ACADEMIE DE MONTPELLIER

UNIVERSITÉ MONTPELLIER II

-SCIENCES ET TECHNIQUES DU LANGUEDOC-

THÈSE

Présentée à l'Université Montpellier II - Sciences et Techniques du Languedoc
pour obtenir le **DIPLÔME DE DOCTORAT**

SPÉCIALITÉ : MATHÉMATIQUES ET APPLICATIONS
FORMATION DOCTORALE : BIOSTATISTIQUES

FACTEURS À MESURES RÉPÉTÉES ET ANALYSES FACTORIELLES : APPLICATIONS À UN SUIVI ÉPIDÉMIOLOGIQUE

par

Didier LEIBOVICI

Soutenue le 33 octobre 1993 devant le Jury composé de :

MM.	ESCOUFIER Yves	Professeur, U.M.II,	Président
	SABATIER Robert	Maîtrede Conférences, U.M.I,	Éxaminateur
	FRANC Alain	CEMAGREF,	Éxaminateur
	DAURÈS Jean-Pierre	Professeur, U.M.I,	Éxaminateur
	ASSELAIN Bernard	INSERM,	Rapporteur
	CHESSEL Daniel	CNRS,	Rapporteur
	SCHWARTZ Daniel	INSERM	Invité

Remerciements

La thèse a été reconstituée à partir d'une serie de fichiers que j'avais sur une sur une archive .sea disquette Mac et je n'y avais pas mis les remerciements … dommage ils étaient assez sympa d'autant que je me rappelle !

Quelques autres problèmes sont survenus lors de ce transfert ; à part les bas de pages et en-têtes, qq figures sont manquantes (n'étant pas à l'origine sous forme électronique) et par ci par là qq coquilles … en plus des coquilles originales !

(ajout Nov2004)

Résumé

On s'intéresse à l'apport de l'analyse factorielle dans le traitement de mesures répétées sur des groupes d'individus.

Une première partie traite, sur la base de modèles linéaires inférentiels classiques, d'ACP particulières permettant une description géométrique des écarts aux hypothèses nulles faites dans ces modèles. Ces modèles couvrent les deux approches traditionnelles : univariée et multivariée.

Une deuxième partie aborde les aspects algébriques définissants le type données étudiées : contiguïté, contrainte ordinale, "cube" de données.
On y généralise notamment la décomposition en valeurs singulières d' une matrice, à un tenseur d'ordre 3 et k quelconque. Ceci nous permet alors de généraliser le théorème d'Eckart et Young et de proposer l'ACP d'un tenseur d'ordre k : l'ATP-kmodes. En conséquence est proposé l'analyse des correspondances de k variables : l'AFC-kmodes.
En rajoutant des contraintes de sous-espaces, l'application de ces généralisations aux mesures répétées sur des groupes d'individus introduit diverses méthodes.

Une troisième partie montre des applications de ces deux premières à un suivi épidémiologique de personnes VIH+ : l'enquête SEROCO.

Abstract

We are interested by the utilization of factor analysis, such as PCA, in repeated measures factors analysis.

A first part deal with the two classical approaches of this problem by linear models : univariate, or multivariate. Particular PCA describe geometricaly the rejection of the null hypothesis of these models.

A second part deal with algebraic aspects of the data : contiguity, ordinal constraints, paralellepipedic form of the data.
Singular value decomposition of a matrix is generalised to a tensor of order 3 and k. So, the generalization of Eckart-Young theorem lead to the PCA of a tensor of order k : the PTA-kmodes (ATP-kmodes in french). In consequence is proposed for exemple the correspondence analysis of k variables : the CA-kmodes (AFC-kmodes in french).
Applied to the context of repeated measures factors, by addition of constraints, these generalisation conduct to different methods.

A third part show applications of the two first part in a cohort study of HIV+ persons : the SEROCO cohort.

Préambule Général
et
Plan du Travail

On peut dire qu'historiquement la statistique s'est d'abord occupée des observations répétées sous forme de séries chronologiques. Le mot série, dans la dénomination série statistique, qui aujourd'hui se rapporte à une collection de valeurs, ne se réfère t-il pas à une suite de valeurs ?

Les statistiques sont à l'origine demandées par les gouvernants des états, d'où le nom de *stat*istique, pour connaître leurs puissances sous divers aspects. Ainsi les premiers recensements seraient, nous disent Droesbeke et Tassi(1990), ceux de la civilisation sumérienne de 5000 à 2000 ans avant notre ère, puis de la civilisation égyptienne qui aurait recensé systématiquement sa population.

Les domaines où apparaissent en premier ces statistiques sont souvent liés à la **chronologie**, comme en astronomie et en démographie. Les **mesures répétées** sont alors le premier intérêt des "statisticiens".

Si le soucis de description statistique par des résumés statistiques tels que : l'étendue, le mode, la moyenne etc...remonte, semble t-il d'après Droesbeke et Tassi(1990), à il y a prés de 2500 ans chez les astronomes Babyloniens, les représentations graphiques pour décrire les données viennent beaucoup plus tard, vers les XVII et XVIII siècle.

Une référence serait l'ouvrage de Playfair William "The Commercial and Political atlas" publié à Londres en 1786, dans lequel 43 des 44 graphiques concernent des séries chronologiques.

L'analyse des données sous son aspect analyse factorielle, date de la fin du XIX^{ème} siècle et du début du XX^{ème}. L'article faisant date est celui de Pearson(1901). D'autres comme ceux de Spearman, Fisher, Hotelling ont largement contribué à l'élaboration des méthodes actuellement courantes comme l'Analyse en Composantes Principales, l'Analyse Canonique, l'Analyse Factorielle des correspondance, l'Analyse Factorielle Discriminante.

L'Analyse en Composantes Principales contient donc, de notre point de vue, le double objectif historique de résumer les données et de les représenter graphiquement.

Mais une première question se pose : **l'ACP, ou toute autre méthode factorielle apparentée, est-elle appropriée pour traiter des observations répétées ?**

De nombreux travaux ont apporté des réponses à cette question dans divers domaines appliqués tels que : l'économie, l'écologie, la psychométrie, l'agronomie et les sciences sociales en général. Pour notre part nous n'avons pas vu d'exemples conséquents en épidémiologie.

Nous ne prétendrons pas faire la synthèse de toutes les approches possibles mais dans le chapitre BII nous en donnerons quelques aspects.

Une deuxième question, nous occupera plus particulièrement dans ce travail : **quelles doivent être les analyses factorielles pour rendre compte de l'évolution de groupes d'individus observés ?**

L'étude de l'évolution de groupes d'unités statistiques ou plus généralement l'étude de la répétition d'observations de groupes d'unités statistiques, porte le nom de problème de **facteurs à mesures répétées**, "repeated measures factors" en anglais.

Une nombreuse littérature en langue anglaise traite du sujet dans le soucis de l'intégrer à une analyse de la variance. Elle souhaite bien entendu faire de la statistique inférentielle pour répondre à des questions fondamentales sur l'étude de ces groupes d'individus.

La **partie A** de ce travail fournira une première réponse à notre question par le biais de ***modèles inférentiels classiques*** . L'objectif est d'associer à ces analyses classiques, des **analyses factorielles pertinentes**. On disposera alors de **tests statistiques** pour étudier les hypothèses faites sur les données, et, de **descriptions géométriques traduisant l'écart à l'hypothèse**. La lecture simultanée des sorties : procédures des tests et ACP particulières proposées, permettront de meilleures conclusions.

Le chapitre **AI** exposera les solutions classiques pour les variables quantitatives tandis que le chapitre **AII** s'occupera du cas des variables qualitatives.

Issues de discussions plus algébriques explicitant ce qu'est une mesure ordonnée, de nouvelles approches ont vu le jour, souvent par l'optique factorielle.

On examinera ces aspects dans la **partie B** que nous avons intitulé ***modèles algébriques*** bien que certains aspects probabilistes et inférentiels y soient reliés, aussi bien que des aspects géométriques sont sous-jacents dans la **partie A** .

D'une part, la **structure chronologique** impliquant, naturellement, une **contiguïté** des observations nous amènera à considérer aussi le temps comme **contrainte ordinale**. Ces contraintes peuvent prendre plusieurs formes que nous discuterons dans le chapitre **BI**. Le temps sera alors pris de façon extrinsèque par rapport à l'objectif fixé, mais nous verrons que l'on peut tout aussi bien le considérer comme intrinsèque au problème. Les choix d'analyses pour des facteurs à mesures répétées y seront abordés.

D'autre part la simple constatation pratique de disposer de plusieurs mesures répétées, signifie que l'on n'a plus simplement un tableau de données, mais un "cube" de données.

Ainsi dans le chapitre **BIII** on s'intéressera à la conception d'analyses factorielles sur **trois modes et non plus deux**. La considération d'une structure à mesures répétées sur cette base y sera aussi abordée. Tandis qu'au chapitre **BII** ce "cube" était vu comme une collection de tableaux à deux modes, l'approche faite, sur le plan algébrique, dans ce chapitre est assez nouvelle pour l'analyse des données.

En effet non seulement elle permet de traiter entièrement un "cube" de données, c'est à dire un **multitableau d'ordre trois**, mais elle permet aussi une généralisation quasi-immédiate aux **multitableaux d'ordre k** quelconque.

Cet aspect nous conduira à introduire de nouvelles analyses basées sur la **décomposition en valeurs singulières d'un tenseur d'ordre k**. Cette dernière généralise strictement la décomposition en valeurs singulières sur deux modes par les relations de transitions généralisées et des relations d'orthogonalités adaptées. La prise en compte de métriques sur les k espaces vectoriels considérés y est immédiate.

On y décrira alors comme nouvelle méthode la généralisation de **l'ACP sur k modes** sous le nom d'**ATP-kmodes** : Analyse en Tenseurs Principaux sur k modes.

Comme principale autre nouvelle méthode nous introduirons l'**analyse des correspondances de k variables**, qui était souvent traitée par une analyse des correspondances multiples, sous la forme d'une ATP-kmodes : l'**AFC-kmodes**.

Notre souci dans tout ce travail a, sans aucun doute, été de résoudre concrètement des problèmes pratiques, de proposer de nouvelles méthodes, même si parfois celles-ci n'ont pas a priori un grand intérêt méthodologique. En effet on a voulu en tout premier lieu traiter des données provenant d'un suivi épidémiologique, mais on a souhaité également, le faire de la façon la plus originale possible.

Si beaucoup de méthodes statistiques utilisées en épidémiologie dérivent de pratiques statistiques des essais cliniques, il semble que certains problèmes notamment d'échantillonnage y soient beaucoup moins maitrisables. Ceci est trivial pour les études de cohortes. Pour cette raison nous pensons que les méthodes descriptives telles que les méthodes factorielles peuvent apporter beaucoup aux épidémiologistes, surtout dans le cas de suivi d'une cohorte.

Ceci nous a occupé dans la **partie C**. Elle illustre les développements des parties précédentes, mais aussi traite des problèmes de recherche épidémiologique sur le SIDA, problèmes que nous nous sommes posés avec l'équipe SEROCO.

Puisse ce travail servir, même très modestement, la recherche épidémiologique en général, ainsi que l'étude des cohortes, mais aussi la compréhension de l'évolution des patients séropositifs au VIH.

Plan et mots clés

INDEX des ABRIÉVIATIONS

Les nouvelles sont soulignées.

A.C.	Analyse Canonique
A.C.G.	Analyse Canonique Généralisée (de Caroll)
A.C.P.	Analyse en Composantes Principales
A.C.P.-3 modes	ACP sur 3 modes
A.C.P.-SPS-3 modes	ACP sur 3 modes, sur modèles SPS
A.C.P.V.I.	ACP par rapport à des Variables Instrumentales
A.C.P.V.I.-3modes	ACPVI sur 3 modes
A.C.V.	Analyse en Composantes de Voisinage
A.F.C.	Analyse Factorielle des Correspondances
A.F.D.	Analyse Factorielle Discriminante
A.F.C.M.	AFC Multiple
A.F.C.-kmodes	AFC sur k modes
A.F.C.3.M.	AFCM avec un troisième modes
A.F.C.k'.B.	AFC-k'modes de k variables k'<k.
A.F.C.C.M.	AFC-(k'+1)modes de k variables k'<k.
A.F.C.V.I.	AFC par rapport à des Variables Instrumentales
AN.O.VA.	ANalyse Of VAriance
A.T.P.-kmodes	Analyse en Tenseurs Principaux sur k modes
A.T.P.V.I.-kmodes	ATP-kmodes par rapport à des Variables Instrumentales
C.P.C.A.	Common Principal Component Analysis
D.V.S.	Décomposition en Valeurs Singulières
D.M.M.	Doubly Multivariate Manova
D.V.S.-3modes	DVS sur 3 modes
G.L.M.	General liear Model
G.M.AN.O.VA	Growth Multivariate ANalyse Of VAriance
G.S.K.	Grizzle Starmer Koch
G.o.M.	Grade of Menbership
I.P.S.N.P.	Indicateur de Pratiques Sexuelles Non-Protégées
Longi	méthode Longi, longitudinales
M.AN.O.VA	Multivariate ANalyse Of VAriance
PRINQUAL	méthode PRINQUAL (dans SAS) sur données qualitatives
Pré-Statis	Statis sur les tableaux
Pré-Statis-Croisé	Modèle, sur tableaux np.qp ou Fictif, association de 2 Pré-Statis
Statis	méthode STATIS, trois indices
S.P.S.	Statis-Pré-Statis, Pré-statis optimal
S.U.R.	Seemingly Unrelated Regression

Modèles inférentiels classiques

AI
Modèles linéaires classiques et ACP

AII
Modèles Linéaires et log-linéaires pour variables qualitatives

Modèles algébriques

BI
Observations répétées :
variable ordinale ou contrainte ordinale ?

BII
Aperçus des méthodes factorielles multitableaux pour des mesures répétées

BIII
Approche tensorielle des méthodes factorielles et applications à l'évolution de facteurs

Applications dans l'enquête SEROCO

CI
Description de l'enquête SEROCO et buts de notre étude

CII
Ré-exposition sexuelle au VIH :

A) construction d'un indicateur

CII
Ré-exposition sexuelle au VIH :

B) Description par l'AFC-kmodes

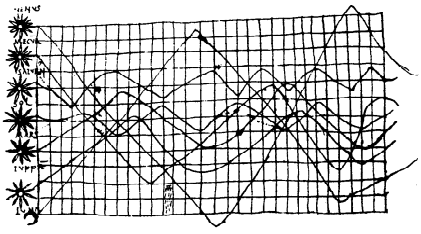
CIII
Evolution du profil biologique en fonction :

A) de facteurs pronostiques et socio-épidémiologiques

Annexes générales

"Imagination is more important than knowledge."
A.EINSTEIN

Manuscrit du X^{ème} XI^{ème} siècle : inclinaison des orbites des planètes en fonction du temps



Funkhauser, H.G. (1936) A note on the 10th century graph Osiris, vol.1, 260-262.

AI
Modèles linéaires classiques et ACP

I . Introduction

Le lien entre l'ACPVI et l'analyse de variance classique a déjà été établi par la décomposition d'un modèle suivant les effets des facteurs, Sabatier(1987), Sabatier et al.(1989), Leibovici et al.(1990). On se propose dans ce chapitre d'étendre ces résultats au cas de facteurs à mesures répétées. Ceci nous a amené à envisager les différentes approches citées dans la littérature et ainsi à examiner l'apport des méthodes de l'analyse factorielle.

La problématique générale est la suivante.

On dispose de q variables mesurées sur n individus repartis en I_G groupes en p occasions (conditions, dates, visites, dosages,...). Winer(1971) distingue deux sortes de facteurs:

- "**between subject factor**" (facteur inter-sujet)
  ou facteur non répété : un individu prend une seule des I_G modalités.
- "**within subject factor**" (facteur intra-sujet)
  ou facteur répété : un individu prend toutes les p modalités.

On retrouve cette présentation dans les logiciels BMDP(1981), SAS(1985) ou il peut y avoir plusieurs facteurs répétés ou non-répétés (i.e plusieurs classes d'individus et/ou par exemple plusieurs dosages et dates de mesures), mais une seule variable mesurée (q=1) pour les procédures classiques.

On sera amené par la suite à distinguer le cas univariable (q=1) et le cas multivariable (q > 1).

Devant de telles données on peut parler :

- **de mesures répétées**

si le nombre de répétitions est voisin du nombre de modalité du facteur groupe, et "très inférieur" au nombre d'individus :

$p \spadesuit I_G << n$,

• d'études longitudinales et/ou courbes de croissances

(où le facteur répété est le temps) si le nombre de répétitions est supérieur au nombre de modalité du facteur groupe, et est voisin ou supérieur au nombre d'individus :

$p > I_G$ et $p \spadesuit$ ou$> n$,

• de séries chronologiques

(où le facteur répété est le temps), si le nombre de répétitions est "très supérieur" au nombre de modalité du facteur groupe, et est "très supérieur" au nombre d'individus :

$p >> I_G$ et $p >> n$.

Lorsque l'on utilise l'analyse de variance on est confronté à deux principales difficultés, que ce soit dans le cas univariable comme dans le cas multivariable :

1) Tout d'abord le choix de l'approche :

- **univariée**, chaque variable mesurée en p occasions sur n individus est analysée comme la donnée d'une variable mesurée sur np observations,
- **multivariée**, chaque variable mesurée en p occasions sur n individus est analysée comme la donnée de p variables mesurées chacune sur les mêmes n observations.

Les Hypothèses sur les distributions, pour la validité des tests, sont plus ou moins restrictives dans l'une ou l'autre approche. En effet la même hypothèse nulle peut être acceptée dans une approche et refusée dans l'autre.

2) Ensuite la prise en compte de l'ordre des mesures pour le facteur répété, c'est à dire de la succession des conditions. Le lissage et la planification ("change-over") peuvent y apporter des solutions.

Dans un article général sur les mesures répétées Koch et al (1980) ont passé en revue les différentes méthodes et techniques utilisées en analyse de la variance dans divers domaines (agriculture, essais cliniques,...) et proposent suivant la planification et les hypothèses nécessaires pour les tests (paramétriques ou non-paramétriques) différentes stratégies:

Randomization Structure	Interval data	Ordinal data	Nominal data
Randomization of experimental units to treatment groups	Repeated measurement multivariate analysis of variance	Rank analysis of variance methods (Kruskal-Wallis procedures) and signed rank methods (generalized sign procedure and Wilcoxon signed rank procedure)	multivariate chi-square tests for homogeneity
Randomization of observational units to conditions	Repeated measurement (split-plot) analysis of variance	Partial association rank methods (Friedman chi-square and aligned rank procedures)	Generalized Mantel-Haenszel methods and Cochran Q-statistic
Stratified random selection of experimental units from possibly different populations	Repeated measurement multivariate analysis of variance with assumed distribution and covariance structure	Multivariate linear model categorical data methods with respect to either maximum likelihood or weighted least squares estimation and hypothesis testing procedures	

Les articles clés, couvrant ces problèmes, en majeure partie dans la littérature en langue anglaise, sont :

- Boik(1988), Thomas(1983), Barcikowski et Robey(1984), Timm(1984) pour l'optique multivariée,
- Potthof et Roy(1964), Rao(1965, 67), Ware(1985) pour les courbes de croissances,
- Winer(1971) et Monlezun et Blouin(1988) pour l'approche split-plot.(i.e univarié)

Dans ce chapitre nous aborderons d'abord l'apport de l'Analyse en Composantes Principales dans le cadre de l'analyse de variance des mesures répétées pour les deux approches classiques citées ci-dessus (parties II et III). On donnera ainsi les ACP à réaliser pour obtenir une décomposition de l'écart à l'hypothèse nulle considérée. Dans la partie IV, une autre approche, liée au modèle linéaire général abordé dans la partie III, apportera des compléments aux deux premières parties.

On peut noter que les autocorrélations prises en compte dans la littérature, lorsque qu'elles sont modélisées, sont la corrélation intra-classe ("compound symetry") pour le cas d'une répétition due à différents traitements, ou la corrélation sérielle pour le cas d'une répétition due au temps :

$$\begin{pmatrix} 1 & \rho & . & \rho \\ \rho & . & . & . \\ . & . & . & \rho \\ \rho & \rho & . & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & \rho & . & \rho^{p} \\ \rho & . & . & . \\ . & . & . & \rho \\ \rho^{p} & \rho^{p-1} & . & 1 \end{pmatrix} .$$

intra-classe sérielle

Ces considérations rentrent souvent dans la dénomination structure de covariance que nous aborderons dans la partie IV.

II . Approche univariée

II.1 split-plot :

Dans le cas univarié, le modèle considéré est un modèle mixte où les "between subject factors" sont considérés comme fixés, le facteur sujet est aléatoire et hiérarchisé par rapport aux "between subject factors" et croisé avec le ou les "within subject factors". Le modèle est alors un split-plot où le "whole plot" (facteur hiérarchisant) est le "between subject factor", et le "sub-plot" ou "split-plot" est le "within subject factor", Winer(1971), Monlezun et Blouin(1988).

exemple :

Soit q variables mesurées sur n individus repartis en I_G groupes et ceci répété sur p visites, on aura alors :

1 facteur groupe G, à I_G modalités (whole plot),
1 facteur sujet S, à n modalités (aléatoire),
1 facteur visite V, p modalités (split-plot).

On considère $Y_{(np),q}$ le tableau des q variables centrées, $V_{(np),p}$ le tableau disjonctif complet des p modalités des visites, $G_{(np),I_G}$ le tableau disjonctif complet des groupes et $S_{(np),n}$ constitué de p blocs identité de dimension n identifiant les sujets.
On peut écrire le modèle suivant pour la $l^{ième}$ variable le $i^{ième}$ individu du $j^{ième}$ groupe pour la $k^{ième}$ visite :

$$\mathbf{y^l_{ijk} = \alpha^l_j + \beta^l_{i(j)} + \gamma^l_k + \delta^l_{jk} + \varepsilon^l_{ijk}}.$$

On a dans l'ordre l'effet groupe, puis l'effet individu (hiérarchisé par rapport au groupe), l'effet visite, l'effet d'interaction groupe et visite et le résidu de ce modèle.

Le split plot est traité pour une variable (q=1) dans Winer(1971). On peut le généraliser à q variables en décomposant la somme totale des carrés des écarts, i.e. l'inertie du nuage, en inertie due à chaque effet dans le modèle ci-dessus :

$$\mathbf{I_{tot} = I_G + I_{S(G)} + I_V + I_{G.V} + I_{Rsd}}.$$

Ainsi par exemple I_G est l'inertie expliquée par G dont la décomposition peut être obtenue par l'ACPVI de (Y , Q , D) par rapport à G, c'est à dire l'ACP de (P_GY , Q , D). P_G est le projecteur D-orthogonal sur le sous-espace engendré par les colonnes de G, Q et D sont les matrices des produits scalaires respectivement sur R^q et $R^{(np)}$ (Sabatier(1987)).

Les choix de Q et D introduiront différentes analyses suivant les buts et contraintes supplémentaires, notamment sur D qui intervient dans le projecteur et l'estimation des coefficients de régressions. On en verra l'importance dans la partie IV pour l'optique "Seemingly Unrelated Regressions".

Lors de la réalisation des tests des effets de chaque structure, ceci de manière analogue à l'analyse de variance split-plot (i.e. l'erreur change avec l'effet testé du fait du facteur aléatoire) Winer(1971), les hypothèses de validité du F (ou une généralisation pour q variables) nécessitent ce que l'on appelle la sphéricité de la matrice d'autocorrélation Σ (voir *II.3*).

Un moyen de palier ce problème de corrélation des mesures est d'effectuer une transformation rendant les mesures orthogonales. C'est ce que propose Snee(1972) à propos d'un article de Church(1966), mais nous en faisons une utilisation assez différente de celui-ci.

L'idée est d'effectuer pour q=1 l'ACP du tableau à n lignes et p colonnes, de vectorialiser les composantes principales normées et de réaliser l'analyse split-plot. Dans cette analyse le facteur split -plot ne peut plus s'appeler facteur Visite mais plutôt facteur modélisation de la visite. En effet l'ordre (succession des visites) est transformé en ordre d'importance des modélisations du suivi (l'ordre des valeurs propres).

On a alors Σ=Id_p (si les composantes sont normées à l'identité) et les F sont valides.
Pour q variables, il faut effectuer une transformation commune pour la construction du facteur modélisation de la visite que l'on peut faire par exemple, avec la méthode de Flury(1984) ou par la méthode STATIS, Lavit(1988).

Remarques:

-1 On peut noter que Snee et al (1979) ont écrit un modèle analogue à un Split-plot, mais effectuent une décomposition en valeur singulière de l'interaction Groupe.Visite pour en garder ensuite les deux premières composantes afin d'avoir un modèle avec moins de degrés de liberté.

-2 Une contrainte assez forte, pour l'écriture du modèle, est l'orthogonalité des facteurs, c'est à dire l'orthogonalité des sous espaces engendrés par les indicatrices des modalités des facteurs. Quand les facteurs ne sont pas orthogonaux, on peut écrire d'autres modèles, par exemple, des modèles à effets conditionnels, c'est ce que fait Winer quand il écrit :

$$\begin{aligned} SS_{total} &= SS_{b.people} + SS_{w.people} \\ &= SS_{b.people} + SS_{b.treat} + SS_{rsd}, \end{aligned}$$

en terme d'inertie ceci revient à décomposer de la façon suivante :

$$\mathbf{I_{tot} = I_S + I_{G/S} + I_{Rsd}}.$$

Pour la construction des sous-espaces sur lesquels on projette on pourra consulter, Rao et Yanaï(1979), Sabatier(1987), Sabatier et al.(1990), Pontier et al.(1990).

-3 On peut faire d'autres approches dans le cas univarié multivariable, plus générales que la vision split-plot, elles sont liées au modèle linéaire général. Il y a l'approche que Boik(1988) appelle "multivariate mixed model" ou encore l'optique S.U.R. (partie IV).

II.2 anacova :

Si dans l'exemple décrit plus haut nous ajoutons que l'on a mesuré aussi à chaque visite z variables explicatives,on entre dans le cadre ANACOVA (analyse de la covariance). En s'inspirant de Monlezun et Blouin (1988) pour écrire le modèle avec des covariables Z, il vient :

$$\mathbf{I_{tot} = I_{P_{G(Z)}} + I_{P_{V(Z)}} + I_{P_{G.V(Z)}} + I_{Rsd}}.$$

Z est la matrice $Z_{(np,z)}$; $I_{P_{G(Z)}}$ est l'inertie dûe à P_GZ décomposée par l'ACP de ($P_{(P_{G(Z)})}$ Y , Q , D); c'est à dire que l'on a le modèle précédent mais "à travers" Z.

On peut noter que les covariables peuvent être considérées comme un "facteur" et écrire des modèles tels que :

$$\mathbf{I_{tot} = I_G + I_{S(G)} + I_V + I_{G.V} + I_{Z/(G\ V)} + I_{Rsd}}.$$

II.3 tests :

Lorsque l'on veut faire des tests avec des modèles de mesures répétées on s'intéresse plus particulièrement aux hypothèses suivantes :

$\mathbf{H_{01}}$: V.G interaction groupe visite,
$\mathbf{H_{02}}$: V effet de la visite,
$\mathbf{H_{03}}$: G effet du groupe.

Les hypothèses sur les distributions, pour obtenir des tests valides, sont très restrictives et l'écart à ces hypothèses semble souvent catastrophique (Keselman HJ et JC(1984)). Ces problèmes ont entrainé une littérature nombreuse pour étudier les validités et les "stratégies" pour les tests, (Koch et al(1980), Looney et Stanley(1989), Rouanet et Lepine(1970), Barcikowsky et Robey(1984), Boik(1981 88), Keselman HJ et JC(1984), Grieve(1984) etc..).

Les discussions et stratégies sont basées sur le F ajusté pour les degrés de liberté car il n'y a pas indépendance entre les p échantillons. On corrige les degrés de liberté du F traditionnel par le facteur multiplicatif :

- ε de Box(54), ou,
- $\hat{\varepsilon}$ de Greenhouse et Geisser(1959), ou,
- $\tilde{\varepsilon}$ fonction de $\hat{\varepsilon}$ de Huyn et Feldt(76).

On a $1/_{(p-1)} \le \varepsilon \le 1$ et ε=1 est la condition de sphéricité où $\varepsilon=[tr({}^tC \bullet C)]^2/_{(p-1)tr[({}^tC \bullet C)^2]}$ avec C le contraste.

La comparaison entre l'optique univariée et multivariée est souvent faite et certains auteurs conseillent même de faire le test pour un niveau $\alpha/_2$ dans chaque optique et de rejeter si chaque test l'est (Looney et Stanley(1989)).

Les conditions de validité pour un F (ou un test approprié pour le multivariable) sont exposées de manière générale dans Boik(1981), Thomas(1983) et Boik(1988) et s'expriment comme suit,

cas univariable :

(a) normalité : vec(e) ~ $N_{np}(0, Id_n \bullet)$, où e est la partie aléatoire du modèle,

(b) hypothèse de sphéricité pour Σ : E = σ^2Id où E = ${}^tC\Sigma C$ et C est un contraste p ∞ (p-1) (i.e pour avoir un Chi-Deux),

(c) homogénéité : si il y a plusieurs groupes $E_j = E_{j'}$ pour j et j' avec $E_j = {}^tC\Sigma_j C$ où Σ_j est la matrice de covariance p ∞ p pour le groupe j,

cas multivariable :

(a) vec(e) ~ $N_{nqp}(0, Id_n \bullet_{qp,qp})$,
(b) $({}^tC\ Id_q) \bullet (C\ Id_q) = Id_{p-1} \otimes \Lambda$ (i.e pour avoir des lois de Wishart), où Λ est semi-définie positive,
(c) $({}^tC\ Id_q) \bullet_j (C\ Id_q) = ({}^tC\ Id_q) \bullet_{j'} (C\ Id_q)$ (homogénéité).

Si la condition de sphéricité **(b)** n'est pas satisfaite, Boik(1988) donne un théorème d'approximation par une loi de Wishart avec un terme ε multiplicatif pour les degrés de liberté qui généralise celui de Box(1954) établi pour le cas univariable. Pour l'approche multivariée la condition **(b)** n'est pas nécessaire ce qui nous amène à dire que cette approche semble moins restrictive.

Remarques:

-1 On peut voir le ε comme le Rv d'Escoufier(1973), au carré entre les opérateurs Id_{p-1} et ${}^tC\Sigma C$. On fait donc le rapprochement entre l'Hypothèse de sphéricité et le fait

que ε=1. Connaissant la loi du Rv^2, donnée par Cleroux et Ducharme (1986), on peut effectuer un test de sphéricité. Il en existe déjà par ailleurs fondés sur d'autres hypothèses, comme ceux de Mauchly (1940), Mendoza(1980) ... Le premier est fourni dans PROC GLM de SAS.

-2 La matrice de corrélation intra-classe ("compound symetry") vérifie l'hypothèse de sphéricité, mais la condition n'est pas nécessaire comme le montre Winer(1971) sur un exemple.

-3 On peut proposer pour le cas split-plot multivariable les mêmes tests que ceux réalisés classiquement en univariable mais donc en remplaçant les variances par les inerties et alors si les conditions de validité sont vérifiées on a des tests F (ou un quotient de deux Wisharts indépendantes suivant la forme de Λ).

-4 Une manière de contourner les difficultés dues aux hypothèses est d'utiliser des tests de permutation comme l'ont exposé Zerbe et Walker(1977) ou Raz(1989).

III . Approche multivariée

III.1 modèle linéaire général :

Les développements abordés dans cette partie sont inspirés de Boik(1988), Timm(1984), Sabatier(communication personnelle).
Le cadre est le modèle linéaire général :

$$\mathbf{Y_{n,q} = X_{n,f}\,\beta_{f,q} + e_{n,q}},$$

où $E(Y) = X\beta$, $e \sim N_q(o,\Sigma)$, $rg(X) = r$.
L'estimation par les moindres carrés ordinaires donne :

$$\mathbf{X\hat{\beta} = P_X Y = X({}^tXX)^{-}{}^tXY}.$$

Le test classique est pour l'hypothèse nulle suivante :

$$\mathbf{H_0 : L\beta M = \Gamma_0 \text{ contre } H_1 : L\beta M = \Gamma}$$

où L : a ∞ f rg(L)≤ r
M : q ∞ u rg(M) = u ≤ q ≤ n-r
Γ : a ∞ u.

Pour ce test on peut utiliser par exemple la statistique de Lawley-Hotelling :

$$tr(E^{-1}H),$$

avec $H = {}^t(L\hat{\beta} M - \Gamma_0)(L({}^tXX)^{-}{}^tL)^{-}(L\hat{\beta} M - \Gamma_0)$,
et $E = {}^tM{}^tY(Id_n - X({}^tXX)^{-}{}^tX)YM = {}^t(YM)P_X{}^{\perp}(YM)$,
qui sont les matrices classiques de l'Hypothèse H_0.

Remarque :

-1 On sait que les quatre statistiques usuelles :
Wilks $\det(E)/_{\det(E+H)}$, Pillaï $tr(H(E+H)^{-1})$, Lawley-Hotelling $tr(E^{-1}H)$ et Roy la plus grande valeur propre de HE^{-1} sont équivalentes dans le sens où aucun des tests basés sur ces statistiques n'est uniformément plus puissant (U.M.P.).

III.2 $tr(HE^{-1})$ ou $tr(H(E+H)^{-1})$ et ACP :

Pour le modèle précédent, après estimation de β si l'on réalise un test de la forme $\beta M = 0$, c'est à dire si l'on effectue des combinaisons linéaires sur Y, et on teste l'adéquation du modèle, alors :

$$tr(HE^{-1}) = tr(P_X Q {}^tY P_X) \text{ avec } Q = M({}^tM{}^tYP_X{}^{\perp}YM)^{-}{}^tM.$$

C'est 'inertie de l'ACPVI de (Y,Q,Id) par rapport à X (Sabatier(1987)), c'est à dire de l'ACP de $(P_XY, M({}^tM{}^tYP_X{}^{\perp}YM)^{-}{}^tM, Id)$.

L'adéquation du modèle est le test $\beta = 0$. Il est décomposé par l'ACP de $(P_XY, ({}^tYP_X{}^{\perp}Y)^{-}, Id)$. La métrique sur les individus est l'inverse de la matrice de variance intra-X.

Si l'on effectue un test de la forme $L\beta = 0$ alors :

$$tr(HE^{-1}) = tr(PYQ{}^tYP),$$

avec $Q = ({}^tYP_X{}^{\perp}Y)^{-}$,
et $P = X({}^tXX)^{-}{}^tL(L({}^tXX)^{-}{}^tL)^{-}L({}^tXX)^{-}{}^tX$, le projecteur sur $X({}^tXX)^{-}{}^tL$.
Pour ce dernier, si $L = Id_f$ alors $P = P_X$, si $L = (1,..,1)_{IG}$ alors $P = P_\Delta$. On peut écrire aussi selon le principe de Wald (Seber (1964)) :

$$P = P_X - P(X \leftrightarrow (L({}^tXX)^{-}{}^tX)^{\perp}).$$

L'inertie est celle de l'ACP de $(PY, ({}^tYP_X{}^{\perp}Y)^{-}, Id)$. On peut remarquer que cette dernière ACP est aussi l'ACP de $(PP_XY, ({}^tYP_X{}^{\perp}Y)^{-}, Id)$ qui est liée à l'ACP de $(P_XY, ({}^tYP_X{}^{\perp}Y)^{-}, Id)$, (l'adéquation du modèle), donc le test $L\beta = 0$ peut être appelé l'adéquation du sous modèle induit par L.

