Statistique et analyse des données

Y. Romain
S. Viguier

COMPCOV : un outil de comparaison de plusieurs matrices de covariance

Statistique et analyse des données, tome 14, n° 3 (1989), p. 37-52
<http://www.numdam.org/item?id=SAD_1989__14_3_37_0>

Statistique et Analyse des Données
1989 - Vol. 14 n°3 - p. 37 - 52

COMPCOV : un outil de comparaison de plusieurs matrices de covariance

Y.Romain & S.Viguier
Laboratoire de Statistique et Probabilités
U. A. CNRS 735 , Université Paul Sabatier
31062 TOULOUSE CEDEX

Résumé :

A la suite des travaux de FLURY (1988), on présente les méthodes et le logiciel COMPCOV relatifs à la comparaison de plusieurs matrices de covariance telles que les tests d'égalité, de proportionnalité, de diagonalisations simultanées (totale ou partielle), ou encore d'égalité de sous-espaces propres et on donne un exemple d'application en orthodontie.

Abstract :

Following FLURY's (1988) works, methods for comparing several covariance matrices and a related software named COMPCOV are presented. These methods include tests of equality, proportionality, total or partial simultaneous diagonalisations and equality of eigenspaces. An example for orthodontics data is given.

Mots-clé : Matrices de covariance, diagonalisation simultanée, Analyse en Composantes Principales.

Keywords : Covariance matrices, k-sample case, simultaneous diagonalization , Principal Component Analysis.

Manuscrit reçu le 25 juillet 1989, révisé le 26 février 1990

Introduction

En statistique multidimensionnelle, les matrices de covariance jouent un rôle prépondérant aussi bien lors de problèmes à approche purement descriptive que lors de problèmes de modélisation ou de type inférentiel. A une famille de matrices de covariance $(\Sigma_i)_{i=1,...,k}$, de même ordre p, peuvent être associés divers tests d'hypothèses telles que l'égalité de ces matrices, notée $\mathfrak{E}$, leur proportionnalité, notée $\mathfrak{P}$, leurs diagonalisations simultanées totale et partielle, notées $\mathfrak{DST}$ et $\mathfrak{DSP}$, ou encore l'égalité de leurs sous-espaces propres, notée $\mathfrak{ESP}$. L'ensemble des récents travaux de Flury (1984,1986b,1987) sur les tests de type $\mathfrak{P}$, $\mathfrak{DST}$, $\mathfrak{DSP}$ et $\mathfrak{ESP}$ complétés notamment par ceux de Schott (1988) pour les tests de type $\mathfrak{ESP}$ et par les travaux plus "classiques" sur les tests d'homogénéité de type $\mathfrak{E}$ (voir Muirhead (1982) par exemple), permet d'appréhender ces problèmes de comparaison de façon globale, hiérarchisée et cohérente.

Toutes ces méthodes, essentiellement basées sur les estimateurs du maximum de vraisemblance et sur les statistiques de test du rapport des vraisemblances, s'appuient sur des hypothèses de multinormalité et d'indépendance des observations. En fait, le modèle paramétrique de base est:

$$\bigotimes_{i=1}^{k} (\mathbb{R}^p, \mathcal{B}_{\mathbb{R}^p}, \mathcal{N}(\mu_i, \Sigma_i))^{\otimes N_i}$$

où les deux paramètres μ_i, élément de $\mathbb{R}^p$ et Σ_i, élément de l'ensemble des matrices symétriques définies positives d'ordre p (on suppose, pour tout i, que N_i est plus grand que p), sont respectivement l'espérance et la matrice de covariance du $i^{ème}$ vecteur gaussien, noté ${}^tX^i = (X^{i1}, X^{i2}, ..., X^{ip})$. On a donc k vecteurs gaussiens indépendants, chacun observé indépendamment N_i fois pour le $i^{ème}$ échantillon : concrètement dans les applications, on a souvent k tableaux (N_i,p) des observations des mêmes p variables quantitatives sur N_i individus. Dans la suite, on note S_i les estimateurs non biaisés classiques des Σ_i, obtenus à partir des estimateurs de maximum de vraisemblance des Σ_i, c'est-à-dire, pour tout i de 1 à k,

$$S_i = \frac{1}{n_i} \sum_{h=1}^{N_i} {}^t(X_h^i - \bar{X}^i)(X_h^i - \bar{X}^i) \quad \text{où } n_i = N_i - 1 \text{ , } \bar{X}^i = \frac{1}{N_i} \sum_{h=1}^{N_i} X_h^i,$$

et où $(X_h^i)_{h=1,...,Ni}$ sont les applications coordonnées usuelles pour le $i^{ème}$ échantillon statistique: on sait alors que les variables aléatoires $(n_i S_i)_{i=1,...,k}$ sont indépendantes, chacune de loi de Wishart de paramètres n_i et Σ_i.

La première partie de l'article est consacrée à une brève revue méthodologique et bibliographique des tests de comparaison cités plus haut. Dans la seconde partie, on présente le logiciel interactif COMPCOV qui permet d'effectuer tous ces tests et dans la troisième partie, on donne un exemple d'application de COMPCOV sur des données d'orthodontie. Divers

éléments de développements possibles de ces méthodes de comparaison et autres remarques sont proposés en guise de conclusion.
Il faut remarquer que ces méthodes peuvent aussi être présentées comme des outils de comparaison d'Analyses en Composantes Principales sur k groupes distincts et qu'elles s'inscrivent en complément de toutes les méthodes étudiant k tableaux (N_i,p) successifs. On doit néanmoins les distinguer des autres travaux d'approche purement descriptive comme ceux de Krzanowski(1979,1984), Kiers et Ten Berge(1989) ou encore de Millsap et Meredith(1989) concernant les analyses simultanées sur plusieurs groupes d'individus.

