QUE FAIT LE PROGRAMME STATIS? SUR QUELLES DONNEES L'UTILISER?

Christine LAVIT
unité de biométrie (INRA-ENSA.M-USTL)
place Viala 34060 MONTPELLIER
tél : 67 61 24 28 - 67 61 24 22

Résumé :

Lorsque les données se présentent sous la forme de plusieurs tableaux de mesures recueillies sur les mêmes individus dans des situations différentes, la méthode STATIS, basée sur le principe de l'Analyse en Composantes Principales, permet de répondre aux objectifs suivants :

- déceler quels sont les tableaux qui se ressemblent,
- fournir un tableau résumé de l'ensemble,
- décrire les différences entre tableaux par rapport à ce tableau résumé : sont-elles dues aux individus ou aux variables?

Pour mettre en oeuvre cette méthode, nous vous proposons les programmes source en Fortran 77 dans la bibliothèque MODULAD, ou une version exécutable sur micro-ordinateur du type IBM PC distribuée par l'unité de biométrie.

1 METHODES PERMETTANT L'ANALYSE CONJOINTE DE TABLEAUX

Pour explorer des données qui se présentent sous la forme d'un tableau contenant les valeurs de caractères (ou variables) prises par un ensemble d'individus, le chercheur dispose de techniques d'analyse de données, comme l'Analyse en Composantes Principales (ACP), développées il y a une vingtaine d'années, et largement utilisées depuis.

Mais il arrive que le protocole de l'expérimentation conduise à considérer que les données recueillies forment, en fait, un ensemble de tableaux (une succession de tableaux lorsqu'ils sont mesurés à des dates différentes, par exemple).

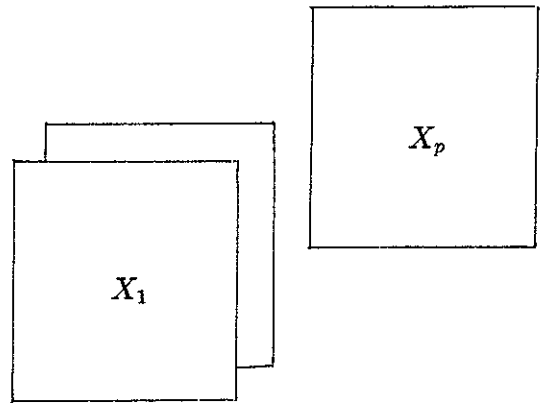
Les individus étant les mêmes, on peut juxtaposer les tableaux et faire l'ACP du grand tableau ainsi obtenu. Mais cette démarche n'est pas satisfaisante car les variables sont le plus souvent fortement autocorrélées d'un tableau à l'autre, et on ne peut pas suivre l'évolution des individus. Si ce sont les mêmes variables qui sont mesurées sur ces individus, on peut également empiler les tableaux pour en faire l'ACP, mais alors si on peut suivre l'évolution des individus, les variables n'interviennent que par leurs corrélations *intra*.

Différentes techniques d'analyse conjointe de tableaux de données sont actuellement proposées. Certaines cherchent à reconstruire le *cube* de données à partir d'un petit nombre d'individus types, de variables latentes et de conditions types. D'autres supposent l'existence d'un modèle sous-jacent, ou d'une structure commune aux tableaux. D'autres enfin tiennent compte de la structure ordonnée du temps pour analyser une succession de tableaux (consulter GLACON 1981, ESCOUFIER et al. 1985, SEMPE et al. 1987, CARLIER et al. 1988 in analysis of multiway data matrices, et KIERS 1989 sur les comparaisons entre méthodes).

Peu de programmes sont diffusés à l'heure actuelle : INDSCAL, TUCKER, CANDECOMP-PARAFAC ... dans les logiciels de langue anglaise (consulter la monographie de KROONENBERG 1983 à ce sujet, et l'ouvrage édité par LAW et al. 1984); l'Analyse Factorielle Multiple, l'Analyse Canonique Généralisée, STATIS ... dans les logiciels français.

2 PRINCIPE DE LA METHODE STATIS

La méthode STATIS est une méthode exploratoire d'Analyse des Données, qui s'applique à des données quantitatives : p tableaux de mesures X_k ont été observés en différentes occasions sur les mêmes individus.