Si l'on effectue le test global $L\beta M = 0$ alors :

$$tr(E^{-1}H) = tr(PYQ{}^tYP),$$

avec $P = X({}^tXX)^{-}{}^tL(L({}^tXX)^{-}{}^tL)^{-}L({}^tXX)^{-}{}^tX$,
et $Q = M({}^tM{}^tYP_X{}^{\perp}YM)^{-}{}^tM$,
on peut donc faire l'ACP de $(PY, M({}^tM{}^tYP_X{}^{\perp}YM)^{-}{}^tM, Id)$ pour décomposer ce test.

Remarques :

-1 La métrique Q est "proche" de l'écriture du projecteur sur M avec la métrique colonne ${}^tYP_X{}^{\perp}Y$ (i.e. variance intra X) qui ferait de M une contrainte colonne sur Y.

-2 La dernière ACP peut s'écrire, en conservant le même opérateur de produits scalaires entre individus, comme celle de $(P(YM), ({}^t(YM)P_X{}^{\perp}(YM))^{-}{}^t, Id)$. C'est l'ACP de l'image par P, des variables transformées par M, avec comme métrique ligne la matrice de variance intra-X.

-3 On peut écrire les ACP en faisant apparaître la métrique D (au lieu de Id) ceci entraîne une modification des écritures de $\hat{\beta}$, H, E, P, P_X, et Q que nous n'avons pas voulu noter par souci de clarté, mais on a :

$X\hat{\beta} = X({}^tXDX)^{-}{}^tXDY$ estimatin des moindres carrés pondérés,
$H = {}^t(L\hat{\beta} M - \Gamma_0)(L({}^tXDX)^{-}{}^tL)^{-}(L\hat{\beta} M - \Gamma_0)$,
$E = {}^tM{}^tY(Id_f - X({}^tXDX)^{-}{}^tXD)YM$,
$= {}^t(YM)P_X{}^{\perp}(YM)$,
$P = X({}^tXDX)^{-}{}^tL(L({}^tXDX)^{-}{}^tL)^{-}L({}^tXDX)^{-}{}^tXD$,
$P_x = X({}^tXDX)^{-}{}^tXD$.

-3 On pourrait utiliser le critère de Pillaï $tr(H(E+H)^{-1})$ et écrire les mêmes analyses, alors la métrique Q est simplement la métrique de Mahalanobis des variables transformées par M.

III.3 choix des contrastes :

Pour appliquer ce cadre à l'analyse des facteurs à mesures répétées et notamment pour effectuer les tests analogues à $\mathbf{H_{01}}$, $\mathbf{H_{02}}$, $\mathbf{H_{03}}$ de la présentation univariée nous allons choisir les matrices M et L.

III.3.a. cas univariable avec un seul groupe :

Le modèle est donc :

$$\mathbf{Y_{n,p} = G_{n,IG}\ \beta_{IG,p} + e_{n,p}}$$

Y n ∞ p p mesures répétées (Traitements ou visites)
G n ∞ I_G 1 facteur groupe à I_G modalités
Les contrastes sont :
L a ∞ I_G

M $p \infty u$.

Un premier test qui correspond à **H_{02}**, mais sans considérer qu'il y a des groupes est le test :

$\mu_1 = \mu_2 = \ldots = \mu_p$

$$\Leftrightarrow \quad \begin{cases} \mu_1 - \mu_2 = 0 \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \mu_{p-1} - \mu_p = 0 \end{cases}$$

$\Leftrightarrow L\beta M = 0$ avec $L = (1, 1, 1, \ldots 1)_{I_G}$

et

$$M_{p,p-1} = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & -1 & & \\ & & & 1 \\ & & & -1 \end{pmatrix}$$

On effectue le test avec $tr(E^{-1}H)$ que l'on décompose par l'ACP de $(P_\Delta Y, M({}^tMtYP_G{}^\perp YM)^{-t}M, D)$ qui n'a qu' un seul axe. On s'intéressera plutôt à l'ACP de $(P_\Delta{}^\perp Y, M({}^tMtYP_G{}^\perp YM)^{-t}M, D)$ qui est une ACPVI orthogonale.

D'autres choix peuvent être fait pour M par exemple :

$$M_{p,p-1} = \begin{pmatrix} -1 & \ldots & -1 \\ 1 & & \\ & 1 & \\ & & \\ & & \cdot \quad \cdot \\ & & 1 \end{pmatrix}$$

où l'on compare le premier traitement (ou visite) aux autres. On peut aussi chercher une évolution (plutôt applicable pour des visites) avec le contraste suivant (exemple pour 4 répétitions) :

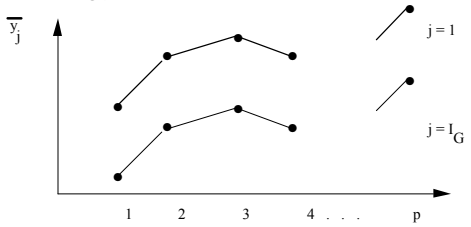
$$M_{4,3} = \begin{pmatrix} -1 & -1/2 & -1/3 \\ 1 & -1/2 & -1/3 \\ & 1 & -1/3 \\ & & 1 \end{pmatrix}.$$

Remarque :

-1 Ce contraste fait penser à la métrique venant de la transformation F_0 (matrice triangulaire de 1) utilisée dans Besse (1987). En effet si l'on fait l'ACP de YF_0 on compare la première visite, la moyenne des deux premières, la moyenne des trois premières, etc ... ici l'on compare chaque visite (par exemple) à la moyenne des précédentes.

Le test qui correspond à l'interaction groupe.visite est le test noté **H_{01}** :

"les profils des I_G groupes sont-ils parallèles ?"



Les profils seront parallèles si les accroissements sont identiques. Le test s'écrit donc :

Erreur !)

et se traduit par LßM = 0 avec

$$\beta_{I_G,p} = \begin{pmatrix} \mu_{1,1} & \cdot & \cdot & \mu_{1,p} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \mu_{I_G,1} & \cdot & \cdot & \mu_{I_G,p} \end{pmatrix},$$

et

$$M_{p,p-1} = \begin{pmatrix} -1 & \ldots & -1 \\ 1 & & \\ & 1 & \\ & & \\ & & \cdot \quad \cdot \\ & & 1 \end{pmatrix} \quad L_{I_G-1,I_G} = \begin{pmatrix} 1 & & -1 \\ & \cdot & \cdot \\ & 1 & -1 \end{pmatrix}.$$

Alors, pour le test **H_{01}**, on effectue l'ACP du triplet $(PY, M({}^tMtYP_G{}^\perp YM)^{-t}M, D)$ qui est aussi l'ACP de $(PYM, ({}^tMtYP_G{}^\perp YM)^-, D)$, du moins pour la représentation des individus.

On peut alors constater que PYM revient à remplacer une valeur observée en une occasion, pour un individu, par la différence de la comparaison (dépendant de L et des effectifs des groupes) de son groupe et avec les (I_A-1) autres groupes, entre cette occasion et l'occasion suivante. La métrique $Q = ({}^tMtYP_G{}^\perp YM)^-$ est l'inverse généralisé de la matrice de variance intra-groupe des p variables transformée par M (i.e des (p-1) différences successives).

Le test **H_{02}**, effectué si le test H_{01} est accepté, est :

" y - a t'il une différence entre les visites ? ".

C'est le même test que le premier que nous avons écrit mais ici il est réalisé simultanément par chaque groupe :

$$\begin{pmatrix} \mu_{1,1} \\ \cdot \\ \cdot \\ \mu_{I_G,1} \end{pmatrix} = \ldots = \begin{pmatrix} \mu_{1,p} \\ \cdot \\ \cdot \\ \mu_{I_G,p} \end{pmatrix} \quad (*)$$

On peut l'écrire de la forme :

$L\beta M = 0,$

avec $L = Id_{I_G,}$

et

$$M_{p,p-1} = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & -1 & & \\ & & & 1 \\ & & & -1 \end{pmatrix}$$

et l'ACP correspondant à **H_{02}** est celle de $(P_GY, M({}^tMtYP_G{}^\perp YM)^{-t}M, D)$.

Si H_{01} n'est pas accepté on récupère l'intéraction. En effet l'égalité (*) implique : $\mu_{1,1} - \mu_{1,2} = \ldots = \mu_{I_G,1} - \mu_{I_G,2} = 0$ ... Si l'on veut tout de même faire le test on prend $L = (1, 1, 1, \ldots 1)_{I_G}$ (premier test décrit) mais l'ACP ne donne qu'un seul axe.

Le dernier test classique que l'on effectue est le test **H_{03}** (si H_{01} est accepté) : **" y - a t'il une différence entre les groupes ? "** :

$$\begin{pmatrix} \mu_{1,1} \\ \cdot \\ \cdot \\ \mu_{1,p} \end{pmatrix} = \ldots = \begin{pmatrix} \mu_{I_G,1} \\ \cdot \\ \cdot \\ \mu_{I_G,p} \end{pmatrix} \quad (**)$$

donc avec $M = Id_p$

et

$$L_{I_G-1,I_G} = \begin{pmatrix} 1 & & -1 \\ & \cdot & \cdot \\ & 1 & -1 \end{pmatrix}$$

Le test **H_{03}** est décrit par l'ACP de $(PY, ({}^tYP_G{}^\perp Y)^-, D)$; le projecteur P est identique à celui utilisé pour le test H_{01}.

De même que pour le test H_{02} l'égalité (**) entraine l'absence d'intéraction. Si H_{01} n'est pas accepté, on pourra alors prendre $M = {}^t(1, 1, 1, \ldots 1)_p$ ou pour q variables $M = {}^t(1,1,1, \ldots 1)_p \quad Id_q$ et réaliser l'ACP de $(PY, M({}^tMtYP_G{}^\perp YM)^{-t}M, D)$

III.3.b. cas multivariable :

Maintenant nous allons étendre les résultats précédents au cas de plusieurs variables (q>1). Le modèle linéaire général s'écrit :

$$\mathbf{Y_{n,pq} = G_{n,IA}\beta_{IA,pq} + e_{n,pq}}$$

avec vec(e) ~ $N_{nqp}(0, Id_n \bullet_{qp,qp})$,

q : variables

p : répétitions

n : individus.

Boik(1988) parle de "Doubly multivariate model". Le test global prend la forme générale :

$$L\beta(M\ Id_q) = 0$$

où L : a ∞ I_A

M : p ∞ u avec rg(C) = v et $^tCC = Id_u$.

La contrainte colonne sur β s'exprime ici sous forme tensorielle ce qui signifie que l'on applique le contraste M à chacune des q variables. Les matrices "between et within" ls'écrivent alors :

$$H = {}^t(M\ Id_q)\hat{\beta}\ L)(L({}^tGG)^{-t}L)^{-}(L\hat{\beta}(M\square Id_q)\ ),$$

$$\text{et } E = {}^t(M\ Id_q){}^tY(Id_n - G({}^tGG)^{-t}G)Y(M\ Id_q)$$

$$= {}^t(Y(M\ Id_q))P_G{}^{\perp}(Y(M\ Id_q)).$$

On écrit toujours la statistique $tr(E^{-1}H)$ et l'on a les mêmes tests et donc les mêmes ACP que pour le cas univariable avec M Id_q à la place de M.

Un théorème de Khatri (1959), rappelé par Boik(1988), nous assure la validité distributionnelle du test :

$$E \sim W_{qu}(n\text{-}I_G, \Psi, 0),$$

$$H \sim W_{qu}(s, \Psi, d),$$

$$\text{avec } \Psi = {}^t(M\ Id_q)\Omega(M\ Id_q),$$

$$\text{et } d = \Psi^{-1}\Phi,$$

$$\text{où } \Phi = {}^t(M\ Id_q)\beta L)(L({}^tGG)^{-t}L)^{-}(L\beta(M\ Id_q)\ ),$$

sous les hypothèses (a) et (c) énoncées dans la partie II.

Remarques :

-1 Si l'on a plusieurs facteurs de classement on peut travailler facteur par facteur, ou bien sur le modèle complet (le design X représente le modèle) et choisir les contrastes L adaptés pour chaque effet.

On peut encore décomposer H suivant un modèle additif entre ces facteurs et garder le même E que celui calculé sur le résidu du modèle. Par exemple avec deux facteurs G_1 et G_2 orthogonaux on écrit le modèle :

$F = F_{G_1} H F_{G_2} H F_{G1.G2} H F_{(G_1\ \ G_2)^{\perp}}$ (où F est l'espace R^n et F_S le sous-espace de F engendré par les colonnes de S), (Sabatier(1987)). On décompose H sur F_{G_1}, F_{G_2} et $F_{G_1.G_2}$ et le E est calculé pour le modèle complet. Alors on a :

$$tr(H_{G_1}E_m^{-1}) + tr(H_{G_2}E_m^{-1}) + tr(H_{G_1.G_2}E_m^{-1}) = tr(H_mE_m^{-1}).$$

On peut remarquer que $tr(H_{G_1}E_m^{-1})$ n' est plus la statistique de Lawley-Hotteling. On peut aussi avoir des covariables Z et les traiter comme un facteur groupe etc... .

-2 Le cas univarié peut, comme nous l'avons déjà signalé, se présenter de cette manière, ce que fait d'ailleurs Boik(1988) mais nous remarquerons que l'approche S.U.R. est relativement semblable (partie IV). On peut noter toutefois que le contraste M, qui joue le rôle dans l'optique multivariée de "contrainte colonne" pour la répétition devra se retrouver dans l'expression de "L". On n'a plus que "L "pour traduire les tests à réaliser et il s'écrira sous forme d'un produit tensoriel de deux contrastes :

$$Y_{np,q} = (Idp\ \ G)_{np,pI_A}\beta_{pI_A,q} + \varepsilon_{np,q},$$

$$\text{et } H_0 : (M\ \ {}^tL)\beta = 0.$$

Pour ce modèle appelé par Boik M.M.M (Multivariate Mixed Model) , on a comme ACP correspondante à l'hypothèse H_0 celle du triplet :

$$(PY, ({}^tYP_{(Idp\ \ G)^{\perp}}Y)^{-}, D)$$

où $P=(Idp\ G)({}^t(Idp\ G)D(Idp\ G))^{-t}(M\ {}^tL)((M\ {}^tL)({}^t(Idp\ G)D(Idp\ G))^{-t}(M\ {}^tL))^{-}(M\ {}^tL)({}^t(Idp\ G)D(Idp\ G))^{-t}(Idp\ G)D.$

On peut alors (devant l'expression de P) préférer l'utilisation du modèle split-plot (partie I) qui pour chaque test se fonde sur un nouveau "design".

III.4 courbes de croissances :

Une autre approche dans le cadre multivarié est le modèle de courbes de croissances. Initialement présenté par Potthoff et Roy(1964) ce modèle est une extension du modèle linéaire général et s'écrit :

$$\mathbf{Y_{n,p} = X_{n,m}\ \beta_{m,f}\ \Gamma_{f,p} + e_{n,p}}$$

avec les mêmes hypothèses que précédemment.

Y représente les n observations en p instants n ∞ p,

X est la planification (design) n ∞ m,

β sont les paramètres inconnus m ∞ f,

Γ est la modélisation f ∞ p.

Dans leur article Potthof et Roy(1964) opèrent sur Y la transformation suivante qui permet de se ramener à MANOVA :

$$Y_0 = YQ_0^{-1t}\Gamma(\Gamma\ Q_0^{-1t}\Gamma)^{-1} \text{ et alors } E(Y_0) = X\beta.$$

Les choix préconisés par les auteurs pour Q_0 sont :

$Q_0 = Id_p$ ou $Q_0 = \Sigma$ ou $\hat{\Sigma}$. On a alors une double contrainte sur Y. Pour le test global LßM' = 0 on effectue :

l'ACP de $(PY_0, M'({}^tM'^tY_0 P_X{}^{\perp}Y_0 M')^{-t}M', D)$.

Cette ACP est équivalente à :

(a) l'ACP de $(PY{}^tP_{t_\Gamma}, ({}^tYP_X{}^{\perp}Y)^{-}, D)$ si M'=Idp

(b) l'ACP de $(PY{}^tP_{t_\Gamma}, M({}^tMP_{t_\Gamma}{}^tYP_X{}^{\perp}Y{}^tP_{t_\Gamma}M)^{-t}M, D)$ si M'= ΓM, (c'est le test global pour la matrice $(Y{}^tP_{t_\Gamma})$).

(a) est une ACPVI double contrainte si on choisit $Q_0 = {}^tYP_X{}^{\perp}Y$ pour la cohérence des métriques dans l'écriture du projecteur P_{t_Γ}. Dans la littérature Lee(1974) et D'Aubigny(1989) mettent l'accent sur la double contrainte. Lee montre l'équivalence avec le modèle de RAO(1965) pour $Q_0 = {}^tYP_X{}^{\perp}Y$ c'est à dire le choix donnant l'ACPVI double contraintes et l'équivalence précédente devient :

l'ACP de $(PY_0, ({}^tY_0 P_X{}^{\perp}Y_0)^{-}, D)$

<=> l'ACP de $(PY{}^tP_{t_\Gamma}, Q_0^{-}, D)$.

D'Aubigny (1989) en écrivant le modèle sous forme vectorialisée :

$\dot{Y} = (X \otimes^{t} \Gamma)\dot{\beta} + \dot{\varepsilon}$, aboutit par simple résolution des moindres-carrés à l'estimation de Y par $P_XY{}^tP_{t_\Gamma}$.

Remarques :

-1 On a le résultat de Khatri(1966) qui montre que :

$({}^tXDX)^{-t}XDYQ_0^{-t}\Gamma(\Gamma Q_0^{-t}\Gamma)^{-}$ est l'estimateur du maximum de vraisemblance de β pour le choix $Q_0 = {}^tYPX{}^{\perp}Y$.

-2 La modélisation (choix de Γ)est en général faite en utilisant des polynômes orthogonaux, mais on pourrait imaginer d'autres familles de fonctions orthogonales par exemple des splines polynômiales ou encore les M-splines ou I-splines. El-Faouzi et Escoufier(1990) donnent un exemple d'utilisation de ces dernières pour des courbes de croissances monotones.

-3 L'aspect double contrainte peut être vu comme un changement de métriques. Le fait de considérer des courbes peut nous amener à l'approche de Besse et Ramsay(1986) et Houiller(1987) qui ont montré que l'ACP des courbes (dans un sous-espace de L_2) est identique à l'ACP des données discrètes muni d'une métrique dépendant du sous-espace dans lequel la base de courbes est plongée (matrice liée au noyau autoreproduisant de l'espace).

-4 Dans le test précédent on peut remarquer qu'une fois la modélisation faite il n'y a plus lieu d'utiliser le contraste M et donc le cadre d'ACPVI double contrainte ou non ne dépend que du choix de Q_0.

-5 On peut écrire le même modèle pour q variables :

$$Y = X\beta(Id_q\ \Gamma) + \varepsilon,$$

si toutes les variables sont modélisées de la même manière (sinon mettre une matrice bloc-diagonale de blocs Γ_i) et écrire les mêmes tests avec les mêmes ACP.

-6 De nombreux tests dans des cas particuliers de structure d'autocorrélation ou bien de design ont été développés dans la littérature et on peut citer Rao(1967), Khatri(1973), Kleibaum(1973) qui étend le modèle de Potthof et Roy au cas de données manquantes en écrivant :

$E(Y_j) = X_j\beta PB_j$ $Y_j : n_j \infty p_j$

$var(Y_j) = Id_{nj}\ {}^tB_j\Sigma B_j$ j= 1..u (les n individus étants regroupés en u classes "homogènes").

-7 Chakravorti(1974) teste l'égalité des courbes de croissances (paramètres) dans la problématique de Behrens-Fisher. Plus récemment certains auteurs se sont intéressés à la prédiction dans les modèles de courbes de croissance, pour des observations nouvelles aussi bien pour des individus nouveaux sur les mêmes occasions que sur les mêmes individus pour des occasions suivantes, on peut citer Rao(1987),Lee(1988).

IV . Une approche univariée et multivariée

Une autre approche introduite en économétrie est celle liée aux équations multiples et/ou simultanées. C'est sous la dénomination de "Seemingly Unrelated Regressions" que Zellner(1962) en parle. Cette méthode se rapproche de l'optique anova split-plot dans le sens où l'on modélise la ou les variables en tant que vecteur à np dimensions, mais le design rend compte d'une optique multivariée.

Stanek et Koch(1985) établissent des liens entre le modèle de Potthof et Roy et les S.U.R. en utilisant les modèles conditionnels de Rao(1965, 66, 67) et Khatri(1966). Cette optique est développée dans Verbyla(1988) et Verbyla et Venables(1988) pour un modèle de somme de profils ($y = \sum_{k=1}^{r} X_k\beta_k P_k + \varepsilon$).

L'intérêt de cette méthode réside dans la souplesse du design données manquantes, changement de groupe etc...) et le calcul du ß par moindres carrés pondérés. Une revue des différentes estimations de ß est faite dans Srivastava et Dwivedi(1979).

IV.1. présentation univariable du S.U.R. :

On dispose de p équations :

$$y_i = x_i\beta_i + \varepsilon_i \text{ , } i = 1....p \text{ ,}$$

y_i ($n \infty 1$), n observations,
x_i ($n \infty l_i$),
βi($l_i \infty 1$).
Chaque ε_i est normal, $E(\varepsilon_i)=0$, $E(\varepsilon_i{}^t\varepsilon_j)=\sigma_{ij}Id_n$ avec $\Sigma_{p,p}=(\sigma_{ij})$.

Alors en re-écrivant le modèle sous la forme :

$$Y_{np,1} = X_{np,(l_1+...+l_p)}\,\beta_{(l_1+...+l_p),1} + \varepsilon_{np,1},$$

on a :
$\hat{\beta}_{mco} = ({}^tXX)^{-1}{}^tXY$

$\hat{\beta}_{mcg} = ({}^tX(\Sigma^{-1} \otimes Id_n)X)^{-1}{}^tX(\Sigma^{-1} \otimes Id_n)Y.$

L'estimateur de Zellner est le $\hat{\beta}_{mcg}$ dans le cas ou Σ est inconnue et estimée en utilisant une première estimation des β_i par moindres-carrés ordinaires :

$\hat{\beta}_z = ({}^tX(S^{-1} \otimes Id_n)X)^{-1}{}^tX(S^{-1} \otimes Id_n)Y$
avec $S=(s_{ij})$,
où $s_{ij} = 1/n\,{}^ty_i[Id_n - x_i({}^tx_ix_i)^{-1}{}^tx_i][Id_n - x_j({}^tx_jx_j)^{-1}{}^tx_j]y_j$.

Remarques:

-1 Si dans l'expression de s_{ij} le dénominateur (i.e n) est remplacé par ($n-l_i-l_j+tr(P_{X_i}P_{X_j})$) l'estimateur est sans biais. Si il est remplacé par ($n-l_i-l_j+tr(P_{Xi}P_{Xj})+2$) alors l'estimateur est de variance minimum (Stroud, Zellner et Chau(1963)).

-2 En supposant que l'erreur est stationnaire auto-régressive du premier ordre :
$\varepsilon_{it} = \delta_i\varepsilon_{i(t-1)} + u_{it}$ $\delta_i<1$ i=1.....p
où $E(u_{it})=0$ et $E({}^tu_{it}u_{j(t+s)})=\theta_{ij}$ si s=0
= 0 sinon,
Parks(1967) propose l'estimateur suivant de β :
$\hat{\beta}_p = ({}^tX(\hat{\theta}^{-1} \otimes D^tD)^{-1}X)^{-1}{}^tX(\hat{\theta}^{-1} \otimes D^tD)^{-1}Y$
avec $D={}^t(D_1,D_2,...D_p)$

$$Di = \begin{pmatrix} \hat{\delta}_i^{-1/2}(1-\hat{\delta}_i)^{-1/2} & 0 & . & 0 \\ \hat{\delta}_i(1-\hat{\delta}_i)^{-1/2} & 1 & 0 & . & 0 \\ \hat{\delta}_i^{2}(1-\hat{\delta}_i)^{-1/2} & \hat{\delta}_i & 1 & . & 0 \\ . & . & & . & . \\ \hat{\delta}_i^{n-1}(1-\hat{\delta}_i)^{-1/2} & \hat{\delta}_i^{n-2} & & . & 1 \end{pmatrix}$$

où $\hat{\delta}_i = \sum_{t=2}^{n} \hat{\varepsilon}_{it}\hat{\varepsilon}_{it-1} / \sum_{t=1}^{n-1} \hat{\varepsilon}_{it}{}^2$,

$\hat{\theta}_{ij} = [\,{}^t\hat{\varepsilon}_i{}^tD_i^{-1}D_j^{-1}\hat{\varepsilon}_i] / \sqrt{(n-di)}\sqrt{(n-dj)}$,

dans laquelle $\hat{\varepsilon}_{it} = (Id_n - x_i({}^tx_ix_i)^{-1}{}^tx_i)y_i$. L'estimateur est consistant et plus efficace que $\hat{\beta}_z$.
Kmenta et Gilbert(1973) construisent un autre estimateur avec l'erreur suivante :

$\varepsilon_{it} = \sum_{j=1}^{p} (\delta_{ij}\varepsilon_{jt-1}) + u_{it}.$

IV.2 cas multivariable et ACP :

On peut généraliser à q variables en écrivant le modèle:

$$Y_{np,q} = X_{np,(l_1+...+l_p)}\,\beta_{(l_1+...+l_p),q} + \varepsilon_{np,q} \text{ ,}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ Y_p \end{pmatrix} = \begin{pmatrix} X_1 & & & \\ & X_2 & & \\ & & . & \\ & & & X_p \end{pmatrix} \beta + \varepsilon \text{ .}$$

On peut alors utiliser les mêmes modèles que précédemment, pour chaque variable, puis juxtaposer les variables ajustées (les $X\beta^1...X\beta^q$) (les β sont soit des $\hat{\beta}_z$ soit des $\hat{\beta}_p$ ...) puis en faire l'ACP.

C'est à dire que l'on effectue l'ACP de ($(X\beta^1...X\beta^q)$,Q,D). Ou bien chercher un compromis sur l'erreur pour construire la matrice $S_{(q)}$ et on choisit comme estimateur de β:

$\hat{\beta}_{z(q)} = ({}^tX(S_{(q)}^{-1} \otimes Id_n)X)^{-1}{}^tX(S_{(q)}^{-1} \otimes Id_n)Y.$
On réalise alors l'ACP : (P_XY , Q , $S_{(q)}^{-1} \otimes Id_n$), ou encore utiliser l'estimation de Parks par compromis sur l'erreur.

En suivant les tests introduits dans la partie III (les choix des contrastes sont évidemment différents) on peut effectuer certains tests sur β et en réaliser les ACP correspondantes en utilisant la métrique D ainsi construite. Celles-ci sont différentes de celles de la partie I ou III de part la métrique D et la structure du design (ou covariables).

En explicitant par rapport à notre exemple théorique, on peut déjà remarquer que l'on peut, avec ce modèle, réaffecter les individus dans les classes du groupe G à chacune des p visites, ainsi on peut gagner en souplesse d'analyse, et, en pratique constater un aspect "dynamique" de l'effet de ce facteur.

Remarques :

-1 On peut remarquer que si l1=...=lp, X est l'effet de Z V où Z est la matrice obtenue en juxtaposant les matrices x_i et Z est le design identifiant la répétition. De plus :

$$P_{\begin{pmatrix} x_1 & & & \\ & x_2 & & \\ & & . & \\ & & & x_p \end{pmatrix}} \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ Y_p \end{pmatrix} = \begin{pmatrix} P_{x_1}Y_1 \\ P_{x_2}Y_2 \\ . \\ P_{x_p}Y_p \end{pmatrix} \neq P_{\begin{pmatrix} x_1 & & & \\ & x_2 & & \\ & & . & \\ & & & x_p \end{pmatrix}} \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ Y_p \end{pmatrix}.$$

L'expression à droite du signe différent est l'estimation de Xß dans le modèle univarié (split-plot).

-2 On peut aussi écrire le modèle suivant, différent du modèle S.U.R. et du modèle multivarié (bien que les hypothèses à faire pour les tests soient les mêmes) :

$$\begin{pmatrix} Y_1 & & & \\ & Y_2 & & \\ & & . & \\ & & & Y_p \end{pmatrix} = \begin{pmatrix} X_1 & & & \\ & X_2 & & \\ & & . & \\ & & & X_p \end{pmatrix} \beta + \varepsilon \text{ .}$$

Alors l'estimation de ce modèle est :

$$P\begin{bmatrix} X_1 & & & \\ & X_2 & & \\ & & . & \\ & & & X_p \end{bmatrix}\begin{pmatrix} Y_1 & & & \\ & Y_2 & & \\ & & . & \\ & & & Y_p \end{pmatrix} = \begin{pmatrix} P_{X_1}Y_1 & & & \\ & P_{X_2}Y_2 & & \\ & & . & \\ & & & P_{X_p}Y_p \end{pmatrix}.$$

On peut noter que les vecteurs propres issus de l'ACP de cette matrice sont des juxtapositions d'un vecteur propre issu d'une ACP d'un $P_{Xi}Y_i$ et de vecteurs nuls ailleurs.

Notons que si il y a des valeurs propres communes on aura une multiplicité de ces valeurs propres. Une combinaison linéaire des vecteurs propres ayant cette valeur commune sera alors aussi vecteur propre, avec comme valeur propre la valeur propre mutiple divisée par la norme de la combinaison linéaire.

V Conclusions et remarques

Avant de conclure on peut noter que nous n'avons pas parlé du cas où il y a plusieurs facteurs répétés ("within subject factor"). Cette approche peut s'inclure dans la partie II ou la partie III, et donner par exemple pour deux facteurs répétés des modèles appelés "split-split-plot" Winer(1971) (partie II). L'on peut aussi imaginer des "doubly doubly mutivariate model" (partie III, il faut faire attention à l'écriture des contrastes !). On peut aussi réaliser une analyse qui amènerait à une double contrainte de répétition (ligne et colonne) en écrivant un modèle qui soit un modèle split-plot pour le premier facteur et "doubly multivariate " pour le second.

Récemment Yoshimasa et al.(1992) ont développé des tests invariants UMP pour des modèles linéaires de mesures répétées où les matrices de covariance ont une propriété (de connaitre la matrice de ces vecteurs propres et leurs multiplicités mais pas les valeurs propres) dont la matrice de compound symétrie fait partie.

Nous avons dans ce chapitre envisagé l'utilisation de l'analyse des données et plus particulièrement de l'ACP dans des modèles statistiques classiques utilisés pour les facteurs à mesures répétées. Mais il existe aussi une littérature nombreuse concernant le traitement de tableaux multiples ou de "cubes" de données que nous discuterons dans la partie B.
Sous certaines hypothèses, on peut tester, avec par exemple avec la statistique de Roy, chaque valeur propre obtenue lors d'une méthode factorielle.

On pourra ainsi après une analyse qui donnera un sens à l'une des trois hypothèses classiques ($\mathbf{H_{01}}$, $\mathbf{H_{02}}$, $\mathbf{H_{03}}$), tester les axes obtenus et retrouver l'objectif du statisticien inférentiel.

Suivant le domaine d'application : épidémiologie, biologie, agronomie, psychologie...mais aussi le type d'étude réalisée, en analyse de variance des mesures répétées, les hypothèses concernant la collecte des données peuvent induire plus une approche plutôt qu'une autre, comme Koch et al.(1980) l'ont souligné (c.f. tableau de l'introduction).

L'utilisation des ACP décrites dans ce chapitre induit non-seulement une illustration des tests effectués, mais aussi une critique de ces tests car on peut détecter ce qui fait "marcher le test" à tort ou à raison avec éventuellement des anomalies. Donc chaque approche peut orienter le statisticien à reconsidérer le modèle choisi ou les données utilisées.

Pour la mise en oeuvre des méthodes décrites dans les parties précédentes il faudra avoir un programme d'ACP avec métriques non-diagonales.
Nous avons, pour notre part, écrit le programme (non presenté en annexe) avec SAS/IML qui bénéficie des convivialités du langage matriciel mais est limité pour la taille des données. Pour l'approche univariée on peut aussi si l'on a que des métriques diagonales, utiliser d'autres logiciels tels que Biomeco ou A.D.E. (Chessel et Doledec(1992)) qui possédent la régréssion multivariable et les ACP avec métriques diagonales. Espérons voir prochainement dans ces logiciels ou sur gros systèmes des ACP à métriques non-diagonales.

VI Références bibliographiques

BARCIKOWSKI, R.S. and ROBEY, R.R. (1983) Univariate and Multivariate repeated measures analysis through PROC's REG, GLM and MATRIX. In: Proceeding of the SUGI conference, SAS Institute, Cary, NC.

BARCIKOWSKI, R.S. and ROBEY, R.R. (1984) Decisions in single group repeated measures analysis: Statistical tests and three computer packages. Am. Stat. 148-150.

BERENBLUT, I.I. and WEEB, G.I. (1974) Experimental design in the presence of autocorrelated errors. Biometrika, 427-437.

BERK, K. (1985) Computing for unbalanced repeated measures experiments. In: Proceeding of the SUGI conference, SAS Institute Cry, NC.

BESSE, P. and RAMSAY, J.O. (1986) Principal components analysis of sampled functions. Psychometrika, 285-311.

BESSE, P. (1987) Choix de la métrique pour l'ACP de series d'évènements discrets. S.A.D. 1-16.

BOIK, R.J. (1981) A priori tests in repeated measures designs : effects of nonsphericity. Psychometrika, 241-255.

BOIK, R.J. (1988) The mixed model for multivariate repeated measures: validity conditions and an approximate test. Psychometrika, 489-486.

BRILLINGER, D.R. (1984) Analysis of variance and problems under time series models. In: Handbooks of Statistics. Krishnaiah, P.R. (ed) North-Holland Publishing Company, Amsterdam, New-York, vol.1.

BURMAN, P. (1991) Regression function estimation from dependent observations. J.Mult.An. 263-279.

BYRNE, P.J. and ARNOLD, S.F. (1983) Inference about multivariate means for a nonstationary autoregressive model. Journal of the American Statistical Association, 850-855.

CAUSSINUS, H. and FERRE, L. (1989) Analyse en Composantes Principales d'individus définis par un modèle. S.A.D. 19-28.

CHAKRAVORTI, S.R. (1974) On some tests of growth curve model under Behrens-Fisher situation. J.Mult.An. 31-51.

CHESSEL, D. et DOLEDEC, S.(1992) A.D.E.software (version 4.3) Multivariate analysis and graphical display for environmental data. Chessel, D et Dolédec, S, URA CNRS 1451, Lyon.

CHURCH, A.Jr. (1966) Analysis of data when the response is curve . Technometrics, 229-246.

