Biom. J. vol. 20 · 1978 · no. 5, pp. 477–483

Université des Sciences et Techniques du Languedoc,
Centre de Recherche en Informatique et Gestion, Montpellier, France

A Propos de la Comparaison Graphique des Matrices de Variance

Y. ESCOUFIER et H. L'HERMIER

Abstract

We compare variance-covariance matrices considered as matrices of similarities between variables. The first part recalls the mathematical bases of the proposed method. The second part treats a set of meteorological data.

Key words: *Principal components analysis, biplots, three way grids, graphical methods, multiple comparison of variances.*

Introduction

Dans un travail récent (1), L. C. A. CORSTEN et K. R. GABRIEL abordent le problème de la comparaison graphique des matrices de variances et covariances. Nous présentons ici une alternative à la solution proposée par ces auteurs.

Notre démarche s'appuie sur les coefficients COVV et RV qui permettent de quantifier la ressemblance de deux tableaux de similarité. Dans une première partie nous rappelons rapidement les bases mathématiques de notre approche. La seconde partie du travail consiste à traiter les données de L. C. A. CORSTEN et de K. R. GABRIEL ce qui permet d'apprécier aussi bien les limites que les avantages de notre solution.

L'essentiel de la méthode proposée a fait l'objet d'une thèse (4) pratiquement épuisée à ce jour. Pour cette raison, l'article a été conçu de manière à être compris par lui-même, sans référence à la thèse en question.

1. Présentation de la Méthode

I_1 – Soit $\boldsymbol{E}$ l'ensemble des matrices réelles symétriques, $n \times n$, $\boldsymbol{E}$ est un espace vectoriel sur R qu'on peut munir du produit scalaire:

$$\forall (\boldsymbol{M}_1, \boldsymbol{M}_2) \in \boldsymbol{E} \times \boldsymbol{E} : \langle \boldsymbol{M}_1, \boldsymbol{M}_2 \rangle = Tr(\boldsymbol{M}_1 \boldsymbol{M}_2).$$

A ce produit scalaire correspond la norme $\|\boldsymbol{M}_1\| = [Tr(\boldsymbol{M}_1^2)]^{1/2}$ et la distance:

$$d^2(\boldsymbol{M}_1, \boldsymbol{M}_2) = Tr[(\boldsymbol{M}_1 - \boldsymbol{M}_2)^2] = Tr(\boldsymbol{M}_1^2) + Tr(\boldsymbol{M}_2^2) - 2Tr(\boldsymbol{M}_1 \boldsymbol{M}_2).$$

On peut aussi parler de l'angle que font deux éléments de $\boldsymbol{E}$; il aura pour cosinus la quantité

$$Tr(\boldsymbol{M}_1\boldsymbol{M}_2)/[Tr(\boldsymbol{M}_1^2)\ Tr(\boldsymbol{M}_2^2)]^{1/2}.$$

Les propriétés du produit scalaire (qu'on peut retrouver par des calculs simples) montrent alors que si $\boldsymbol{M}_3 = k\boldsymbol{M}_1$ où k est un nombre réel quelconque, on a:

$$\|\boldsymbol{M}_3\| = |k|\ \|\boldsymbol{M}_1\|$$

et

$$Tr(\boldsymbol{M}_3\boldsymbol{M}_2)/[Tr(\boldsymbol{M}_3^2)\ Tr(\boldsymbol{M}_2^2)]^{1/2} = kTr(\boldsymbol{M}_1\boldsymbol{M}_2)/|k|\ [Tr(\boldsymbol{M}_1^2)\ Tr(\boldsymbol{M}_2^2)]^{1/2}.$$

Dans certaines applications, on peut souhaiter considérer comme équivalentes des matrices $n \times n$ qui sont proportionnelles. Ceci conduit à travailler avec $\boldsymbol{M}_1^* = \boldsymbol{M}_1/[Tr(\boldsymbol{M}_1^2)]^{1/2}$. On a alors

$$d^2(\boldsymbol{M}_1^*, \boldsymbol{M}_2^*) = 2\ (1 - Tr(\boldsymbol{M}_1\boldsymbol{M}_2)/[Tr(\boldsymbol{M}_1^2)\ Tr(\boldsymbol{M}_2^2)]^{1/2}).$$