I Revue des techniques de tests de COMPCOV

I 1. Test de multinormalité

Toutes les méthodes de tests possibles dans COMPCOV présupposent la multinormalité et l'indépendance des observations.
Pour éprouver l'hypothèse de multinormalité de chaque vecteur X^i, on peut trouver de tels outils chez Cox et Small(1978), Royston(1983), puis récemment chez Jaiswal et Jain(1988). C'est le test proposé par ces deux derniers auteurs qui a été intégré dans COMPCOV.

I.2. Test d'égalité

Ce test, aussi appelé test d'homogénéité des variances, est celui de l'hypothèse H_0 : $\Sigma_1=\Sigma_2=...=\Sigma_k=\Sigma$, avec Σ non spécifiée, contre sa négation. Sa forme la plus classique, telle qu'on peut la trouver chez Anderson(1984) ou Muirhead(1982), est fondée sur le rapport des vraisemblances maximales des $(\Sigma_i)_{i=1,...,k}$ et la statistique du test, notée E, est:

$$E = \frac{\prod_{i=1}^{k} (\det n_i S_i)^{N_i/2} N^{pN/2}}{(\det \sum_{i=1}^{k} n_i S_i)^{N/2} \prod_{i=1}^{k} N_i^{pN_i/2}} \qquad \text{avec } N = \sum_{i=1}^{k} N_i.$$

En fait, lorsque les N_i sont différents, on modifie E pour rendre le test sans biais (cf. Perlman, 1980) en remplaçant N_i par n_i et N par N-k et on note E* la nouvelle statistique. De plus, on sait que, sous H_0, -2logE et -2logE* suivent asymptotiquement, c-à-d. pour $\min_i N_i \to +\infty$, la

loi du chi-deux à $\frac{1}{2}p(p+1)(k-1)$ ddl et que, pour améliorer la précision du test, plusieurs types d'ajustements, dits "de Bartlett", ont été proposés par divers auteurs dont on peut avoir une synthèse dans les travaux de Møller (1986) et de Barndorf-Nielsen & Hall (1988). Le test implanté dans COMPCOV est celui basé sur E*, auquel on a intégré l'amélioration b_3 citée par Møller.

Concernant ce test sans hypothèse de multinormalité, on peut se référer aux travaux de Sen & Puri (1968) et à l'algorithme de calcul de ce test publié par Parhizgari & Prakash (1989).

I.3. Test de proportionnalité

L'hypothèse nulle est : pour tout i de {2,...,k}, $\Sigma_i=c_i\Sigma_1$ avec c_i réel strictement positif. Une étude détaillée concernant la propriété de proportionnalité entre matrices de covariance est donnée par Flury (1986b) et on a repris l'algorithme de résolution des équations de vraisemblance proposé par cet auteur, l'existence et l'unicité de ces solutions ayant été montrées par Jensen et Johansen (1987) et par Eriksen (1987).

Soient $(\lambda_i)_{i=1,...,p}$ et $\beta=(\beta_i)_{i=1,...,p}$ les valeurs et vecteurs propres de Σ_1. En notant, pour tout i, $\hat{c}_i$, $\hat{\lambda}_i$, $\hat{\beta}_i$ les solutions de l'algorithme de résolution des équations de vraisemblance puis $\hat{\Sigma}_i$ les estimateurs associés à ces solutions ($\hat{\Sigma}_i=\hat{c}_i\,\hat{\Sigma}_1$, ce qui implique $\hat{c}_1=1$), la statistique du test, notée P, est :

$$P=\sum_{i=1}^{k} n_i \mathrm{Log}\left(\frac{\det\hat{\Sigma}_i}{\det S_i}\right) = n\sum_{i=1}^{p}\mathrm{Log}\hat{\lambda}_i+\sum_{i=1}^{p} n_i\,(p\,\mathrm{Log}(\hat{c}_i - \mathrm{Log}(\det S_i))\,,$$

qui suit asymptotiquement, sous H_0, la loi du chi-deux à $(k-1)(p^2+p-2)/2$ d.d.l..

I.4. Test de diagonalisation simultanée totale

L'hypothèse nulle est : pour tout i de {1,...,k}, il existe Ω_i matrice diagonale et B matrice orthogonale d'ordre p telles que ${}^tB\Sigma_iB=\Omega_i$. C'est donc l'hypothèse associée à la recherche d'une décomposition simultanée des Σ_i sur les mêmes axes principaux (cf. Flury 1987,1988).

Si, pour tout i de 1 à k, on note $(\lambda_{ij})_{j=1,...,p}$ et $\beta=(\beta_1,...,\beta_p)$ les valeurs et vecteurs propres de Σ_i, le système d'équations de vraisemblance à résoudre est, pour tout m et j de {1,...,p}, m différent de j :

(*) $${}^t\beta_m\left(\sum_{i=1}^{k} n_i\frac{\lambda_{im}-\lambda_{ij}}{\lambda_{im}\lambda_{ij}}\right)\beta_j=0,\quad m,j=1,...,p,\ m\neq j,$$ sous les contraintes ${}^t\beta_m\beta_j=\delta_{mj}$, où δ_{mj} désigne le symbole de Kronecker.