CLEROUX, R. and DUCHARME, G. (1986) Vector correlation for elliptical distributions. Rapport n° 586, Département Informatique et Recherche Opérationnelle, Université de Montreal.

CORNELIUS, P.L. (1979) Curve fitting by regression on smoothed singular vectors. Biometrics, 849-859.

DEVILLE, J.C. (1974) Méthodes statistiques et numérique de l'analyse harmonique. Ann. INSEE , 3-101.

D'AUBIGNY, G. (1989) L'analyse multidimensionnelle des tableaux de dissimilarité. , Thèse de doctorat Grenoble.

DUNCAN, G.M. (1983) Estimation and inference for heteroscedastic systems ofequations. Int.Eco.Rev. 559-566.

EL FAOUZI, N. and ESCOUFIER, Y. (1991) Modélisation de courbes de croissance par les I-splines. R.S.A. XXXIX, 51-64.

ESCOUFIER, Y. (1973) Le traitement de variables vectorielles. Biometrics (29), 4, 751-760.

ESTEVE, J. and SHIFFLERS, E. (1976) Discussion et illustration de quelques méthodes d'analyse longitudinale. In : Proceeding of the 9th International biometric conference, biometric society, Boston 1976, 463-358.

GABRIEL, K.R. (1961) The model of ante-dependence for data of biological growth. Bull. I.S.I. 253-264.

GABRIEL, K.R. (1962) Ante-dependence analysis of an ordered set of variables. Ann. Math. Stat. 201-212.

GOLDSTEIN, H. and McDONALD, P. (1988) A general method for the analysis of multilevel data. Psychometrika, 445-467.

GRIEVE, A.P. (1984) Tests of shericity of normal distributions and the analysis of repeated measures designs. Psychometrika, 257-267.

GRIZZLE, J.E. and ALLEN, D.M. (1969) Analysis of growth and dose response curves. Biometrics, 357-381.

HAND, D. J. and TAYLOR, C.C. (1987) Multivariate analysis of variance and repeated measures. Chapman and Hall (eds), London, New-York.

HANNAN, E.J. (1967) Canonical correlation and multiple equation systems in economics. Econometrica (35), 123-138.

HOULLIER, F. (1987) Comparaison de courbes et de modèles de croissance choix d'une distance entre individus. S.A.D. 17-36.

HUYNH, H. and FELDT, L.S. (1970) Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. J.A.S.A. 1585-1589.

JORESKROG, K.G. (1970) A general method for analysis of covariance structures. Biometrika, 239-251.

JOERESKOG ,K G. (1986) Analysis of longitudinal data with LISREL. In: Statistical Software (3rd Conference on the use of statistical software) , Lehmacher, W et Hoermann, A (eds), New-York.

KANG, G. and BATES, D.M. (1990) Approximate inferences in multiresponse regression analysis. Biometrika, 231-331.

KENWARD, M.G. (1987) A method for comparing profiles of repeated measurements. Appl. Stat. 296-308.

KESELMAN, H.J. and KESELMAN, J. C. (1984) The analysis of repeated measures designs in medical research. Stat.in Med. 185-195.

KHATRI, C.G. (1966) A note on MANOVA model applied to problems in growth curves. Stat.Math. 75-86.

KHATRI, C.G. (1973) Testing some covariance structure under a growth curve model. J.Mult.An. 102-116.

KOCH, G.G. AMARA, I.A. STOKES, M.E. and GILLINGS, D.B. (1980) Some views on parametric and non-parametric analysis for repeated measurements and selected bibliography. Int.Eco.Rev. 249-265.

LAVIT, C. (1988) Analyse conjointe de tableaux quantitatifs.Masson, Paris.

LEBRETON, J.D. ROUX, M. BANCO,G. et BACOU, A.M. (1992) logiciel BIOMECO (version 4.1) Analyse statistique et modélisation des processus écologiques. CEFE-CNRS, Montpellier.

LECOUTRE, B(1991) A Correction for ε (Huynh et Feldt) approximate test in repeated measures designs with two or more independent groups . J.Edu.Stat. 371-372.

LECOUTRE, B. and ROUANET, H. (1981) Deux structures statistiques fondamentales en analyse de la variance univariée et multivariée. Math.Sci.Hum. 71-82.

LEE, J.C. (1988) Prediction and estimation of growth curves with special covariance structures. Journal of the American Statistical Association 83(402), 432-440.

LEE, Y.K. (1974) A note on Rao's reduction of Potthoff & Roy'sgeneralized linear model. Biometrika, 349-352.

LEIBOVICI, D. BUCQUET, D. SABATIER, R. CURTIS, S. and COLVEZ, A. (1990) Data analysis with dependent structure applied to epidemiological survey data in gerontology. In: 11th Meeting of the International Society for Clinical Biostatistics, Volume d'abstracts, 18-21 Sept Nîmes.

LEIBOVICI, D. (1992) Modèles Linéaires et Analyses Factorielles pour l'Analyse de Facteurs à Mesures Répétées. In : A.S.U. XXIVe Journées de Statistique Bruxelles.

LOONEY, S.W. and STANLEY, W.B. (1989) Exploratory repeated measures analysis for tow or more groups, review and update. Am. Stat. 220-225.

MAUCHLY, J.W. (1940) Significannce test for sphericity of a n-variate distribution. An.Math.Stat. 204-209.

MENDOZA, J.L. (1980) A significant test for multisample sphericity. Psychometrika, 495-498.

MONLEZUN, C. and BLOUIN, D. (1988) A general nested split-plot analysis of covariance. J.A.S.A. 818-823.

MORRISON, D.F. (1970) The optimal spacing measure. Biometrics, 281-290.

MORRISON, D.F. (1972) The analysis of single sample of repeated measurements. Biometrics, 55-71.

PATEL, H.I. ROY, T. and HUQUE, M.F. (1985) Some preliminary tests in repeated measures design. In: . Proceeding of the SUGI conference, SAS Institute, Cary, NC.
PATEL, H.I. (1986) Analysis of repeated measures designs with changing covariates in clinical trials. Biometrika, 707-715.
PONTIER, J. DUFOUR, A.B. and NORMAND, M. (1990) Le modèle euclidien en analyse des données. Editions Ellipses, Paris.
POTTHOF, R.F. and ROY, S.N. (1964) A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 313-626.
RAO, C.R. (1964) The use and interpretation of principal component analysis in applied research. Sankhya A, 329-359.
RAO, C.R. (1965) The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. Biometrika, 447-458.
RAO, C.R. (1967) Least square theory using an estimated dispersion matrix and its application to measurement of signals. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics, Universi Press, B, 355-372.
RAO, C.R. (1972) Recent trends of research work in multivariate analysis. Biometrics, 3-22.
RAO, C.R. and YANAI, H. (1979) General definition and decomposition of projectors and some applications to statistical problems. J.Stat.Plan.Infe., 3, 1-17.
RAO, C.R. (1987) Prediction of future observations in growth curve models. Stat.Sci. 434-471.
RAZ, J. (1989) Analysis of repeated measurements using nonparametric smoothers and randomization tests. Biometrics, 851-871.
REINSEL, G. (1982) Multivariate repeated-measurement or growth curve model with multivariate random-effects covariance structure. J.A.S.A. 190-195.
REVANKAR, N.S. (1974) Some finite sample results in the context of two seemingly unrelated regression equations. J.A.S.A. 187-190.
ROBERTS, J.S. and LAUGHLIN, J.E. (1983) Analyzing anlysis of covariance models in univariante repeated measures designs using SAS. In: . Proceeding of the SUGI conference, SAS Institute, Cary, NC.
ROUANET, H. and LEPINE, D. (1970) Comparison between treatments in a repeated-measurement design : anova and multivariate methods. Brit.J.Math.Stat.Psychol. 147-163.
SABATIER, R. (1987) Méthodes factorielles en analyse des données : approximations et prise en compte de variables concomitantes. Thèse doct.ès-sciences, U.S.T.L., Montpellier
SABATIER, R. LEBRETON, J. and CHESSEL, D. (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. In: Multiway data analysis, Coppi, R et Bolasco, S (eds), North-Holland, Amsterdam, 341-352.
SCHATZOFF, M. (1966) Sensitivity comparisons among tests of general linear hypothesis. J.A.S.A. (part 1), 415-437.
SINGH, S. and ULLAH, A. (1974) Estimation of seemingly unrelated regressions with random coefficients. J.A.S.A. 191-195.
SNEE, R.D. (1972) On the analysis of response curve data. Technometrics 14 (1) ,47-62.
SNEE, R.D. , ACUFF, S.K. and GIBSON, J.R. (1979) A useful method for tha analysis of growth studies. biometrics, 835-848.
SPECTOR, P. (1985) Repeated Measures Analysis in PROC GLM. In: . Proceeding of the SUGI conference, SAS Institute, Cary, NC
SRIVASTAVA, V.K. and DWIVEDI, T.D. (1979) Estimation of seemingly unrelated regression equations. Jo.Eco. 15-32.
SRIVASTAVA, M.S. and CARTER, E.M. (1983) An introduction to applied multivariate statistics. Elsevier Science Pub Co, New York.
STANEK, E.J. and KOCH, G.J. (1985) The equivalence of parameter estimates from growth curve models and seemingly unrelated regression models. Am. Stat. 149-152.
STANEK, E.J. and DIEHL, S.R. (1988) Growth curve models of repeated response. Biometrics, 973-983.
TANDON, P.K. and MOESCHBERGER, M.L. (1983) The SAS macros for nonparametric analysis of repeated measures designs. In: Proceeding of the SUGI conference, SAS Institute, Cary, NC

THOMAS , D.R. (1983) Univariate repeated measures techniques applied to multivariate data. Psychometrika, 451-464.
TIMM, N H. (1984) Multivariate analysis of variance of repeated measurements.In: Handbooks of Statistics Krishnaiah, P R (ed) North-Holland Publishing Company, Amsterdam, New-York, vol.1, 41-87.
VERBYLA, A.P. (1988) Analysis of repeated measures designs with changing covariates. Biometrika, 172-174.
VERBYLA, A.P. and VENABLES, W.N. (1988) An extension of the growth curve model. Biometrika, 129-138.
WARE, J.H. (1985) Linear models for the analysis of longitudinal studies. Am. Stat. 95-101.
WINER, B.J. (1971) Statistical principle in experimental design. Mc Graw-Hill Book Company, New-York, Toronto.
WISHART, J. (1938) Growth rate deteermination in nutrition studies with the baccon pig and their analysis. Biometrika, 16-28.
WRIGHT, P.S. (1982) Repeated measures ANOVA : The multivariate approach. In: Proceeding of the SUGI conference, SAS Institute, Cary, NC.
YATES, F. (1982) Regression models for repeated measurement. Biometrics, 850-853.
YOSHIMASA, U. KAZUO, N. and ETSUO, M. (1992) UMP invariants tests for generalized linear model. J.Mult.An. 1-12.
ZELLNER, A. (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J.A.S.A. 348-368.
ZERBE, G.O. and WALKER, S.H. (1977) A randomization test for comparison of groups of growth curves with different polynomius design curve. Biometrics, 653-657.
ZERBE, G.O. and JONES R, H. (1980) On application of growth curve techniques to time series. J.A.S.A. 507-509.
ZUPKIS, R.V. and SHARP, M. (1982) SAS macro "REP2" for Unweighted means analysis of repeated measures. In: Proceeding of the SUGI conference, SAS Institute, Cary, NC.

AII
Modèles Linéaires et log-linéaires pour variables qualitatives

I . Introduction

Souvent en épidémiologie, les variables traitées sont qualitatives ou qualitatives ordinales.

D'une part les mesures biologiques sont très fluctuantes ; des étalonnages permettent les diagnostics et les catégorisations des individus. La réponse est donc qualitative. Ayant une catégorisation "fixe" des individus, on souhaite alors comparer ces deux groupements que sont la réponse et le facteur "fixe", et ceci soit lors d'une expérience soit lors d'expériences répétées.

D'autres part la pratique de questionnaires en épidémiologie est très courante dans les enquêtes transversales et de plus en plus lors d'enquêtes longitudinales.

On a donc une, ou plusieurs, réponses qualitatives d'expériences répétées et une réponse fixe (ou supposée fixe) sur la période étudiée qui constitue le facteur groupe. On souhaite toujours tester les trois hypothèses fondamentales H_{01}, H_{02}, H_{03} .

Une importante littérature traite des modèles log-linéaires, la référence classique étant Bishop et al.(1975). Imrey et al.(1981) les ont discutés avec notamment un examen de la régression logistique fréquemment utilisée en épidémiologie.

Dans ce chapitre, nous allons examiner plus particulièrement le modèle linéaire pour des données qualitatives présenté par Grizzle et al.(1969) pour appliquer celui-ci aux données répétées. Koch et al.(1977) ont abordé le sujet en proposant les tests adéquats aux trois hypothèses précédentes uniquement pour des proportions.

Nous verrons comment mettre en oeuvre cette méthodologie lorsque l'on dispose de plusieurs variables qualitatives, pour envisager une formulation factorielle sous forme d'ACPVI.

On rappellera alors les liens et approches similaires des modèles log-linéaires et de l'AFC qui ont été discuté notamment par Goodman(1986).

D'autres méthodes comme les chaînes de Markov sont fréquemment utilisées pour des variables qualitatives répétées, nous en discutons quelque peu dans le chapitre BII et l'on pourra aussi consulter aussi Ware et al.(1988) pour un aperçu de des diverses techniques utilisées en épidémiologie.

II . Méthode G.S.K. et ACP

II.1 Modèle général pour une variable qualitative :

L'approche générale pour aborder l'analyse de variables qualitatives établie par Grizzle Starmer et Koch (1969) a prit le nom de méthode G.S.K.. Cette méthode est fondée sur les modèles linéaires et la "delta-method" pour approximer de la matrice de covariances de fonctions non-linéaires.

Cette dernière que l'on peut consulter plus précisément dans Bishop et al.(1975) assure la consistance de l'estimation de la matrice de variances et covariances de fonctions des proportions.

Le problème général se pose de la façon suivante :

-on mesure une variable qualitative Y à r modalités sur une population groupée en s sous-populations,

-on observe alors pour chaque sous-population un vecteur r-dimensionnel $n_i={}^t(n_{i1},\ldots,n_{ir})$ contenant les nombres de réponses de chaque modalité,

-chaque vecteur observé est supposé suivre une loi multinomiale à r probabilités $\pi_{i1},\ldots,\pi_{ir}$ (qu'un individu de la i ème sous-population réponde l'une des modalités de Y).

Le modèle global est donc un produit de loi multinomiales. Les proportions observées dans chaque sous-population sont les estimations par maximum de vraisemblance des probabilités π_{ij} i=1,...,s et j=1,...,r.

Le modèle étudié est alors de la forme :

soit ${}^tF(\pi)=(f_1(\pi),f_2(\pi),\ldots,f_u(\pi))$ où les f_k sont des fonctions réelles deux fois continuement différentiables) et $u \le (r-1)s$

X un tableau u × t exprimant le plus souvent le "design " du modèle .
On souhaite trouver par estimation des moindres carrés β tel que

$$F(\pi)=X\beta$$

F(p) étant l'espérance asymptotique de F(π), on effectue une estimation par les moindres carrés pondérés de β par :

$$\hat{\beta}=[{}^tXS^{-1}X]^{-1}{}^tXS^{-1}F(p)$$

où S est l 'estimation de la matrice de variance et covariance de F (π) :

$S=HV_p{}^tH$ avec la matrice u × sr , $H=\left(\frac{\partial f_k(\pi)}{\partial \pi_{ij}}\right)_{\pi_{ij}=p_{ij}}$ $k=1,\ldots,u$

et l'estimation de la matrice de variance et covariance de π

$$V_p=\begin{pmatrix} V_{p_1} & 0 & 0 \\ 0 & O & 0 \\ 0 & 0 & V_{p_s} \end{pmatrix} \text{ où } V_{p_i}=\frac{1}{n_{i.}}\left(\begin{pmatrix} p_{i1} & 0 & 0 \\ 0 & O & 0 \\ 0 & 0 & p_{ir} \end{pmatrix}-p_i{}^tp_i\right).$$

$\hat{\beta}$ est le meilleur estimateur asymptotiquement normal (BAN) de β.

Le test d'adéquation du modèle est $SS_0=({}^tF(p)S^{-1}F(p)-{}^t\hat{\beta}{}^tXS^{-1}X\hat{\beta})$ qui suit asymptotiquement une loi de χ^2 à (u-t) degrés de liberté.

Le test général de l'hypothèse :

$$H_C : C\beta=0$$

où C est un contraste de rang c se fait par la statistique $SS_C={}^t(C\hat{\beta})[C({}^tXS^{-1}X)^{-1}{}^tC]^{-1}(C\hat{\beta})$ qui suit asymptotiquement une loi de χ^2 à (c) degrés de liberté.

Remarques :

-1 La situation est assez différente d'un modèle linéaire classique car en combinant les fonctions f_k et le design X (souvent un "faux" design) on sera pratiquement toujours amené à tester β=0. Ceci est très caractéristique lorsque l'on a des mesures répétées comme on va le voir maintenant.

On peut parfois, en choisissant bien les fonctions f_k avoir le modèle F(π)=0 et n'utiliser que le test global d'adéquation. C'est à dire la statistique $SS_{00}=({}^tF(p)S^{-1}F(p))$ qui suit un χ^2 à (u) degrés de liberté.

-2 Pour faire une analogie avec le chapitre AI où l'on a distingué l'approche univariée et multivariée on peut dire que l'on a ici une approche univariée.

II.2 Modèle général pour une variable qualitative répétée :

Pour une variable qualitative répétée le modèle est le même que précédemment avec la variable qualitative définissant les profils observables.

C'est à dire qu'une variable à l modalités observées d fois constitue l^d profils de réponses possibles. Le modèle général utilisé par Koch et al.(1977) est donc le modèle GSK avec la variable à $r=l^d$ modalités.

Remarques :

-1 On a donc une approche univariée des mesures répétées d'une variable qualitative, c'est à dire un Split-Plot comme le notent Koch et Reinfurt(1971). L'analyse restera faisable pour des nombres de modalités de la variable et de répétitions assez faibles, sinon le design peut devenir une matrice très importante ainsi que S.

Le problème de zéros structurels peut alors arriver, mais il sera ici encore facile de les écarter par le choix de F.

Pour les trois hypothèses $\mathbf{H_{01}}$, $\mathbf{H_{02}}$, $\mathbf{H_{03}}$, on va choisir X et le contraste C pour effectuer les tests voulus. Ces choix sont décrits dans Koch et al.(1977). Ils dépendent fortement de la structure des données et nous ne les donnerons pas ici. Les hypothèses se traduisent de la façon suivante:

$\mathbf{H_{01}}$: $\pi_{1(t+1)f}-\pi_{1tf}=\ldots=\pi_{s(t+1)f}-\pi_{stf}$ pour t=1,...(d-1) f=1,...,L,

$\mathbf{H_{02}}$: $\pi_{i1f}=\pi_{i2f}\ldots=\pi_{idf}$ pour i=1,...s et f=1,...,L si $\mathbf{H_{01}}$ est vrai,

$\mathbf{H_{03}}$: $\pi_{1j}=\pi_{2j}\ldots=\pi_{sj}$ pour j=1,...,r si $\mathbf{H_{01}}$ est vrai,

Remarque :

-1 La procédure CATMOD de SAS permet de réaliser ces analyses en utilisant les moindres carrés pondérés mais permet aussi l'estimation par maximum de vraisemblance.

II.3 Cas de plusieurs variables qualitatives répétées :

Si l'on souhaite analyser q variables qualitatives répétées par cette méthode, le nombre de profils distincts va croître de façon vertigineuse : $r=L_1{}^dL_2\ldots{}^dL_q{}^d$, et la présence de zéros structurels va augmenter ce qui ici sera plus difficile à surmonter.

Pour décomposer factoriellement les tests comme au chapitre AI il faut donc ici à priori plusieurs profils mesurées : $r_1, r_2, ..., r_q$.

On est alors obligé de considérer, d'une part que les transformations définies par F sont les mêmes pour toutes les q variables, et surtout pour calculer les estimateurs des moindres-carrés généralisés d'avoir un compromis des q matrices de covariances V_p.

On retrouve alors le modèle linéaire général avec une estimation en moindres-carrés généralisés (III.1 du chapitre AI). On peut alors effectuer une ACP décomposant la statistique de Lawley-Hotelling.

Remarques :

-1 La démarche est parfois analogue à celle employée pour le modèle S.U.R. du chapitre AI comme le montrent Imrey et al.(1982).

-2 La loi de la statistique de Lawley Hotelling n'est qu'asymptotique.

III . Modèle log-linéaire et AFC

III.1 Extensions de GSK aux modèles log-linéaires :

Bonnet et al.(1985) ont présenté des généralisations du modèle GSK d'une part à des distributions généralisant le produit de lois multinomiales. Ils écrivent le modèle linéaire sur des fonctions des effectifs plutôt que des proportions observées. La matrice de covariance des effectifs est alors pour des lois de Poisson conditionnelles :

$$V_f = D_\theta - D_\theta K({}^tKD_\theta K)^{-1}{}^tKD_\theta$$

où D_θ est la matrice diagonale contenant $\theta = E(f)$ et K dépend de la distribution choisie. Le terme contenant K disparait si la distribution est de Poisson non-conditionnelle et dans le cas d'un produit de lois multinomiales, $K=Id_s \ 1_r$.

Ils proposent aussi des estimations de β sans contraintes ou avec des contraintes lineaires diverses : exactes, $C\beta=c$, ou stochastiques, $R\beta + \varepsilon = r$, que l'on peut tester comme précédemment avec une statistique de Wald.

D'autres part ils étendent le modèle au cas ou u=rs pour avoir le modèle log-linéaire. S^{-1} est remplacé par S^-, un inverse généralisé de S, dans les calculs précédents.

III.2 AFCVI et modèle log-linéaire :

La technique d'ACPVI brièvement décrite au chapitre AI peut donner naissance en considérant l'AFC comme l'ACP de $(D_I^{-1} P D_J^{-1} - 1_{I,J}, D_J, D_I)$, à une AFCVI qui sera alors l'ACPVI de ce triplet par rapport à la structure définissant la contrainte.

Plusieurs auteurs ont établi des liens entre le modèle log-linéaire et l'AFC. Grizzle et Williams (1972) par exemple font remarquer que si pour une table à deux entrées (généralisable à k entrées) il n'y a pas indépendance marginale (i.e. l'hypothèse $p_{ij} = p_{i.}p_{.j}$ rejetée) le modèle log-linéaire avec l'effet constant, ligne, et colonne est rejeté. Plus récemment une importante

discussion s'est faite sous la forme de commentaires à un article de Goodman(1986).

Ainsi faire l'AFC est "relativement" équivalent à décomposer le résidu, qui d'ailleurs contient l'intéraction ligne colonne, du modèle log-linéaire précédent.

On peut alors proposer pour un modèle log-linéaire différent de réaliser l'approximation de la table de contingence par le modèle GSK, puis décomposer l'AFC totale de la table en l'AFCVI par rapport à cette approximation et l'AFCVI orthogonale .

Une démarche analogue a été entreprise par Sabatier et Lebreton(1990) pour les tables à deux entrées. Ceci rentre dans le cadre d'analyse par rapport à un modèle introduit par Escofier(1984).

L'intérêt de cette démarche est alors la modélisation de tables de contingences multiples (à plus de deux entrées) par un modèle log-linéaire. On reconstitue la table multiple par ce modèle puis on effectue des décompositions des AFC, par exemple du tableau croisant une variable et le "croisé des autres" par des AFCVI.

III.3 Application aux mesures répétées :

On peut utiliser directement les assertions précédentes pour une table à trois entrées i.e. le temps en troisième entrée. Lorsque l'on a plusieurs variables on peut aussi imaginer un modèle log-linéaire avec seulement les effets d'une des variables (le facteur groupe) et le temps...

Avec plusieurs variables, une autre approche, appréciable dans le cas de présence de beaucoup de zéros, est l'AFCMVI. Nous avons déjà, Leibovici et al.(1990), traité un exemple de cette technique, notamment en privilégiant l'aspect analyse de la variance, pour des données non-répétées. On peut de façon analogue au modèle Split-plot étudié au chapitre AI décomposer les effets groupe, visite, groupe.visite par les AFCMVI correspondantes.

Remarque :

-1 Pour une seule variable répétée on pourra encore se placer dans le cadre multivariable multivarié du AI.III, mais les tests seront valides que pour des gros échantillons.

IV . Conclusions

Dans ce bref chapitre on a voulu simplement montrer que les méthodes décrites dans le chapitre AI pouvaient s'adapter aux données qualitatives. Pour cela on a été amené à expliciter brièvement les modèles log-linéaires qui s'y rattachent.

Il s'avère donc que surtout des modèles univariés sont utilisés classiquement pour des mesures répétées qualitatives.

En se plaçant dans le cadre AFCM ou AFCMVI, on peut toutefois appliquer le modèle multivarié multivariable DMM (AI.III.3.b) où chaque modalité jouera le rôle d'une variable, mais pour plusieurs variables ce modèle peut devenir inextricable.

La pratique de simples χ^2, pour le classement étudié et la variable qualitative à chaque date, est en fait souvent la plus courante. Cette pratique suggére qu'une décomposition factorielle de l'écart à l'indépendance marginale entre chaque variable, le temps, et le facteur groupe étudié serait intéressante. Nous aborderons ce fait dans le chapitre BIII (BIII.V.2).

V . Références bibliographiques

AGRESTI, A. (1988) Logit models for repeated ordered categorical response data. Sas Sugi 997-1005.
BISHOP, Y.M. FIENBERG, S.E. and HOLLAND, P.W. (1975) Discrete multivariate Analysis : Theory and Practice. M.I.T. Press, Cambridge, Massachussets.
ANDREWS, D.M. and DAVID, H.A. (1990) Nonparametric analysis of unbalanced paired-comparison or ranked data. J.A.S.A. 1140-1146.
BECKER, M.P and AGRESTI, A (1992) Log-linear modelling of pairwise interobserver agreement on a categorical scale. Stat.Med. 11, 101-114.
BHAPKAR, V.P. and KOCH, G.G. (1968) On the Hypotheses of 'NO Interaction' in Contingency Tables. Biometrics, 567-594.
BONETT, D.G. WOODWARD, J.A. and BENTLER, P.M. (1985) Some extension of a linear model for categorical variables. Biometrics, 745-750.
CLAYTON, D.G. (1974) Some odds ratio statistics for the analysis of ordered categorical data. Biometrika, 525-531.
COX , D.R. (1972) Regression Models and Life Tables. J. R. S. S. B,187-220.
DAUDIN, J.J and TRECOURT,P. (1980) Analyse factorielle des correspondances et modèle log-linéaire : comparaison des deux méthodes sur un exemple. R.S.A. XXVIII, 5-24.
ESCOFIER, B. (1984) Analyse factorielle en référence à un modèle : application à l'analyse de tableaux d'échanges. Rapport Technique n°337. INRIA, Rocquencourt.
FISHER, G.H. (1989) An IRT-Based Model for Dichotomous Longitudinal data. Psychometrika, 599-624.
GRIZZLE, J.E. STARMER, C.F. and KOCH, G.G (1969) Analysisof categorical data by linear models. Biometrics, 489-504.
GRIZZLE, J.E. and WILLIAMS, D.O (1972) Log linear models and tests of independance for contingency tables. Biometrics, 137-156.
GOODMAN, L.A. (1986) Some useful extensions of the usual correpondence analysis approach an the usual log-linear models approach in the analysis of contingency tables. In.Stat.Rev. 243-309.
IMREY, P.B. KOCH, G.G. and STOKES, M.E. et al. (1981) Categorical data analysis : Some reflections on the log-linear model and logistic regression. Part I :Historical and Methodoligical overview. Int.Stat.Rev. 265-283.
IMREY, P.B. KOCH, G.G. and STOKES, M.E. et al.(1982) Categorical data analysis : Some reflections on the log-linear model and logistic regression. Part II : Data analysis. Int.Stat.Rev. 35-63.
JÖRESKOG, K.G. (1970) A general method for analysis of covariance structures. Biometrika, 239-251.
KOCH, G.G. LANDIS, J.R. FREEMAN, J.L. FREEMAN, D.H. AND LEHNEN, R.G. (1977) A general methodology for analysis of experiments with repeated measurement of categorical data. Biometrics, 133-158.
KOCH, G.G. and REINFURT, D.W. (1971) The analysis of categorical data from mixed models. Biometrics, 157-173.
KOCH, G.G. STOKES, M.E. and BROCK, D. (1980) Applications of weighted least squares methods for fitting variational models to health survey data. In : Proceedings of the American Statistical Association.Section on Survey Research Methods. 218-223.
LANDIS, J.R. and KOCH, G.G. (1977) The measurement of observer Agreement for categorical data. Biometrics, 159-174.
LEIBOVICI, D. BUCQUET, D. SABATIER, R. CURTIS, S. and COLVEZ, A. (1990) Data analysis with dependent structure applied to epidemiological survey data in gerontology. In:

11th Meeting of the International Society for Clinical Biostatistics, Volume d'abstracts, 18-21 Sept.
MCLEAN, R.A. SANDERS, W.L. and STROUP, W.W. (1991) A unified approach to mixed linear models. Am. Stat. 55-64.
SABATIER, R. and LEBRETON,J.D. (1990) Comparaisons d'Analysesfactorielles sous-contraintes linéaires et desmodèles log-linéaires pour l'analyse de tables de contingenes. In: XXII journnées de satistique . ASU. Tours mai 1990.
VAN Der HEIJDEN, P.M. De FALGUEROLLES, A. and De LEEUW, J. (1989) A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. Appl. Stat. 249-292.
WARE, J.H. LIPSITZ, S. and SPEIZER, F.E. (1988) Issues in the analysis of repeated categorical outcomes. Stat.Med. 95-107.

BI
Observations répétées : variable ordinale ou contrainte ordinale ?

I . Introduction

L'analyse multidimensionnelle de variables ordinales est encore un vaste champ de recherches des statisticiens et psychométriciens. L'analyse des dissimilarités y est souvent attachée et diverses approches sont apparues dans la littérature.

Notre propos, dans ce chapitre, ne sera pas d'expliciter ce côté riche de l'analyse des données que l'on pourra consulter dans la thèse de d'Aubigny(1989) pour un exposé des différents problèmes avec une importante bibliographie.

La recherche d'un codage optimal, par exemple par transformation isotone des variables, apparaît dans les méthodes. Car si pour des variables numériques le codage apparaît imposé, pour les variables ordinale celui-ci est semble-t-il arbitraire.
On parle ainsi d'un codage qui en pratique se présente comme une transformation d'un codage initial. Ceci, par la suite, incitera les statisticiens à l'étude de codages des variables numériques pour donner naissance à l'idée d'analyse non-linéaire des données.

Le concept sous-jacent à cette optique est en fait la définition même de la mesure d'un phénomène. Ainsi par exemple une chose aussi simple, que de mesurer la taille des individus, peut se faire de toutes sortes de façons (ex: en ordinal : grand, moyen, petit, en numérique : la longueur de la tête aux pieds, en référence à un objet : i.e. plus grand que ... ou une autre échelle que le mètre etc....).

Mais le statisticien ne se contente pas de mesurer, il veut comparer, estimer, trouver des liens, représenter. Pour cela il est amené à chercher les paramètres pour optimiser des critères, ainsi la mesure et les critères sont interdépendants. C'est ce que prônent Young(1981) et De Leeuw, à travers Gifi(1990), dont le principe est mis en oeuvre dans les algorithmes de type ALSOS (pour Alternating Least Squares and Optimal Scaling).

Ainsi une première approche, pour une variable répétée est d'appliquer une méthode non-linéaire. On a à notre disposition diverses méthodes d'analyses des données généralisant l'ACP linéaire à l'ACP non-linéaire. En particuliers citons la méthode PRINQUAL de Tenenhaus et Vachette(1977), l'ACPVI-Spline de Durand(1990) et les FPA (Fonctions Principales Additives) de El Faouzi(1992).

Lorsque l'on parle de variable ordinale on pense à un ordre, à une succession, à une contiguïté comme pour la mesure du temps. Donc on peut souhaiter par exemple, aux vues des observations d'une ou plusieurs variables sur des individus, rechercher :

- quelle a été l'évolution, les phases de croissances, décroissances stagnations, cycles etc....,
- comment se sont structurés les individus entre eux, ressemblances oppositions etc... .

Dans cette recherche on peut éventuellement la contrainte de contiguïté temporelle de l'ordre imposé du temps.

Ce chapitre s'articule autour du soucis du statisticien, face à des données de facteurs à mesures répétées, de prendre en compte l'aspect évolutif dans le temps, de retrouver un ordre préfixé. Cette optique est évidemment très intéressante pour l'épidémiologiste qui est souvent confronté à des données ordinales ou à des données qualitatives qu'il voudrait bien ordonner. Il arrive souvent par exemple qu'il cherche des poids (un codage) pour une telle variable. C'est par exemple ce que nous avons fait dans le chapitre CII de ce travail où l'ordre des poids a suivi l'ordre préfixé sans ajouter de contraintes au problème. Ceci peut souvent arriver en AFC avec le fameux effet Guttman.

Ce sont ces différents aspects qui vont nous intéresser dans ce chapitre, nous considérerons :

- d'une part le problème d'une contrainte ordinale sur les solutions des analyses factorielles,

- d'autre part la question de structure de voisinage linéaire dernièrement synthétisée avec les opérateurs de voisinage par Meot(1992).

Cette dernière sera vue par la considération de comparaisons par paires, ce qui nous amènera à introduire d'autres métriques. Ainsi nous généraliserons ce problème à l'optique de contraintes définies par des matrices de dissimilarités, dont l'intérêt sera la prise en compte de dissimilarités non-euclidiennes comme contraintes.

Notons que le premier problème a été abordé par Tenenhaus(1988) et Nishisato(1980) dans le cadre de l'AFC ou la contrainte est explicite. Young et al.(1978) ont construit l'algorithme PRINCIPALS, et Saito et Otsu(1988) proposent la méthode OSMOD où l'on cherche plutôt à combiner codage isotone des variables et "composantes principales". Winsberg et Ramsay(1983) en utilisant des transformations splines monotones, adoptent la même optique

II . Contrainte ordinale

II.1 régression isotone, monotone :

La régression isotone est souvent le centre d'intérêt des psychométriciens qui travaillent avec des échelles ordinales. On entend par transformation isotone, une transformation respectant l'ordre ou le préordre déterminé par le régresseur. Le critère de la régression isotone est le critère des moindres-carrés.

Cette transformation isotone peut se définir mathématiquement comme l'espérance conditionnelle d'une variable aléatoire de $L^2(\Omega,a,P)$, par rapport à un sous σ-treillis de la tribu a. Les développements mathématiques de cette présentation peuvent être consultés dans Barlow et al.(1972).

Plusieurs algorithmes de régression isotone existent, le plus célèbre est certainement le "Pool Adjacent Violators" dont on peut trouver une description dans l'ouvrage cité précédemment (ainsi que d'autres algorithmes). Tenenhaus(1988) le présente comme une projection sur un cône convexe polyédrique.