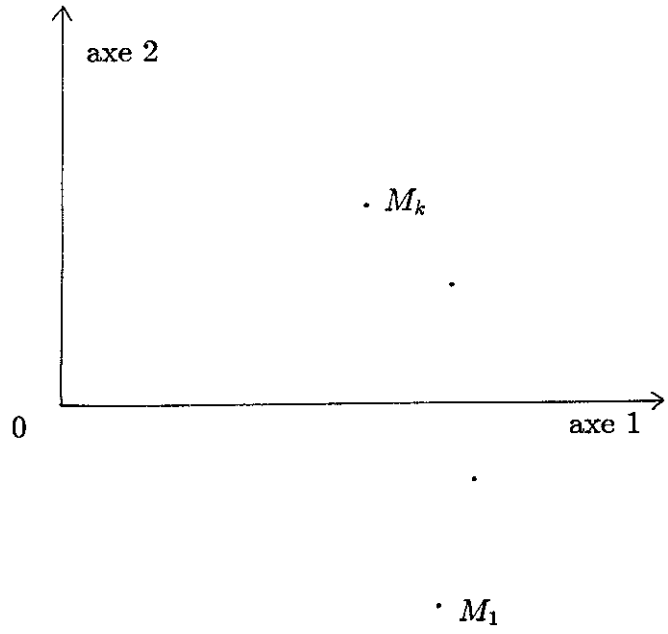


L'idée essentielle de la méthode est la recherche d'une structure commune aux tableaux, qu'on appelle *intrastructure*. Pour le tableau X_k, cette structure est décrite par les distances mutuelles entre individus, déduites du tableau de produits scalaires $W_k = X_k X_k'$.

2.1 Analyse globale des relations entre tableaux

On compare les tableaux au moyen des *objets* W_k. Par opposition au terme *intrastructure* qui décrit la structure des individus à l'intérieur d'un tableau, on appelle *interstructure* les relations entre tableaux, décrites par les distances entre W_k. Ces distances sont déduites du produit scalaire de Hilbert-Schmidt entre applications linéaires.

A partir de ces produits scalaires, on construit une image euclidienne plane des tableaux. Soit $M_1, \ldots M_p$ le nuage des points-tableaux :



Sur ce graphique, le cosinus de l'angle entre les vecteurs OM_k et OM_l est l'approximation du produit scalaire normé entre W_k et W_l, appelé coefficient RV. Un coefficient RV proche de 1 signifie qu'on a la même structure des individus à l'intérieur des tableaux X_k et X_l, et que par conséquent les positions mutuelles des individus sont les mêmes dans les conditions k et l.

2.2 Positions compromis des individus

A partir de l'image euclidienne des tableaux, on construit un *objet* compromis W en prenant la moyenne des W_k pondérés par les coordonnées des points-tableaux sur le premier axe. W peut être considéré comme un tableau de produits scalaires *moyens* entre individus. L'image euclidienne des individus, associée à ces produits scalaires, représente les positions mutuelles *moyennes* des individus.

Lorsque les distances entre les *objets* W_k définies dans l'interstructure sont faibles, on peut affirmer qu'il existe bien une structure des individus, commune aux tableaux. Cette structure est alors décrite par les distances compromis entre individus.

On peut expliquer les positions compromis des individus par les variables en considérant que les coordonnées des individus sur un axe sont les valeurs d'une variable fictive, appelée *composante principale*. Pour interpréter les positions des individus le long de l'axe, on calcule les corrélations de la composante principale avec les variables des différents tableaux, ou avec des variables exogènes dont on connaîtrait les valeurs sur les individus.

2.3 Trajectoires des individus

Dans l'image euclidienne compromis des individus, on trace la trajectoire de chaque individu, en utilisant la technique des points supplémentaires. L'interstructure a mis en évidence, sans les expliquer, les écarts entre tableaux. Les trajectoires permettent de déceler quels sont les individus responsables de ces écarts.

Lorsque sur le graphique des corrélations des variables avec les axes du compromis, les points se regroupent nettement par variable, on peut donner un nom aux axes et interpréter le sens de parcours des trajectoires. Ce cas est fréquent, car les variables sont souvent fortement autocorrélées dans les différentes études.

Enfin, il faut noter que lors de l'analyse d'un phénomène évolutif, la méthode donnerait les mêmes résultats si on intervertissait l'ordre des tableaux. Par conséquent, lorsque les tableaux sont indicés par le temps, ce n'est qu'implicitement que la structure ordonnée du temps intervient dans l'interprétation des trajectoires.

3 QUELQUES PRECISIONS POUR BIEN UTILISER STATIS

Dans vos données, est-ce que les corrélations entre colonnes ont un sens ?