En notant $\hat{\lambda}_{ij}, \hat{\beta}$ les solutions de ce système et en posant $\hat{\Sigma}_i = \hat{\beta}(\mathrm{diag}\hat{\lambda}_{ij})^t\hat{\beta}$, la statistique du test, notée DST, est $\sum_{i=1}^{k} n_i \mathrm{Log}\left(\frac{\det\hat{\Sigma}_i}{\det S_i}\right)$; elle suit asymptotiquement, sous H_0, la loi du chi-deux à (k-1)p(p-1)/2 d.d.l.. Le test basé sur DST dans COMPCOV utilise l'algorithme proposé par Flury et Gautschi (1986), appelé le FG-algorithme, amélioré par les remarques de Clarkson (1988).

I.5. Test de diagonalisation simultanée partielle

On teste l'hypothèse que les ensembles de vecteurs propres des Σ_i ont une partie commune (cf. Flury, 1987) et donc H_0 s'écrit: pour tout i de {1,...,k}, $\Sigma_i = \beta_i \Omega_i {}^t\beta_i$ où $\Omega_i = (\mathrm{diag}\lambda_{ij})_{j=1,...,p}$ et $\beta_i = (\beta_1^i, ..., \beta_p^i) = (\beta_c, \beta_s^i)$ où β_c est une partie commune à tous les Σ_i, et β_s^i est une partie spécifique à Σ_i. L'hypothèse est appelée d'ordre q si β_c est de dimension (p,q). Le système des équations de vraisemblance, plus complexe que celui du cas précédent, contient notamment les mêmes équations que (*), où m et j sont inférieurs ou égaux à q, plus d'autres contraintes d'orthogonalité. La statistique de test, notée DSP(q), qui suit asymptotiquement, sous H_0, la loi du chi-deux à (k-1)q(2p-q-1)/2 d.d.l., est définie de manière analogue à DST. En fait, on ne sait obtenir que des solutions approchées du système d'équations de vraisemblance et, dans COMPCOV, on utilise la procédure de résolution approchée proposée par Flury (1987).

I.6. Test d'égalité de sous-espaces propres

La recherche de sous-espaces propres communs à plusieurs matrices de covariance est directement associée au problème de comparaison de plusieurs Analyses en Composantes Principales sur k groupes ou échantillons.

Plusieurs auteurs (Levin (1966), Corsten et Gabriel (1971), Krzanowski (1979,1984), Millsap et Meredith (1989), Kiers et Ten Berge (1989)) ont proposé des approches différentes méthodologiquement de celles reprises dans ce paragraphe. Les caractères profondément différents de ces méthodes nous semblent plaider en faveur de leur complémentarité plutôt que de leur opposition que soulignent Kiers et Ten Berge.

I.6.1 Le test proposé par Flury (1987)

Pour tout i de {1,...,k}, chaque β_i se partitionne en (β_1^i, β_2^i) de dimensions respectives (p,q) et (p,p-q) et l'hypothèse nulle, d'ordre q, s'écrit : pour tout i et m de {1,...,k}, ${}^t\beta_1^i \beta_2^m$ est la matrice nulle de dimension (q,p-q). Comme précédemment, les solutions du système des équations de vraisemblance sont approchées, la statistique de test, notée $ESP_F(q)$, suit asymptotiquement, sous H_0, la loi du chi-deux à (k-1)q(p-q) d.d.l. et on a utilisé dans COMPCOV la procédure proposée par l'auteur.

1.6.2. Le test proposé par Schott (1988)

Le test précédent n'oblige en rien à ce que les q vecteurs propres considérés pour chacune des matrices de covariance soient, par exemple, ceux associés aux q plus grandes valeurs propres, et, dans le contexte de l'Analyse en Composantes Principales sur plusieurs échantillons, on peut le compléter avantageusement par la procédure proposée par Schott (1988) qui est cependant restreinte au cas de deux échantillons. L'hypothèse nulle est donc ici que les sous-espaces propres associés aux q plus grandes valeurs propres de Σ_1 et Σ_2 sont identiques. On note $(\lambda_i(V))_{i=1,...,p}$, la suite décroissante des valeurs propres, supposées distinctes, d'une matrice de covariance V. Si H_0 est vraie, on a :

$$\sum_{i=1}^{q} \lambda_i(\Sigma_1+\Sigma_2)=\sum_{i=1}^{q} \lambda_i(\Sigma_1)+\sum_{i=1}^{q} \lambda_i(\Sigma_2)$$

et la statistique de ce test, notée $ESP_S(q)$, est :

$$\sum_{i=1}^{q} [\lambda_i(n_1S_1)+\lambda_i(n_2S_2)-\lambda_i(n_1S_1+n_2S_2)] ,$$

qui suit asyptotiquement, sous H_0, la loi du chi-deux ajusté, noté $c\chi^2(d)$ avec $c=\frac{\hat{\sigma}^2}{2\hat{\mu}}$ et d entier le plus proche de $\frac{2\hat{\mu}^2}{\hat{\sigma}^2}$, où $\hat{\mu}$ et $\hat{\sigma}$ sont les estimations de la moyenne et de l'écart-type de $ESP_S(q)$ données par l'auteur.

II Le logiciel COMPCOV

Les tests présentés dans la partie précédente ont tous été programmés et rassemblés dans le logiciel COMPCOV. Le programme pour tester l'hypothèse $\mathfrak{DST}$ a déjà été écrit en Fortran par Flury et Constantine (1985); nous en avons adopté une version améliorée donnée par Clarkson (1988). Les algorithmes utilisés pour les autres tests sont : pour le test de proportionnalité, celui écrit par Flury (1986b), et pour les test des hypothèses $\mathfrak{DSP}$ et $\mathfrak{ESP}_F$, ceux publiés dans Flury (1987) avec des initialisations tenant compte des remarques de Krzanowski (1984).