La projection sur un cône convexe joue un rôle central pour la prise en compte de contraintes ordinales, ou d'inégalité, et divers algorithmes de projection ont été proposés par Bremner(1982) ou Dykstra(1983) par exemple. Ils portent souvent le nom de NNLS(Non-Negative Least Square) ou moindres-carrés restreints .

Nous allons, dans ce paragraphe, rappeler les définitions et propriétés essentielles de la projection sur un cône pour décrire la régression monotone par l'algorithme des "Pool Adjacent Violators". Nous nous plaçons dans le cas le plus simple où les produits scalaires sont définis par la métrique

identité. L'introduction d'une autre métrique produit un léger changement d'écriture.

définitions :

i) un cône C est un ensemble de vecteurs stables par homothéties positives.

ii) un cône convexe C est un cône stable par combinaisons convexes, c'est à dire :

soit $\{x_1, \ldots, x_f\}$ un nombre fini de vecteurs du convexe C et soit $\{\theta_1, \ldots, \theta_f\}$ des coefficients positifs de somme 1, alors $\sum_i \theta_i x_i \in C$.

iii) un cône polyédrique convexe est cône convexe engendré (combinaisons linéaires positives) par un système générateur fini.

iv) un cône polyédrique convexe est de la forme $C = \{x / {}^t Ax \leq 0\}$.

v) le cône polaire de C est $C^p = \{y / {}^t xy \leq 0\}$ est un cône convexe polyédrique engendré par les colonnes de A.

Les propriétés essentielles de la projection sur un cône convexe polyédrique sont :

propriétés :

soit C un cône convexe polyédrique engendré par l'ensemble de vecteurs G_c, d'un espace E quelconque, alors :

a. Si x est la projection d'un vecteur z sur C, alors l'ensemble des vecteurs R_z de G_c orthogonaux à (z-x) engendrent un sous-espace $L(R_z)$ tel que x soit la projection z sur ce sous-espace.

b. La projection x de z sur un sous-espace $L(R_z)$ pourra être la projection de z sur le cône si et seulement si ${}^t(z-x)g \leq 0$ pour tout g de G_c.

c. Tout vecteur z peut se décomposer de manière unique de la façon suivante $z = x + x^p$ où x appartient à C et x^p à C^p et ${}^t xx^p = 0$. x et x^p sont les projections de z sur C et C^p.

d. L'opérateur de projection sur un cône convexe polyédrique est linéaire et continu sur tout l'espace. La projection sur l'intersection du cône et de la sphère unité est continue sur tout l'espace, sauf sur le polaire du cône.

La propriété a) est très importante pour la recherche de la projection dans les algorithmes. Les démonstrations peuvent être consultées dans Tenehaus(1988).

Une transformation monotone d'une variable ordinale à m modalité va consister à trouver m nombres positifs (codages) qui respectent l'ordre des modalités. La nouvelle variable se trouve être une combinaison positive des variables sur-indicatrices des modalités (i.e. valent 1 si la modalité prise est supérieure ou égal à la "modalité" de l'indicatrice et 0 sinon). Cette nouvelle variable est un vecteur du cône polyédrique convexe engendré par les sur-indicatrices.

Ainsi, la régression isotone consistera à projeter une variable numérique sur un cône convexe polyédrique déterminé par le régresseur ordinal.

La structure particulière du cône polyédrique construit à partir de la variable ordinale conduit naturellement à l'algorithme du "Pool Adjacent Violators". Ceci a été montré par Tenenhaus(1988). En effet, la recherche de la "meilleure" partition croissante (i.e. le plus grand élément de h^{ème} partie est plus petit que le plus petit élément de la (h+1)^{ème} partie), de l'ensemble des modalités de la variable ordinale, est induite par le sous-ensemble optimal R_z solution du problème de projection.

Les algorithmes recherchent donc souvent cet ensemble optimal. C'est par exemple le cas de l'algorithme de Bremner(1982) qui part d'un R_z vide et l'augmente pas à pas. La forme de l'algorithme classique du "Pool Adjacent Violators" qui cherche lui cette partition croissante, adopte une démarche inverse :

1) calculer la projection sur l'espace entier,

2) remplacer, de proche en proche, les valeurs impliquées dans le non respect de l'ordre, par leur moyenne, en rassemblant ainsi deux modalités, et continuer jusqu'à obtention de l'ordre.

Remarques :

-1 Cet algorithme rapide est très pratique, mais les algorithmes de projection sur un cône sont évidemment plus généraux, et donc dans certains cas seront plus intéressants. C'est par exemple le cas si le cône n'est pas simplement défini par un ordre mais aussi par des valeurs. C'est ce qui se passe si l'on a des mesures répétées non-espacées de façon régulière.

-2 On peut souhaiter que la projection obtenue ait certaines propriétés comme par exemple dans la régression spline, et utiliser la régression sur splines monotones de Ramsay(1988).

II.2 Analyse canonique de cônes convexes :

L'analyse canonique de deux cônes convexes polyédriques se présente comme une généralisation de l'analyse canonique de deux sous-espaces vectoriels. Ainsi sur cette même base on peut effectuer une analyse canonique entre un sous-espace vectoriel et un cône convexe polyédrique.

Tenenhaus(1988) a présenté un algorithme aux moindres-carrés alternés pour résoudre ce problème. Il assure l'optimalité de la solution en vérifiant les conditions de Khun et Tucker ainsi que la convergence vers la solution.

Nous présentons ici de façon succincte son approche et son application à l'AFC de variables ordinales.

Le problème (P) est donc :

- maximiser le cosinus carré de vecteurs normés, chacun dans un cône convexe polyédrique.

On a donc un problème de maximisation sous contraintes d'inégalités. Les propriétés de la projection sur un cône convexe polyédrique permettent d'écrire la propriété suivante.

Propriété d'analyse canonique de deux cônes convexes polyédriques:

Soit deux cônes convexes polyédriques C_1 et C_2 engendrés par G_{c1} et G_{c2} respectivement, alors :
Il existe un sous-ensemble R_{c1} de G_{c1} et un sous-ensemble R_{c2} de G_{c2} tels que le problème (P) d'analyse canonique des deux cônes se ramène au problème d'analyse canonique des deux sous-espaces $L(R_{c1})$ et $L(R_{c2})$.
La solution est alors donnée par le meilleur couple canonique, c'est à dire pour la plus grande corrélation canonique.

L'algorithme consiste à projeter alternativement sur chacun des cônes. Plus précisément on se donne un y_0 dans C_2- $C_1{}^P$, on construit les suites :

$$x_n = P_{C_1 \cap S}(y_{n-1}),\ y_n = P_{C_2 \cap S}(x_n),\ \rho_n = \cos(x_n, y_n),$$

où S est la sphère unité.

Tenenhaus(1988) énonce alors le résultat :

Propriété de la convergence de l'algorithme alterné :

Sous l'hypothèse de nombre fini de solutions stationnaires (i.e. couple (x,y) de C_1 6 C_2 qui au signe prés sont projections l'un de l'autre sur l'intersection du cône et de la sphère unité), et l'hypothèse que C_2 n'est pas contenu dans le polaire de C_1 alors les suites précédentes vérifient :

i. ρ_n est croissante et convergente vers ρ

ii. $\lim_{n \to +\infty} \|x_n - x_{n+1}\| = 0$ et $\lim_{n \to +\infty} \|y_n - y_{n+1}\| = 0$

Tenenhaus propose également, pour accélérer la convergence de l'algorithme, de chercher au bout de quelques itérations, les sous-espaces vectoriels tels que la solution soit donnée par l'analyse canonique de ceux-ci.

Les projections seront bien entendu obtenues par les algorithmes auxquels nous avons fait références plus haut.

Remarques :

-1 Le théorème précédent ne nous assure toutefois pas de l'optimalité de la solution.

-2 La solution stationnaire obtenue vérifie les relations de l'analyse canonique ordinaire pour les sous-espaces optimaux obtenus par les algorithmes de projection. On peut même écrire que la solution vérifie, si ρ est positif : $P_{C_1}(P_{C_2}(x)) = \rho^2 x$. Ainsi l'interprétation d'une analyse canonique de deux cônes convexes polyédrique se fait de la même façon que dans le cas ordinaire.

Le cas de l'analyse canonique de deux convexes polyédriques peut donc s'appliquer à deux variables qualitatives dont l'une ou les deux sont ordinales. C'est à dire que l'on souhaite effectuer une AFC avec une ou deux contraintes ordinale.

Tenenhaus(1988) décrit comment mettre en oeuvre une telle analyse, et traite un exemple déjà analysé par Nishisato(1980) par sa méthode SDM, et par Bradley et al.(1962,exemple 3). Il souligne alors que la méthode décrite dans l'article de Bradley et al. est supérieure car elle fournit une solution optimale alors que celle de Nishisato et la sienne ne fournissent que des solutions stationnaires.

La méthode de Nishisato consiste à itérer des AFC après avoir remplacé deux colonnes, ne respectant pas l'ordre souhaité, chacune par la moyenne des deux, (SDM : Successive Data Modifications), jusqu'à obtention de l'ordre souhaité sur le premier axe fourni par l'AFC. Elle a l'avantage d'être très simple à mettre en oeuvre et correspond à l'algorithme du "Pool Adjacent Violators".

On effectue donc, dans cette méthode, à chaque étape, une transformation isotone du tableau. Par exemple si la contrainte ordinale est sur les lignes, la transformation est sur les lignes indépendamment des colonnes. Cela peut être très réducteur et l'on pourrait effectuer des transformations sur les lignes conditionnellement aux colonnes. C'est à dire que dans le premier cas, les transformations en tant qu'opérateurs sur R^I R^J, sont décomposables en un opérateur linéaire sur R^I et l'identité sur R^J. Alors que dans le second cas, sur R^J on n'a pas forcement l'identité. On sort alors du cadre de l'analyse canonique de deux convexes mais pas de l'AFC sous contrainte.

Nous appellerons cette nouvelle approche, la méthode SDM-L, L pour Local, dont voici une comparaison avec SDM.

exemple 3 de Bradley et al.(1962) :

J **Traitement**

I échelle	I	II	III	IV	V
s1	9	7	14	11	0
s2	5	3	13	15	2
s3	9	10	6	3	10
s4	13	20	7	5	30
s5	4	4	0	8	2

Les deux solutions calculée pour une contrainte d'ordre sur les lignes avec la méthode SDM et la méthode SDM-L sont :

	AFC	SDM	SDM-L
	ρ^2 0,533	0,493	0,498
s1	1,043	1,27	1,29
s2	1,30	1,27	1,22
s3	-0,54	-0,55	-0,51
s4	-1,06	-0,86	-0,68
s5	0,47	-0,86	-1,19

Les deux transformations dans la méthode SDM ont porté sur la quatrième et dernière ligne d'abord puis sur la première et deuxième ligne. Pour la méthode SDM-L les transformations ont porté donc sur les mêmes

lignes, mais uniquement pour le traitement V lors de la première transformation, et uniquement pour le traitement II pour la dernière transformation. On remarquera que l'ordre conditionnel est plus satisfaisant, on pourra chiffrer l'effet d'ordre en effectuant l'AFCVI du tableau de départ par rapport au tableau modifié.

Remarques :

-1 Pour une deuxième composante dans l'analyse canonique de deux convexes polyédriques, il faut donc rajouter une contrainte linéaire au problème pour avoir l'orthogonalité avec la première solution. Dans le cas d'un cône convexe polyédrique engendré par un système libre, deux vecteurs sont orthogonaux si, les coefficients étants positifs, ils ne sont pas combinaisons linéaires positives des mêmes vecteurs.

II.3 ACP sous contrainte ordinale :

Lorsque l'on fait une ACP où les variables représentent des dates (par exemple les mois, des années, etc....), ou des dosages on aimerait avoir une représentation des individus par rapport à une évolution temporelle ou une augmentation de dosage. C'est à dire que l'on souhaite avoir une interprétation simple de l'axe factoriel comme la "gravité" dans le cas d'une variable d'états. Cet axe ne servira qu'à étalonner cette "gravité" ou pour les dates à exprimer les variations observées.

On peut noter que le problème des naissances dans les familles observées par Deville et relaté dans Besse(1987) n'est pas très éloigné de ce souci. En effet ils utilisent une transformation linéaire définie par une matrice carrée de dimension p qui est une matrice triangulaire de 1, pour obtenir un tableau des naissances jusqu'à la date d'observation. Son ACP donne un axe principal (la première "harmonique") satisfaisant la condition de contrainte ordinale .

Dans le cas général, on introduit une contrainte d'ordre sur les coordonnées des axes factoriels :

soit φ un tel axe factoriel, c'est un vecteur de R^p vérifiant :

$$\varphi_1 \le \varphi_2 \ldots \le \varphi_{p-1} \le \varphi_p$$

qui peut s'écrire

$$\begin{pmatrix} 1 & -1 & 0 & . & . & 0 \\ 0 & 1 & -1 & 0 & . & 0 \\ 0 & 0 & 1 & -1 & . & 0 \\ . & . & . & . & . & . \\ 0 & . & . & . & 1 & -1 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ . \\ \varphi_{p-1} \\ \varphi_p \end{pmatrix} \le \begin{pmatrix} 0 \\ 0 \\ . \\ 0 \\ 0 \end{pmatrix}$$ c'est à dire de la forme ${}^tA\varphi \le 0$.

On impose donc aux axes factoriels d'être dans un cône polyédrique convexe C_A. On peut alors effectuer une analyse canonique entre le sous-espace engendré par le tableau analysé et le cône C_A.

Techené(1980) a décrit, dans un cadre plus général d'ACP de fonction aléatoire, comment réaliser une ACP sous contrainte ordinale. Cet auteur parle de contrainte d'intérieur relativement à un cône convexe polyédrique car il se place dans un Hilbert quelconque. Pour réaliser cette analyse, on a besoin de conditions de continuité sur les fonctions convexes définissant le cône, ce qui ici est vérifié.

Ses propriétés s'énoncent de façon analogue à la propriété d'analyse canonique de deux cônes convexes polyédriques :

Propriété d'ACP sous contrainte ordinale :

La première solution d'ACP sous contrainte relativement à un cône C_A est obtenue par une ACPVI par rapport au sous-espace optimal $L(R_{CA})$. On recherche la solution en calculant les vecteurs propres des $(2^{|G_{CA}|}-1)$ opérateurs de covariances projetés sur les sous-espaces engendrés chacun par une partie non-vide de G_{CA}. On prend alors la plus grande valeur propre pour laquelle le vecteur propre est dans C_A.

Pour continuer l'ACP on doit ajouter une contrainte d'orthogonalité et l'on n'a pas forcement de solution i.e. si l'intersection du cône et du sous-espace orthogonal à la première solution est vide. On peut alors continuer l'ACP sans contrainte ordinale ou encore définir une autre contrainte

ordinale. Techené dans cet esprit propose d'imposer des contraintes multiples sur une famille de cônes.

Remarques :

-1 Ce qui nous a intéressé jusqu'à présent, c'est de contraindre les solutions de l'ACP pour avoir un ordre. C'est à dire que l'on s'est placé dans l'optique, de choisir l'explication, puis d'observer son impact. On a été amené à forcer les données à suivre ce choix, comme par exemple dans l'AFC sous contrainte ordinale réalisée par la SDM ou SDM-L.

-2 Si l'on a plusieurs variables répétées on pourra utiliser des méthodes qui traitent plusieurs tableaux et imposer une contrainte ordinale comme précédemment. On peut aussi faire une analyse canonique de q convexes polyédriques dont on trouvera une description dans Tenenhaus(1983). Dans le cadre de l'ACP des tableaux concaténés, sous contrainte ordinale, le cône sera un produit tensoriel de cônes polyédriques.

-3 L'optique, que l'on va aborder maintenant, est plutôt de rechercher ce que peut expliquer la structure sous-jacente à la collecte des données. C'est à dire une connaissance a priori du système. Donc le souci est d'expliciter au mieux la structure de répétition, de succession.

III . Utilisation des comparaisons par paires

III.1 idée générale :

Le domaine de l'analyse des comparaisons par paires est assez vaste en statistique, ainsi bon nombre d'indices de comparaisons par paires existent. Nous allons plutôt ici tirer profit de l'idée qui est de se soucier des paires d'observations dans les analyses.

Lorsque l'on a des mesures répétées dans le temps on peut se poser la question : "Qu'est ce qui est comparable ?". On peut certainement répondre de différentes manières.

Tout d'abord, pour une variable, si l'on ne s'occupe que de la trajectoire d'un seul individu, on peut dire que toutes les mesures sont comparables. Ou bien, dire qu'une mesure est seulement comparable avec celles qui lui sont contiguës. C'est à dire que les comparaisons ne peuvent être que locales.

L'idée de graphe lié à une relation binaire décrivant la structure temporelle ou plus généralement spatiale ou spatio-temporelle est assez naturelle. Meot(1992) fait le bilan de trente ans de diverses méthodes relatant cet aspect, depuis l'indice Geary en 1954 en passant par les analyses de Lebart en 1969, et synthétise cette approche par un schéma de dualité modifié. Son utilisation en ACP sous le nom d'Analyse en Composantes de Voisinages (ACV) est originale.

Enfin si l'on s'occupe de deux individus différents, on peut dire qu'ils ne peuvent être comparable qu'à la même date. Dans les approches de contraintes de graphe ceci est rarement évoqué car tout simplement il y a en général soit qu'un seul individu mesuré sur plusieurs variables répétées, soit qu'une seule variable répétée pour plusieurs individus.

Cette vision de l'utilisation des comparaisons par paires rejoint dans sa mise en oeuvre, l'idée de pondérations des distances en ACP introduite par Le Foll(1982). On est alors amené à faire une ACP avec une semi-métrique D (sur les colonnes) sous la forme $D = A_P - V_P$ où V_P est la matrice symétrique contenant les pondérations des paires et A_P est la matrice diagonale contenant la marge de V_P. Carlier(1985) a appliqué une technique analogue, pour des données temporelles multivariables. Nous expliciterons son approche dans le chapitre BII.

III.2 Une ACPVI spéciale :

Une application simpliste, des hypothèses de comparaisons possibles, peut être mis en oeuvre pour les mesures répétées de la manière suivante.

On dispose de q variables mesurées p fois sur n individus. Si l'on considère l'approche univariée multivariable du chapitre AI, on a un opérateur de produits scalaires qui peut se représenter par une matrice carrée à np dimensions. On peut alors "trouer" cette matrice de façon à obtenir la matrice symétrique de la forme suivante :
par exemple pour p = 3,

$$\begin{pmatrix} W_{11} & D_{21} & D_{31} \\ D_{12} & W_{22} & D_{32} \\ D_{13} & D_{23} & W_{33} \end{pmatrix} \overset{\text{not}}{=} W\_Id ,$$

où les $W_{ij} = X_i {}^tX_j$,avec X_k le tableau (n,q) des mesures à la date k et les matrices D_{ij} les diagonales des opérateurs W_{ij}.

Dans ce cas, on dit que les individus sont comparables à la même date, et qu'à des dates différentes un individu n'est comparable qu'à lui même. Si l'on fait $D_{13}=D_{31}=0$ alors on se place dans les comparaisons locales pour un même individu. On peut aussi ne considérer que les diagonales des W_{ij}, alors un individu n'est comparable qu'à lui-même.

L'opérateur W_Id est Id-symétrique mais n'est pas forcement défini positif. Il le sera si les diagonales D_{ij} sont positives ou nulles. On peut alors

l'utiliser pour faire une ACP qui tient compte des hypothèses de comparaisons possibles. On peut également effectuer l'ACPVI de X par rapport à W_.

La démarche parait peu orthodoxe algébriquement mais correspond à un souhait empirique acceptable. L'optique algébrique liée aux graphes de voisinages va palier à ce manque de justifications.

III.3 Contraintes de graphe , ACV :

On considère une relation binaire symétrique sur l'ensemble des unités statistiques, c'est notre relation de voisinage. Le graphe de la relation est associé à la matrice M où donc m_{il}=1 si irl et 0 sinon.

Momentanément nous nous plaçons dans un cadre où l'on a n unités statistiques mesurées sur q variables. M est associée au graphe. N est la marge de M (i.e. nombre de voisins) sous forme diagonale. D_n^* est la matrice diagonale contenant la somme des poids contenus dans D_n des voisins de l'individu considéré. Et, Δ est le vecteur de coordonnées égales à 1.

Alors soit x l'une des q variables observées on peut écrire la variance de x :

$$v_{xx} = {}^t(P_{\Delta^\perp}x)D_n(P_{\Delta^\perp}x) = {}^txD_nP_{\Delta^\perp}x = \sum_{i=1}^{n} p_i(x_i-\bar{x})^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n} p_i p_l (x_i-x_l)^2$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n} p_i p_l m_{il}(x_i-x_l)^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n} p_i p_l (1-m_{il})(x_i-x_l)^2$$

$$= v_{xx_S} + v_{xx_{\bar{C}S}}$$

$$= {}^txD_n(D_n^* - MD_n)x + {}^txD_n(P_{\Delta^\perp} - D_n^* + MD_n)x$$

$$= {}^txD_nEx + {}^txD_nSx.$$

On décompose donc la variance d'une variable en une partie due au graphe de voisinage l'autre partie due au graphe complémentaire. Meot(1992) introduit ici les opérateurs E et S associés au graphe M et les appellent opérateurs de voisinage :

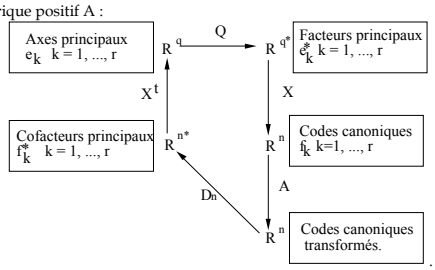
$E = D_n^* - MD_n$ opérateur d 'écart de voisinage ,

$S = P_{\Delta^\perp} - D_n^* + MD_n$ opérateur de lissage de voisinage .

$(Ex)_i$ est l'écart pondéré de la i ème valeur de x à la moyenne de ses voisins, et $(Sx)_i$ correspond à un centrage et à un lissage de la i ème valeur de x.

Plusieurs propriétés intéressantes sont à noter pour ces opérateurs, notamment ce sont des opérateurs D_n-symétriques, positifs, contractants, ayants les mêmes vecteurs propres, dont les valeurs propres associées sont de somme 1.

Les propriétés de tels opérateurs poussent Meot à proposer un schéma de dualité général prenant en compte la donnée d'un opérateur D_n-symétrique positif A :



Les e_k sont les vecteurs propres Q-normés de l'opérateur Q-symétrique :
$C = {}^tXD_nAXQ$.
On note le schéma (X,Q,D_n,A).

Remarque:

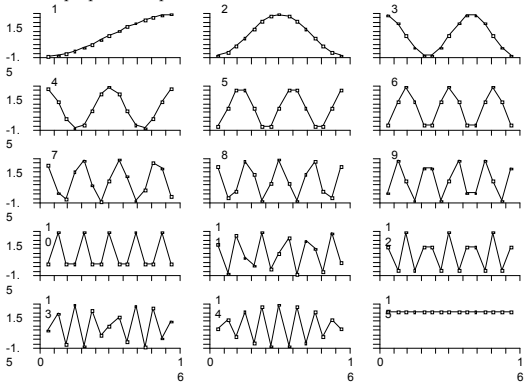
-1 Si A est un projecteur on retrouve l'ACPVI, et si l'on a une relation d'équivalence on obtient une ACP inter-classes, c'est à dire une ACPVI par rapport à une structure de groupe.

A étant défini positif, les vecteurs propres de A constituent, de part leurs propriétés de maximalité, un décryptage intéressant de la contrainte et peuvent alors servir à modéliser les données.

C'est cette optique qu'adopte Meot dans l'Analyse en Composantes de Voisinages (ACV) qui consiste, pour la première étape, à diagonaliser l'opérateur de lissage de voisinage associé au graphe de voisinage. La deuxième étape consiste à retenir, quelques premiers vecteurs propres qui traduisent une autocorrélation positive, et donnerons des codes de variance locale faible (codes lisses), et quelques derniers vecteurs propres qui traduisent une autocorrélation négative (codes antilisses).

Ces deux groupes de vecteurs constituent la structure de modélisation. On décrit le modèle sur X par les deux ACPVI correspondantes.

Pour la relation temporelle sur un individu, définie plus haut, les vecteurs propres de l'opérateur S sont :



Représentation des 15 vecteurs propres de l'opérateur de lissage de voisinage S pour une relation de voisinage linéaire à 15 points.

Ceci nous fait certainement penser à la modélisation de courbes par des polynômes telle que nous l'avons vu dans le AI avec le modèle de Potthof et Roy. En fait, on peut mettre en évidence une solution analytique sous forme de fonction trigonométrique.

Dans le cas de plusieurs individus observés, si la relation de voisinage ne fait pas intervenir un lien entre les différents individus pour la même date ceci reste vrai (l'opérateur sera factorisable tensoriellement par l'identité d'ordre n).

Mais dans notre optique qui considère ce lien, l'approche par les graphes de voisinages et par la modélisation de courbes par des polynômes sera différente. On peut même penser que si on a peu d'individus la différence sera peu importante alors que dans le cas contraire les résultats seront très différents.

Remarque :

-1 Si on s'intéresse à np unités statistiques représentant les individus à des dates différentes, on pourra, comme il est classique de faire, considérer que l'on a une réunion de graphes connexes. Chacun étant défini par un individu et la relation de contiguïté i.e. l'individu à la date t est en relation avec l'individu à la date (t-1) et (t+1). C'est l'approche de Carlier(1985).

Mais on peut aussi, compléter le graphe en reliant tous les individus à une même date (nous discuterons ceci dans le BII).

-2 On peut se demander si, des graphes non-symétriques ne traduiraient pas mieux l'ordre chronologique. Mais, alors les opérateurs de voisinages n'ont peut-être pas les bonnes propriétés enoncées dans le cas symétrique ?

-3 Citons la contribution Mom(1988) aux analyses locales dans le cadre de reseaux de transports, ainsi que celle de Traissac(1985) pour un peuplement forestier.

IV . Distances de Hölder

IV.1 présentation :

En analyse des données, on utilise presque exclusivement des distances euclidiennes, à cause d'une part de la facilité de représentations, et d'autre part des bonnes propriétés d'optimisations dans les espaces euclidiens. Toutefois certaines distances ont de l'intérêt pour des situations spéciales comme dans le cas de mesures longitudinales. Marcotorchino et Michaud(1979) ont présenté des utilisations des métriques de Hölder pour l'agrégation d'échelles ordinales. Ils présentent quelques propriétés de ces métriques appréciables pour des données longitudinales.

La formule générale d'une métrique de Hölder (appelée aussi métrique de Minkowski) est :

$$d_h(x,y) = \sqrt[h]{\sum_i |x_i - y_i|^h} \quad \text{pour un} \quad h \geq 1.$$

Le cas h=1 correspond à la métrique d_1 , distances des écarts absolus appelé aussi distance L_1 en référence à l'espace portant le même nom ou encore distance de type m_1. Le cas h=2 correspond à la distance euclidienne classique.

Notons également la distance définie par :

$$d_\infty(x,y) = \lim_{h \to +\infty} d_h(x,y) = \max_i |x_i - y_i|.$$

Lors du problème d'agrégation d'échelles, résolu dans Marcotorchino et Michaud(1979), l'utilisation de ces différentes métriques nous parrait intéressante, car la solution obtenue, prend en compte diverses caractéristiques de la distribution des valeurs prises par un individu.

Disposant d'un tableau de p mesures ("p juges") sur n individus, le problème est de trouver une mesure X_v ("juge virtuel") agrégeant les

"opinions" des juges dans le sens ou elle minimise la somme des distances aux p mesures :

$$X_v = \arg\min_{X \geq 0} \sum_{k=1}^{p} d(X, Y_k).$$

Les caractéristiques de la distribution des valeurs prise par un individu intervenant dans la solution sont répertoriées dans l'ouvrage précité. Nous en donnons ici les résultats :

Caratéristiques distributionnelles intervenant dans la solution X_v

Métriques	Tendances centrales		Dispersions				
	médiane	moyenne	écart moyen	écart type	coefficient de dissymétrie	autre	étendue
d_1	Oui						
d_2		Oui					
d_3		Oui	Oui				
d_4		Oui		Oui	Oui		
d_h		Oui				Oui	
d_∞							Oui

On sait que, l'utilisation de métriques non-euclidiennes en ACP est un problème assez délicat, source de nombreuses études actuelles, ne serait-ce que l'idée de représentation géométrique sur notre feuille A4 euclidienne d'une figure de points d'un espace affine non-euclidien !

Alors, on peut utiliser ces métriques, dans le cas de mesures répétées de façon assez simple, et, tout d'abord, dans l'esprit où elles ont été étudiées. En effet, pour chaque variable, on peut choisir d'agréger les p mesures au cours du temps, en choisissant l'une des métriques précédentes puis d'effectuer l'ACP du tableau des mesures agrégées.

Remarque :

-1 Pour la métrique d_2 cette méthode correspond à une ACPVI par rapport au facteur individu (i.e. l'effet sujet). Ainsi, si on effectue une ACPVI

des variables agrégées sur le facteur groupe, on obtient l'effet groupe du modèle Split-Plot.
On peut procéder de façon identique, mais en utilisant une autre métrique, comme par exemple d_4. On calcule alors un effet groupe sur des variables synthétisant une évolution sur p instants, qui tient compte de paramètres de dispersions de la série des instants.

-2 Une autre façon d'agréger les p mesures est de prendre la première composante principale, mais le critère d'agrégation n'est pas le même. En ACP le critère est la somme des carrés des covariances.
Cette façon de procéder par des ACP successives où emboîtées est justifiée par la méthode Pré-Statis (chapitre BII). L' emboîtement d'ACP est aussi décrit dans D'Alessio (1988).

IV.2 positionnement multidimensionnel :

Il peut paraître curieux, lorsque l'on dispose des mesures, de chercher à représenter graphiquement les objets étudiés par des techniques de positionnement multidimensionnel (en anglais "Multidimensionnal scaling "). Car justement ces méthodes sont faites pour les cas où l'on n'a pas les mesures mais seulement les dissimilarités entre objets. Mais ici, notre problème est d'utiliser les distances de Hölder non-euclidiennes pour particulariser le temps.

Tout d'abord, rappelons la technique la plus répandue dans ce domaine est la méthode de Torgerson(1958) :
Soit une dissimilarité d qui vérifie les axiomes suivants :
i. $d_{ij}\geq 0$,
ii. $d_{ii}=0$,
iii. $d_{ij} = d_{ji}$.
Rappelons que, si on a **iv.** $d_{ij} \leq d_{ik} + d_{kj}$, on a alors une distance et si l'on a
v. $d_{ij}\leq \max(d_{ik} , d_{kj})$, on a une distance ultramétrique.

On construit la matrice de Torgerson de la façon suivante.

Étants donnés les d_{ij} on construit la matrice W, associée en plaçant l'origine de l'espace affine au centre de gravité du nuage de points, de terme général :

$$W_{ij} = \frac{1}{2}(d^2_{i.} + d^2_{j.} - d^2_{ij} - d^2_{..})$$

$$\text{où } d^2_{i.} = \frac{1}{n}\sum_{j=1}^{n} d^2_{ij} \text{ et } d^2_{..} = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} d^2_{ij}$$

$$\text{en notant } H_{ij} = -\frac{1}{2}d^2_{ij} \text{ on a } W = P_{\Delta^i} H\, {}^t P_{\Delta^i}$$

.

Un théorème du à Shöenberg en 1938, dit que, l'ensemble des n points munis de la dissimilarité d est représentable dans un espace euclidien si et seulement si W est semi-définie positive. Escoufier(1975) relate un résultat, de Holman en 1972 et de Gower en 1974, qui affirme que si d est une distance ultramétrique alors W est semi-définie positive.

On effectue alors l'ACP associée à W. Si elle n'est pas semi-définie positive on ne conserve que les valeurs propres positives dont les vecteurs fournissent une représentation des points. Escoufier(1975) remarque alors que ceci revient à approcher W par une matrice W^+ semi-définie positive au sens des moindres-carrés. On peut noter que d'autres techniques pour trouver une approximation de W semi-défini-positive existent. Citons celui de la constante additive avec notamment la solution donnée par Cailliez(1983).

On trouvera dans D'Aubigny(1989) un approfondissement de ces problèmes de plongement euclidien et d'approximations d'une forme semi-définie positive. Cet auteur cite d'ailleurs le fait que les distances de Hölder (pour h≠2) ne vérifient pas le théorème de Schöenberg.

On pourra utiliser ces distances malgré tout en positionnement multidimensionnel avec la restriction précédente.

Lorsque l'on a q variables répétées p fois dans le temps sur n individus, on pourra alors calculer les distances entre individus par exemple par :

$$d(i,i') = \sqrt{\sum_{j=1}^{q} (d_h(X_{ij}, X_{i'j}))^2} \text{ pour un h donné ,}$$

où X_{ij}est un p-vecteur.
Pour q>1, c'est une sorte de généralisation de la distance d_2 euclidienne classique. En effet, pour h=2 on a d=d_2 sur les individus qui sont alors vu comme des vecteurs de R^{qP}, mais aussi pour q=1 on a d=d_h.

IV.3 distances et ACPVI :

Puisqu'en général, la distance précédente ne donnera pas par la formule de Torgerson une matrice W semi-définie positive, on pourra l'utiliser comme contrainte dans les ACPVI de chaque tableau.

Dans la parfaite analogie avec l'ACPVI, on peut imaginer un problème qui semble difficile à résoudre. Il s'agit de trouver une métrique ou semi-métrique Q_s, sur les lignes d'un tableau structure S à n lignes et s colonnes, telle que l'opérateur de produits scalaires sur S soit proche au sens de la norme trace du W construit précédemment :

$$\exists ? \tilde{Q} / \tilde{Q} = \arg\min_{Q_s} \left\| WD - SQ_s {}^t SD \right\|_{tr} .$$

Naturellement si W est semi-défini positif on a la solution semi-métrique de l'ACPVI classique après avoir fait l'ACP associé à WD.

Remarque :

-1 S sera par exemple la matrice identifiant la structure des groupes d'individus. Si la solution existe, on effectue alors l'ACP de (S,Q_s,D).

V . Conclusions

Dans ce chapitre, on a soulevé les problèmes sous-jacent à toute analyse de mesures chronologiques.
Doit-on considérer que l'on a des variables indépendantes ? Y-a t-il une contrainte ordinale ? Doit-on considérer que la variable mesuré s'associe avec le temps pour donner une variable ordinale que l'on quantifie de façon floue ?

La réponse, après cet exposé de divers aspects utilisables pour ce type de données, n'est certainement pas catégorique et nous ne le voudrions surtout pas. Il est certain toutefois, que les choix préconisés pour rendre compte du "temps", assurent chacun des aspects importants, lorsque l'on est amené à traiter les données.

Il semble important de distinguer :
- la contrainte d'ordre qui a fixé l'explication du temps pour nous permettre une interprétation du passé vers le futur,
- et l'aspect dépendance locale par les graphes de voisinage.