Pour simplifier l'écriture et également pour faire le lien avec d'autres travaux (2,5), nous adopterons dans la suite les notations suivantes

$$COVV(\boldsymbol{M}_1, \boldsymbol{M}_2) = Tr(\boldsymbol{M}_1\boldsymbol{M}_2)$$
$$VAV(\boldsymbol{M}_1) = Tr(\boldsymbol{M}_1^2)$$
$$RV(\boldsymbol{M}_1, \boldsymbol{M}_2) = Tr(\boldsymbol{M}_1\boldsymbol{M}_2)/[Tr(\boldsymbol{M}_1^2)\ Tr(\boldsymbol{M}_2^2)]^{1/2}.$$

I_2 – Supposons donné $(\boldsymbol{M}_1, \ldots, \boldsymbol{M}_k)$ un sous-ensemble de $\boldsymbol{E}$ formé de k matrices *semi définies positives* et définissons la matrice $\boldsymbol{S}$, $k \times k$, d'éléments $S_{ij} = COVV(\boldsymbol{M}_i, \boldsymbol{M}_j)$, $(i = 1, \ldots, k; j = 1, \ldots, k)$. On a les résultats suivants.

Proposition 1: S est semi définie positive

Le résultat s'obtient en calculant la norme de $\sum_{i=1}^{k} \alpha_i \boldsymbol{M}_i$ où $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_k)$ est un vecteur quelconque de R^k.

Proposition 2: Tous les éléments de $\boldsymbol{S}$ *sont non négatifs*

Si $\boldsymbol{M}_i$ et $\boldsymbol{M}_j$ sont semi-définies positives, il existe $\boldsymbol{U}_i$ et $\boldsymbol{U}_j$ telles que $\boldsymbol{M}_i = \boldsymbol{U}_i\boldsymbol{U}_i'$ et $\boldsymbol{M}_j = \boldsymbol{U}_j\boldsymbol{U}_j'$. Alors on a:

$$\begin{aligned} COVV(\boldsymbol{M}_i, \boldsymbol{M}_j) &= Tr(\boldsymbol{U}_i\boldsymbol{U}_i'\boldsymbol{U}_j\boldsymbol{U}_j') = Tr(\boldsymbol{U}_i'\boldsymbol{U}_j\boldsymbol{U}_j'\boldsymbol{U}_i) \\ &= Tr((\boldsymbol{U}_i'\boldsymbol{U}_j)\ (\boldsymbol{U}_i'\boldsymbol{U}_j)') \geqq 0. \end{aligned}$$

Proposition 3 (corollaire). Le premier vecteur propre $\boldsymbol{\alpha}$ *de* $\boldsymbol{S}$ *a tous ses éléments de même signe.* On choisira le signe positif.

Ce résultat est connu sous le nom de théorème de FROBÉNIUS.

Proposition 4 (corollaire). Soit $\boldsymbol{\alpha}$ le premier vecteur propre à éléments positifs de $\boldsymbol{S}$. On pose $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_k)$. *La matrice* $\boldsymbol{M} = \sum_{i=1}^{k} \alpha_i \boldsymbol{M}_i$ *est semi-définie positive* en tant que combinaison linéaire positive de matrices semi-définies positives.

Théorème: *Pour tout vecteur* $\boldsymbol{\beta}$ *tel que* $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_k)$ et $\sum_{i=1}^{k} \beta_i^2 = \sum_{i=1}^{k} \alpha_i^2$, on a:

$$\sum_{j=1}^{k} \left[COVV\left(\sum_{i=1}^{k} \beta_i \boldsymbol{S}_i, \boldsymbol{S}_j \right) \right]^2 \leqq \sum_{j=1}^{k} \left[COVV\left(\sum_{i=1}^{k} \alpha_i \boldsymbol{S}_i, \boldsymbol{S}_j \right) \right]^2 .$$

Démonstration du théorème: Formons la quantité

$$\varphi(\boldsymbol{\beta}) = \sum_{i=1}^{k} COVV\left(\sum_{i=1}^{k} \beta_i \boldsymbol{S}_i, \boldsymbol{S}_j \right)^2 - \lambda \sum_{i=1}^{k} \beta_i^2 .$$

La linéarité de l'opérateur Trace et donc de $COVV$ permet d'écrire:

$$\varphi(\boldsymbol{\beta}) = \boldsymbol{\beta}' \boldsymbol{S}^2 \boldsymbol{\beta} - \lambda \boldsymbol{\beta}' \boldsymbol{\beta} .$$

On sait que le maximum de cette fonction sera obtenu en prenant pour $\boldsymbol{\beta}$ le vecteur propre de $\boldsymbol{S}^2$ (et donc de $\boldsymbol{S}$) associé à la plus grande valeur propre de $\boldsymbol{S}^2$ (et donc de $\boldsymbol{S}$ puisque $\boldsymbol{S}$, semi-définie positive, a ses valeurs propres non négatives).