Pour l'écriture de COMPCOV, on a utilisé le logiciel GAUSS (1988) qui sert à la fois de langage de programmation et de logiciel interactif et qui présente un certain nombre d'avantages pour les programmes statistiques :

- existence de sous-programmes de calcul matriciel,
- traitements statistiques utilisables par des commandes déjà intégrées,
- grande précision de calcul (par exemple, utilisation de l'algorithme QR pour la diagonalisation des matrices), etc...

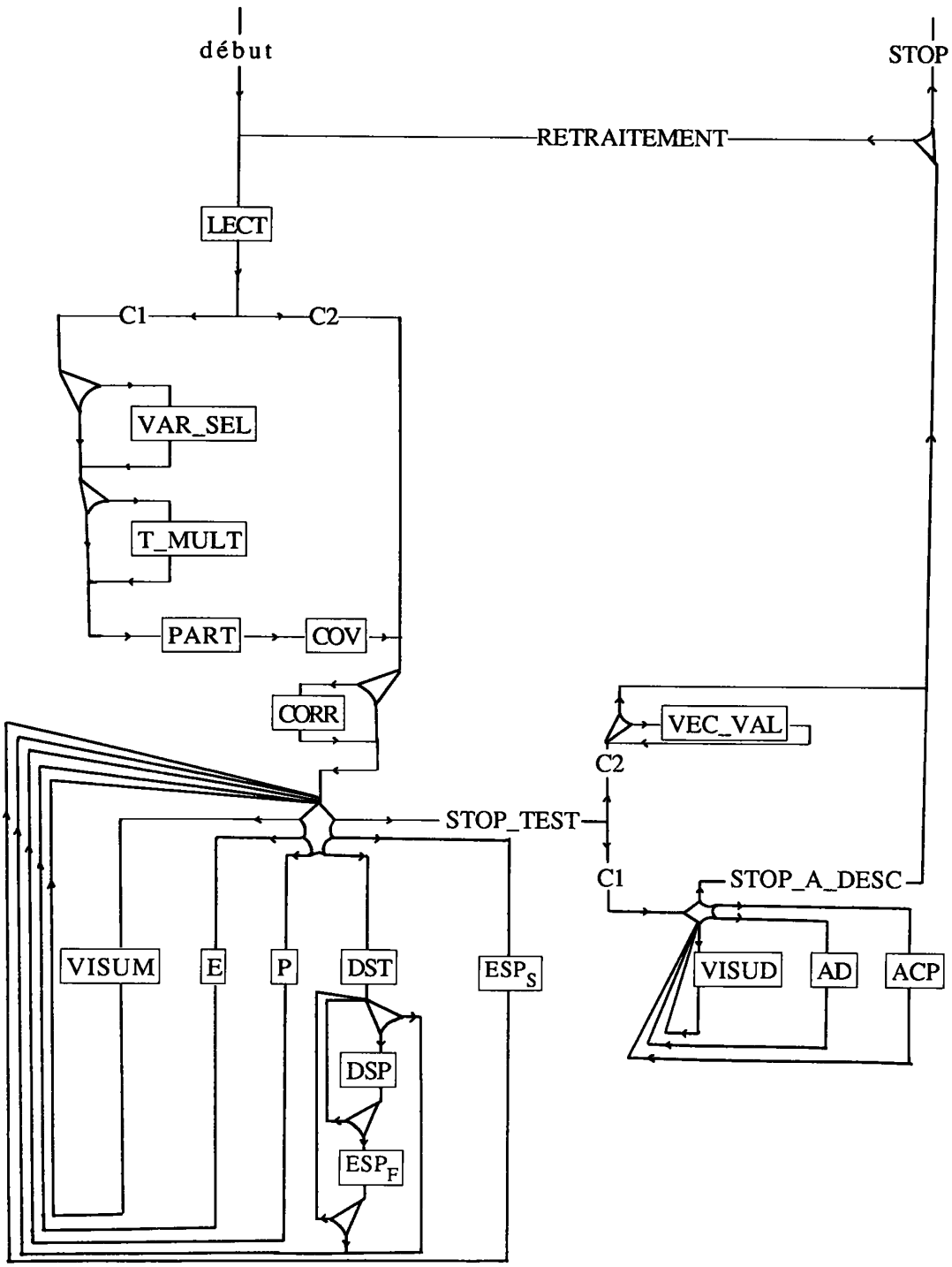
COMPCOV travaille interactivement avec menus, l'utilisateur n'ayant qu'une suite de réponses à donner aux questions successives qui lui sont posées (soit oui ou non, soit une valeur numérique entière, soit des données de tableaux).

Enfin, outre les tests déjà vus dans la première partie de l'article, COMPCOV offre quelques possibilités auxiliaires. Pour la gestion des données, toutes les formes d'entrées sont acceptées (brutes, en fichiers, simulées, ...) et on peut, à tout moment, refaire une étude sur une autre partition de l'ensemble des données disponibles.

Dans la partie des traitements statistiques, des modules d'Analyse en Composantes Principales et d'Analyse Discriminante ont été intégrés pour permettre une étude descriptive préalable des données.