Le problème est donc de choisir, d'avoir une optique extrinsèque ou intrinsèque de la répétition.
Le § II a décrit une approche double, le § III a été vraiment extrinsèque, et le§ IV a présenté un aspect intrinsèque. Un aspect sous-jacent à ces deux approches est la notion de distance entre courbes.

Notons aussi que dans la littérature anglo-saxone et américaine comme par exemple dans Agresti(1984) on s'occupe plus directement de trouver des statistiques, des indices d'associations pour des données ordinales, ce dont nous n'avons pas explicitement parlé ici.

VI . Références bibliographiques

AGRESTI, A.G. (1984) Analysis of Ordinal Categorical Data. Wilegard, New-York
d'AUBIGNY, G. (1989) L'Analyse Multidimensionnelle des données de dissimilarité. Thèse d'état, Grenoble.
BARLOW, R.E. BARTHOLOMEW, D.J. BREMNER,J.M. and BRUNK, H.D. (1972) Statistical Inference under order restrictions : the theory and applications of isotonic regression. Wiley, New-York.
BREMNER, J.M. (1982) An algorithm for NonNegative Least Squares and projection into cones. In : Data Analysis and Informatics . COMPSTAT 1982 , 155-160. Caussinus,H. Pagès, J.P. and Tomassone, R. (eds). Wien : Physica-Verlag.
BRADLEY, R.A. KATTI, S.K. and COONS, I.J. (1962) Optimal scaling for ordered categories. Psychometrika, 355-374.
CAILLIEZ, F.(1983) The analytical solution of the additive constant problem. Psychometrika, 305-310.
CLAYTON, D.G. (1974) Some odds ratio statistics for the analysis of ordered categorical data. Biometrika, 525-531.
CROON, M. (1990) Latent class Analysis with ordered latent classes. Brit.J.Math.Stat.Psychol. 171-192.
D'ALESSIO, G (1988) Multistep Principal Components Analysis : a new approach for the analysis of contingency tables series. In : Classification and related Methods of Data Analysis. H.H.Bock(ed), Elsevier Science Publishers B.V. North-Holland.
DURAND, J.F. (1989) Analyse en composantes principales par rapport à des variables instrumentales via les transformations splines univariées. Rapport Technique n° 8902, INRA, Unité de Biométrie, Montpellier.
DYKSTRA, R.L. (1983) An algorithm for restricted least squares regression. J.A.S.A. 837-842.
El FAOUZI, N.E (1992) Extensions non-linéaies de l'analyse en composantes principales. Thèse de doctorat Biostatistiques, Montpellier II.
ESCOUFIER, Y. (1975) Le Positionnement Multidimensionnel. R.S.A., XXIII, 4, 5-14.
GIFI, A. (1990) Non-linear Multivariate Analysis . DSWO Press, Leiden.
GILULA, Z. and RITOV, Y. (1990) Inferential ordinal correspondence analysis: motivation, derivation and limitations. Int.Stat.Rev. 99-108.
HEISER, W.J. (1989) The city-block model for three-way multidimensionl scaling. In: (R.Coppi and S.Bolasco). Multiway Data Analysis, North-Holland, Amsterdam.
JANSEN, M. and VAN DUIJN, M. (1992) Extensions of Rasch's Multiplicative poisson model. Psychometrika, 405-414.
KRUSKAL, J.B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 1-27.
KRUSKAL, J.B. (1964) Nonmetric Multidimensional scaling : a numerical method. Psychometrika, 1-27.
KRUSKAL, J.B. (1965) Analysis of factorial esperiments by estimating monotone transformations of the data. J.R.S.S. B,251-263.
KRUSKAL, J.B. (1971) Monotone Regression : continuity and differentiability propertie. Psychometrika, 57-62.
KRUSKAL, J.B. , SHEPARD, R.N. (1974) A nonmetric variety of linear factor analysis. Psychometrika, 123-157.

LEBART, L. (1969) Analyse statistique de la contiguité . Publication de l'institut de statistiques universitaire.Paris XVIII , 81-112.
Le FOLL, Y. (1982) Pondération des distances en analyse factorielle. S.A.D. 13-31.
MARCOTORCHINO, F. and MICHAUD,P. (1979) Optimisation en analyse ordinale des données. Masson, Paris.
MEOT, A. (1992) Explicitation de contraintes de voisinage en analyse multivariée. Thèse de doctorat Biométrie, Lyon I.
MOM, A. (1988) Méthodologie statistique de la classification des réseaux de transports. Thèse de doctorat, USTL Montpellier II
NISHISATO, S. (1980) Analysis of categoical data : Dual scaling and its applications. University of Toronto Press, Toronto.
PCHÉNITCHNY, B. and DANILINE, Y. (1977) Méthodes Numériques dans les Problèmes d'Extrémum. MIR, Moscou.
RAMSAY, J.O. (1988) Monotone Regression Splines in Action. Stat.Sci. 425-461.
SAITO, T. and OTSU, T. (1988) A method of optimal scaling for multivariate ordinal data and its extensions. Psychometrika, 5-25.
SCHRIEVER, B.F.(1983) Scaling of order dependent categorical variables with correspondence analysis. Int.Stat.Rev, 225-238.
TECHENÉ, J.J. (1980) Reductions optimales d'opérateurs, application aux analyses factorielles. Thèse de 3éme cycle Toulouse.
TENENHAUS, M. and VACHETTE, J.L. (1977) PRINQUAL : un programme d'analyse en composantes principales de variables nominales ou numeriques. Cahier de recherche du CESA, n°68, Jouy-en-Josas.
TENENHAUS, M. (1983) L'analyse de données qualitatives par des méthodes de codage optimal. Cahier de recherche du CESA, n°229, Jouy-en-Josas.
TENENHAUS, M. (1988) Canonical analysis of two convex polyhedral cones and aplications. Psychometrika, 503-524.
TRAISSAC, P. (1985) Analyse en composantes principales par rapport à une structure (partition, hierarchie, graphe) : Application à l'étude d'une population d'Hevea. Mémoire de D.E.A., USTL Montpellier II.
TORGERSON , W. (1958) Theory and Methods of Scaling. Wiley, New York.
TUTZ, G. (1990) Sequential item response models with an ordered response. Brit.J.Math.Stat. Psychol. 39-55.
WELLS, A.J. (1991) Optimal presentation orders for the method of paired comparisons. Brit. J. Math. Stat. Psychol. 181-193.
WINSBERG, S. and RAMSAY, J.O. (1983) Monotone spline transformations for dimension reduction. Psyhometrika, 575-595.
WINSBERG, S. and CARROLL, J. (1989) A quasi-nonmetric method for multidimensional scaling of multi-way data via a restricted case of an extended INDSCAL model. In: (R.Coppi and S.Bolasco). Multiway Data Analysis,
YOUNG, F.W (1981) Quantitative Analysis of qualitative Data. Psychometrika, 357-388.
YOUNG, F.W. TAKANE, Y. De LEEUW, J. (1978) The Principal components of mixed measurement level data; An alternating least squares method with optimal scaling. Psychometrika, 279-282.

BII
Aperçus des méthodes factorielles multitableaux pour des mesures répétées

I . Introduction

Lorsque nous avons parcouru les différentes approches pour l'analyse de facteurs à mesures répétées dans le chapitre AI, nous avons constaté que, outre les problèmes de modèles de distributions et de tests, on était amené à

considérer, lors du passage à des ACP où ACPVI, que l'on avait plusieurs tableaux (un pour chaque date). Ces tableaux étaient juxtaposés pour l'analyse Split-plot ou concatenés pour l'analyse multivariée, afin d'effectuer l'analyse factorielle.

Donc on peut dire que les analyses des facteurs à mesures répétées, par le biais de méthodes traitant plusieurs tableaux à la fois, sont des méthodes multitableaux d'ordre trois. Le terme multitableau est la traduction de l'anglais multi-array, c'est à dire un tableau à plusieurs entrées, on parle aussi de tableau à 3 entrées.

Ceci nous amène donc, dans ce chapitre, à examiner quelques méthodes multitableaux générales, applicables aux mesures répétées, ainsi que des méthodes plus spécifiquement crées pour ce type de données. Une brève description de ces méthodes, nous permettra dans le chapitre suivant d'envisager, l'apport de la vision tensorielle du problème qui pourra alors s'étendre à plusieurs entrées (plus seulement 3).

De nombreux auteurs se sont prêtés à la comparaison des diverses méthodes aussi bien du point de vue, algébrique et statistique, que du point de vue pratique sur un jeu de données. On peut citer par exemple Janssen et al. (eds) (1987) où la plupart des méthodes multitableaux sont testés sur le même jeu de données. Kiers(1988) a proposé des comparaisons par rapport aux fonctions objectifs, c'est à dire les critères optimisés.

Nous distinguerons, les méthodes qui généralisent une analyse à deux tableaux et traitent effectivement plusieurs tableaux, et les méthodes qui se ramènent à deux tableaux. Ces dernières sont en général plus portées sur la structure des données longitudinales.

Enfin, dans l'esprit d'arranger les tableaux étudiés de façon adéquate et empirique pour se ramener à une analyse classique, nous proposerons deux façons simples de traiter des répétitions. L'une cherchera à déceler des évolutions stables à longs termes, l'autre, les conséquences à courts termes à posteriori.

Avant tout énonçons le cas idéal pour traiter simultanément p tableaux de q mesures sur n individus. On souhaiterait trouver une décomposition en valeurs singulières commune aux p tableaux :

U et V unitaires tels que ${}^tUX_kV = \Delta_k \quad \forall k = 1,..p,$

avec Δ_k diagonale positive .

Malheureusement déjà pour k=2 ceci n'est vrai que si et seulement si ${}^tX_iX_j$ et $X_i{}^tX_j$ (i,j=1,2) sont auto-adjointes, Rao et Mitra(1971). Ces auteurs montrent également que pour p matrices définies positives, il existe une matrice non-singulière V tel que ${}^tVV_kV$ (k=1,..,p) est diagonale si et seulement si il existe un i tel que ${}^tV_jV_i^{-1}V_h={}^tV_hV_i^{-1}V_j \quad \forall j, h$.

L'objectif de ce chapitre est donc de rappeler brièvement les méthodes les plus originales traitant de plusieurs tableaux, et comment elles peuvent s'appliquer au cas qui nous occupe. Une multitude d'autres méthodes utilisant l'analyse factorielle existent, nous en donnons quelques références à la fin de ce chapitre.

II . Généralisations de méthodes à deux tableaux

II.1 Analyses canoniques généralisées , ACPVI généralisée :

L'analyse canonique est à la base de nombreuses analyses et a un caractère général. Sa généralisation à plus de 2 tableaux entraîne diverses méthodes comme le relate Sabatier(1991) en introduisant divers critères et contraintes pour le problème générique de l'ordination. La plus répandue est certainement l'Analyse Canonique Généralisée (ACG) de Caroll(1968).

De nombreux ouvrages traitent de la méthode, nous reprenons la présentation de Saporta(1981) qui la décrit dans le même contexte que le notre.

On dispose de p tableaux de n lignes et q colonnes. Dans le cas p=2 l'analyse canonique consiste à trouver des couples (c1,c2) de nouvelles variables "latentes", que l'on appellera variables canoniques. Ces variables sont des combinaisons linéaires des colonnes de X_1 et X_2, respectivement, ayant la plus grande corrélation r. Les éléments de R^q (f1,f2) réalisant les combinaisons linéaires sont appelées facteurs canoniques.

En notant P_1 et P_2 les projecteurs orthogonaux respectivement sur les espaces engendrés par les applications linéaires associées à X_1 et X_2, les propriétés des solutions vérifient :

i. $\begin{cases} P_1P_2c_1 = r^2c_1 \text{ et } P_2P_1c_2 = r^2c_2, \end{cases}$

ii. $\begin{cases} c_1 = \frac{1}{\sqrt{r^2}} P_1c_2 \text{ et } c_2 = \frac{1}{\sqrt{r^2}} P_2c_1, \end{cases}$

iii. $\begin{cases} \text{si } z = c_1 + c_2 \text{ alors } (P_1 + P_2)z = (1+\sqrt{r^2})z, \\ \text{et donc } c_1 = \frac{1}{1+\sqrt{r^2}} P_1z \text{ et } c_2 = \frac{1}{1+\sqrt{r^2}} P_2z, \end{cases}$

iv. $\begin{pmatrix} P_1 & P_1P_2 \\ P_2P_1 & P_2 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = (1+\sqrt{r^2})\begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$

v.

$\begin{cases} \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}\begin{pmatrix} V_{11}^{-1} & 0 \\ 0 & V_{22}^{-1} \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = (1+\sqrt{r^2})\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \text{ où } V_{ij}={}^tX_iX_j \text{ et } c_i = X_if_i. \\ \text{C'est à dire que les vecteurs formés des facteurs canoniques empilés sont} \\ \text{les axes principaux de l'ACP du triplet } ((X_1\ X_2), \begin{pmatrix} V_{11}^{-1} & 0 \\ 0 & V_{22}^{-1} \end{pmatrix}, Id_n). \end{cases}$

La propriété **iii** est à la base de la généralisation proposée par Caroll(1968) car on a aussi la propriété suivante :

iii⁺. Les vecteurs propres z de $P_1 + P_2$ maximisent la somme des carrés des corrélation multiples : $R^2(z; X_1) + R^2(z; X_2)$.

L'analyse canonique généralisé des p tableaux sera la recherche de z :

$$z = \underset{u}{\operatorname{argmax}} \sum_{k=1}^{p} R^2(u; X_k).$$

La solution est donc donnée par la généralisation des propriétés **iii** et **v** c'est à dire :

iii^g. z est vecteur propre de $\sum_{k=1}^{p} P_k$ et est appelé composante canonique ,

v^g.

$\begin{cases} z = \sum_{k=1}^{p} X_kf_k \text{ , où les vecteurs formés des facteurs canoniques } f_k \text{ empilés sont} \\ \text{les axes principaux de l'ACP du triplet } ((X_1 \dots X_p), \begin{pmatrix} V_{11}^{-1} & & 0 \\ & \ddots & \\ 0 & & V_{pp}^{-1} \end{pmatrix}, Id_n). \end{cases}$

Les variables canoniques c_k sont alors données par , $c_k = \frac{1}{\lambda_z} P_kz = X_kf_k.$

La contrainte sur les c_k est en fait : $\sum_{k=1}^{p} \| c_k \|^2 = 1.$

On a aussi une généralisation de la propriété **iv** qui sera à relier à une analyse que nous verrons au BIII sous le nom de Pré-Statis-Croisé- modèle :

$$\begin{pmatrix} P_1 & P_1P_2 & \dots & P_1P_p \\ P_2P_1 & P_2 & & P_2P_p \\ \vdots & \vdots & \ddots & \vdots \\ P_pP_1 & P_pP_2 & \dots & P_p \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = \lambda_z \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix}.$$

Donc l'analyse canonique généralisée sur des tableaux répétés dans le temps peut nous permettre d'interpréter l'évolution globale des variables. Cette analyse sera, bien naturellement, plus facilement interprétable qu'une analyse canonique généralisée sur des tableaux quelconques, car les variables sont les mêmes dans chaque tableau.

Il est important de constater que l'introduction de la métrique dans l'ACP de la propriété **v^g** nous permet d'avoir des projecteurs et donc un lien simple entre les composantes principales (composante canoniques) et les variables canoniques.

En présence d'une structure de groupe sur les individus on peut adopter plusieurs procédures. Soit :

a) en partant de l'ACP équivalente à l'analyse canonique généralisée, on cherche l'ACPVI par rapport à la structure,

b) on cherche l'analyse canonique généralisée des p+1 tableaux,

c) on fait une analyse canonique généralisée sous contrainte. Des liens très étroits existent entre ces trois procédures.

Liens entre les analyses a) b) c) :

i. L'analyse a) est équivalente à l'analyse c) dans le sens où les composantes canoniques de c) sont les composantes principales de a).

ii. L'analyse b) est équivalente à l'analyse a) seulement dans le cas ou p=1(voir Sabatier(1983,1987)).

On notera évidemment qu'avec l'analyse b) on perd toute logique de mesures répétées avec une structure de groupe sur les individus.

Sabatier(1987) a proposé plusieurs généralisations de l'ACPVI d'un tableau par rapport à un tableau structure, en l'ACPVI d'une famille de tableaux par rapport à un tableau structure. Nous y renvoyons pour plus de détails, mais notons que les généralisations proposées se rapprochent soit de l'analyse canonique généralisée sous contraintes, soit de la méthode Statis sur les triplets d'ACPVI.

II.2 Statis, Pré-Statis, CPCA :

La méthode Statis adopte directement l'idée de réduction des données de façon à avoir une représentation des individus comme en ACP. Cette méthode peut être vue en détail dans le livre de Lavit(1988).

L'idée générale est de construire le meilleur **compromis** (au sens des moindres-carrés) des opérateurs d'Escoufier (i.e. de produits scalaires sur les unités statistiques) pour fournir une représentation des individus optimale.
Pour cela on diagonalise la matrice des Rv des opérateurs W_kD (i.e. les cosinus pour le produit scalaire $(u,v)\mathrm{a}\ tr(uv^*)$) qui correspond à une matrice de corrélation mais dont toutes les valeurs sont positives (les W_kD sont semi-définis positifs D-symétriques). Le premier vecteur propre τ est peut être pris positif donc la "première composante principale", $\sum_{k=1}^{p}\tau_k W_k D$,est un opérateur semi-défini positif et correspond au **compromis** cherché.

On peut alors effectuer l'ACP associée à ce compromis, ce qui défini l'**inter-structure,** et l'ACPVI de chaque tableau par rapport à ce compromis ce qui défini l'**intra-structure**. Une représentation des p instants à partir des vecteurs propres de la matrice des Rv donne aussi une analyse de l'**inter-structure**.

Notons que si pour chaque triplet, on a choisi comme métrique sur les individus la métrique de Mahalanobis on retrouve, en diagonalisant le compromis, l'ACG, mais de façon optimale puisque cette opérateur

maximise un critère de moindres-carrés. C'est à dire que dans l'ACG on diagonalise le compromis, non-optimal, défini à partir du vecteur $^t(1,1,...,1)$.

La méthode Pré-Statis consiste, à travailler sur les tableaux qui doivent avoir les mêmes dimensions (ce qui est le cas pour des mesures répétées), et à construire des compromis de tableaux fondés sur la diagonalisation de la matrice des produits scalaires des tableaux.
Jaffrenou(1978) a utilisé cette méthode sous le nom d'ACP d'une famille finie de vecteurs aléatoires réels de dimension m, pour aboutir à une analyse triadique. Nous lui donnons cet autre nom de part les liens évident avec la méthode Statis, et du fait de la différence des objets étudiés.
L'avantage par rapport à la méthode Statis est donc que l'on peut avoir plusieurs compromis et que l'on a aussi la représentation des variables pour chaque compromis. Chaque compromis représente donc une tendance, ils sont indépendants dans le sens où les combinaisons des tableaux sont des vecteurs propres d'un même opérateur. On a ainsi une représentation des instants dans l'espace des tendances (i.e. l'inter-structure), avec une description de chaque tendance par une ACP.

Cette analyse peut être très intéressante, dans le cas de données longitudinales (i.e. p assez grand), car les autres méthodes peuvent être difficiles du point de vue temps de calcul informatique. Ici on peut montrer, (voir BIII), que le Pré-Statis précédent est équivalent à une ACP sur le tableaux à nq lignes et p colonnes. Les axes principaux fournissent une représentation des instants et les composantes principales sont les tendances que l'on analyse en "devectorialisant" le vecteur à nq lignes pour donner la matrice (n,q).
On peut aussi chercher des compromis des variables, c'est à dire des compromis des q tableaux (n,p). On a alors des variables vectorielles résumant au mieux, au sens des moindres-carrés, les évolutions des individus. Évolutions que l'on analyse par l'ACP de chaque variable vectorielle compromis.

Avoir une contrainte sur les lignes dans Statis ou Pré-Statis peut poser des problèmes de choix. Ils dépendront bien évidemment de la

problématique. Dans Pré-Statis, où l'on souhaite exhiber les tendances, on se posera deux questions :

- "En quoi le classement par le facteur groupe explique la tendance ?", on effectue alors une ACPVI de chaque compromis représentant la tendance par rapport à la structure de groupe,
- "Quelle est la tendance construite par les groupes ?", on fait alors une ACP du compromis construit par les tableaux projetés sur la structure.

Ce dernier problème correspond à une ACPVI généralisée de Sabatier(1987).

CPCA est la méthode, "Common Principal Component Analysis" élaborée par Flury(1984,1987). Il souhaite, en quelque sorte, contourner le théorème de Rao et Mitra cité en introduction. La méthode est en fait mise en oeuvre pour des observations indépendantes. L'idée est de trouver des composantes principales communes à plusieurs vecteurs aléatoires indépendants. Donc ici en toute rigueur elle n'est pas applicable puisque les individus sont les mêmes à des instants différents.

Soit p variables q - vectorielles indépendantes $X_k \sim N_q(\mu_k,\Sigma_k)$,
l' hypothèse testée est $H_b : {}^tV\Sigma_k V = \Lambda_k \ ,k = 1\!\!L\ p$
avec V orthogonale et Λ_k diagonale ,
alors $U_k = X_k V$ est la composante principale commune pour X_k.

Comme les matrices de covariances empiriques suivent des lois de Wishart, Flury estime les q composantes principales communes (ou plutôt les q axes principaux communs) par maximum de vraisemblance.
Flury(1987) généralise cette méthode à la recherche moins stricte de r<q composantes principales communes, ce qui est moins restrictif.

En fait, on peut facilement comparer la méthode CPCA et la méthode Statis, en constatant que Statis consiste à estimer au sens des moindres-carrés des composantes principales communes sous l'hypothèse :

$$H_b - : {}^tV\Sigma_k V = \frac{1}{p\ \tau_k}\Lambda \ ,k = 1\!\!L\ p$$

Cette Hypothèse est moins générale que celle de CPCA, mais la solution est plus aisée à obtenir. Le modèle Tuckals2 de Tucker(1966) et Kroonenberg(1983) semble plus proche de CPCA et même du Partial-CPCA de Flury(1987) puisqu'il n'impose pas de forme particulière à Λ_k et que l'on recherche aussi un nombre r<q de composantes.
Ces points mériteraient d'être approfondis mais dans la description rapide que nous faisons ici, il nous importe seulement de faire le tour de quelques méthodes et d'envisager globalement leurs différences conceptuelles.

II.3 l'AFCM:

Cette méthode a une longue histoire depuis Richarson et Kuder(1933), Horst(1935), Hirschfeld(1935), Guttman(1941) etc... pour arriver aux descriptions actuelles de Benzecri et al.(1973), Lebart(1977), Escofier et Pages(1982,1984), Greenacre(1984) et Escoufier(1985). On trouvera un historique intéressant ainsi que toutes les références anciennes citées ici et d'autres encore dans Nishisato(1980).

L'analyse des correspondances multiples est certainement l'outil statistique le plus largement utilisée en analyse des données de variables qualitatives. Elle se veut une généralisation de l'analyse des correspondances, mais d'ores et déjà notons que les écarts à l'indépendance observés sont pris deux à deux. Au BIII on proposera une analyse factorielle des correspondances de k variables modélisant l'écart à l'indépendance prise entre les k variables.

En fait on peut voir l'AFC comme une analyse canonique des tableaux des indicatrices des deux variables. L'AFCM est alors l'ACG des tableaux des indicatrices de chaque variable étudiée. Cette généralisation paraissait donc naturelle.

Ainsi l'étude d'une variable qualitative répétée dans le temps peut être menée par une AFCM des p variables la représentant à différentes dates.

Pour l'étude d'une structure de groupe sur les lignes, on effectue alors une AFCVI où une AFCMVI, Sabatier(1987). Van der Heijden(1987) décrit suivant les différents tableaux rencontrés en sociologie, psychométrie, éthologie etc... les façons de faire des AFC ou des AFCM en reliant celles-ci aux modèles log-linéaires.

III . Quelques méthodes spécifiques

III.1 Longi :

Cette méthode est sans doute celle qui correspond le mieux à une problématique générale d'étude de données longitudinales. Elle a été élaborée spécifiquement pour ce genre de données par Pernin(1986), Pontier et Pernin(1987), Pontier et al.(1990).

L'idée est assez proche des ACPVI avec le modèle Split-plot vu au chapitre AI. En effet, elle tire parti de l'analyse canonique et de la méthode que les auteurs ont appelé l'analyse canonique complète qui consiste à identifier toutes les valeurs propres issues de l'analyse canonique :
si l'on diagonalise l'opérateur $P_X P_Y$, en notant L(A) l'espace engendré par les colonnes de A,

. le sous-espace propre correspondant à la valeur propre λ =1 est l'intersection des deux sous-espaces L(X) et L(Y).

. le sous-espace somme directe des sous-espaces propres de valeur propre $0<\lambda<1$ est appelé l'espace des obliques de L(X) par rapport à L(Y).

. le sous-espace somme directe des sous-espaces propres de valeur propre λ =0 est l'intersection de L(X) et de l'orthogonal de L(Y).

On considère le tableau Z à q colonnes et np lignes obtenu par juxtaposition des p tableaux observés. On construit alors le tableau V des indicatrices des dates (visites) et le tableau S des indicatrices des individus (sujets).

Le but de l'analyse est d'une part de caractériser l'évolution globale, et d'autre part de différencier les trajectoires des individus.

Le premier objectif est obtenu en effectuant l'analyse canonique entre Z et $L(V)\cap L(S)^{\perp}$. Ce sous-espace est obtenue préalablement par analyse canonique complète entre V et S. On souhaite donc obtenir un effet de la

visite indépendamment de l'effet sujet; pour améliorer l'analyse comme en ACP classique on centrera et réduira le tableau Z avant l'analyse.

Le deuxième objectif est obtenu en effectuant l'analyse canonique entre Z et $L(S)\cap L(V)^{\perp}$. On souhaite donc avoir un effet sujet indépendamment de la visite; préalablement on centrera et réduira Z par date.

III.2 ACP d'évolution :

L'objectif poursuivi ici semble le même que dans la méthode précédente. Ici on tire parti de la structure sous-jacente de la collecte des données et l'on effectue des analyses locales. Nous avons déjà décrit la méthodologie générale dans le chapitre BI. Carlier(1985) sur les traces de Le Foll(1982) est à l'origine de ce cas particulier d'analyse en composante de voisinage de Meot(1992).

Comme nous l'avons déjà évoqué au chapitre BI le graphe utilisé par Carlier est une réunion des n graphes connexes, de chaque individu pour la relation temporelle de contiguïté. Donc dans ce cas on ne tient pas compte de la variabilité inter-individuelle à une même date.

On propose donc de rajouter au graphe précédent les arêtes joignant les individus à une même date. Les vecteurs propres de l'opérateur de lissage de voisinage dépendront donc non seulement du nombre de dates mais aussi du nombre d'individus.
Cette façon de procéder nous semble plus appropriée car on effectue un lissage local et par date simultanément.

Si il n'y a pas de contiguïté (i.e. la matrice M du graphe est le produit tensoriel de $1_n{}^t1_n$ par l'identité d'ordre p), en effectuant l'ACV on aura un effet date. Si la relation temporelle est étendue à tous les instants on aura l'effet total (i.e. l'ACV est l'ACP totale).

Alors que dans le cas de la simple réunion des graphes de chaque individus. Si il n'y a pas de contiguïté on a aucune inertie à analyser, et si la relation temporelle est totale on aura l'effet individu.

Dans le graphe complété, il sera intéressant de pondérer différemment les arêtes liées à la contiguïté et celles liées à la "coexistence". On retrouve le cadre défini par Le Foll(1982) puisque ce sont les arêtes qui sont pondérées et non les unités statistiques. L'ACP est celle de (Z,Q,A-V) où A-V est une semi-métrique. V est la matrice des poids des arêtes et A la matrice diagonale de terme la marge de V.

III.3 Analyse de processus, AFC et chaînes de Markov :

De nombreux auteurs français ont travaillé dans ce domaine, comme Escoufier(1970), Deville(1974), Dauxois et Pousse(1976) pour les plus importants et plus récemment Besse et Ramsay(1986) et Pardoux(1989). Pour un exposé détaillé et clair des analyses de processus on consultera la thèse de Saporta(1981) qui traite des processus scalaires vectoriels, et aussi des processus qualitatifs. Le nom générique adopté par Deville et Saporta pour leurs analyses factorielles est l'analyse harmonique.

L'outil de base est la décomposition de Karhunen-Loève. Certaines analogies avec l'analyse spectrale dans des cas particuliers, sont décrites par Saporta(1981).

Soit X_t un processus centré de $L^2([0;p])$. L'opérateur de covariance C défini par :

$$C(f(t)) = \int_0^p C(t,s)f(s)ds = \int_0^p E(X_t X_s)f(s)ds,$$

est compact, auto-adjoint et positif. La famille de ses fonctions propres $f_k(.)$ nous permet d'écrire le processus sous la forme d'une somme de processus quasi-déterministes :

$$X_t(\omega) = \sum_{k=1}^{\infty} \xi_k(\omega) f_k(t)$$

où les $\xi_k(\omega) = \int_0^p X_s(\omega) f_k(s)ds$ sont des variables aléatoires réelles non-corrélées d'espérance nulle, de variance λ_k..

Les sommes finies effectuées suivant l'ordre décroissant des valeurs propres, fournissent les meilleures approximations en somme de processus quasi-déterministes au sens des moindres-carrés.

Dans la pratique on n'a qu'un nombre fini de d'observations, alors des propriétés :

- d'équivalences, Besse et Ramsay(1986), entre l'ACP de fonctions plongées dans des espaces de fonctions à noyaux auto-reproduisants et, l'ACP des données discrétisées avec une métrique définie par le noyau auto-reproduisant,
- et de convergence dans Dauxois et Pousse(1976),

assurent la cohérence entre l'ACP de variables aléatoires et l'ACP sur un échantillon.

Pour des données qualitatives, un processus vectoriel est difficilement analysable par ce genre d'analyse. Pour un processus ponctuel qualitatif par contre des solutions existent.

Yousfate(1986) s'est intéressé au lien entre la décomposition d'un processus qualitatif Markovien et l'AFC. Dans le cas d'un processus stationnaire du premier ordre, il propose l' ACP de :

(A , D^{-1} , D) où A est la transposée de la matrice de transition, D est la matrice diagonale des probabilités. Cette ACP décompose donc le processus suivant l'ordre de prévisibilité.

L'estimation de A est donnée à l'aide de la matrice M des temps de moyens de passage, M :

$$\hat{A} = I - (DM(D1_k{}^t1_k - I) + I)^{-1}(I - D1_k{}^t1_k),$$

où $D = \mathrm{diag}(M)^{-1}$.

Cet auteur donne un cas intéressant d'estimation de la matrice A=(a_{ij}) : celui où l'on connaît les tableaux de contingence de t_0 à t.

Cette estimation est :

$$\hat{A}_{ij} = \frac{\hat{p}_{ij}}{\hat{p}_{.j}} \text{ pour i , j = 1,...,g,}$$

$$\text{avec } \hat{p}_{ij} = \frac{1}{t-t_0}\sum_{s=t_{0+1}}^{t}\left(\frac{n_{ij}(s)}{n} - \frac{n_i(s)n_j(s-1)}{n^2} + \frac{(n_i(s)+n_i(s-1))(n_j(s)+n_j(s-1))}{4n^2}\right),$$

$$\text{et } \hat{p}_{.j} = \sum_{i=1}^{g}\hat{p}_{ij},$$

où $n_{ij}(s)$ = nombre en i à l ' instant s et en j à l ' instant s - 1,

$n_i(s)$ = nombre en i à l ' instant s ,

n = nombre total d ' individus ,

g est le nombre de modalité de la variable qualitative.

L'ACP présentée par Yousfate est différente d'une AFC sur la matrice des probabilités conjointes.

III.4 Grade of Membership (GoM) :

Ce modèle est présenté par Woodbury et Manton(1982) qui en sont les principaux instigateurs, comme une technique de discrimination "factorielle". La méthode est introduite pour des variables qualitatives mesurées sur des individus. Mais, Manton et al.(1986) utilisent un modèle GoM pour des données qualitatives répétées.

Le modèle initial est le suivant.

Soit la variable j qualitative à l_j modalités alors pour le $i^{ème}$ individu on souhaite écrire:

$$P(x_{ijl} = 1) = p_{ijl} = \sum_{k=1}^{r} g_{ik}\lambda_{kjl},$$

où , x_{ijl} = 1 si le ième individu pour la j variable prend la modalité l ,

avec les contraintes $0 \le g_{ik}$ et $\sum_k g_{ik} = 1$,

et l ' on a par construction $\sum_l \lambda_{kjl=1}$.

Les coefficients g sont les poids chiffrants les degrés d'appartenance à un groupe k (Grade of Membership), et les coefficients λ sont les probabilités

qu'un individu purement du groupe k (g_{ik}=1), prenne la modalité l de la variable j.

L'estimation des paramètres se fait par maximum de vraisemblance . La vraisemblance multinomiale est

$$L = \prod_i \prod_j \prod_l \left(\sum_{k=1}^{r} g_{ik}\lambda_{kjl}\right)^{x_{ijl}}.$$

Une comparaison de ce modèle avec le "latent class model", de Goodman et Lazarsfeld exprimé dans Everitt(1984), et avec le modèle d'analyse factorielle, a été faite par Manton et Stallard(1991). Pour ce dernier ils soulignent que, la maximisation se fait en n'utilisant seulement les moments d'ordre 2, alors que dans le GoM tous les moments sont pris en compte, et que les contraintes sur les paramètres sont très différentes.

GoM est plus proche du "latent class model" par la maximisation d'une vraisemblance multinomiale. On retrouve dans les deux modèles les paramètres λ, mais les coefficients g sont propres au modèle GoM. Ils représentent l'hétérogénéité au sein de chaque groupe.

Lorsque l'on a des mesures répétées pour une variable Manton et al.(1986) introduisent des distributions conditionnelles Poissoniennes pour les observations avec comme espérance le modèle GoM :

$$P(x_{i(t+1)u} \,/\, x_{its} = 1) = (\lambda_{itsu})^{x_{i(t+1)u}} \exp(-\lambda_{itsu}) \Big/ x_{i(t+1)u}!,$$

$$\text{avec } \lambda_{itsu} = \sum_{k=1}^{r} g_{ikt}\lambda_{ktsu}.$$

Ils introduisent aussi des covariables, qui suivent des distributions de Poisson avec comme espérance le modèle GoM.

Supposant d'abord l'indépendance entre, la variable étudiée, et les covariables conditionnellement aux g, ils se placent dans un contexte de processus de Markov (ou semi-Markovien) afin d'estimer les transitions conditionnellement aux groupes flous.

Nous avons ici décrit un modèle très spécifique qui s'éloigne quelque peu du soucis premier de l'analyse factorielle de réduction et représentation

des données. La méthode GoM se rapproche plutôt du soucis de classification et discrimination.

IV . Méthodes Ad Hoc avec des "super-matrices"

IV.1 Analyse de comportements stables à long terme :

Ici on se place dans le cas où l'on a q mesures numériques évaluées sur n individus en p instants successifs, et p est assez grand. Éventuellement on aura un facteur groupe de classement des individus obtenu indépendamment des q variables.