I$_3$ – De ces résultats découle la procédure suivante. Soit à comparer k matrices $(\boldsymbol{M}_i, i = 1, \ldots, k)$ de variances et covariances portant sur les mêmes n variables. Ces matrices sont considérées comme des matrices de similarité entre variables.

Première étape: Construire la matrice $\boldsymbol{S}$ d'éléments

$$S_{ij} = Tr(\boldsymbol{M}_i \boldsymbol{M}_j) = COVV(\boldsymbol{M}_i, \boldsymbol{M}_j) .$$

Seconde étape: Calculer les valeurs propres et les vecteurs propres de $\boldsymbol{S}$. Ceci permet une représentation (3) des matrices par des points de R^k qui sont tels que la distance entre les points P_i et P_j associés aux matrices $\boldsymbol{M}_i$ et $\boldsymbol{M}_j$ est $d(\boldsymbol{M}_i, \boldsymbol{M}_j)$. On obtient donc à cette étape une interprétation globale des ressemblances et des différences entre les matrices $\boldsymbol{M}_i$.

Nous appellerons cette étape, l'étape de l'*interstructure*.

Troisième étape: Calculer $\boldsymbol{M} = \sum_{i=1}^{k} \alpha_i \boldsymbol{M}_i$ où α d'éléments $(\alpha_1, \ldots, \alpha_k)$ est le vecteur propre de $\boldsymbol{S}$ associé à la plus grande valeur propre. Puisque $\boldsymbol{M}$ est semi-définie positive, sa factorisation fournit une représentation des variables qui est le *meilleur compromis* (au sens du théorème) entre les diverses représentations que pourraient donner chacune des matrices prises séparément.

Quatrième étape: Appelons $\boldsymbol{U}$, la matrice dont les colonnes sont les vecteurs propres de $\boldsymbol{M}$, de normes égales à la racine carré des valeurs propres correspondantes. $\boldsymbol{U}$ est la matrice qui a permis la représentation des variables à la troisième étape. Appelons $\boldsymbol{U}_i$ la matrice associée de façon analogue à $\boldsymbol{M}_i$ $(i = 1, \ldots, k)$.

Considérons les colonnes de $\boldsymbol{U}$ et de $\boldsymbol{U}_i$ $(i = 1, \ldots, k)$ comme des vecteurs de R^n. Les cosinus des vecteurs colonnes de $\boldsymbol{U}$ avec les vecteurs colonnes des $\boldsymbol{U}_i$ $(i = 1, \ldots, k)$ sont donnés par les matrices.

$$\boldsymbol{Q}_i = (\boldsymbol{U}'\boldsymbol{U})^{-1/2} \boldsymbol{U}'\boldsymbol{U}_i (\boldsymbol{U}_i'\boldsymbol{U}_i)^{-1/2} \quad (i = 1, \ldots, k) .$$

Il en découle que les matrices $\boldsymbol{U}_i\boldsymbol{Q}_i'$, $(i=1, \ldots, k)$ sont les coordonnées des n variables telles qu'elles sont vues dans les matrices $\boldsymbol{M}_i$, $(i=1, \ldots, k)$ exprimées dans la base du compromis. Une superposition des différentes représentations permet alors d'étudier les évolutions des variables d'une matrice à l'autre. Cette étape est appelée étape des *intrastructures*.

2. Application

II_1 – L. C. CORSTEN et K. R. GABRIEL étudient trois matrices de variances et covariances portant sur huit variables ($k=3$; $n=8$). Nous ne reproduirons pas ces matrices. Rappelons simplement que l'étude porte sur la comparaison des pluies journalières observées dans huit régions d'Israël "There were three northern regions: coast, interior and east and three corresponding central regions. There was also a narrow buffer region between north and centre, and a small region in the south".