Un mode d'emploi précis et une présentation détaillée, notamment au niveau des algorithmes utilisés, ont été écrits par Viguier (1989). Le schéma suivant présente l'organisation générale du logiciel qui peut être scindé en trois parties : gestion des données,traitement statistique inférentiel et traitement statistique descriptif.

Organisation générale de COMPCOV



Description des noms des modules et des chemins désignés dans cet organigramme

LECT : Lecture des données à traiter; ces données peuvent être entrées:
- au clavier (matrices de covariance et tailles des échantillons, ou données brutes),
- par le simulateur de loi normale de GAUSS,
- par fichier (mêmes possibilités qu'au clavier),
- en les récupérant du traitement précédent par le chemin RETRAITEMENT.

C1 et C2 désignent les types de données entrées : - C1 : matrice (individus x variables),
- C2 : matrices de covariance.

VAR_SEL : fait la sélection des variables à traiter.

T_MULT : test de multinormalité globale des variables sélectionnées.

PART : partitionne l'ensemble des individus.

COV : calcule les matrices de covariance.

CORR : change les matrices de covariance en matrices de corrélation.

VISUM : visualisation des matrices de covariance à traiter.

E : test d'égalité de tout ou partie des matrices de covariance.

P : test de proportionnalité de tout ou partie des matrices de covariance.

DST : test de diagonalisation simultanée totale de tout ou partie des matrices de covariance (l'ordre des vecteurs propres est celui des estimateurs des valeurs propres qui correspondent aux élément diagonnaux de chaque matrice de covariance quasi-diagonalisée).

DSP : test de diagonalisation simultanée partielle, c-à-d. d'égalité de sous-ensembles de vecteurs propres.

$ESP_F(q)$: test d'égalité des sous-espaces propres engendrés par les ensembles de q vecteurs propres choisis dans la DSP.

$ESP_S(m)$: test d'égalité des sous-espaces propres de deux matrices associés aux m plus grandes valeurs propres.

STOP_TEST : arrêt des tests sur les matrices de covariance.

VISUD : visualisation des matrices de données brutes.

AD : analyse discriminante .

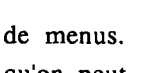
ACP : analyse en composantes principales : visualisation des vecteurs et valeurs propres de la matrice de covariance correspondante, des composantes principales, et représentation graphique des individus sur un plan au choix.

STOP_A_DESC : arrêt analyse descriptive.

VEC_VAL : visualisation des vecteurs et valeurs propres des matrices de covariance.

RETRAITEMENT : on recommence un traitement de données.

STOP : arrêt du programme.

Remarque : les choix sont affichés sous forme de menus. Les symboles  ,  et  sont des "ou", c'est-à-dire qu'on peut choisir l'un ou l'autre des chemins, tandis que les séparations de chemin en  correspondent à des choix conditionnés par les types de données de départ.

III Un exemple d'application

Dans cette partie, on présente un exemple sur des données d'odontologie recueillies par le professeur Charron, de l'Université Paris V, et le docteur Delage (1989) de Toulouse. Il s'agit de mesures dentofaciales d'enfants du même âge. L'échantillon global comporte 380 individus séparés en deux échantillons indépendants : le premier groupe d'enfants polyrégionnaux français non basques comporte 281 individus, le second groupe d'enfants français basques comporte 99 individus. Les cinq variables étudiées, toutes exprimées en millimètres, sont les suivantes: X_1 mesure le surplomb incisif, X_2 la protrusion labiale supérieure dans le profil, X_3 la divergence squelettique de la hauteur faciale, X_4 la hauteur de la branche montante de la mandibule selon Ricketts et X_5 la longueur de la branche horizontale de la mandibule selon Ricketts.

Le tableau suivant donne les moyennes et écart-types partiels et globaux:

		X_1	X_2	X_3	X_4	X_5
Moyenne	1er groupe	4.23	0.96	104.3	37.4	61.1
	2ème groupe	4.04	2.88	107.3	39.2	63.1
	ensemble	4.18	1.46	105.1	37.9	61.6
Ecart-type	1er groupe	3.0	2.34	6.92	4	4.28
	2ème groupe	2.8	3.01	6.65	3.01	4
	ensemble	3	2.66	6.97	3.84	4.29

Les matrices de covariance empiriques sont respectivement :

$$S_1=\begin{pmatrix} 9.18 & -2.24 & 3.9 & 0.34 & 0.03 \\ -2.24 & 5.48 & -0.19 & -0.13 & 1.32 \\ 3.9 & -0.19 & 47.91 & 9.72 & 14.58 \\ 0.34 & -0.13 & 9.72 & 15.97 & 3.3 \\ 0.03 & 1.32 & 14.58 & 3.3 & 18.31 \end{pmatrix}, \quad S_2=\begin{pmatrix} 8.16 & -4.11 & 1.41 & 0.91 & -0.96 \\ -4.11 & 9.06 & -2.58 & -0.55 & 1.62 \\ 1.41 & -2.58 & 44.2 & 5.51 & 11.85 \\ 0.91 & -0.55 & 5.51 & 9.06 & 5.2 \\ -0.96 & 1.62 & 11.85 & 5.2 & 15.97 \end{pmatrix}$$

pour les premier et second échantillons, et

$$S=\begin{pmatrix} 8.9 & -2.79 & 3.13 & 0.42 & -0.3 \\ -2.79 & 7.11 & 0.33 & 0.44 & 2.14 \\ 3.13 & 0.33 & 48.63 & 9.69 & 15.03 \\ 0.42 & 0.44 & 9.69 & 14.79 & 4.5 \\ -0.3 & 2.14 & 15.03 & 4.5 & 18.44 \end{pmatrix}$$ pour l'ensemble des 380 individus.