Au §II, nous avons pu voir que la plupart des méthodes voulant traiter plusieurs tableaux simultanément se ramenaient souvent à une ACP particulière d'une configuration des tableaux. Cette ACP s'effectuait éventuellement après une optimisation, comme dans Statis ou Pré-Statis.
Au §III, on a examiné des aspects spécifiques aux mesures répétées, et les choix d'analyses qui en découlaient.

Dans ce paragraphe on va essayer de se placer entre ces deux approches, et nous allons envisager de façon empirique la constitution des matrices que l'on va soumettre à l'ACP pour analyser les mesures répétées.

Tout d'abord examinons l'objectif d'analyse de comportements stables sur un long terme.

On peut remarquer que concatener des tableaux obtenus pour différentes dates revient à considérer les individus sur un long terme. Deux individus seront proches s'ils ont eu une évolution semblable, plus le nombre d'instants concaténés sera grand, plus l'évolution devra être similaire.
De la même façon si l'on considère les tableaux juxtaposés on est amené par exemple à comparer un individu à des dates différentes. La représentation factorielle de deux unités statistiques ou plus sera proche si le comportement est semblable dans le temps et donc une idée de stabilité.

On propose alors pour analyser et pour chercher des comportements stables sur un long terme de concatener et juxtaposer les tableaux. Par exemple si l'on dispose de 5 répétitions :

$$\begin{pmatrix} X_1 & X_2 & X_3 \\ X_2 & X_3 & X_4 \\ X_3 & X_4 & X_5 \end{pmatrix}.$$

L'ACP d'une telle matrice nous donnera donc des composantes principales expliquant soit une stabilité sur trois instants, soit une évolution semblable sur trois instants, soit les deux. Les représentations des points dates-individus nous renseigneront sur l'interprétation. Disposant d'un nombre suffisant de répétitions on pourra choisir le nombre de mesures définissant le long terme et le nombre de mesures définissant la stabilité, par exemple :

$$\begin{pmatrix} X_1 & X_2 & X_3 \\ X_2 & X_3 & X_4 \\ X_3 & X_4 & X_5 \\ X_4 & X_5 & X_6 \\ X_5 & X_6 & X_7 \end{pmatrix} \text{ ou } \begin{pmatrix} X_1 & X_2 & X_3 & X_4 & X_5 \\ X_2 & X_3 & X_4 & X_5 & X_6 \end{pmatrix}$$

On peut aussi imaginer que les décalages ne soient pas forcement de 1, en ligne pour par exemple être moins exigeant sur la stabilité, ou en colonne pour le long terme :

$$\begin{pmatrix} X_1 & X_2 & X_3 \\ X_3 & X_4 & X_5 \\ X_5 & X_6 & X_7 \\ X_7 & X_8 & X_9 \\ X_9 & X_{10} & X_{11} \end{pmatrix} \text{ ou } \begin{pmatrix} X_1 & X_3 & X_5 \\ X_2 & X_4 & X_6 \\ X_3 & X_5 & X_7 \\ X_4 & X_6 & X_8 \end{pmatrix}$$

Ces analyses simples sont donc susceptibles de donner des résultats satisfaisants, que d'autre méthodes ne nous fourniraient pas devant un nombre important de répétitions.

Remarque :

-1 Les choix sont multiples, et on peut aussi choisir des métriques appropriées comme par exemple les semi-métriques de voisinages.

IV.2 ACPVI sur un décalage :

Une autre approche qui peut être complémentaire ou encore s'appliquer aux tableaux précédents est, dans l'évolution d'une variable sur

une durée définie, expliquer ce qui est due à une évolution antérieure sur une même durée.

On effectue alors une ACPVI par rapport à un décalage, par exemple l'ACPVI de $X_{2\text{-}4}$ par rapport à $X_{1\text{-}3}$ où :

$$X_{2-4} = \begin{pmatrix} X_2 \\ X_3 \\ X_4 \end{pmatrix} \text{ et } X_{1-3} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

le décalage peut être libre autant que la longueur de la durée ainsi que sur le décalage d'observation...

V . Conclusions et remarques

Nous avons pu voir dans ce chapitre que l'on pouvait adapter les méthodes multitableaux aux problèmes de mesures répétées de différentes façons.
Un survol de quelques méthodes factorielles spécifiques nous a permis d'appréhender la variété des problèmes pratiques et des méthodes employées. D'autres méthodes sont citées dans les références. Et enfin, une vision simpliste de l'approche des buts recherchés ne nous n'est pas apparu sans intérêt.

Nous n'avons pas parlé ici des méthodes multitableaux qui généraliserait l'ACP à trois ou plusieurs entrées telles que l'ACP-3modes Tucker(1966) et Kroonenberg(1983), Candecomp, Parafac, ou aussi avec des contraintes comme Candelinc de Caroll et al.(1980).
Ce sera l'objet du prochain chapitre.

Le fait de vouloir particulariser le temps dans les analyses, nous a amené souvent à réaliser des compromis sur le mode temps. Il n'en est pas moins, de part la pratique d'ACP de tableaux à une seule variable répétée, que nous souhaitons vouloir que le temps joue un rôle factoriel dans la décomposition.

VI . Références bibliographiques

ARAGON, Y. and CAUSSINUS, H. (1980) Une analyse en composantes principales pour des unités statistiques corrélées. In : Data Analysis and Informatics. Diday, E et al. (eds), North Holland,121-131.

BENALI, H and ESCOFIER, B. (1989) Smooth factorial analysis and factorial analysis of local differences. In: Multiway Data Analysis.R.Coppi and S.Bolasco (eds). North-Holland. 327-341.

BENZECRI, J.P. (1983) Analyse de l'inertie intraclasse par l'analyse d'un tableau de correspondance. C.A.D. 351-358.

BESSE,P. and THOMAS-AGNAN, C. (1989) Le lissage apr fonctions splines en statistique : revue bibliographique. S.A.D. 55-84.

BESSE, P. and RAMSAY, J.O. (1986) Principal components analysis of sampled functions. Psychometrika, 285-311.

BOX, G.P. and TIAO, G.C. (1977) A canonical analysis of multiple time series. Biometrika, 355-365.

CARLIER, A. (1985) Application de l'analyse factorielle des évolutions et de lanalyse intra-périodes. S.A.D. 27-53.

CARLIER, A. LAVIT, C. PAGES, M. PERNIN, M. and TURLOT, J. (1989) A comparative review of methods which handle a set of indexed data tables.In: Multiway data analysis, Coppi, R et Bolasco, S (eds), North-Holland, Amsterdam. 85-103.

CARROLL, J.D. PRUZANSKY, S. and KRUSKAL, J.B. (1980) Candelinc: a general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. Psychometrika, 3-24.

CAZES, P. (1977) Etude de quelques propriétés extrémales des facteurs issus d'un sous-tableau de Burt. C.A.D. 143-160.

CAZES, P. (1986) Une généralisation des correspondances multiples et des correspondances hiérarchiques. Cahiers du BURO, n°46 et47, 37-64.

CHESSEL, D. and MERCIER, P. (1991) Couplage de triplets statistiques et liaisons espèces-environnement. In : Journées "Biométrie et Environnement". A.S.U. XXIII^{e} journées de Statistiques, Strasbourg.

CHURCH, A.J. (1966) Analysis of data when the response is curve. Technometrics, 229-246.

DARROCH, J.N. and MOSIMANN, J.E. (1985) Canonical and principal components of shape. Biometrika, 241-252.

DE LEEUW, J.D. (1988) Multivariate analysis with linearizable regressions. Psychometrika, 437-454.

DEVILLE, J.C. (1974) Méthodes statistiques et numérique de l'analyse harmonique. Ann.INSEE. 15, 3-101.

DURAND, J.F. (1989) Analyse en composantes principales par rapport à des variables instrumentales via les transformations splines. rapport INRA, Unité de Biométrie, Montpellier.

EL FAOUZI, N. and ESCOUFIER, Y. (1991) Modélisation de courbes de croissance par les I-splines. R.S.A. XXXIX, 51-64.

ESCOFIER, B. PAGES, J. (1982) Comparaison de groupes de variables définies sur le même ensemble d'individus. IRISA, Publication interne n°166. Rennes.

ESCOFIER, B. PAGES, J. (1984) L'analyse factorielle multiple. cahiers du B.U.R.O. n°42.

ESCOUFIER, Y. (1985) L'Analyse des Correspondances : ses propriétés et ses extensions.Proceeding of 45th Session ISI,1985,28.2.1-28.2.16.

EVERITT, B.S. (1984) An introduction to latent variable models. Chapman and Hall (eds), London, New-York.

FLURY, B.N. (1987) Two generalizations of the common principal component model. Biometrika, 59-69.

FLURY, B.N. (1984) Common principal componenets in k groups. J.A.S.A. 892-898.

GLACON, F. (1981) Analyse conjointe de plusieurs tableaux de données. Thèse 3 e cycle Grenoble

GREENACRE, M.J. (1984) Theory and applications of correspondence analysis. Academic Press, New-York.

HANNAN, E.J. (1973) Multivariate Time Series Analysis. J.Mult.An. 395-407.

HOULLIER, F. (1987) Comparaison de courbes et de modèles de croissance choix d'une distance entre individus. S.A.D. 17-36.

IZENMAN, A.J. and WILLIAMS, J.S. (1989) A class of linear spectral models and analyses for the study of longitudinal data. Biometrics, 831-849.

JAFFRENOU, P.A. (1978) Sur l'analyse des familles finies de variables vectorielles. Bases algébriques et application à la description statistique. Thèse 3e cycle Lyon I

JANSSEN, J. MARCOTORCHINO, F. and PROTH, J.M. (Eds)(1987) Data Analysis : The ins and outs of solving real problems. Plenum Press, New York.

JOERESKOG, K.G. (1986) Analysis of longitudinal data with LISREL. In : Statistical Software, 3rd conference on the use of statistical software, Lehmacher, W and Hoermann, A (eds), New-York

KIERS, H.L. (1988) Comparison of Anglo-saxon and french three-mode methods. S.A.D. 14-32.

KROONENBERG, P.M. (1983) Three mode principal component analysis. DSWO Press, Leiden.

KRZANOWSKI, W.J. (1982) Between-group comparison of principal components-some sampling results. J.Stat.Comp.Sim. 141-154.

KRZANOWSKI, W.J. (1984) Principal component analysis in the presence of group structure. Appl.Stat.64-168.

LAVIT, C. (1988) Analyse conjointe de tableaux quantitatifs. Masson, Paris.

LEBRETON, J. CHESSEL, D. RICHARDOT-COULET, M. and YOCCOZ, N. (1988) L'analyse des relations espèces-milieu par l'Analyse Canonique des correspondances II. variables de milieu qualitative. Acta.Oecol. [Oecol. Gen.], 137-151.

LEBRETON, J. CHESSEL, D. PRODON, R. and YOCCOZ, N. (1988) L'analyse des relations espèces-milieu par l'Analyse Canonique des correspondances I. variables de milieu quantitative. Acta.Oecol. [Oecol.Gen.], 53-67.

MANTON, K.G. STALLARD,E and VAUPEL, J.W. (1986) Alternative models for the heterogeneity of mortality risks among the aged. J.A.S.A. 635-644.

MANTON, K.G. STALLARD,E WOODBURY, M.A. and YASHIN, A.I (1986) Applications of the grade of menbership technique to event history analysis : extensions to multivariate unobserved heterogeneity. Math.Mod. 1375-1391.

MANTON, K.G. STALLARD, E. (1991) Cross-sectional estimates of active life expectancy for U.S. elderly and oldest-old populations. J.Geron: Social Sciences, s170-s182.

MEREDITH, W. and TISAK, J. (1982) Canonical analysis of longitudinal and repeated measures data with stationary weights. Psychometrika, 47-67.

MEREDITH, W. and MILLSAP E, R. (1985) On component analysis. Psychometrika, 495-507.

MILLSAP, R.E. and MEREDITH, W. (1988) Component analysis in cross-sectional and longitudinal data. Psychometrika, 123-134.

MOLENAAR, P. De GOOIJER, J. and SCHMITZ, B. (1992) Dynamic factor analysis of nonstationary multivariate time series. Psychometrika, 333-349.

PARDOUX, C. (1989) Apport de l'analyse factorielle à l'étude d'un processus. R.S.A. XXXVII, 41-60.

PERNIN, M.O. (1986) Contribution à la méthodologie d'analyse de données longitudinales.Exemple de la croissance chez l'être humain. Thèse de doctorat, Université Claude Bernard, Lyon.

PONTIER, J. and PERNIN, M.O. (1986) Multivariate and longitudinal data on growing children : solution using LONGI. In :Janssen et al.(eds)(1987),49-65.

PONTIER, J. DUFOUR, A.B. and NORMAND, M. (1990) Le modèle euclidien en analyse des données. Editions Ellipses, Paris.

RAMSAY, J.O. (1988) Monotone regression splines in action. Stat. Sci. 3(4), 425-461.

RAO, C.R. (1964) The use and interpretation of principal component analysis in applied research. Sankhya A, 329-359.

RAO, C.R. and MITRA, S.K. (1971) Generalized Inverse of Matrices and its Applications. John Wiley and Sons,Inc., New York.

ROBERT, P. and ESCOUFIER, Y. (1976) A unifying tool for linear-multivariate statistical methods : the rv-coefficient. Appl.Stat. 25-265.

ROBINSON P.M. (1973) Generalized canonical analysis for time series. J.Mult.An. 141-160.

ROMAIN, Y. and VIGUIER, S. (1989) COMPOV : un outils de comparaison de plusieurs matrices de covariance. S.A.D. 37-52.

SABATIER, R. (1987) Méthodes factorielles en analyse des données : approximations et prise en compte de variables concomitantes. Thèse d'état , USTL Montpellier.

SABATIER, R. (1991) Critères et contraintes pour l'ordination simultanée de k tableaux. In: Biométrie et Environnement. Lebreton,J.D. et Asselain,B. (eds). Masson, sous-presse.

SAPORTA, G. (1981) Méthodes exploratoires d'Analyse de données Temporelles.Thèse d'état, Paris VI.

TUCKER, L.R (1966) Some mathematical notes on three-mode factor analysis. Psychometrika, 279-311.

VAN Der HEIJDEN, P.M. (1987) Correspondence analysis of longitudinal categorical data. Dswo Press, Leiden.

WOODBURY, M.A. and MANTON, K.G. (1982) A new procedure for the analysis of medical classification. Meth.Info.Med. 210-220

YOUSFATE, A. (1986) Décomposition canonique d'un processus qualitatif de type Markovien Stationnaire. S.A.D.64-89.

BIII
Approche tensorielle des méthodes factorielles et applications à l'évolution de facteurs

0 . Introduction

Nous avons pu voir dans la partie précédente que la littérature est assez fournie en méthodes multitableaux. Elles sont pour la plupart issues de problématiques statistiques spécifiques et ensuite traduites dans le langage mathématique utilisé habituellement par les auteurs de la méthode. Ce langage est le calcul matriciel et l'algèbre linéaire qui s'y rapporte. Les optimisations algébriques des méthodes et des problèmes rencontrés font appel souvent à des techniques fondamentales telles que la recherche de valeurs singulières et la décomposition en valeurs singulières.

Lorsque l'on parle de tableaux à plus de deux entrées on sort du cadre matriciel et une approche algébrique nouvelle est à considérer.

Le cadre mathématique que nous allons introduire ici est le calcul tensoriel. En effet nous verrons que les tenseurs généralisent de façon naturelle les applications linéaires entre deux espaces vectoriels. L'utilisation de la propriété universelle de construction du produit tensoriel nous permettra d'aborder les problèmes fondamentaux dans ce cadre.

Divers auteurs, Tucker(1966), Kroonenberg et De Leeuw(1980), Kruskal(1977), Kapteyn et al(1986), ont abordé un cadre étendant le calcul matriciel en introduisant le produit de Kronecker de deux matrices pour résoudre leurs problèmes statistiques de tableaux à plusieurs entrées.

Franc(1992) a effectué une présentation algébrique complète, dans un cadre cohérent, de certaines méthodes à trois entrées qui se sont alors naturellement généralisées à un nombre quelconque d'entrées. C'est ce cadre introduit par Franc qui va nous intéresser dans les paragraphes suivants. Mais nous allons faire un usage important du diagramme tensoriel pour rendre compte des problèmes fondamentaux. L'introduction des schémas de dualité s'y rapporte de façon naturelle.

Ce cadre algébrique nous permet de redécrire dans le paragraphe I, les problèmes d'ACP et de DVS (Décomposition en Valeurs Singulières) et dans le paragraphe II ceux d'analyses sous contraintes linéaires telle que l'ACPVI.

Sur cette base nous envisagerons leurs généralisations à trois modes (au lieu de deux) dans les paragraphes III et IV On introduira notamment d'autres algorithmes pour effectuer l'ACP-3modes, sous la forme de décomposition en valeurs singulières, qui se généralisent aisément à k modes sous le nom de ATP-kmodes : Analyse en Tenseurs Principaux sur k modes. Cette analyse nous donne une généralisation du théorème d'Eckart et Young.

On verra certains liens entre les méthodes Pré-Statis et Statis, ce qui nous amènera à considérer une troisième analyse du genre : le Statis-Pré-Statis, et une autre analyse, que nous avons appelé Pré-Statis-Croisé. Cette dernière méthode sera introduite sur la base des différents schémas de dualité décrivants un tableau à trois entrées.

Dans le paragraphe V nous présenterons en application de l'ATP-kmodes, deux méthodes pour des variables qualitatives : l'AFC de k variables en tant que généralisation stricte de l'AFC de deux variables (différente de l'AFC multiple) et l'AFC3M, une AFC de trois variables dont les deux premiers modes sont identiques à l'optique d'AFCM.

On conclura par des remarques méthodologiques et pratiques notamment pour l'analyse de mesures répétées sous contraintes. On trouvera des exemples traités dans la partie C et les programmes écrits avec SAS/IML dans les annexes.

I . Diagramme tensoriel et ACP

I.1 préliminaires :

I.1.a. produit tensoriel

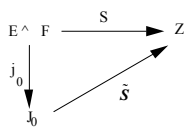
Le produit tensoriel peut être vu comme solution d'un problème universel. Cette approche est par exemple décrite dans Schwartz(1975), Chambadal et Ovaert(1968) ou Charles et Allouch(1984). Nous allons rappeler quelques propriétés utiles à notre démarche avant de les utiliser pour exposer divers aspects de l'analyse de tableaux à trois entrées.

On trouvera une description plus algébrique de l'utilisation du produit tensoriel en analyse des données, notamment avec une généralisation à plusieurs modes de procédures connues, dans la thèse de Franc(1992). On peut citer encore dans l'approche tensorielle la thèse de Jaffrenou(1978), les travaux de D'aubigny(1987,1989) et Kroonenberg et De Leeuw(1980).

Soient E et F deux espaces vectoriels (qui seront en général de dimension finie) sur un corps K (R en général) et S une application bilinéaire de E^F dans Z, ou Z est un espace vectoriel sur K .

Le problème universel est le suivant :

existe-t-il J_0 et j_0 respectivement espace vectoriel et application bilinéaire de E^F dans J_0 tel que le diagramme suivant soit commutatif, avec $\tilde{S}$ application linéaire de J_0 dans Z ?



$S = \tilde{S} \circ j_0$

En fait J_0 et j_0 ne dépendent que de E et F d'après le théorème suivant :

théorème du diagramme tensoriel :

E et F étants deux espaces vectoriels sur K, il existe J_0, qui sera noté E F, et j_0 de E^F dans E F tels que :

$\forall$ S : E^F $\varnothing$ Z application bilinéaire dans l'espace vectoriel Z,

$\exists! \tilde{S}$: E F $\varnothing$ Z application linéaire / $S = \tilde{S} \circ j_0$.

De plus E F et j_0 sont uniques à un isomorphisme près, ainsi si J_0' et j_0' sont solutions du problème universel, alors :

$\exists! j$: E F $\varnothing$ J_0' bijection linéaire / $j_0' = j \circ j_0$ et $\tilde{S}' = \tilde{S} \circ j^{-1}$.

La solution est construite en utilisant le diagramme universel de l'espace quotient :

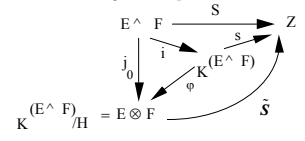
J_0 = E F= $K^{(E \wedge F)} / H$ où $K^{(E \wedge F)}$ est l'ensemble des applications de E6F dans K. Et, H est engendré par les éléments de $K^{(E \wedge F)}$ de la forme :

$i(\alpha\, e + \alpha' e', f) - \alpha i(e,f) - \alpha' i(e',f)$

et $i(e, \alpha f + \alpha' f') - \alpha i(e,f) - \alpha' i(e,f')$

avec α et α' K , e et e' E , f et f' F, et i est l'injection canonique de E^F $\varnothing$ $K^{(E \wedge F)}$.

On a alors le diagramme qui résume la construction de E F :



$S = s \circ i = \tilde{S} \circ \varphi \circ i$

donc $j_0 = \varphi \circ i$

et j_0 est aussi notée □.

Nous renvoyons à Chambadal et Ovaert(1968) pour une démonstration complète, ainsi que pour les quelques propriétés suivantes qui nous seront utiles .

propriétés fondamentales du produit tensoriel :

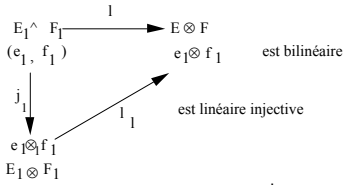
- E F + F E même s'ils sont isomorphes.
- (E F)*= E* F* si l'un d'eux est de dimension finie :

ce qui nous permettra d'écrire $\langle , \rangle_{E\,F} = \langle , \rangle_E \,.\, \langle , \rangle_F$ qui est parfois prise comme définition des tenseurs.

- on se servira aussi de la définition du produit tensoriel de deux applications linéaires (qui donc généralise l'assertion précédente) :
soit $A_1 : E_1 \varnothing F_1$ et $A_2 : E_2 \varnothing F_2$ alors $A : E_1 \ E_2 \varnothing F_1 \ F_2$ définie par $A(e_1 \ e_2) = A_1(e_1) \ A_2(e_2)$ est unique, et l'on note $A = A_1 \ A_2$.
- $E^* \ F \sim l(E;F)$ application linéaire de E dans F (Schwartz(1975)).
- est associative.
- Les tenseurs décomposables (i.e. de la forme e f) engendrent E F.
- si $\{e_i\}$ et $\{f_j\}$ sont des bases de E et F alors $\{e_i \ f_j\}$ est une base de E F canoniquement associée aux précédentes.
- soit E_1 et F_1 deux sous-espaces vectoriels respectivement de E et F alors $E_1 \ F_1 \wp E \ F$

dont une démonstration est la suivante (vu l'importance du résultat dans les applications)

soit $e_1 \ E_1$ et $f_1 \ F_1$ on a $e_1 \ f_1 \ E \ F$ et $e_1 \ f_1 \ E_1 \ F_1$ (ce produit tensoriel devrait être noté $_1$) par la propriété universelle du diagramme tensoriel on a le schéma :



.

L'injectivité peut être montrée par l'utilisation de bases ou par produit de deux applications injectives (Chambadal et Ovaert(1968)) alors $E_1 \ F_1$ peut être identifié à un sous-espace vectoriel de E F.

Toutes ces propriétés s'étendent facilement au cas de 3, ou plus espaces vectoriels et le passage au produit tensoriel transformera une application p-linéaire en une application linéaire. On pourra consulter cette extension dans les ouvrages cités précédemment.

-Une opération intéressante qui généralise l'image d'un vecteur par une application linéaire est la contraction d'un tenseur par un vecteur que nous noterons "..", nous allons en donner un exemple.
Soient A un tenseur de E F G, $\{e_i\}$, $\{f_j\}$ et $\{g_k\}$ des bases de E, F et G :

$A = \sum_{ijk} A_{ijk} e_i \otimes f_j \otimes g_k$ et un vecteur $z^* \in G^*$

alors on définit $A..z^* = \sum_{ijk} A_{ijk} e_i \otimes f_j < g_k, z^* >$

$= \sum_{ijk} A_{ijk} e_i \otimes f_j < g_k, \sum_{k'} z_{k'} g_{k'} > = \sum_{ijk} A_{ijk} z_k e_i \otimes f_j$.

$A..z^*$ est un élément de E F. On peut remarquer qu'il s'obtient en transformant A en une matrice à dim(E F) lignes et dim(G) colonnes c'est à dire une matrice notée :

$$\overrightarrow{A}^{P} \text{ ou } A_G \text{ ou } \overrightarrow{A}^{G},$$

à qn lignes et p colonnes. Calculant l'image de z par cette matrice on a :

$$\overleftrightarrow{A..z^*} = \overrightarrow{A}^{P} z^* = A_G z^* \ .$$

Nous rappelons que l'opération notée flèche à double sens signifie la vectorialisation complète (i.e. on identifie par exemple A B C... Z à $l(R^*;A \ B \ C... \ Z)$. Nous définissons l'opération simple flèche indicée comme la vectorialisation, sauf sur l'espace mis en indice (i.e. on identifie A B C... Z à $l(A^*;B \ C... \ Z)$.

Avec z élément de G, on notera parfois de la même façon A..z, pour exprimer une contraction dans le sens du produit scalaire. Dans l'expression de la contraction, $<g_k, z^*>$ est remplacé par $<g_k, z>_G$.

I.1.b. schéma de dualité

Le schéma de dualité introduit par Cailliez et Pagès(1976) est l'outil favori du statisticien pratiquant l'Analyse des Données. La notion d'opérateurs de produits scalaires de métriques s'y trouve de façon naturelle et son utilisation pour l'ACP, l'ACPVI est commode.

Sa description et sa construction sont les suivantes.
Soit q variables mesurées sur n individus, posons $E = R^q$ et $F = R^n$ alors la description par les q variables d'un individu correspond à un vecteur de E, et de même l'ensemble des observations sur une variable correspond à un vecteur de F. Pour l'individu i' et la variable j' on a :

$$X_{i'} = \sum_{j=1}^{q} X_{i'j} e_j \qquad X^{j'} = \sum_{i=1}^{n} X_{ij'} f_i$$

.

Soit E^* le dual de E muni de la base canonique $\{e_j^*\}$ définie par :
$e_j^*(e_{j'}) = <e_{j'}, e_j^*> = \delta_{jj'}$ alors $\forall j \ e_j^*(X_i) = X_{ij'}$ c'est à dire que e_j^* fait correspondre à tout individu i sa valeur pour la variable j. Donc e_j^* est une représentation de la variable j dans E^*. De même la base canonique de F^* fournit une représentation des individus. Par construction la matrice X à n

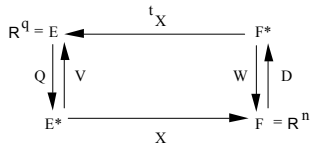
lignes et q colonnes d'éléments $X_{ij'}$ est la matrice de l'application linéaire notée aussi X de E^* dans F qui synthétise le passage des représentations des variables aux valeurs. De même, pour les individus on a une application linéaire de F^* dans E qui n'est autre que la transposée de X.

Pour mesurer les distances dans E, l'on choisit une métrique Q matrice à q lignes et q colonnes symétrique et définie positive (pour avoir un produit scalaire) qui est considérée donc comme élément de $l(E;E^*)$ et de même, on introduit, D appartenant à $l(F;F^*)$. On cherche alors des métriques V et W à mettre sur E^* et F^* qui respectent respectivement les distances dans E, entre vecteurs individus, et dans F, entre les vecteurs variables. Pour F et W on veut :

$$\text{W tel que } \left\| X_i - X_{i'} \right\|_Q = \left\| f_i^* - f_{i'}^* \right\|_W$$

alors on a $W = XQ^tX$ et de même on a $V = {}^tXDX$ pour E et V.

Tout ceci se résume donc sur le schéma commutatif :



Les opérateurs WD et VQ ont la propriété intéressante d'être auto-adjoints positifs et d'avoir les mêmes valeurs propres car WD= uu* et VQ =u*u, où u= XQ. Les vecteurs propres de l'un se déduisent des vecteurs propres de l'autre par application de u (connaissant les vecteurs propres de VQ) ou par application de u* (connaissant les vecteurs propres de WD). Cette propriété du schéma de dualité est très importante pour l'ACP puisque l'on pourra obtenir une représentation conjointe des éléments de E et de F.

Par la suite l'on pourra confondre aisément la matrice X (élément de m(n;q)), l'application linéaire X (élément de $l(E^*;F)$) et le tenseur X (élément de E F).

I. 2 recherche des valeurs singulières et diagramme tensoriel :

I.2.a. propriétés classiques

Le concept de valeurs singulières apparaît en fait indissociable de l'ACP. En effet, cette dernière n'apparaît souvent qu'à travers la décomposition, en valeurs singulières et le théorème d'Eckart et Young(1936), d'approximation par une matrice de rang inférieur. Greenacre(1984) retrace l'historique de la DVS ("Singular Value Décomposition") dans la statistique multidimensionnelle, où elle apparaît donc comme une reconstruction de plein rang d'une matrice à l'aide de deux matrices unitaires, ψ et φ et d'une matrice diagonale Λ:

$$X_{ij} = \sum_{s=1}^{r} \sigma_s \psi_{is} \varphi_{js}$$

avec ${}^t\psi D \psi = Id_r$ et ${}^t\varphi Q \varphi = Id_r$, et r=rang(X). Ce résultat peut s'écrire matriciellement :

$$X = \psi S {}^t\varphi.$$

S est la diagonale des σ_s. Les σ_s^2 sont les valeurs propres de XQ^tXD ou ${}^tXDXQ$ dont les vecteurs propres sont respectivement les colonnes de ψ et φ.

Plusieurs propriétés bien connues sont à noter :

i $S = {}^t\psi DXQ\varphi$,

ii $\begin{cases} XQ\varphi = S\psi \\ {}^tXD\psi = S\varphi \end{cases}$ les formules de transition,

iii $\begin{cases} {}^tXD\psi {}^t\psi DXQ\varphi = S^2\varphi \\ XQ\varphi {}^t\varphi Q {}^tXD\psi = S^2\psi \end{cases} <> \begin{cases} {}^tX {}^tP\psi DP\psi XQ\varphi = S^2\varphi \\ XP\varphi Q {}^tP\varphi {}^tX\psi = S^2\psi \end{cases}$
$<> \begin{cases} {}^t(P\psi X)D(P\psi X)Q\varphi = S^2\varphi \\ (XP\varphi)Q {}^t(XP\varphi)D\psi = S^2\psi \end{cases}$

où Pφ est le projecteur sur le sous-espace engendré parles colonnes de φ.

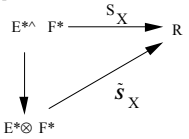
Remarques :

-1 Les premières équations de **iii** présentent ψ et φ comme solutions simultanées, d'équations aux vecteurs propres, des opérateurs de produits scalaires, avec comme semi-métrique ("dans l'autre espace") la reconstruction des produits scalaires (par exemple : $D\psi {}^t\psi D$). Les deuxièmes et troisièmes équations les présentent comme solutions d'ACPVI.

-2 Cette dernière remarque sert de base à l'ACP-3mode de Kroonenberg(1980, 1983), et devient donc une ACPVI sous contrainte "tensorielle" alternée décrite par Franc(1992) comme solution explicite du modèle de Tucker (voir III.3).

I.2.b. approche tensorielle

On peut formuler la recherche d'une valeur singulière, ici avec des métriques identité, en utilisant le diagramme tensoriel :
soit S_X : $E^* \wedge F^* \otimes R$ l'application bilinéaire définie par $S_X(e_j^*, f_i^*) = X_{ij}$ avec $\{e_i^*\}_{1..q}$ $\{f_j^*\}_{1..n}$ les bases canoniques des espaces considérés. D'après la propriété universelle du produit tensoriel on a le schéma suivant :



Alors pour tout α^* et β^* de E^* et F^* on a :
$S_X(\alpha^*,\beta^*) = S_X(\sum_j \alpha_j e_j^*, \sum_i \beta_i f_i^*) = \sum_j \sum_i \alpha_j \beta_i X_{ij} = {}^t\beta X \alpha$

$= {}^t(\alpha \otimes^k \beta) \overleftrightarrow{X} = \tilde{S}_X(\alpha^* \otimes \beta^*) = X..(\alpha^* \otimes \beta^*)$

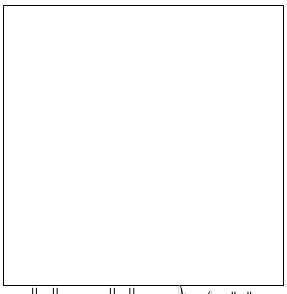
où $\otimes^k$ désigne le produit de Kronecker, et $\overleftrightarrow{X}$ le vectorialisé, Henderson et Searle(1979), de la matrice X. On peut remarquer déjà que l'on utilise l'isomorphisme $E \otimes F \sim l(E^*;F) \sim m(n;q;R)$.
On a donc la propriété suivante :

propriété d'une valeur singulière :

La recherche de la première valeur Singulière peut s'écrire :
$\sigma_1 = \max \tilde{S}_X(\alpha^* \otimes \beta^*) = \max < \alpha^* \otimes \beta^*, X > = \max < \alpha \otimes \beta, X >_{E \otimes F} = \max (X..(\alpha \otimes \beta))$

$\|\alpha^*\|_{E^*} = 1$	$\|\alpha^*\|_{E^*} = 1$	$\|\alpha\|_E = 1$	$\|\alpha\|_E = 1$
$\|\beta^*\|_{F^*} = 1$	$\|\beta^*\|_{F^*} = 1$	$\|\beta\|_F = 1$	$\|\beta\|_F = 1$

où l'on voit que l'on peut exprimer cette maximisation de différentes façons. Or on maximise une forme linéaire continue sur l'ensemble :



$$\left\{ (\alpha \otimes \beta) / \|\alpha\|_E = 1 \text{ et } \|\beta\|_F = 1 \right\} \subset \left\{ z / \|z\|_{E \otimes F} = 1 \right\}$$

qui est un ensemble fermé dans un compact (la sphère unité) donc un compact, d'où l'existence de σ_1.

On a unicité, mais en fait à un élément de l'orthogonal de X dans E F près. En effet, on maximise une forme linéaire, deux tenseurs solutions ne différeront que d'un élément annulant cette forme linéaire qui de plus devra faire rester la solution dans l'ensemble d'optimisation.

De la même manière, on atteint les autres valeurs singulières en imposant des contraintes d'orthogonalité aux solutions

On construit alors de proche en proche, des sous-espaces vectoriels de E F engendrés par les solutions des maximisations successives. Sous-espaces qui sont tels que la projection de X sur chacun soit optimale par rapport aux valeurs singulières.

Ces sous-espaces engendrés par $\{ \varphi_l \quad \psi_l ; l=1..k\}$ pour $k \leq rang(X)$ sont des sous-espaces vectoriels de $E_k \quad F_k$ engendrés par $\{ \varphi_i \quad \psi_j \; i,j=1..k\}$.Mais ici, grâce à la dualité des solutions, la projection de X est la même sur les deux espaces.