La méthode STATIS est basée sur le principe de l'Analyse en Composantes Principales, et sur les notions de moyenne et de corrélations entre variables : les ressemblances entre tableaux s'interprètent en termes de corrélations, les trajectoires décrivent l'évolution de chaque individu par rapport à l'évolution de l'individu *moyen*. Il faut donc que ces notions aient un sens.

La méthode a été conçue pour des données quantitatives, mais on peut à la rigueur l'appliquer à des données qualitatives *en échelle* . L'interprétation se fait alors en termes de corrélations de rang.

Les tableaux portent-ils sur les mêmes variables?

Puisque on travaille sur les *objets* W_k, les variables peuvent être différentes d'un tableau à l'autre. Cela permet de conserver l'ensemble des données même si une variable est manquante dans un tableau.

Tous les individus sont-ils présents dans tous les tableaux?

Chaque tableau porte sur les mêmes individus. Mais si un individu n'est pas présent dans tous les tableaux, on peut toutefois effectuer l'analyse en le traitant en individu *supplémentaire*, et tracer la partie de la trajectoire qui le concerne.

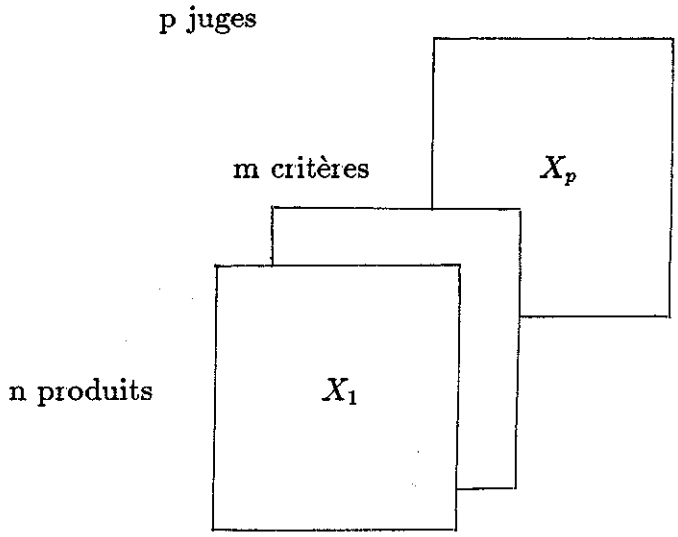
Trop de tableaux?

Si le nombre de tableaux est important, le nuage de points- tableaux peut faire apparaître une structure : par exemple deux sous-nuages distincts, ou un tableau *orthogonal* aux autres. Alors le compromis n'est pas un bon résumé de l'ensemble, et les trajectoires des individus ne sont pas interprétables. Il est préférable de recommencer l'analyse sur un ensemble homogène de tableaux, ou de mettre le tableau *génant* en élément supplémentaire. Les résultats seront plus clairs.

A partir des coordonnées des points-tableaux, on peut également chercher l'existence de groupes à l'aide d'un programme de classification.

4 EXEMPLES D'UTILISATION

4.1 Chaque tableau correspond à l'opinion d'un juge



L'interstructure permet de repérer quels sont les juges dont les réponses sont voisines, et quels sont ceux dont les opinions divergent. Le compromis fournit une répartition des produits qui résume l'opinion *majoritaire* qui ressort de l'enquête. Enfin, l'examen des trajectoires, qu'il vaudrait mieux dans ce cas nommer *écarts* entre les juges, donne une idée des produits qui ont fait l'unanimité, et de ceux pour lesquels il y a désaccord entre les juges.

4.2 Chaque tableau correspond à la hauteur de prélévement du latex sur le tronc d'un hévéa

Objectif de l'étude[1]

Le caoutchouc naturel est extrait du latex, qui est un cytoplasme cellulaire situé en particulier dans l'écorce de l'hévéa. Le latex est prélevé par saignée, c'est-à-dire par écoulement le long d'une encoche découpée dans l'écorce de l'arbre. Une bonne compréhension des mécanismes biologiques permet d'optimiser la