L'étude de ces données ne fait pas apparaître de différences significatives sur les moyennes des deux groupes, tant du point de vue descriptif (ACP et Analyse Discriminante sur l'échantillon total) que du point de vue inférentiel (avec le test d'homogénéité des moyennes avec matrices de covariance inconnues). On peut donc s'interroger quant au degré d'homogénéité des variances et quant à la possibilité de réduction simultanée des données de chacun des deux groupes.
Les matrices des vecteurs propres de S_1 et S_2 ainsi que les valeurs propres sont:

$$B_1=10^{-2}\begin{pmatrix} 7.5 & -17.1 & 40.7 & -78.7 & -42.5 \\ 0.2 & 16.1 & -20.8 & 34.4 & -90.1 \\ 89.7 & -5.1 & 34.7 & 26.8 & 1.5 \\ 24.4 & -72.16 & -63.4 & -13.1 & -3.2 \\ 36.1 & 64.9 & -51.8 & -41.6 & 7.7 \end{pmatrix} \text{ et } (\lambda_i(S_1))_{i=1,\dots,5}=(56.7,13.8,13.1,8.9,4.3)$$

$$B_2=10^{-2}\begin{pmatrix} 3.2 & -41.6 & -51.0 & -18.7 & 72.8 \\ 5.0 & 53.1 & 45.5 & 21.3 & 68.0 \\ 92.1 & -24.5 & 29.0 & 7.2 & 4.1 \\ 17.0 & 22.6 & -53.9 & 79.2 & -5.0 \\ 34.5 & 65.9 & -39.7 & -53.5 & -5.4 \end{pmatrix} \text{ et } (\lambda_i(S_2))_{i=1,\dots,5}=(49.8,15.2,11.2,5.7,4.4)$$

Les principaux résultats des tests effectués par COMPCOV sur ces données sont présentés dans le tableau suivant:

HYPOTHESE	s=stat. du $\chi 2$ exacte ou corrigée	d.d.l.	$p(\chi 2 \geqslant s)=x$
$\mathcal{E}$	39.28	15	$\leqslant 10^{-3}$
$\mathcal{P}$ (avec $\hat{c}$=0.93)	38.65	14	$\leqslant 10^{-3}$
$\mathcal{DST}$	18.63	10	$4.5.10^{-2}$
$\mathcal{ESP}_S(4)$	6.75	3	8.10^{-2}
$\mathcal{ESP}_S(3)$	19.6	1	$\leqslant 10^{-3}$
$\mathcal{ESP}_S(2)$	1.66	2	0.43
$\mathcal{ESP}_S(1)$	2.25	3	0.52

L'égalité et la proportionnalité sont largement rejetées. La diagonalisation simultanée totale ou l'égalité des sous-espaces propres d'ordre 4 ne sont pas décidées franchement au niveau usuel. Par contre, l'égalité des sous-espaces propres de dimension 1 et de dimension 2 sont des hypothèses largement acceptables. En fait, de l'examen des matrices B_1 et B_2, on peut clairement noter la similitude des premiers vecteurs propres de S_1 et S_2, puis une certaine différence des deuxièmes vecteurs propres et une plus nette différence des troisièmes vecteurs propres, ce qui fait rejeter nettement l'hypothèse $\mathcal{ESP}_S(3)$.

Nous n'avons pas reporté sur ce tableau les résultats des hypothèses $\mathcal{ESP}_F(q)$ car, dû à la proximité de chacune des valeurs propres λ_2 et λ_3 de S_1 et S_2, les matrices quasi-diagonalisées fournies par le FG algorithme n'ont pas leurs valeurs propres rangées dans le même ordre, ce qui perturbe l'étude prise dans un contexte d'Analyses en Composantes Principales conjointes. On peut donc ici ne pas rejeter l'hypothèse d'un premier sous-espace principal commun et, dans ce contexte, les deux nuages des individus se projettent sur leur premier plan principal de façon très similaire avec des qualités respectives globales presque identiques (72% pour le groupe 1, 75% pour le groupe 2) alors que leur "non-homogénéité" n'apparaît que sur les axes qui ont le moins d'importance.

IV Remarques

IV 1. Un des intérêts du logiciel COMPCOV est de pouvoir disposer de l'ensemble des tests pris séquentiellement. Par exemple, Schott (1988) a proposé une comparaison de deux matrices de covariance et n'a pas rejeté l'existence d'un sous-espace propre commun de dimension 3. En

traitant ces données avec COMPCOV, il est rapide de constater que ni l'hypothèse d'égalité, ni celle de proportionnalité ne sont refusées et donc, a fortiori, les autres hypothèses non plus. La hiérarchisation des différentes hypothèses et la décomposition du "chi-deux total", associé à la statistique E, en différents chi-deux successifs, associés aux autres hypothèses emboîtées, est une présentation intéressante tant théoriquement que concrètement : voir Flury (1988), chap.7. Elle peut être représentée sur le schéma suivant:

$$\mathcal{E} \subset \mathcal{P} \subset \mathcal{DST} = \mathcal{DSP}(p\text{-}1) \subset \mathcal{DSP}(i)_{i=1,\ldots,p-2} \subset \mathcal{DSP}(k)_{k=1,\ldots,i} = \bigcap_{j=1}^{k} \mathcal{ESP}_{\mathcal{F}}(j) \subset$$

$$\mathcal{ESP}_{\mathcal{F}}(i)_{i=1,\ldots,k} \not\supset \mathcal{ESP}_{\mathcal{F}}(j)_{j\neq i}$$

et $\mathcal{E} \subset \mathcal{P} \subset \bigcap_{i=1}^{j} \mathcal{ESP}_{\mathcal{S}}(i)_{,j=1,\ldots,p-1}$ = égalité des j premiers vecteurs propres 2 à 2 pris dans le même ordre des valeurs propres $\subset \mathcal{ESP}_{\mathcal{S}}(i)_{i=1,\ldots,j} \not\supset \mathcal{ESP}_{\mathcal{S}}(h)_{h\neq j}$.