Remarques :

-1 A priori on a le choix, pour les contraintes d'orthogonalités, introduit dés la deuxième valeur singulière. Elles pourraient être à la fois dans E et F ou bien seulement dans le produit tensoriel. En effet :

$$E \otimes F = (E_1 \overset{\perp}{\oplus} E_1^{\perp}) \otimes (F_1 \overset{\perp}{\oplus} F_1^{\perp}) = (E_1 \otimes F_1) \overset{\perp}{\oplus} \underbrace{(E_1^{\perp} \otimes F_1^{\perp}) \overset{\perp}{\oplus} (E_1 \otimes F_1^{\perp}) \overset{\perp}{\oplus} (E_1^{\perp} \otimes F_1)}_{\text{14 4 4 4 4 4 4 2 4 1 4 4 4 1 4 4 43}}$$

$$= (E_1 \otimes F_1) \overset{\perp}{\oplus} (E_1 \otimes F_1)^{\perp}$$

avec donc $E_1^{\perp} \otimes F_1^{\perp} \subset (E_1 \otimes F_1)^{\perp}$.

Grâce à la propriété de dualité des solutions, la solution de maximisation, avec comme contrainte supplémentaire d'être dans l'ensemble $(E_1 \otimes F_1)^{\perp}$, se trouve en fait dans $E_1^{\perp} \otimes F_1^{\perp}$. Nous appellerons ce sous-espace **l'orthogonal tensoriel** des sous-espaces E_1 et F_1.

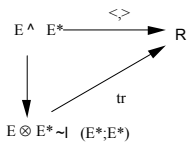
L'apparition de la fameuse matrice noyau lorsque l'on est avec 3 modes provient de cette constatation et de la perte de dualité dans les solutions. C'est à dire que la solution dans l'orthogonal de la première solution n'est pas forcément dans l'orthogonal tensoriel de la première solution. On verra au III dans l'optique de DVS que en fait la matrice noyau dans le cas de trois modes n'a pas une forme quelconque, c'est la "Rocket form" décrite par Denis et Dhorne(1989).

-2 L'écriture de la propriété est valable avec des métriques. Les normes sur E et F venant de produits scalaires, on se donne en fait les métriques Q et F respectivement éléments de l(E;E*) et l(F;F*). De façon canonique, du fait des propriétés fondamentales du produit tensoriel en dimension finie, E F est muni du produit scalaire défini par la métrique Q D qui est un produit tensoriel d'applications linéaires (voir I.1).

-3 On peut remarquer enfin que munir E* F* de sa base canonique, implique que la matrice de l'application linéaire $\tilde{S}_X$ est la transposé du vectorialisé de la matrice X, $\overleftrightarrow{X}$.

I. 3 opération trace et ACP:

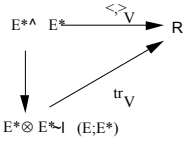
A l'aide du produit tensoriel on peut définir naturellement la trace d'un opérateur linéaire, comme application linéaire associée, par le diagramme tensoriel, à l'application bilinéaire crochet de dualité :



soit $A = \sum_{i,j} A_{ij} \, e_i \otimes e_j^* \in l(E^*;E^*)$, on a $tr(A) = \sum_{i,j} A_{ij} < e_i, e_j^* > = \sum_j A_{jj}$

Historiquement la première composante principale d'une matrice X est définie comme la colonne, issue de la transformation linéaire unitaire des colonnes de la matrice X, de variance maximale.

En se donnant un produit scalaire sur F, par la métrique D, on se donne une application bilinéaire $< , >_D$ qui introduit naturellement sur E^* la métrique $V = {}^tXDX$ et on peut écrire alors :
$<X\varphi^*, X\varphi^*>_D = <X\varphi^*, DX\varphi^*> = <\varphi^*, \varphi^*>_V$.



Donc toute transformation linéaire unitaire des colonnes de X (combinaison linéaire dans E*), de carré scalaire maximum (sa variance et la variance de la transformée), maximisera également l'application linéaire tr_V.

propriété des composantes principales :

La norme sur E^* canoniquement associée à Q étant Q^{-1}, la recherche de la première composante principale peut s'écrire :
$\varphi_1^* = \text{argmax} \; (\; tr_V(\varphi^* \otimes \varphi^*)) \;$, alors la composante principale est $\psi_1 = \dfrac{X\varphi_1^*}{\sqrt{\lambda_1}}$

$\|\varphi^*\|_{Q^{-1}} = 1$ où $\lambda_1 = tr_V(\varphi_1^* \otimes \varphi_1^*)$

On maximise donc une forme linéaire sur un compact d'où l'existence de φ_1^*.

On atteint les autres composantes principales en rajoutant des contraintes d'orthogonalité.

On a aussi :
$tr_V(\varphi^* \otimes \varphi^*) = <\varphi^*, \varphi^*>_V = <\varphi^*, (VQ)\varphi> = <\varphi, (VQ)\varphi>_Q$, que l'on veut maximiser avec la contrainte $<\varphi, \varphi>_Q = 1$.

On retrouve ainsi la propriété d'optimalité pour un opérateur auto-adjoint. C'est à dire par exemple, que le maximum est la première valeur propre (la plus grande) de VQ obtenue avec le vecteur propre correspondant (i.e. théorème de Courant(1920) et Fisher(1905)).

Remarques :

-1 ψ_1 est alors de norme 1 et l'on peut voir que :
$\langle\varphi_1^*, \varphi_1^*\rangle_V = \langle X\varphi_1^*, X\varphi_1^*\rangle_D = \sqrt{\lambda_1}\,\langle X\varphi_1^*, \psi_1\rangle_D = \sqrt{\lambda_1}\,\langle X\varphi_1^*, D\psi_1\rangle = \sqrt{\lambda_1}\,\langle X\varphi_1^*, \psi_1^*\rangle = \sqrt{\lambda_1}\ X..\varphi_1^* \square \psi_1^* = \lambda_1$. On retrouve ainsi la solution du problème de la première valeur singulière associée à X avec $\sigma_1 = \sqrt{\lambda_1}$.

-2 On vérifie facilement que la composante principale associée est un vecteur propre de WD avec la même valeur propre. Pour le détail des différentes propriétés (théorème de Sturm, Okamoto, Hottelling ...) on peut consulter Sabatier(1987), Escoufier(1988) ou Okamoto(1969).

-3 On retrouve l'expression de la trace de l'opérateur de variance associée à l'application identité considérée comme élément de E* E*.
En effet, soit $\{\varphi_1,...,\varphi_p\}$ une base Q-orthogonale de E alors on a :

$$\mathrm{tr}_V\Big(\sum_j \varphi^*_j \otimes \varphi^*_j\Big) = \sum_j \langle\varphi^*_j \varphi^*_j\rangle_V = \sum_j \langle\varphi^*_j (VQ)\varphi_j\rangle = \mathrm{trace}(VQ).$$

Si on note φ la matrice obtenue en juxtaposant les φ_j j= 1,.., k (k<p), et P_φ le projecteur sur φ, alors :

$$\mathrm{tr}_V\Big(\sum_{j=1}^{k} \varphi^*_j \otimes \varphi^*_j\Big) = \sum_{j=1}^{k} \langle\varphi^*_j \varphi^*_j\rangle_V = \sum_{j=1}^{k} \langle\varphi^*_j (VQ)\varphi_j\rangle$$

$$= \sum_{j=1}^{k} \langle Q\varphi_j (VQ)\varphi_j\rangle = \sum_{j=1}^{p} \langle QP_\varphi\varphi_j, (VQ)\, P_\varphi\varphi_j\rangle = \sum_{j=1}^{p} \langle {}^tP_\varphi\varphi^*_j, (VQ)\, P_\varphi\varphi_j\rangle$$

$$= \sum_{j=1}^{p} \langle\varphi^*_j, (P_\varphi VQ\, P_\varphi)\varphi_j\rangle = \mathrm{trace}(P_\varphi VQ\, P_\varphi).$$

On reconnaît la part d'inertie reconstruite par les k premiers vecteurs de la base considérée. L'opérateur défini par $\Pi_S(Z) = P_S\, Z\, P_S$ sur l(E*;E) est un projecteur orthogonal, Sabatier (1987). Si cette base est la base des vecteurs propres de VQ, pour un k fixé cette part d'inertie reconstruite est maximale par rapport à une autre base quelconque.

I. 4 S.V.D. et A.C.P. :

Comme on l'a vu l'ACP peut se définir indépendamment de la recherche des valeurs singulières et nous amène directement au problème d'optimalité des vecteurs propres.

Une autre façon historique d'approcher l'ACP est apparue avec le théorème d'Eckart et Young en 1936. C'est la recherche d'une matrice $\tilde{X}$ de rang plus petit que celui de X, la plus proche de X, au sens d'un certain critère. Pour le théorème d'Eckart et Young, ce critère est la norme trace (norme de Schur) mais ce théorème se généralise à une classe de normes plus générales, appelées les Normes Unitairement Invariantes. Elles sont décrites par des fonctions particulières des valeurs singulières. Pour les détails des propriétés des N.U.I. et leurs applications on pourra consulter Rao(1979), Claret(1987) et Sabatier(1987).

En fait, souligne Yoshisawa(1987), la recherche des valeurs singulières et le soucis d'approcher une matrice par une matrice de rang inférieur s'avèrent liés par les mêmes équations de résolution appelées par Hill(1974) "reciprocal averaging", et plutôt en France, formules de transitions. Ces équations mènent éventuellement aux diagonalisations des opérateurs VQ ou WD, et donc à l'ACP.

De part l'existence et l'unicité des solutions on peut construire pour toute matrice X des matrices unitaires U et V dont les vecteurs colonnes sont les solutions du problème de la recherche des valeurs singulières. Alors, ${}^tUXV$ est une matrice diagonale contenant les valeurs singulières, ${}^tUXV = S$.
On peut écrire $X = US{}^tV$, et toute matrice peut se décomposer de la sorte : c'est la Décomposition en Valeurs Singulières ("Singular Value Decomposition ").

Enfin pour X donnée, application linéaire de E* dans F (l(E*;F)) considérée comme élément de E F, le sous-espace engendré par les k premières (k<rang(X)) solutions de valeurs singulières est optimal, pour la recherche d'une application linéaire de rang k proche de X au sens des normes unitairement invariantes :

$$\tilde{X} = \mathrm{proj}(X)_{\left[\varphi_j \otimes \psi_i,\, i \text{ et } j=1,..,k\right] \subset l_k(E^*;F)}$$

On peut écrire alors l'expression de cette projection orthogonale par calcul tensoriel. On l'écrit ici sans métriques pour ne pas alourdir et on note φ^κ et ψ^κ les matrices, de colonnes les k premiers φ_j et k premiers ψ_j, solutions des problèmes de valeurs singulières :

$$\begin{aligned}
\tilde{X} = \mathrm{proj}(X) &= P_{\varphi^\kappa \otimes \psi^\kappa} \overleftrightarrow{X} \\
&= \left(\left(\varphi^\kappa \otimes \psi^\kappa\right)\left(\left(\varphi^\kappa \otimes \psi^\kappa\right)\left(\varphi^\kappa \otimes \psi^\kappa\right)\right)^{-1}\left(\varphi^\kappa \otimes \psi^\kappa\right)\right) \overleftrightarrow{X} \\
&= \left(\left(\varphi^\kappa \otimes \psi^\kappa\right)\left({}^t\varphi^\kappa \varphi^\kappa \otimes {}^t\psi^\kappa \psi^\kappa\right)^{-1} {}^t\varphi^\kappa \otimes {}^t\psi^\kappa\right) \overleftrightarrow{X} \\
&= \left(\varphi^\kappa \left({}^t\varphi^\kappa \varphi^\kappa\right)^{-1} {}^t\varphi^\kappa \otimes \psi^\kappa \left({}^t\psi^\kappa \psi^\kappa\right)^{-1} {}^t\psi^\kappa\right) \overleftrightarrow{X} = \left(\varphi^\kappa {}^t\varphi^\kappa \otimes \psi^\kappa {}^t\psi^\kappa\right) \overleftrightarrow{X} \\
&= \left(P_{\varphi^\kappa} \otimes P_{\psi^\kappa}\right) \overleftrightarrow{X} \\
&= P_{\psi^\kappa} X {}^t P_{\varphi^\kappa}
\end{aligned}$$

Remarques:

-1 Du fait que l'on projette sur un sous-espace vectoriel de dimension k^2 de E F que l'on peut compléter avec les φ_j et ψ_i suivants, on a des solutions de l'ACP sur des sous-espaces emboîtés donc des solutions emboîtées.

D'Aubigny(1989) et D'Aubigny et Polit(1989) arrivent au même résultat pour la norme trace. Ils obtiennent que φ et ψ sont les solutions alternées de recherche de vecteurs propres, en maximisant la projection de X sur un sous espace de E F. Ceci est décrit au paragraphe 1.2 a.. Par contre ces auteurs ne considérant pas que lorsque E E' et F F' alors E F E' F', ne peuvent vérifier directement la propriété d'emboîtement des solutions.

-2 L'expression de $\tilde{X}$, bien connue, vient naturellement, sachant que la DVS de X est $X = \psi S {}^t\varphi$, et les φ^κ et ψ^κ les k premières colonnes de U et V. En utilisant l'expression de $\tilde{X}$ on a $\tilde{X} = \psi^\kappa {}^t\psi^\kappa \psi S {}^t\varphi \varphi^\kappa {}^t\varphi^\kappa = \psi^\kappa S^\kappa {}^t\varphi^\kappa$.

On peut expliciter encore le lien de la DVS et de l'ACP de la manière suivante :
soit φ_1, $\psi_1$et σ_1 la première solution de la recherche de valeurs singulières de X, alors φ_1, $\psi_1$et σ_1 est aussi la première solution de ${}^1X = \sigma_1 \varphi_1 \square \psi_1$.
$\hat{\boldsymbol{S}}_{(\sigma_1\varphi_1 \otimes \psi_1)}(\alpha^* \otimes \beta^*) = \langle \alpha \otimes \beta, \sigma_1 \varphi_1 \otimes \psi_1 \rangle_{E \otimes F} = \sigma_1 \langle \alpha, \varphi_1 \rangle_E \langle \beta, \psi_1 \rangle_F$
qui est maximum (avec les contraintes unitaires) si et seulement si $\alpha = \varphi_1$ et $\beta = \psi_1$.

De plus, soit ${}^1X = \sigma_1 \varphi_1 \square \psi_1$ alors la première solution de la recherche de la composante principale de ${}^1X$ est φ_1 avec comme maximum $\sigma_1{}^2$, la composante principale est alors ${}^1X..\varphi_1 / \sigma_1 = \psi_1$.

Donc X et ${}^1X$ ont les mêmes solutions, pour la première valeur singulière et pour la première composante principale. De plus comme nous l'avions vu dans la remarque précédente la solution du problème de la première composante principale est aussi solution de la première valeur singulière. On appellera alors DVS jusqu'au rang r la matrice $\tilde{X}$ pour k valant r qui sera aussi l'ACP d'ordre r de X.
On peut alors écrire le théorème général suivant.

théorème général liant l'ACP et la DVS :

Soit deux espaces vectoriels E et F (de dimensions finies) et X un élément de E F alors :

i. la recherche de la $k^{ème}$ composante principale de X est équivalente à la recherche de la $k^{ème}$ valeur singulière de X.

ii. l'ACP d'ordre r est donnée par la projection orthogonale de X sur l'espace solution des r premières valeurs singulières, c'est à dire par la DVS jusqu'au rang r.

iii. se donner la matrice des produits scalaires des éléments de E définis par X (de E F) et une matrice de produits scalaires des éléments de F définis par X permet de réaliser l'ACP et la DVS de l'élément X par diagonalisation de ces matrices.

Remarques :

-1 L'emboîtement des solutions (ACP d'ordre r et d'ordre r+1) tient simplement au fait que l'on complète des bases orthogonales des sous-espaces vectoriels.

-2 Nous verrons que le **iii** a une grande importance dans la pratique et également dans la généralisation à 3 ou n modes, il sert de base aux méthodes Pré-Statis et Statis.

II . ACPVI et double ACPVI

II.1 contraintes de sous-espaces :

On peut dire que l'ACPVI (Analyse en Composantes Principales par rapport à des Variables Instrumentales) prend sa source dans l'article de Rao(1964), puis en France elle est complètement décrite par les articles Robert et Escoufier(1976), Sabatier(1983, 1987), Sabatier et al.(1989). Elle a l'avantage de présenter un cadre général qui permet d'englober beaucoup de méthodes factorielles multidimensionnelles telles que l'analyse canonique, l'analyse discriminante, l'analyse des correspondances et ainsi que d'autres analyses anglo-saxonnes fondées sur l'analyse canonique et une contrainte exprimée par les mots : partielle, bi-partielle etc.... Sabatier(1987) les récapitule et donne leurs formes d'ACPVI.

La dénomination générique de toutes ces analyses prend souvent le nom d'analyse sous contraintes linéaires, ce qui est explicite par rapport aux problèmes posés dans la pratique.

L'ACPVI fait le lien entre l'analyse de variance et l'ACP. C'est la base de notre travail réalisé dans la partie A. Nous n'allons pas ici la redéfinir (voir les auteurs précédents) mais en donner une écriture nouvelle par rapport au cadre que l'on s'est défini. Sa généralisation sera alors facile à des "dimensions" supérieures, c'est à dire pour les mesures répétées trois modes : (individus ^ variables ^ répétitions), ou k modes. Ayant deux espaces de travail pour les données (individus ^ variables), on peut imposer des contraintes linéaires sur un espace et l'on a l'ACPVI, ou sur deux espaces et l'on a l'ACPVI double contrainte ou double ACPVI. Avec donc trois espaces l'on pourra avoir une triple contrainte etc...

Prenons directement deux contraintes, soit S_l la matrice des vecteurs engendrant la structure sur les lignes de X (contrainte dans F) et S_c celle engendrant la structure sur les colonnes (contrainte dans E).

Le problème est de chercher une solution au problème de maximisation de $\tilde{S}_X$ (recherche de la première valeur singulière...) non pas dans E F tout entier mais dans le sous-espace vectoriel $im(S_c)$ $im(S_l)$.

Remarquons tout de suite que lors de la recherche des valeurs singulières, on s'est placé dès la recherche de la deuxième solution dans ce cas. La solution doit être dans l'orthogonal du sous-espace engendré par la première.

Il suffit alors maintenant de remarquer que si (φ_j, Ψ_j) sont solutions du problème sous contrainte alors on a :
$P_{im(Sc)}\,\varphi_j = \varphi_j$ et $P_{im(Sl)}\,\psi_j = \psi_j$ et donc

$$\sigma_1 = \max_{\substack{\|\alpha^*\|_{E^*}=1,\ \alpha\in im(Sc)^* \\ \|\beta^*\|_{F^*}=1,\ \beta\in im(Sl)^*}} \tilde{S}_X(\alpha^*\otimes\beta^*) = \max_{\substack{\|\alpha\|_E=1,\ \alpha\in im(Sc) \\ \|\beta\|_F=1,\ \beta\in im(Sl)}} <\alpha\otimes\beta, X>_{E\otimes F} = \max_{\substack{\|\alpha\|_E=1,\ \alpha\in im(Sc) \\ \|\beta\|_F=1,\ \beta\in im(Sl)}} <P_{Sc}\alpha\otimes P_{Sl}\beta, X>_{E\otimes F} =$$

$$= \max_{\substack{\|\alpha\|_E=1,\ \alpha\in im(Sc) \\ \|\beta\|_F=1,\ \beta\in im(Sl)}} <(P_{Sc}\otimes P_{Sl})(\alpha\otimes\beta), X>_{E\otimes F}$$

$$= \max_{\substack{\|\alpha\|_E=1,\ \alpha\in im(Sc) \\ \|\beta\|_F=1,\ \beta\in im(Sl)}} <(\alpha\otimes\beta), (P_{Sc}\otimes P_{Sl})^*(X)>_{E\otimes F} = \max_{\substack{\|\alpha\|_E=1,\ \alpha\in im(Sc) \\ \|\beta\|_F=1,\ \beta\in im(Sl)}} <(\alpha\otimes\beta), (P_{Sc}\otimes P_{Sl})(X)>_{E\otimes F}$$

$$= \max_{\substack{\|\alpha\|_E=1,\ \alpha\in im(Sc) \\ \|\beta\|_F=1,\ \beta\in im(Sl)}} <\alpha\otimes\beta, P_{Sl}X^t P_{Sc}>_{E\otimes F} = \max_{\|\alpha^*\|_{E^*}=1,\ \|\beta^*\|_{F^*}=1} \tilde{S}_{(P_{Sl}X^tP_{Sc})}(\alpha^*\otimes\beta^*)$$

Plusieurs propriétés interviennent dans la suite des égalités précédentes notamment la définition du produit tensoriel de deux applications.

On est donc ramené à la DVS de $P_{Sl}\,X\,{}^tP_{Sc}$, donc l'ACPVI double contrainte est l'ACP de cette matrice.

L'expression de l'ajustement X par une application de rang k donné (k<p), avec les contraintes linéaires données par S_l et S_c, est obtenue par :

$$\tilde{X}=proj(X)_{\left[\varphi_j^{Sc}\otimes\psi_i^{Sl},\ i\text{ et }j=1,..,k\right]\subset\ l_k(im(S_c)^*;im(S_l))}$$

$$=proj(P_{Sl}X\,{}^tP_{Sc})_{\left[\varphi_j^{Sc}\otimes\psi_i^{Sl},\ i\text{ et }j=1,..,k\right]\subset\ l_k(E^*;F)}$$

$$=P_{\left(\psi^{Sl}_{\kappa}\right)}P_{Sl}X\,{}^tP_{Sc}\,{}^tP_{\left(\varphi^{Sc}_{\kappa}\right)}$$

Remarques:

-1 Pour l'expression de $\tilde{X}$ on peut arriver au même résultat en supposant qu'il y ait une solution (restriction d'une application linéaire continue à un sous-espace) puis en effectuant la projection. En exprimant ensuite que $P_{im(Sc)}\,\varphi_j = \varphi_j$ et $P_{im(Sl)}\,\psi_j = \psi_j$, on reconnaît alors l'expression d'une approximation de $P_{Sl}\,X\,{}^tP_{Sc}$.

II.2 approximations d'opérateurs :

Prenons tout d'abord le cas de l'ACPVI simple contrainte. On peut par des considérations analogues aux précédentes, en utilisant la propriété des composantes principales optimisant tr_V, avec la contrainte de sous-espace, montrer que la solution optimise l'application linéaire $tr_{P_{Sc}VQP_{Sc}}$. Donc la décomposition de l'opérateur de variance et covariance par l'ACPVI, est la décomposition par l'ACP de sa projection orthogonale sur l(im(Sc);im(Sc)*) 1 l(E;E*). De même si l'on a une contrainte ligne S_l (contrainte dans F) la solution est obtenue par l'ACP de la projection orthogonale de l'opérateur de produits scalaires de X sur l(im(Sl);im(Sl)*) 1 l(F;F*) i.e. la solution qui optimise l'application linéaire $tr_{P_{Sl}WDP_{Sl}}$.

Pour les deux contraintes la solution doit vérifier les deux précédentes et optimise donc l'application linéaire $tr_{P_{Sc}{}^tX{}^tP_{Sl}DP_{Sl}XQP_{Sc}}$ qui est l'application linéaire associée au problème d'optimisation déjà rencontré de $\tilde{X} = P_{Sl}\,X\,{}^tP_{Sc}$. On remarquera que $\tilde{X}$ est la projection orthogonale de X sur l(im(Sc)*;im(Sl)) (Sabatier(1987)) où l'opérateur de projection orthogonale est $\Pi_{Sl;Sc}(Z) = P_{Sl}\,Z\,{}^tP_{Sc}$.

Remarques :

-1 On a vu aussi dans la partie AI de ce travail que $\tilde{X}$ est l'ajustement par le modèle de courbes de croissances de Potthof et Roy(1964).

-2 Une autre présentation de l'ACPVI faite par Sabatier(1987) qui est plus liée à l'approximation d'opérateurs, consiste à rechercher une métrique à mettre sur l'espace considéré. Elle doit être telle que les études (X,Q,D) et (Sl, $\tilde{Q}$,D) soient proches, au sens de la proximité des opérateurs d'Escoufier. La solution (une semi-métrique) s'obtient par exemple pour la minimisation en utilisant le projecteur Π_{Sl} et le théorème de Pythagore.

Si l'on a deux contraintes ou plutôt une contrainte double on doit mener simultanément une ACPVI et une ACPVI duale (i.e. recherche d'une métrique $\tilde{D}$ et critère sur les opérateurs de variance VQ) dont la justification de la solution optimale n'apparaît pas aisée.

III . Généralisations à 3 modes (à k modes)

III.1 D.V.S. sur 3 modes :

On a vu aux I et II de cette partie que les valeurs singulières jouent un rôle essentiel pour les méthodes factorielles. Il semble donc naturel de chercher à étendre ce concept pour ensuite étendre l'ACP sur ce principe.

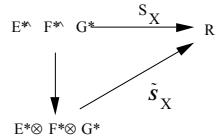
Ceci a beaucoup préoccupé les statisticiens (et algébristes), bien qu'ils se soient surtout intéressés directement à des approximations, avec les modèles de Tucker(1966), Kroonenberg et De Leeuw(1980), Kapteyn et al.(1986), puis les modèles PARAFAC de Harshman(1970), Kruskal(1977, 1984, 1989), CANDECOMP de Caroll et Chang(1970).

Yoshisawa(1987) a tenté une généralisation de la décomposition en valeurs singulières, l'algorithme que nous proposons au paragraphe III.3 c s'en est inspiré.

Franc(1992) décrit algébriquement et analytiquement la résolution des problèmes CANDECOMP et PARAFAC.

On peut directement étendre à 3 ou k modes la recherche des valeurs singulières :

soit S_X : E*^F*^G*∅ R l'application trilinéaire définie par $S_X(e_j^*,f_i^*,g_k^*) = X_{ijk}$ avec $\{e_i^*\}_{1..q}\ \{f_j^*\}_{1..n}\ \{g_k^*\}_{1..p}$ des bases duales canoniques des espaces considérés. On peut en utilisant le diagramme tensoriel écrire :



alors pour tout α*, β*, γ* de E*, F*, G* on a :

$$S_X(\alpha^*,\beta^*,\gamma^*) = S_X\Big(\sum_j \alpha_j e_j^*, \sum_i \beta_i f_i^*, \sum_k \gamma_k g_k^*\Big) = \sum_j \sum_i \sum_k \alpha_j\beta_i\gamma_k X_{ijk}$$

$$= {}^t(\alpha \otimes^k \beta \otimes^k \gamma)X$$

$$= \tilde{S}_X(\alpha^* \otimes \beta^* \otimes \gamma^*) = X..(\alpha^* \otimes \beta^* \otimes \gamma^*)$$

où $\otimes^k$ désigne le produit de Kronecker et $\vec{\vec{X}}$ le vectorialisé du "cube" de données X qui est vu alors comme un tenseur de E F G.

propriété d'une valeur singulière sur le "cube":

La recherche de la première valeur singulière peut donc s'écrire :

$$\sigma_1 = \max \tilde{S}_X(\alpha^* \otimes \beta^* \otimes \gamma^*) = \max\ <\alpha^* \otimes \beta^* \otimes \gamma^*, X> = \max\ <\alpha \otimes \beta \otimes \gamma, X>_{E \otimes F \otimes G}$$

$$\begin{matrix} \|\alpha^*\|_{E^*} = 1 \\ \|\beta^*\|_{F^*} = 1 \\ \|\gamma^*\|_{G^*} = 1 \end{matrix} \qquad \begin{matrix} \|\alpha^*\|_{E^*} = 1 \\ \|\beta^*\|_{F^*} = 1 \\ \|\gamma^*\|_{G^*} = 1 \end{matrix} \qquad \begin{matrix} \|\alpha\|_E = 1 \\ \|\beta\|_F = 1 \\ \|\gamma\|_G = 1 \end{matrix}$$

$$= \max\ (X..(\alpha \otimes \beta \otimes \gamma))$$

$$\begin{matrix} \|\alpha\|_E = 1 \\ \|\beta\|_F = 1 \\ \|\gamma\|_G = 1 \end{matrix}$$

On maximise donc une forme linéaire continue sur :

$$\left\langle (\alpha \otimes \beta \otimes \gamma) / \|\alpha\|_E = 1, \|\beta\|_F = 1 \text{ et } \|\gamma\|_G = 1 \right\rangle \subset \left\langle z / \|z\|_{E \otimes F \otimes G} = 1 \right\rangle$$

qui est un ensemble fermé dans un compact, donc compact, d'où l'unicité et l'existence de la solution avec la restriction décrite pour deux modes (cf. I.2.b).

On atteint les autres valeurs singulières en ajoutant une certaine contrainte d'orthogonalité qui sera discutée plus loin.

On construit alors des sous-espaces vectoriels emboîtés de E F G, en complétant successivement par les solutions des maximisations, qui sont tels que X et sa projection sur chaque sous -espace ont les mêmes premières valeurs singulières.

Remarques :

-1 Attention, il faut remarquer que ce n'est pas équivalent, comme pour le cas de deux modes, au fait de compléter de proche en proche, par exemple jusqu'à k, à l'intérieur de chaque mode. On ne considère ici que le sous-espace de E F G engendré par les k premières tenseurs solutions.

-2 Les normes sur E, F, G viennent des produits scalaires définis à partir des métriques respectivement Q, D, P et de façon canonique on munit E F G de la métrique Q D P.

-3 La résolution directe de cette maximisation ressemble à un modèle Candecomp d'ordre 1, décrite dans Franc(1992) pour un cas plus général de somme de r tenseurs décomposables (modèle CANDECOMP). On notera d'ailleurs que le problème de candecomp que résoud Franc est beaucoup plus compliqué que celui-ci car aucune contrainte d'orthogonalité et de norme des vecteurs n'est demandée. La seule contrainte est la norme de la somme des r tenseurs.

La DVS sera alors l'écriture de X sous la forme :

$$X = \sum_{s=1}^{rg(X)} \sigma_s(\varphi_s \otimes \psi_s \otimes \phi_s).$$

L'égalité est vérifié pour toute norme unitairement invariante. Il faut noter que si les solutions (les tenseurs) sont orthogonales, il arrive en général, qu'une solution pour un mode se retrouve dans une solution ultérieure (dans la suite des valeurs singulières) par exemple $\phi_1=\phi_2$.

On peut de façon analogue à l'ACP projeter X sur l'espace engendré par les solutions. On dira que l'on fait une ACP d'ordre r si l'on prend r vecteurs dans chaque sous-espace :

$$\tilde{X} = \underset{[\varphi_s \otimes \psi_s \otimes \phi_s,\ \text{où } s=1,\dots,r]}{\text{proj}(\tilde{X})} = \sum_{s=1}^{r} \sigma_s(\varphi_s \otimes \psi_s \otimes \phi_s)$$

$$= (\varphi \otimes^k \psi \otimes^k \phi) \circ \sigma$$

Le choix de r depend du "rang" du tenseur X, bien que ce dernier soit quasiment incalculable et est en fait dans sa définition lié à la décomposition cherchée (DVS d'ordre r). On pourra consulter Franc(1992) à ce sujet.

Le problème est donc de trouver les solutions car on a seulement montré qu'elles existent et sont uniques. Remarquons que l'on ne sait pas pour le moment, si on peut appeler cette approximation une ACP car on ne sait pas encore si les φ, ψ, ϕ "ressemblent" à des composantes principales

III.2 différents schémas de dualité :

Lorsque l'on a écrit le diagramme tensoriel universel du paragraphe III.1, on peut donner des écritures de $\tilde{S}_X$ différentes, en utilisant

l'associativité de $\otimes^k$ ou l'associativité du produit tensoriel des trois espaces E, F, G. On peut ainsi écrire alors :

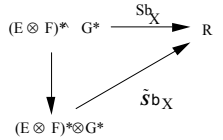
$$(E \otimes F \otimes G) \cong_l (G^*;(E \otimes F)) \cong_m (qn\ ;\ p)$$

$$\cong_l ((F \otimes G)^*;E) \cong_m (q\ ;\ np)$$

et

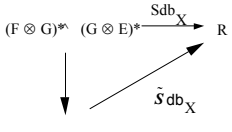
$$(F \otimes G \otimes E) \cong_l (F^*;G \otimes E) \cong_m (pq\ ;\ n)$$

En fait, ces considérations viennent des "passages en bilinéaire" de S_X. Du diagramme initial on peut en opérant ces associativités écrire trois autres diagrammes, réellements différents, ou S_X devient bilinéaire, par exemple on aura :



On retombe alors sur le problème déjà vu de la DVS et de l'ACP d'une matrice : c'est la méthode Pré-Statis que nous détaillerons au III.4.

Un dernier modèle découlant de ceclui-ci est de coupler un des trois espaces avec les deux autres :



avec par exemple $\tilde{S}db_X(f_i^*\ \ g_k^*\ ;\ g_{k'}^*\ \ e_j^*) = (X_{ijk'} + X_{ijk})_{/2'}$ ou d'autres que nous envisagerons au paragraphe III.5 sous le nom de Pré-Statis-Croisé.

On peut suivant chaque équivalence écrire un schéma de dualité dont chacun peut servir pour effectuer une ACP.

Nous les avons rassemblés dans les pages suivantes. La figure **P1** est la forme complète, **P2** et **P3** en sont les détails. La description de la figure **P1** est la suivante :

. en choisissant Q, D, P les métriques sur E, F, G on a naturellement par dualité sur le produit tensoriel, les métriques D_F, Q_E, P_G sur les espaces E G, G F et F E.

. les applications linéaires qui portent le nom générique X, se rapportent au "cube" X à n lignes, q colonnes et p fuyantes ; leurs appartenances à des espaces de matrices sont valables via un isomorphisme.

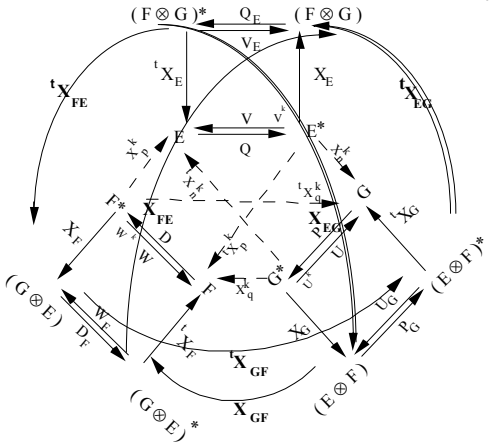
. de ces données et par application du schéma de dualité les métriques sur les espaces duaux sont définies de manière analogue à V et W lorsque l'on a deux espaces.

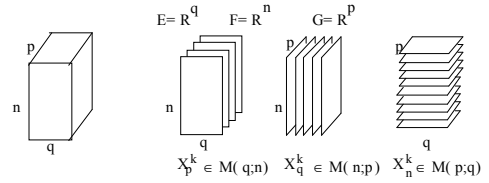
Au centre on a pour chaque k variant soit de 1 à p soit de 1 à q soit de 1 à n, les schémas de dualité sur deux espaces. En considérant les trois espaces, on a trois schémas de dualité prenant en compte l'un des trois et le produit tensoriel des deux autres, ceux-ci correspondent aux Pré-Statis. Puis on peut complêter par trois autres schémas de dualité qui correspondent aux Pré-Statis-Croisés.