La donnée originale étant la hauteur de pluie journalière moyenne sur la région, on a trois matrices de variances et covariances:

$\boldsymbol{N}$ qui correspond à un ensemencement des nuages dans la région nord

$\boldsymbol{C}$ qui correspond à un ensemencement des nuages dans la région centre

$\boldsymbol{P}$ qui correspond à une situation pré-expérimentale.

Le problème est de comparer $\boldsymbol{N}$ et $\boldsymbol{C}$ entre elles et à $\boldsymbol{P}$.

II_2 – Pour l'application les matrices sont prises dans l'ordre $\boldsymbol{P}$, $\boldsymbol{N}$, $\boldsymbol{C}$. Les variables sont notées comme en (1), ($\boldsymbol{NC}$, $\boldsymbol{NI}$, $\boldsymbol{NE}$, $\boldsymbol{B}$, $\boldsymbol{CC}$, $\boldsymbol{CI}$, $\boldsymbol{CE}$, $\boldsymbol{S}$).

Étapes 1 et 2

On trouve d'abord:

$$\boldsymbol{S} = 10^6 \times \begin{bmatrix} 0.434 & 0.443 & 0.563 \\ 0.443 & 0.526 & 0.621 \\ 0.563 & 0.621 & 0.800 \end{bmatrix}$$

Les éléments propres de $\boldsymbol{S}$ sont donnés par le tableau suivant.

Valeur propre		1703975.0	35015.4	21487.5
Vecteur	P	0.492	−0.688	−0.533
propre	N	0.544	0.721	−0.429
normé	C	0.721	−0.079	0.729

La figure n° 1 fournit la représentation de l'interstructure limitée aux deux premiers axes. Les coordonnées des points sont obtenues en multipliant les vecteurs propres par la racine carré de la valeur propre qui leur correspond. L'interprétation de cette figure doit se faire à partir des normes des vecteurs $\overrightarrow{OP}$, $\overrightarrow{ON}$ et $\overrightarrow{OC}$ d'une part, des angles de ces vecteurs d'autre part.

La comparaison des normes nous montre que l'ensemencement des nuages a pour effet une augmentation des normes des matrices N et C, c'est à dire globalement des variances et des covariances des variables.

Des angles, nous déduisons qu'il y a plus de différences entre N et P qu'entre C et P c'est à dire, la référence étant P, les positions des variables sont plus modi-

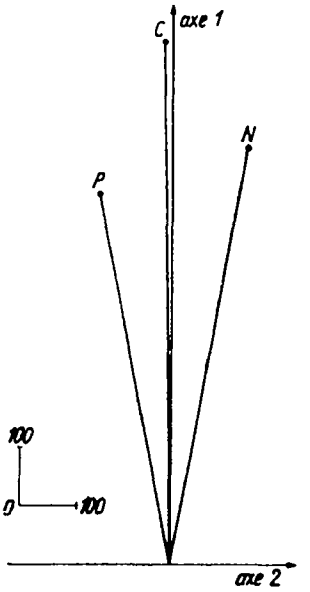

Fig. n° 1. Interstructure

fiées par l'ensemencement en N que par l'ensemencement en C. Autrement dit l'ensemencement des nuages en N perturbe plus les covariances des pluies journalières que ne le fait l'ensemencement en C.

La faiblesse de la seconde valeur propre par rapport à la première (et donc celle des angles) oblige à conclure que l'effet majeur de l'ensemencement est un effet d'augmentation des variances. Les modifications sur les covariances sont réelles mais mineures.

Étapes 3 et 4

La figure n° 2 fournit la représentation du meilleur compromis limité aux deux premiers axes obtenu comme indiqué en I_3. On y retrouve la répartition des régions. Dans des études où k est grand, l'étude du compromis en lui-même présente un intérêt comme vision globale des objets. Ici, ce n'est pas le cas : l'intérêt du compromis est surtout de permettre les figures n° 3 et n° 4 où la superposition des variables telles qu'elles sont vues par P et N d'une part, P et C d'autre part autorise l'étude comparée des matrices.