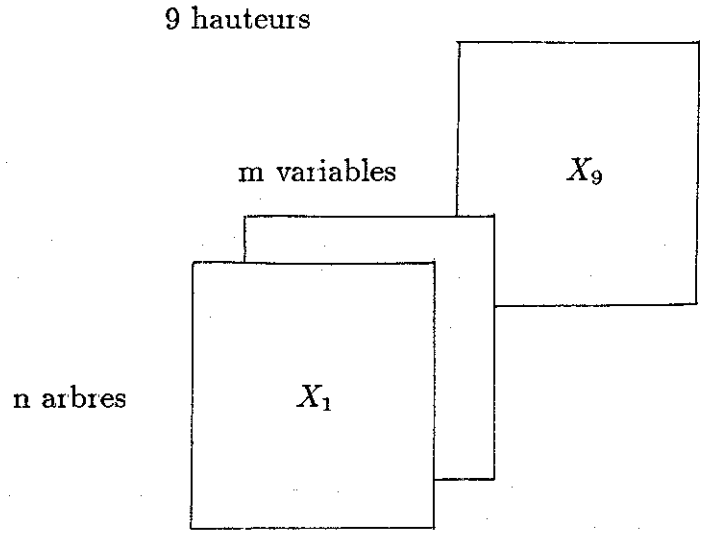
[1] expérimentation réalisée par l'Institut de Recherche sur le CAoutchouc, CIRAD

production sans épuiser les réserves de l'arbre, ce qui pourrait nuire à sa rentabilité future. On sait, par exemple, que la fréquence de saignée et la stimulation de l'arbre par application de produits chimiques sur l'encoche, influencent la quantité de latex produite.

Description des données

L'expérimentation que nous allons décrire a pour but d'étudier l'évolution de la composition du latex lorsqu'il est prélevé à différentes hauteurs sur le tronc de l'arbre, et également l'influence d'un traitement consistant en une greffe de couronne sur cette évolution.

On a donc considéré un lot d'arbres témoins, et un lot d'arbres dont la couronne foliaire résulte d'une greffe. Le latex a été prélevé à neuf hauteurs différentes s'échelonnant le long du tronc. Le premier prélèvement est effectué sous l'encoche de saignée, tandis que les trois derniers sont situés, dans le cas des arbres greffés, au dessus de la greffe. Sur le latex récolté, on a mesuré les taux de magnésium, de phosphore, de thiols, de glucides et d'extrait sec. Les valeurs de ces variables reflètent l'état physiologique de l'arbre.



Résultats

Lorsqu'on applique la méthode STATIS à ces données, un premier résultat global met en évidence les écarts entre tableaux, c'est-à-dire entre hauteurs de prélèvement. En particulier, les prélèvements situés autour de l'encoche qui correspondent à des zones très activées, se distinguent des autres.

Sur un deuxième graphique, chaque arbre est représenté par un point *compromis*. Les arbres greffés se séparent nettement des arbres témoins, ce qui signifie que l'influence de la greffe semble se faire sentir de façon globale sur l'arbre.

Trajectoires de points moyens

La dernière étape de la méthode trace la trajectoire de chaque arbre, c'est-à-dire situe les différentes positions de l'arbre décrit par chaque tableau. Ces trajectoires individuelles ne nous intéressent pas particuliérement. Par contre, on peut visualiser la différence entre les deux traitements en créant un arbre *moyen* fictif pour chaque traitement, et en traçant la trajectoire de ces deux arbres supplémentaires. Les deux trajectoires se séparent nettement, et montrent que l'évolution de la composition physiologique du latex le long du tronc diffère suivant le traitement. De plus, le sens de parcours des trajectoires s'interprète en fonction de la teneur en extrait sec et en glucides, et donne une explication de la différence entre les traitements.

4.3 Chaque tableau correspond à un examen médical

Le logiciel STATIS permet de trier les individus selon un critère qualitatif, et de ne tracer que les trajectoires du groupe d'individus ainsi sélectionnés. Au vu des trajectoires, on peut caractériser ce groupe, mais aussi le comparer à un autre groupe puisque les graphiques sont réalisés avec la même échelle.

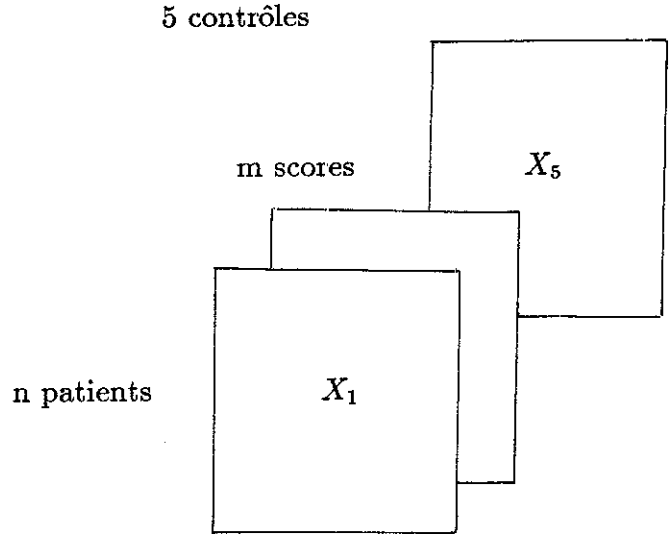
Prenons comme exemple le fichier proposé par l'Association pour la Méthodologie de la Recherche en Psychiatrie, comme base commune pour la confrontation de différentes méthodes d'analyse, aux Journées de Statistique 1985 et 1986.

