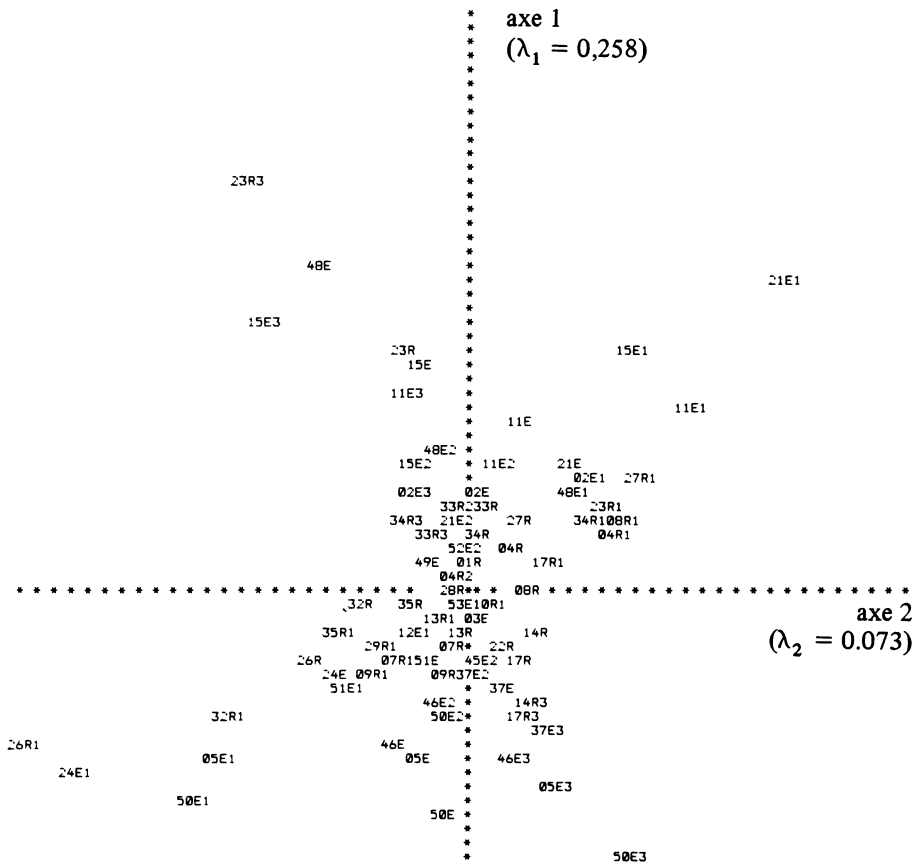
Interstructure des opérateurs de produit scalaire.

La représentation graphique des points définis par le compromis est celle d'une analyse en composantes principales. Elle met surtout en évidence les entreprises dont les ratios, au cours des trois années, prennent des valeurs particulières : leurs points représentatifs sont éloignés du centre de gravité du nuage et l'on peut opposer deux groupes d'entreprises : 48E, 23R, 15E, 11E, 21E d'une part et 50E, 05E, 46E, 24E, 26R d'autre part. On peut remarquer que ces cas particuliers sont surtout des entreprises qui disparaîtront définitivement. Pour interpréter les axes à l'aide de ces deux groupes d'individus, il faudrait examiner les données initiales. Ce n'est pas notre but dans cet article.

La représentation graphique des infrastructures nous apporte des informations intéressantes. Tout d'abord, elle précise la particularité de certaines entreprises : la position du point 48E s'explique par des ratios très particuliers de l'entreprise correspondante dans l'opérateur SCA 3 (le point 48E est barycentre de 48E 3, 48E 2 et 48E 1). Même explication des positions des points 23R, 15E, 11E. L'entreprise 21E se différencie de ce groupe puisque c'est la seule dont la position du point représentatif sur le schéma s'explique par des ratios particuliers dans l'opérateur SCA 1. Le deuxième groupe se scinde en deux : les trois entreprises 50E, 05E et 46E ont des ratios particuliers dans chaque tableau alors que les deux autres : 26R et 24E se différencient surtout dans l'opérateur SCA 1.