Sur la figure n° 3, on voit que l'ensemencement dans la région nord provoque une augmentation des variances des variables NC, B, NI et le pincement du faisceau NC, B, NI, NE. C'est dire que l'ensemencement au nord provoque une

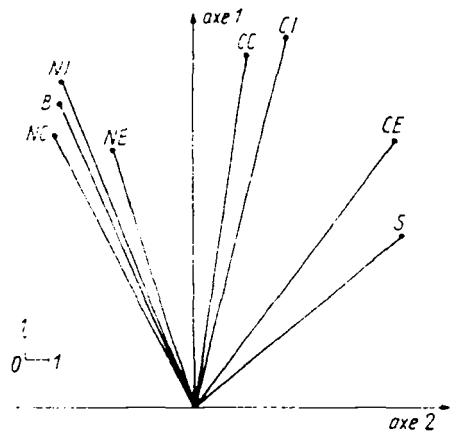

Fig. n° 2. Compromis

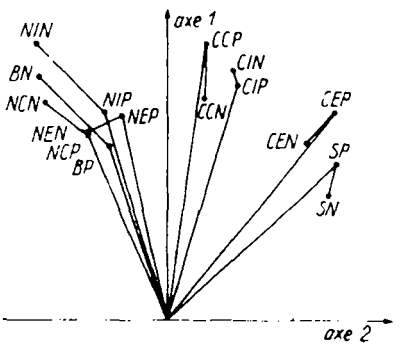

Fig. n° 3. Comparaison des matrices **P** et **N**

augmentation des variances et des covariances des régions du nord. Parallèlement les variables ***S***, ***CE***, ***CC*** voient leurs variances diminuer. De plus, le déplacement vers la gauche du faisceau des régions nord indique une diminution des covariances entre les régions nord et les autres.

L'étude de la figure n° 4, montre que l'ensemencement au centre a un effet global d'augmentation des variances. Seules les régions **S** et ***NE*** y échappent. Le faisceau des régions du centre se pince légèrement, tandis que les régions du nord

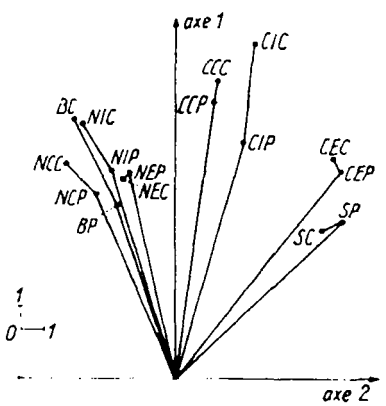

Fig. n° 4. Comparaison des matrices **P** et **C**

s'écartent globalement mais légèrement vers la gauche. On peut dire que l'ensemencement au centre a un effet plus global que l'ensemencement au nord: il augmente les variances sans trop modifier les covariances (on a là une vision détaillée de l'information contenue dans la figure n° 1 où l'angle $\widehat{POC}$ est inférieur à l'angle $\widehat{PON}$).

Remarque finale

Il est certainement apparu clairement au lecteur, que la méthode que nous avons appliquée ici à trois matrices de variances et covariances peut être appliquée sans changement à tout ensemble de k matrice de similarités portant sur les mêmes n objets pourvu que ces matrices soient semi-définies positives.

On doit également noter que la même démarche peut-être faite en substituant le coefficient RV au coefficient $COVV$. Ce choix, qui conduit à considérer comme équivalentes des matrices proportionnelles, n'était pas judicieux ici puisque la norme des matrices avait une importance pratique.

Résume

On compare des matrices de variances et covariances considérées comme matrice de similarités entre variables. Une première partie rappelle les bases mathématiques de la méthode proposée. La seconde partie traite des données météorologiques.

Références

(1) Corsten, L. C. A., et K. R. Gabriel, 1976: Graphical Exploration in Comparing Variance Matrices. Biometrics **32**, 851–863.
(2) Escoufier, Y., 1973: Le traitement des variables vectorielles. Biometrics **29**, 751–760.
(3) Gower, J. C., 1966: Some Distances Properties of latent root and vector methods used in multivariate analysis. Biometrika **58**, 325–338.
(4) L'Hermier, H., Thèse 1976: Structuration des tableaux à trois indices de la statistique: théorie et application d'une méthode d'analyse conjointe. Thèse présentée à l'Université des Sciences et Technique du Languedoc, Montpellier.
(5) Robert, P., et Y. Escoufier, 1976: A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. Appl. Statist. **25**, 257–265.

Received: VIII/4/1977

Author's address:
Université des Sciences et Techniques
du Languedoc, Centre de Recherche
en Informatique et Gestion
Avenue d'Occitanie
34075 Montpellier Cedex, France