Description des données

Le fichier décrit le suivi, sur huit mois, de 230 patients atteints de dépression nerveuse. Pendant la première période s'étalant sur deux mois, ces patients sont traités par un anti-dépresseur. Pendant la deuxième période de six mois, les patients sont divisés en trois groupes. Le premier groupe continue à être traité par l'anti-dépresseur pendant toute la période, le second groupe est traité par l'anti-dépresseur pendant deux mois puis reçoit un placébo, le troisième groupe reçoit un placébo pendant toute la période.

Un contrôle est effectué au début et à la fin de la première période, et tous les deux mois pour la deuxième période. Les données recueillies sont de plusieurs types. La situation familiale, professionnelle, sociale, psychologique du patient

est décrite par un certain nombre de variables qualitatives. Lors de chaque contrôle, la dépression est évaluée à l'aide d'une vingtaine de scores. Enfin, on dispose des réponses de chaque patient à un questionnaire sur les événements intercurrents, survenus entre deux contrôles.



Tri des trajectoires

Dans cet exemple, le sens de parcours de la trajectoire d'un patient renseigne sur l'évolution de sa dépression. Une trajectoire dirigée vers le bas et la gauche du graphique reflète une aggravation de la dépression. A l'opposé, une trajectoire évoluant vers le haut et la droite du graphique indique une amélioration de l'état du patient. Le nombre de patients étant très élevé, la sélection des trajectoires selon un ou plusieurs critères qualitatifs s'avère très utile pour infirmer ou confirmer des hypothèses.

On peut s'intéresser, par exemple, à la comparaison des trois traitements, sur les femmes entre 40 et 50 ans, souffrant d'un type particulier de dépression. On peut juger de l'impact de certains événements intercurrents sur l'état du patient, etc.

Contrairement à l'exemple précédent, nous connaissons un grand nombre d'informations sur chaque patient. La trajectoire moyenne du groupe des femmes de 40 à 50 ans, souffrant d'une dépression réactionnelle apporterait peu d'informations. Par contre, l'examen des trajectoires individuelles de ces patientes donne une idée de la disparité des évolutions, et incite à en rechercher l'explication au moyen d'autres critères.

Suivi des patients sortis de l'étude prématurément

Les patients qui n'ont pas subi les derniers contrôles sont traités en éléments supplémentaires. Le logiciel offre la possibilité de ne tracer que la partie connue de la trajectoire, qui s'interprète avec les mêmes règles que la trajectoire des autres individus. On peut ensuite, de la même manière que précédemment, trier les patients suivant la raison de sortie de l'étude : refus d'un traitement prolongé, effets secondaires nécessitant l'arrêt du traitement, patients perdus de vue, résultat thérapeutique jugé insuffisant.

5 METHODE "STATIS DUALE"

Lorsqu'on observe les mêmes variables mesurées sur des groupes d'individus différents, la méthode STATIS permet de rechercher une structure commune aux tableaux, décrite cette fois par les matrices $V_k = X_k' \, D \, X_k$ de covariances (ou de corrélations) entre variables.

Comme dans le cas précédent, l'interstructure fournit les relations entre tableaux décrites par les distances entre V_k.

Le compromis est une matrice de covariances (ou de corrélations) entre les variables. Les trajectoires permettent de déceler quelles sont les variables responsables des écarts entre V_k. C'est-à-dire celles dont les covariances (ou les corrélations) changent fortement d'un tableau à l'autre.

Les individus peuvent être différents d'un tableau à l'autre mais malgré cela il se peut que la notion d'individus *moyens* ait un sens. Dans ce cas, on obtient une information supplémentaire en projetant les trajectoires de ces individus moyens sur le plan du compromis.

Les résultats sont plus délicats à interpréter que dans le cas précédent, et n'apportent de l'information que sur les liens entre variables.