On pourrait, selon le schéma des emboîtements ci-dessus qui inclut les tests de Flury, décomposer la statistique du chi-deux pour tester l'égalité des matrices comme suit:

$$\chi^2_{\mathcal{E}} = \chi^2_{\mathcal{DSP}(1)} + (\chi^2_{\mathcal{DSP}(2)} - \chi^2_{\mathcal{DSP}(1)}) + \ldots + (\chi^2_{\mathcal{P}} - \chi^2_{\mathcal{DST}}) + (\chi^2_{\mathcal{E}} - \chi^2_{\mathcal{P}})$$

où $\chi^2_{\mathcal{H}}$ est la statistique du $\chi 2$ de test de $\mathcal{H}$ contre non $\mathcal{H}$, et $\chi^2_{\mathcal{H}_1} - \chi^2_{\mathcal{H}_2}$ est la statistique de test de $\mathcal{H}_1$ contre $\mathcal{H}_2$, sachant $\mathcal{H}_2$ acceptable.

Cette manière de décomposer les tests est la méthode descendante, qui n'est pas exhaustive car il n'y a pas emboîtement systématique dans les hypothèses $\mathcal{ESP}(i)$. Dans notre logiciel, nous avons donc adopté la méthode ascendante, en ne faisant que des tests d'hypothèses du type $\mathcal{H}$ contre non $\mathcal{H}$.

Par ailleurs, concernant les procédures de tests simultanés, il est intéressant ici de citer les travaux de Gabriel (1969), Holm (1979), Hommel (1988) et Edwards et Havranek (1987).

IV.2. Dans COMPCOV, on peut aisément supprimer des variables, des individus, des groupes d'individus et faire l'ensemble des tests sur la totalité des individus ou sur certains des k groupes. Pour donner une idée sur les différences de moyennes et sur la répartition des individus nous avons intégré la visualisation par Analyse en Composantes Principales et par Analyse Discriminante des groupes d'individus sur plan. Cette étude descriptive sera avantageusement complétée dans une version prochaine du logiciel par l'implémentation de tests d'homogénéité des moyennes.

IV.3. Les tests proposés dans COMPCOV reposent sur la comparaison de matrices de covariance. Dans une optique d'Analyse en Composantes Principales, on est souvent amené à travailler sur les matrices des corrélations. Les récents travaux de Larntz et Perlman (1986), Manly et Rayner (1987) et de Millsap et Meredith (1989) semblent prometteurs dans ce domaine.

IV.4. Les méthodes exposées ici nécessitent l'hypothèse de normalité, et leur manque de robustesse est bien établi dans certains cas. Des travaux comme ceux d'Anderson et al. (1986) donnent des résultats intéressants pour chercher à étendre ces travaux à certaines familles elliptiques.

IV.5. L'un des aspects les plus intéressants des études de Krzanowski et de Flury a été d'élargir le contexte de l'Analyse en Composantes Principales au cas de k échantillons. Cette idée peut se répercuter sur les différents types d'ACP rencontrés dans la littérature, sur les autres analyses factorielles ou dans les analyses de type en facteurs communs et spécifiques (cf. le tout récent article de Chen et Robinson (1989) dans ce dernier cas).

Les auteurs remercient H. Caussinus, J. Dauxois et les rapporteurs pour leurs remarques et commentaires.

BIBLIOGRAPHIE

T.W. ANDERSON (1984) : *An Introduction to Multivariate Statistical Analysis.* 2nd edition, J.Wiley, New-York.

T.W. ANDERSON, K.T. FANG, H. HSU (1986) : *Maximum Likelihood Estimates and Likelihood Ratio Criteria for Multivariate Elliptically Contoured Distributions.* Canadian Jour. of Statistics vol.14, n°1, *pp* 55-59

O.E. BARNDORFF-NIELSEN, P. HALL (1988) : *On the level-error after Bartlett adjustment of the likelihood ratio statistic.* Biometrika vol.75, n°2, *pp*374-378

K.H. CHEN, J. ROBINSON (1989), : *Comparison of Factor Spaces of Two Related Populations* Jour. Mult. Analysis, vol 28, *pp* 190-203

D.B. CLARKSON (1988) : *A Remark on Algorithm AS211 : the FG-Diagonalization Algorithm.* Applied Statistics, vol 37, *pp* 147-151

L.C.A. CORSTEN, K.R. GABRIEL (1976) : *Graphical Exploration in Comparing Variance Matrices.* Biometrics, vol.32, *pp*851-863

D.R. COX, N.J.H. SMALL (1978) : *Testing Multivariate Normality.* Biometrika , vol 65, *pp* 263-272

H. DELAGE (1989) : *Etude céphalométrique et dentaire d'un échantillon d'enfants dysmorphiques basques en âge orthodontique. Comparaison avec une série équivalente d'origine polyrégionale.* Thèse d'Etat en Odontologie, Paris V, à paraître