Cette représentation graphique met en évidence aussi des propriétés générales : les points représentatifs des entreprises dans l'opérateur SCA 2 n'apparaissent pas sur le schéma : cela est dû à leur concentration autour du centre de gravité et correspond à la diminution de la variance des ratios dans l'opérateur COV 2. Autre caractère général : on retrouve la même propriété d'homothétie rotation, et les points représentatifs d'une même entreprise se suivent autour du centre de gravité dans le sens des aiguilles d'une montre ; le long de la première bissectrice, on trouve les entreprises caractérisées par leurs ratios à la fin de la période (opérateur SCA 1) alors que le long de la troisième, on trouve les mêmes entreprises mais caractérisées par leurs ratios au début de la période (opérateur SCA 3).

Le compromis donne toute l'information classique d'une analyse en composantes principales : groupes de variables, interprétation des axes, groupes d'individus lorsqu'il s'agit d'opérateurs d'Escoufier. Il permet de positionner des points caractérisant les variables ou les individus des opérateurs de la famille et de répondre avec précision aux questions qui ont été posées lors de l'étude de l'interstructure. Il met en évidence les éléments d'évolution particulière et c'est au spécialiste de revenir aux sources pour une étude plus approfondie.



Compromis et infrastructures des produits scalaires

5. CONCLUSION

L'application de la méthode STATIS que nous avons réalisée et dont nous avons interprété les résultats dans les pages précédentes met en évidence un certain nombre d'avantages.

En effet, les résultats que nous avons obtenus sont facilement interprétables : l'interstructure décrit les opérateurs dans leur ensemble et le compromis et les infrastructures détaillent les propriétés de chaque variable ou individu. L'interprétation est analogue à celle d'une analyse en composantes principales, et l'analyse réalisée montre bien le lien entre l'interstructure et les infrastructures, la propriété barycentrique de ces dernières étant issue de la première.

On peut voir un autre avantage à la méthode STATIS telle qu'elle est exposée dans cet article : nous avons proposé une représentation des infrastructures qui ne nécessite pas la diagonalisation de chaque opérateur, ce qui réduit très sensiblement le temps de calcul en ordinateur. Précisions dans le même ordre d'idée que le logiciel existe en BASIC sur une configuration à base de TRS 80 Model I.

Nous terminerons la conclusion par une question : dans l'exemple traité, les tableaux ont mêmes individus et mêmes variables. Quelle est la liaison entre les opérateurs de covariance et les opérateurs de produit scalaire ? Puisqu'il n'existe pas de relation fonctionnelle entre les deux familles d'opérateurs, on peut envisager de mesurer la liaison en calculant le cosinus des opérateurs caractéristiques des familles, mais il reste à en déterminer les propriétés pour pouvoir l'interpréter.

BIBLIOGRAPHIE

- [1] Y. ESCOUFIER. — Opérateur associé à un tableau de données. *Annales de l'INSEE*, n° 22-23, Paris 1976.
- [2] J.C. DEVILLE. — Méthodes statistique et numériques de l'analyse harmonique. *Annales de l'INSEE*, n° 15, Paris 1974.
- [3] L. SCHWARTZ. — Analyse hibernienne. *Hermann, coll. Méthodes*, Paris 1979.
- [4] H. L'HERMIER DES PLANTES. — Structuration des tableaux à trois indices de la statistique. *Thèse 3^e cycle, Université des Sciences et techniques du Languedoc*, Montpellier 1976.
- [5] P. CAZES *et al.* — Codage et analyse des tableaux logiques — introduction à la pratique des variables qualitatives. *Cahiers du BUREAU série Recherche*, n° 27, Paris 1977.
- [6] B. ESCOPIER et J. PAGES. — Comparaison de groupes de variables définies sur le même ensemble d'individus. *Rapport IRISA*, n° 166, Mai 1982, Rennes.