6 LOGICIEL

Pour mettre en oeuvre la méthode STATIS, nous vous proposons les programmes source dans MODULAD, ou une version exécutable sur micro-ordinateur du type IBM PC.

Ce logiciel n'est pas conversationnel. En effet, nous avons préféré la création d'un fichier de paramètres à la solution des questions posées à l'écran, car le nombre d'informations à fournir est important. Mais nous avons simplifié au

maximum la saisie de ces informations : les paramètres sont définis à l'aide de mots-clés sans ordre, les options standard sont prises par défaut, et nous fournissons un fichier de paramètres modèle pour un jeu d'essai.

Actuellement, les graphiques sont assez grossiers, car fournis sous la forme de nuages de points tracés sur imprimante. Mais il vous sera possible de récupérer à chaque fois le fichier des coordonnées des points, afin d'utiliser un outil graphique plus performant.

STATIS est formé de trois programmes qu'on lance successivement. Ces trois programmes correspondent aux étapes 2.1., 2.2. et 2.3. décrites dans le paragraphe précédent.

Dans la version exécutable, la première disquette contient les programmes exécutables notés INTER.EXE pour l'analyse globale, INTRA.EXE pour le compromis et TRAJECT.EXE pour le tracé des trajectoires (pour un jeu de données dont les dimensions globales sont raisonnables : 150 individus, 10 variables, 10 tableaux, ou 120 individus, 20 tableaux, 30 variables par exemple), ainsi qu'un exemple test. La deuxième disquette contient les notices d'utilisation des trois programmes, ainsi que les sorties pour le jeu test, et la troisième disquette contient les programmes source.

Le nombre de tableaux n'est pas limité mais si le volume des données est important, il se peut que les dimensions initialisées dans le programme exécutable soient insuffisantes. Le programme s'arrête avec un message d'erreur précisant les dimensions requises. Dans ce cas, il est nécessaire d'utiliser le programme source en remplissant correctement l'instruction *parameter* du programme principal qui gère l'allocation dynamique de la mémoire. Le programme source peut être compilé avec la version 4.10 de Microsoft, par exemple.

Certaines erreurs sont diagnostiquées par le programme et provoquent l'arrêt du traitement en affichant un message d'erreur. Le plus souvent, les erreurs diagnostiquées par le système proviennent d'enregistrements non conformes dans le fichier des paramètres.

Dans le fichier ASCII des données, les tableaux doivent être à la suite les uns des autres. A l'intérieur d'un tableau, chaque enregistrement doit contenir l'identificateur (alphanumérique) et les valeurs des variables d'un individu. Le programme n'accepte pas qu'une variable soit constante ou nulle sur un tableau.

REFERENCES BIBLIOGRAPHIQUES

Sur les fondements mathématiques de la méthode STATIS, consulter ESCOUFIER 1977, ESCOUFIER 1980, LAVIT 1988. Des indications sur les calculs effectués sont données en commentaires dans le programme source. Des exemples détaillés sont traités dans ESCOUFIER et al. 1985, SEMPE et al. 1987, LAVIT 1988. Sur les autres méthodes d'analyse de données à trois modes, consulter KROONENBERG 1983, LAW et al. 1984 et COPPI et al. 1989.

COPPI R., BOLASCO S. (editors) 1989 - Analysis of multiway data matrices. Elsevier.

ESCOUFIER Y. , 1977 - Operators related to a data matrix. Recent developments in statistics, North Holland, p.125-131.

ESCOUFIER Y. , 1980 - L'analyse conjointe de plusieurs matrices de données. Biométrie et temps, p.59-76.

GLACON F. , 1981 - Analyse conjointe de plusieurs matrices de données : comparaison de différentes méthodes. Thèse de troisième cycle, Grenoble.

KIERS H. , 1989 - Comparison of anglo-saxon and french three mode methods. A paraître dans Statistique et Analyse de Données.

KROONENBERG P. , 1983 - Three mode principal component analysis. DSWO, Leiden, Pays-Bas.

LAVIT C. , 1988 - Analyse conjointe de tableaux quantitatifs. Masson.

LAW H.G., SNYDER Jr., HATTIE J.A., Mc DONALD R.P. (editors) 1984 - Research methods for multimode data analysis, Praeger, New-York.

SEMPE M. et al. , 1987 - Multivariate and longitudinal data on growing children. Data analysis : the ins and outs of solving real problems, Plenum, London.