D. EDWARDS, T. HAVRANEK (1987) : *A Fast Model Selection Procedure for Large Families of Models.* Jour. Amer. Stat. Assoc., vol.82, *pp*205-213

P.S. ERIKSEN (1987) : *Proportionality of Covariance Matrices.* Annals of Stat., vol 15, n°2 *pp* 732-748

B.N. FLURY (1983) : *Some Relations Between the Comparison of Covariance Matrices and Principal Component Analysis.* Comp. Statist. and Data Anal., n°1, *pp* 97-109

B.N. FLURY (1984) : *Common Principal Components in k Groups.* J.Amer.Stat.Assoc., vol. 79, n° 388, *pp* 892-898

B.N. FLURY (1986a) : *Asymptotic Theory for Common Principal Component Analysis.* Ann. Math. Statist. vol 14, *pp* 418-430

B.N. FLURY (1986b) : *Proportionality of k Covariance Matrices.* Stat. and Prob. Letters. vol 4 *pp* 29-33

B.N. FLURY (1987) : *Two Generalizations of the Common Principal Component Model.* Biometrika, vol 74, n°1, *pp* 59-69

B.N. FLURY (1988) : *Common Principal Components and Related Multivariate Models.* J.Wiley, New-York.

B.N. FLURY, G. CONSTANTINE (1985): *Algorithm AS211 : The FG-diagonalization Algorithm.* Applied Statistics, vol 34, *pp* 177-183

B.N. FLURY, W. GAUTSCHI (1986) : *An Algorithm for Simultaneous Orthogonal Transformation of Several Positive Definite Symetric Matrices to nearly Diagonal form.* SIAM. Jour. Sci. Statis. Comput. vol.7, *pp* 169-184.

K.R. GABRIEL (1969) : *Simultaneous Test Procedures - Some Theory of Multiple Comparisons.* Ann. Math. Statist., vol. 40, *pp*224-250

S. HOLM (1979) *: A Simple'Sequentially Rejective Multiple Test Procedure.* Scand. J. Statist., vol.6, *pp*65-70

G. HOMMEL (1988) : *A stagewise rejective multiple test procedure based on a modified Bonferroni test.* Biometrika, vol.75,n°2, *pp*383-386

U.C. JAISWAL, J.P. JAIN (1989) : *An Approximate Method for Assessing Multivariate Normality*, Jour. Ind. Soc. Ag. Statist. vol.XL, n°2, *pp* 164-168

S.T.JENSEN, S. JOHANSEN (1987) : *Estimation of Proportional Covariances.* Stat. and Prob. Letters, 6, *pp* 83-85

B.P. KORIN (1969) : *On testing the equality of k covariance matrices.* Biometrika, vol.56, *pp*216-218

H.A.L. KIERS, J.M.F. TEN BERGE (1989) : *Alternating Least Squares for Simultaneous Components Analysis with equal Component Weight Matrices in two or more Populations.* Psychometrika, vol.54, n°3, *pp*515-521

W.J. KRZANOWSKI (1979) : *Between-group Comparison of Principal Components*. Jour. Amer. Stat. Assoc. vol. 74, n°367, *pp* 703-707

W.J. KRZANOWSKI (1984) : *Principal Component Analysis in the presence of Group Structure*. Applied Statistics, vol.33, n°2, *pp* 164-168

K. LARNTZ, M.D.PERLMAN (1986) : *A simple Test for the Equality of Correlation Matrices*. Stat. Decision Th. and Rel. Topics IV, n°2, *pp*289-298

J. LEVIN (1966) : *Simultaneous Factor Analysis of Several Gaussian Matrices*. Psychometrika vol.31, *pp* 413-420

B.F.J. MANLY, J.C.W. RAYNER (1987) : *The Comparison of Sample Covariance Matrices using Likelihood Ratio Tests*. Biometrika, vol.74, n°4, *pp* 841-847

R.E. MILLSAP, W. MEREDITH (1989) : *Component Analysis in Cross-Sectionnal and Longitudinal Data*. Psychometrika, à paraître

J. MØLLER (1986) : *Bartlett adjustments for Structured Covariances*. Scand. Jour. Stat., vol.13, *pp* 1-15

R.J. MUIRHEAD (1982) : *Aspects of Multivariate Statistical Theory*. J.Wiley, New-York.

A.M. PARHIZGARI, A.J.PRAKASH (1989) : *Test of the Equality of Dispersion Matrices, Algorithm AS 250*. Applied Statistics, *pp*553-564

M.D. PERLMAN (1980) : *Unbiaisedness of the Likelihood Ratio Tests for Equality of Several Covariance Matrices and Equality of Several Multivariate Normal Populations*. Ann. Stat., vol.8, n°2, *pp* 247-263

J.P.ROYSTON (1983) : *Some Techniques for assessing Multivariate Normality based on Shapiro-Wilk W*. Applied Statistics, vol.32, *pp* 121-133

J.R. SCHOTT (1988) : *Common Principal Components Subspaces in two Groups*. Biometrika , vol.75, n°2, *pp* 229-236

P.K. SEN, M.L. PURI (1968) : *On a Class of Multivariate Multisample Rank Order Tests II. Tests for Homogeneity of Dispersion Matrices*. Sankhya, vol.30, *pp*1-22

S. VIGUIER (1989) : *Notice d'utilisation du logiciel COMPCOV*. Note interne, Labo. Stat. et Prob., UPS Toulouse.