Sur la figure **P2** on a les approches multitableaux chaque tableau impliquant un schéma de dualité avec pour chaque k des métriques Q, D, et P identiques, qui induisent les métriques V^k, W^k, U^k. Nous avons d'ailleurs noté qu'un seul indice k mais en fait, pour W^k par exemple, il y a les $W^{k'}$ k'= 1 à p provenant des p shémas de dualité avec les espaces E et F, et les $W^{k''}$ k'' = 1 à q provenant des q shémas de dualité avec les espaces G et F.

Sur **P3** on a la tensorialisation de la première optique multitableaux (il y en deux autres), il y a deux shémas de dualité que nous avons appelé Pré-Statis et un troisième construit sur ces deux le Pré-Statis-Croisé. Remarquons que par exemple la matrice X_F dans cette considération est obtenue par concatenation des p matrices X_p^k, alors que dans la configuration qui met en jeu l'optique multitableaux avec les espaces F et G, X_F est obtenue par concaténation des q matrices ${}^tX_q^k$, les opérateurs sont égaux mais les matrices construites sont égales à une permutation des lignes prés.

On peut constater que dans le cas où les métriques P et Q sont diagonale, W est une somme pondérée (poids issus de P) des p $W^{k'}$ ou avec la deuxième façon d'écrire X_F, somme pondérée des q $W^{k''}$et ces deux écritures sont égales.





$X_F \in M(pq;n)$ $D_F = P \otimes Q$

$X_G \in M(qn;p)$ $P_G = Q \otimes D$

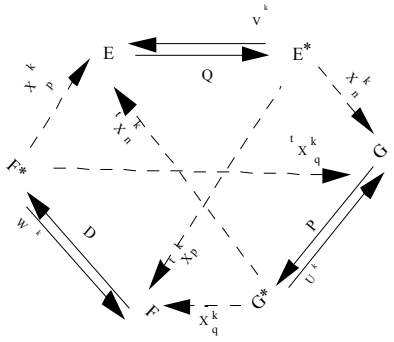
$X_E \in M(np;q)$ $Q_E = D \otimes P$

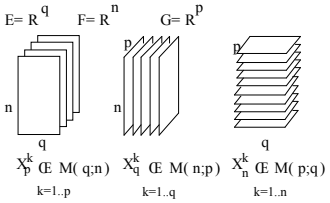
${}^tX_{FE} \in M(pq;np)$

${}^tX_{EG} \in M(np;qn)$

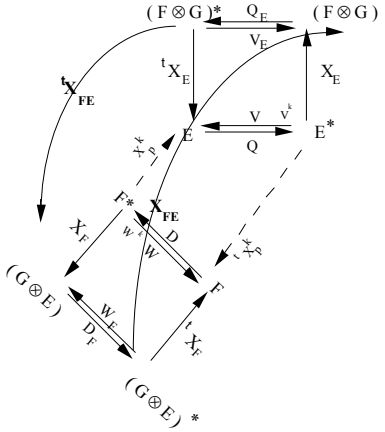
${}^tX_{GF} \in M(qn;pq)$

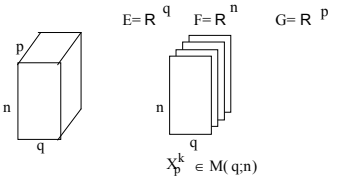
P1.Les différents schémas de dualité pour un "cube" de données à n lignes q colonnnes et p fuyantes





P2.Les différentes optiques multitableaux





$X_F \in M(pq;n)$ $D_F = P \otimes Q$

$X_E \in M(np;q)$ $Q_E = D \otimes P$

${}^tX_{FE} \in M(pq;np)$

P3. Pour chaque optique multitableaux trois schémas de dualité se déduisent de l'approche tensorielle

III.3 ACP-3 modes et DVS-3modes , ATP-kmodes :

La solution pour deux modes de la recherche des valeurs singulières est comme nous l'avons vu indissociable de la recherche des composantes principales. Nous allons dans ce paragraphe généraliser le théorème du lien de l'ACP et de la DVS à 3 modes. Nous présenterons aussi l'ACP-3modes telle qu'elle est décrite par Kroonenberg(1983), et on fera le lien avec les différents schémas de dualité.

III.3.a. La solution naturelle d'ACP sur critère de variance:

On peut généraliser l'ACP de la façon suivante, et cette extension est naturelle :
. on cherche un élément normé (pour Q^{-1}) φ^* de E* tel que la contraction de X, élément de E F G, par φ^* soit de carré scalaire maximum :

$$\langle X..\varphi^*, X..\varphi^*\rangle_{F\ G} = \langle X..\varphi^*, X..\varphi^*\rangle_{D\ P} = \langle X_E\varphi^*, X_E(D\ P)\varphi^*\rangle =$$
$$\langle \varphi^*, {}^tX_E(D\ P)\ X_E\varphi^*\rangle = \langle \varphi^*, \varphi^*\rangle_V = \langle \varphi, VQ\varphi\rangle_Q$$
avec $V = {}^tX_E(D\ P)\ X_E$

. de même la recherche de ψ^* de F* tel que la contraction de X par ψ^* soit de carré scalaire maximum :

$$\langle X..\Psi^*, X..\Psi^*\rangle_{G\ E} = \langle X..\Psi^*, X..\Psi^*\rangle_{P\ Q} = \langle X_F\Psi^*, X_F(P\ Q)\psi^*\rangle =$$
$$\langle \psi^*, {}^tX_F(P\ Q)\ X_F\psi^*\rangle = \langle \psi^*, \psi^*\rangle_W = \langle \psi, WD\psi\rangle_D$$
avec $W = {}^tX_F(P\ Q)\ X_F$

. et de même pour ϕ : $\langle X..\phi^*, X..\phi^*\rangle_{Q\ D} = ... = \langle \phi^*, \phi^*\rangle_U = \langle \phi, UP\phi\rangle_P$
avec $U = {}^tX_G(Q\ D)\ X_G$

Ces trois problèmes d'ACP (identiques à l'approche I.3) reviennent donc à des ACP sur les matrices X_E, X_F et X_G. C'est la solution Tucker1 décrite par Kroonenberg(1983). Elles constituent comme nous le verrons au III.4 pour chacune des ACP, la méthode Pré-Statis.

Le problème est donc d'exprimer les liens entre φ, ψ et ϕ. Autrement dit, ces solutions sont-elles celles de la DVS sur 3 modes ? Ce qui exprimerait la "dualité" entre les solutions comme dans le cas de deux modes !

III.3.b. Liens entre Tucker1, Pré-Statis et la DVS-3modes :

En partant du problème initial de maximisation on peut écrire :

$$S_X(\alpha^*,\beta^*,\gamma^*) = S_X(\sum_j \alpha_j e_j^*, \sum_i \beta_i f_i^*, \sum_k \gamma_k g_k^*) = \sum_j \sum_i \sum_k \alpha_j\beta_i\gamma_k X_{ijk}$$
$$= {}^t(\alpha \otimes^k \beta \otimes^k \gamma)\tilde{X} = {}^t(\beta \otimes^k \gamma)X_E\alpha.$$

Ici la commutativité étant possible on a les expressions avec X_F et X_G

$$= {}^t(\gamma \otimes^k \alpha)\ X_F\ \beta$$
$$= {}^t(\alpha \otimes^k \beta)\ X_G\ \gamma$$

Donc par inclusions des ensembles sur lesquels on maximise on a :

$$\sigma_1 = \max \tilde{S}_X(\alpha^*\otimes\beta^*\otimes\gamma^*) = \max\ \langle \alpha\otimes\beta\otimes\gamma, X\rangle_{E\otimes F\otimes G} = \max\ {}^t(\alpha\otimes^k\beta\otimes^k\gamma)\tilde{X}$$
$$\|\alpha^*\|_{E^*}=1 \quad \|\alpha\|_E=1 \quad \|\alpha\|_E=1$$
$$\|\beta^*\|_{F^*}=1 \quad \|\beta\|_F=1 \quad \|\beta\|_F=1$$
$$\|\gamma^*\|_{G^*}=1 \quad \|\gamma\|_G=1 \quad \|\gamma\|_G=1$$

$$\check{S}\sigma_1^E = \max \tilde{S}_X(\alpha^*\otimes\xi^*_{\beta\gamma}) = \max\ {}^t(\alpha\otimes^k\xi_{\beta\gamma})\tilde{\tilde{X}} = \max\ {}^t\xi_{\beta\gamma}X_E\alpha = {}^t\varphi_{1\beta\gamma}X_E\varphi_1$$

$$\|\alpha^*\|_{E^*}=1 \quad \left\|\xi^*_{\beta\gamma}\right\|_{(F\otimes G)^*}=1 \quad \|\alpha\|_E=1 \quad \|\xi_{\beta\gamma}\|_{F\otimes G}=1 \quad \|\alpha\|_E=1 \quad \|\xi_{\beta\gamma}\|_{F\otimes G}=1$$

où φ_1 est le premier vecteur propre de VQ (i.e une composante des solutions de Tucker1) et $\varphi_{1\beta\gamma}$ est la composante principale associée (élément de F G) qui admet la décomposition en valeurs singulières :

$$\varphi_{1\beta\gamma} = \sum_s^{\min(p\,;\,n)} \sqrt{\lambda_s}\,\beta_s^{\varphi 1}\otimes\gamma_s^{\varphi 1}$$

De même, on a :

$$\sigma_1\check{S}\sigma_1^F \text{ et } \sigma_1\check{S}\sigma_1^G, \text{donc } \sigma_1 \le \min(\sigma_1^E,\sigma_1^F,\sigma_1^G).$$

Si les solutions Pré-Statis donnent une valeur de σ supérieure, ce ne sont pas des valeurs singulières du "cube". C'est dû uniquement au fait que la solution, pour σ_1^E par exemple, dans F G n'est pas de rang 1. On pourrait alors comparer σ_1 avec $X..(\varphi_1\otimes\beta_1^{\varphi_1}\otimes\gamma_1^{\varphi_1})$.

Remarques:

-1 La recherche de la meilleure approximation au sens des valeurs singulières semble compromise sur le plan algorithmique !

-2 On peut déjà remarquer que si $\xi_{\beta\gamma}$ est de rang 1 on a $\sigma_1 = \sigma_1^E = \sigma_1^F = \sigma_1^G$ et la solution d'ACP sur 3 modes (solution Pré-Statis) est alors la solution

de la DVS-3modes (première valeur singulière). On peut alors, dans ces conditions, si $\xi_{\beta\gamma}$ n'est pas de rang 1, en extraire sa meilleure approximation de rang 1, mais alors ce n'est pas la solution de la DVS-3modes (première valeur singulière) car on aura trois jeux de solutions. Une alternative, qui ne sera pas solution, sera de prendre la première composante principale des solutions pour chaque composante. La maximisation se fait en deux étapes et on ne peut la comparer avec la solution Tucker1.

On peut aussi constater à ce stade que rien ne prouve que la solution Tucker1 n'est pas la meilleure ! mais rien ne prouve non-plus que ce soit la bonne car on a en fait :

$$\sigma_1^{\text{Tucker1}}\ \check{S}\ \sigma_1\ \check{S}\ \sigma_1^E.$$

Une propriété intéressante donnée par Jaffrenou(1978) est que l'inertie des trois modèles Pré-Statis (inertie aussi de Tucker1) est la même :

$$\text{Trace}(VQ) = \text{Trace}(WD) = \text{Trace}(UP) = {}^t\tilde{\tilde{X}}\ (P\otimes Q\otimes D)\tilde{\tilde{X}}.$$

La vérification est immédiate mais les valeurs propres des trois opérateurs d'Escoufier ne sont pas forcément les mêmes et donc la dualité est menacée ici.

Il semble difficile de trouver une condition nécessaire et suffisante sur les tableaux pour avoir l'égalité des valeurs propres, mais en passant à l'écriture tensorielle on peut trouver la condition :

Propriété d'égalité des valeurs propres des différents opérateurs d'Escoufier :

Une condition suffisante d'égalité des valeurs propres des opérateurs d'Escoufier pour un tableau à trois dimensions est que le tenseur s'écrive comme :

$$X = \sum_r s_r\ z_r\otimes y_r\otimes x_r \text{ avec les } z_r \text{ orthogonaux deux à deux dans G},$$

ainsi que les y_r et les x_r dans E et F .

c'est à dire que le "cube noyau" est "diagonal". Ceci est évidemment valable en généralisant à un tableau à k dimensions.

La démonstration est triviale après écriture des dits opérateurs en utilisant la réécriture adéquate du tenseur X pour chacun d'eux.

Remarques :

-1 Le problème précédent est en fait de trouver s'il y a une relation entre les valeurs propres de par exemple (uu* + vv*) et (u*u + v*v) où u et v sont des applications linéaires. On connaît les liens seulement entre uu* et u*u ; c'est ce qui fait marcher l'ACP.

-2 On peut aussi écrire une autre condition suffisante d'égalité des valeurs propres, car étant donné deux applications linéaires u et v:

$$uu^* + vv^* \neq (u+v)(u+v)^*.$$

L'égalité est vraie si $uv^* + vu^* = uv^* + (uv^*)^* = 0$ (l'application nulle) et on sait que $(u+v)^*(u+v)$ a les mêmes valeurs propres que $(u+v)(u+v)^*$.

III.3.c. L'ACP-3modes sur critère de "rang" :

La solution proposée par Kroonenberg et De Leeuw(1980), Kroonenberg(1983), repose sur les trois modèles possibles Pré-Statis (ou les métriques sur les espaces sont l'identité). Ils proposent un algorithme alternant ces trois modèles par des ACP avec métriques particulières qui sont en fait des ACPVI. Chacun des modèles s'écrivant comme la DVS-2modes à un rang donné, on a une approximation des moindres-carrés. La solution est donc calculée par moindres-carrés alternés.

L'écriture de leur méthode est la suivante, ils posent le modèle suivant pour X_E :

$$X_E = (\psi^t\ {}^k\Phi^u)\ C\ {}^t\varphi^s + e.$$
$$np,q \quad n,t \quad p,u \quad tu,s \quad s,q \quad np,q$$

Par la suite pour simplifier on ne notera pas la différence entre le produit tensoriel et le produit de Kroneker. On a deux autres modèles analogues pour X_F et X_G. En s'aidant d'un lemme de Penrose en 1955 (référence donnée dans leur article), ils montrent qu'alors le C réalisant la solution des moindres-carrées a la forme :

$$C = {}^t(\psi^t\ \Phi^u)\ X_E\ \varphi^s$$

(qui si t = s = u = 1 est en fait l'écriture de $\tilde{S}_X(\varphi \otimes \psi \otimes \Phi)$) et que la solution est exacte si :

$$X_E = (\psi^t \Phi^u)\,{}^t(\psi^t \Phi^u)\, X_E\, \varphi^s\, {}^t\varphi^s$$

alors en prenant cette dernière forme comme approximation, ils montrent que φ^s maximise ${}^t\varphi^s\, {}^tX_E(\psi^t \Phi^u)\,{}^t(\psi^t \Phi^u)\, X_E\, \varphi^s$. Ceci correspond comme nous l'avons décrit pour deux modes dans le paragraphe I.2.a à une ACPVI par rapport à $\psi^t \Phi^u$.

En effet l'expression à maximiser est :

$${}^t\varphi^s\, {}^tX_E(\psi^t \Phi^u)\,{}^t(\psi^t \Phi^u)\, X_E\, \varphi^s = {}^t\varphi^s\, {}^tX_E(\psi^t\, {}^t\psi^t \Phi^u\, {}^t\Phi^u)\, X_E\, \varphi^s$$
$$= {}^t\varphi^s\, {}^tX_E P_{(\psi^t \Phi^u)} X_E\, \varphi^s = {}^t\varphi^s\, {}^t(P_{(\psi^t \Phi^u)} X_E)(P_{(\psi^t \Phi^u)} X_E)\varphi^s.$$

C'est à dire que φ^s est solution de l'ACP de $P_{(\psi^t \Phi^u)} X_E$.

Avec les deux autres modèles on a l'ACP de $P_{(\Phi^t \varphi^u)} X_F$, et l'ACP de $P_{(\varphi^t \psi^u)} X_G$.. Ils alternent donc ses trois ACPVI, avec comme solution initiale préconisée, celle de l'ACP-3modes solution sur citère de variance (i.e. le modèle Tucker1).

On remarque que cette analyse correspond à, contracter X par un ensemble de vecteurs (t.u vecteurs) de F G de la matrice np 6 tu $\psi^t \Phi^u$ et garder les s premières composantes principales du tableau q 6 tu obtenues pour avoir φ^s, et ensuite à alterner les calculs entreψ, Φ et φ.

Les différentes propriétés de la solution et la convergence de l'algorithme sont donnés dans Kroonenberg(1983) ou Franc(1992). On remarquera que Franc(1992) part directement de l'écriture d'un "cube" de données sous la forme d'un tenseur de E F G, cette présentation devient plus aisée pour les démonstrations.

Cherchant les sous-espaces vectoriels de E, F, G solutions du problème des moindres-carrés, le tenseur cherché, élément du produit tensoriel de ces sous-espaces, est la projection du tenseur X sur ce sous-espace de E F G. La solution pour un des sous-espaces (les autres étant fixés) revient à une ACP de la projection de X sur l'espace produit tensoriel des trois espaces.
Pour effectuer l'ACP, X est considéré comme une application linéaire, par exemple de l(E*; F G). Il suffit alors d'initialiser et d'alterner la recherche des différents sous-espaces.

De plus la présentation de Franc(1992) montre que l'extension à un produit tensoriel d'un nombre quelconque d'espaces ne posent aucun problème supplémentaire (l'ACP-kmodes). Le passage à des métriques quelconques sur les espaces ne pose non plus aucun problème.

Remarques:

-1 Une constatation simple est qu'aucun lien entre les solutions constituant les trois sous-espaces n'est évidente. Le but est une interprétation conjointe comme en ACP.

-2 On pourrait aussi modifier l'algorithme en réintroduisant la solution d'ordre k pour un des trois sous-espaces dès que possible. C'est à dire dans l'étape k, l'algorithme est alors de type Gauss-Seidel et convergera plus rapidement.

III.3.d. Solutions compromis entre variance et "rang":

III.3.d.1 La DVS jusqu'au rang r :

Nous avons vu que la solution de Kroonenberg revenait à contracter le tenseur X par l'ensemble des vecteurs $\psi^t \Phi^u$ etc....On voit donc que si s=u=t=1, cette approche s'apparente à la méthode du "reciprocal averaging", nous proposons tout simplement d'alterner les contractions sur les trois vecteurs. La (n+1) ème étape de l'algorithme proposé est :

1. $(X..\phi_n)..\varphi_n = {}^{\psi}\sigma_{n+1}\psi_{n+1}$
2. $(X..\psi_{n+1})..\phi_n = {}^{\varphi}\sigma_{n+1}\varphi_{n+1}$
3. $(X..\varphi_{n+1})..\psi_{n+1} = {}^{\phi}\sigma_{n+1}\phi_{n+1}$

Le maximum de $\tilde{S}$ est défini par les équations de l'algorithme, et ceci est aussi équivalent à chercher un tenseur de rang 1 le plus proche de X au sens des moindres carrés.
En effet, on a :

$$\left\|X - \sigma_1(\varphi_1 \otimes \psi_1 \otimes \phi_1)\right\|^2 = \|X\|^2 - 2\sigma_1\tilde{S}(\varphi_1 \otimes \psi_1 \otimes \phi_1) + \sigma_1^2.$$

L'algorithme étant donc en fait l'algorithme TUCKALS3 de Kroonenberg et De Leeuw(1980), on a convergence et estimation en moindres-carrés alternés du tenseur X par un tenseur de rang 1.

On a $\sigma_1 = \tilde{S}(\varphi_1 \otimes \psi_1 \otimes \phi_1)$, le point d'accumulation de l'algorithme correspond au maximum de $\tilde{S}$.
C'est aussi l'algorithme de Franc(1992) pour r=1 de CANDECOMP, nous l'appelerons : Recherche de la Première Valeur Singulière par Contractions Complètes (RPVSCC), qui se généralise naturellement à k modes. Le j ème pas de la (n+1) ème étape est :

$$j.\ X..(s_n^k \otimes \ldots \otimes s_{n+1}^{j-1} \otimes \ldots \otimes s_{n+1}^1) = {}^j\sigma_{n+1} s_{n+1}^j$$

En fait on a la propriété suivante de maximisation de la fonction objectif de recherche de la première valeur singulière :

Propriétés de l'algorithme RPVSCC :

i. Les équations de l'algorithme de recherche de la première valeur singulière par contractions complètes (RPVSCC) sont les équations du problème de Lagrange associé à la fonction $\tilde{S}$:

$$L(\alpha,\beta,\gamma,{}^{\alpha}\sigma,{}^{\beta}\sigma,{}^{\gamma}\sigma) = X..(\alpha \otimes \beta \otimes \gamma) - \frac{1}{2}{}^{\alpha}\sigma(\|\alpha\|_E^2 - 1) - \frac{1}{2}{}^{\beta}\sigma(\|\beta\|_F^2 - 1) - \frac{1}{2}{}^{\gamma}\sigma(\|\gamma\|_G^2 - 1)$$

où ${}^{\alpha}\sigma, {}^{\beta}\sigma$ et ${}^{\gamma}\sigma$ sont les multiplicateurs de Lagrange .

avec une écriture analogue pour k modes.

ii. La hessienne de l'équation normale de **i.** pour chaque paramètre est une matrice diagonale négative et proportionelle à l'identité.

iii. Les propriétés i et ii sont naturellement généralisables pour k modes.

On prouve par exemple sur α : $\frac{\partial}{\partial\alpha}L(\alpha,\beta,\gamma,{}^{\alpha}\sigma,{}^{\beta}\sigma,{}^{\gamma}\sigma) = X..(\beta \otimes \gamma) - {}^{\alpha}\sigma\alpha$, où l'on a choisi la métrique identité sur E pour coller à la présentation de l'algorithme.

Pour chercher la deuxième solution, il faut se placer dans l'orthogonal de la première. En fait on se placera dans un sous-espace de l'orthogonal car on peut écrire, en notant φ, ψ et ϕ la première solution :

$$E \otimes F \otimes G = (\varphi \overset{\perp}{\oplus} \varphi^{\perp}) \otimes (\psi \overset{\perp}{\oplus} \psi^{\perp}) \otimes (\phi \overset{\perp}{\oplus} \phi^{\perp})$$
$$= (\varphi \otimes \psi \otimes \phi) \overset{\perp}{\oplus} (\varphi^{\perp} \otimes \psi^{\perp} \otimes \phi^{\perp})$$
$$\overset{\perp}{\oplus} (\varphi \otimes \psi^{\perp} \otimes \phi^{\perp}) \overset{\perp}{\oplus} (\varphi^{\perp} \otimes \psi \otimes \phi^{\perp}) \overset{\perp}{\oplus} (\varphi^{\perp} \otimes \psi^{\perp} \otimes \phi)$$
$$\overset{\perp}{\oplus} (\varphi \otimes \psi \otimes \phi^{\perp}) \overset{\perp}{\oplus} (\varphi \otimes \psi^{\perp} \otimes \phi) \overset{\perp}{\oplus} (\varphi^{\perp} \otimes \psi \otimes \phi)$$
et
$$P_{E\otimes F\otimes G}(X) = \sigma(\varphi \otimes \psi \otimes \phi) + P_{(\varphi^{\perp} \otimes \psi^{\perp} \otimes \phi^{\perp})}(X)$$
$$+ P_{(\varphi \otimes \psi^{\perp} \otimes \phi^{\perp}) \overset{\perp}{\oplus} (\varphi^{\perp} \otimes \psi \otimes \phi^{\perp}) \overset{\perp}{\oplus} (\varphi^{\perp} \otimes \psi^{\perp} \otimes \phi)}(X).$$

La dernière égalité est due au fait que contracter par deux éléments de la solution nous donne le troisième élément normé à σ. Donc la projection sur l'un des trois sous-espaces faisant intervenir deux éléments de la solution et l'orthogonal de la troisième, est nulle.

La recherche de la décomposition de la projection de X sur la somme directe des trois sous-espaces sera obtenue pour chaque sous-espace en contractant par l'élément de la solution intervenant et en effectuant la décomposition en valeurs singulières de la matrice obtenue dont on ne retiendra pas la première (cas de trois modes) solution.

Nous appelerons par la suite ces solutions **les solutions associées aux solutions k-modes**, ici à la première solution trois modes. Elles sont un peu l'équivalent des éléments extra-diagonaux du fameux cube noyau du modèle Tucker3 de Kroonenberg(1983).

Donc pour la deuxième solution trois modes, il suffira de projeter le tenseur X sur le sous-espace $(\varphi^{\perp} \otimes \psi^{\perp} \otimes \phi^{\perp})$, que nous appellerons **l'orthogonal tensoriel** de la première solution, et de recommencer l'algorithme RPVSCC sur cette projection notée ${}_1\tilde{X}$.

Dans la recherche successive des valeurs singulières, on a donc deux types de solutions :

- les **solutions k-modes**, obtenues par projection sur un orthogonal tensoriel, (le premier est l'espace entier), et maximisation de $\tilde{S}$ par le RPVSCC,

• les **solutions associées aux solutions k-modes**, ce sont des solutions (k-1)-modes que l'on obtient aprés une contraction par un élément de la solution k-modes.

On peut donc obtenir la décomposition en valeurs singulières d'un "cube" de données, mais aussi d'un tenseur d'ordre k quelconque (k>3) par ces étapes écrites de façon naturelle par l'algorithme ci-dessous noté DVS-kmodes :

α. recherche des solutions k-modes

on réitère ,

1. RPVSCC

2. projection de X pour la deuxième solution (ou $\tilde{X}_{s-2}$ pour la s ème solution) sur l'orthogonal tensoriel de la première solution (ou de la (s-1) ème solution).

β. recherche des solutions associées aux solutions k-modes

pour chaque solution k-modes,

pour chaque élément de la solution,

1. contraction par cet élément

2. décomposition en valeurs singulières sur (k-1) modes

Remarques importantes :

-1 Cette décomposition en valeurs singulières est unique puisque à chaque étape l'algorithme RPVSCC réalise l'estimation des moindres-carrés et que la projection orthogonale sur l'orthogonal tensoriel d'une solution est aussi unique.

-2 On aboutit en fait pour trois modes à ce que Denis et Dhorne(1989) ont appelé la décomposition en "Rocket Form", bien que l'image soit fausse ou trompeuse. En effet, elle suggère que les solutions 3-modes se combinent comme dans le modèle Tucker3. Toutefois la forme de fusée suggère aussi que toutes les combinaisons ne sont pas possibles, ce qui fait écho aux solutions associées aux solutions k-modes. En fait l'image pourrait être une "Rocket form", mais une fusée "trouée". L'élément de la supra-diagonale est nul lorsqu'apparait un élément en dehors de cette supra-diagonale.

Ces derniers auteurs n'ont pas l'air de reconnaitre la décomposition en valeurs singulières car en fait pour eux la décomposition en valeurs singulières est "diagonale" (i.e comme pour l'écriture de CANDECOMP) et montrent à juste titre sur un exemple que CANDECOMP ne peut fournir cette décomposition.

On peut en fait remarquer que c'est le cas 2-modes qui est particulier et n'offre pas par construction de termes hors-diagonaux. C'est à dire, par exemple, que la projection de X sur l'espace $\varphi_1 \otimes \psi_1^{\perp}$ est réduite au vecteur nul. Pour 3 modes on a déjà remarqué que c'est par exemple la projection sur $\varphi_1 \otimes \psi_1^{\perp} \otimes \phi_1$ qui est réduite au vecteur nul etc...

-3 Yoshisawa(1987) a tenté une décomposition en valeurs singulières, mais ne donne que les solutions 3-modes, (ou k-modes) et ne fournit donc pas de décomposition complète.

-4 Dans l'algorithme DVS-kmodes, pour la recherche des solutions associées aux solutions k-modes, il faut noter qu' au lieu de projeter (ce qui est envisageable) on contracte. Ceci a pour effet de retrouver dans la première solution par DVS-(k-1)modes, le reste de la solution k-modes. Il faudra donc écarter cette solution. On pourrait alors après la contraction projeter sur l'orthogonal tensoriel du reste de la solution. Le choix entre ces stratégies sera dicté par la résolution informatique.

-5 Pour trois modes, on peut modifier l'agorithme RPVSCC. Aprés une première contraction, il suffit d'effectuer la DVS. On réitère alors ces deux étapes. On a alors la meilleure approximation de rang 1 de X et de plus on a la relation :

$$X..\phi_1 = \psi_1 \sigma_1 {}^t\varphi_1 + \varepsilon,$$

et deux autres analogues avec ε orthogonal au premier terme. En fait cette solution est la même que celle obtenue par l'algorithme RPVSCC.
Pour la recherche de la solution suivante on projette X sur l'orthogonal tensoriel de la première solution etc...

La décomposition en valeurs singulières d'un tenseur d'ordre k peut alors s'écrire :

• $k = 3$,

$$X = \sum_s \sigma_{sss}\, z_s \otimes y_s \otimes x_s + \sum_s \sum_{t=s+1} \sigma_{stt} z_t^{x_s} \otimes y_t^{x_s} \otimes x_s + \sum_s \sum_{t=s+1} \sigma_{tst} z_t^{y_s} \otimes y_s \otimes x_t^{y_s} + \sum_s \sum_{t=s+1} \sigma_{tts} z_s \otimes y_t^{z_s} \otimes x_t^{z_s},$$

• $k > 3$,

$$X = \sum_s \sigma_s\, \varphi_s^k \otimes \mathsf{L} \otimes \varphi_s^1,$$

où pour i quelconque et $r < r'$, φ_r^i et $\varphi_{r'}^i$ sont tels que :

. $\varphi_r^i = \varphi_{r'}^i$, alors la r' solution est une solution associée à la solution r pour le r 'ème tenseur solution ceci arrivera au plus (k - 2) fois,

. $\varphi_r^i \perp \varphi_{r'}^i$, si pour le r 'ème tenseur solution ceci arrive k fois , on a une solution k - modes .

III.3.d.2. Approximation par un tenseur de rang inférieur :

Si il apparait difficile de calculer le rang d'un tenseur comme le relate Franc(1992), on peut toutefois construire aisement un tenseur ayant un rang donné . Pour une description détaillée nous renvoyons aux àla thèse de Franc(1992) pour la définition du rang d'un tenseur.
En utilisant des minorations du rang d'un tenseur, établies par Franc(1992) notamment, on pourra toujours arriver à choisir un r qui soit inférieur au rang du tenseur.

Dans un soucis de coller à une généralisation du théorème d'Eckart et Young(1936) de meilleure approximation pour un rang donnée, on supposera donc que ce rang choisi pour l'approximation est evidemment inférieur à celui de X. Notons d'ailleurs que la DVS-k propose en fait une autre définition du rang.

Théorème d'Eckart et Young généralisé :

La meilleure approximation, de rang r, d'un tenseur d'ordre k, d'un espace produit tensoriel de k espaces vectoriels de dimensions finies, au sens d'une norme unitairement invariante, est donnée par le tenseur somme des r premières solutions de la décomposition en valeurs singulières sur k modes (DVS-k) ordonnées suivant les valeurs singulières décroissantes.

On remarquera que les r tenseurs de rang 1 solutions ne sont pas forcément les solutions kmodes : par exemple une solution associée à la première solution peut avoir une valeur singulière supérieure à celle de la deuxième solution k-mode.

La preuve du théorème est indépendante de l'algorithme RPVSCC dont la propriété de maximum global n'est pas prouvée, mais tient uniquement à l'existence et l'unicité des solutions et de l'unicité de la projection sur l'orthogonal tensoriel d'une solution ainsi que de la contraction par un vecteur.

L'obtention de cette approximation de rang r pourra être appelée ACP-kmodes par DVS-k, ou encore, Analyse en Tenseurs Principaux sur k modes , notée : ATP-kmodes. Cette appélation peut être justifiée par la propriété suivante.

Propriété des diverses solutions d'ACP-3modes :

On peut comparer la solution Tucker1 et la solution DVS-3 par les trois critères suivants pour des vecteurs unitaires :
-critère CP de composantes principales :
carré scalaire maximum pour X..φ, X..ψ et X..ϕ
-critère TDP de tenseurs décomposables principaux :
carré scalaire maximum pour X..(φ □ ψ), X..(φ □ ϕ) et X..(ψ □ ϕ)
-critère VS de valeur singulière :
maximum de X..(φ □ ψ □ ϕ)
alors
i. La solution Tucker1 ne vérifie que le critère CP.
ii. La solution DVS-3 vérifie les critères TDP, VS,
et le critère CP si l'on rajoute la contrainte que la composante principale doit être de rang minimun.
iii. La solution fournie par l'algorithme Tuckals3 (lorsque l'on demande s,t,u > 1) ne vérifie aucun des critères et serait plutôt un compromis entre les trois !

Remarques fondamentales :

-1 La suite des carrés des valeurs singulières donne la décomposition de l'inertie de façon analogue à l'ACP et on a :

$$X..X = {}^t\grave{X}\grave{X} = \sum_s (\sigma_s)^2$$

et l'on pouura exprimer un pourcentage d'inertie reconstitué par chaque tenseur principal de l'ATP-kmodes .

-2 Pour une analyse avec des métriques M_k, ..., M_1 sur les modes correspondants, on opère comme avec deux modes :

- on transforme X préalablement en :

$$(M_k^{1/2} \otimes \ldots \otimes M_2^{1/2} \otimes M_1^{1/2}) \ldots X,$$

- on effectue la DVS-k de ce tenseur, la solution est tranformé par l'inverse de la transformation préalable.

-3 Les représentations graphiques peuvent être conjointes grâce à la relation, qui existe entre les éléments d'une solution. C'est une généralisation des formules de transition.

III.4 Pré-Statis et Statis :

Nous avons déjà décrit ces deux méthodes dans le chapitre BII. On va ici faire leurs descriptions par l'approche tensorielle et établir le lien entre ces deux méthodes et celui de la méthode Pré-Statis avec la méthode Tucker1.

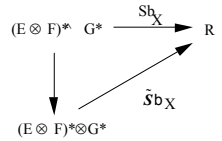
La modèle Pré-Statis tel qu'on l'a décrit dans le chapitre BII peut être vu, sous l'angle du paragraphe I, comme une ACP d'un tenseur de E_{PS} F_{PS} où $E_{PS}=R^p$ et $F_{PS}=l(R^q; R^n)$ par exemple. Ceci est, grâce au point iii du théorème liant l'ACP et la DVS.

L'axe principal fournie est l'axe principal de l'ACP de X mis sous la forme d'une matrice à p colonnes et nq lignes (la matrice X_G). C'est à dire une des ACP -3modes sur critère de variance, ou encore une des ACP de la méthode Tucker1.

En effet les matrices de produits scalaires des éléments de F coïncident dans les deux cas car :

$${}^t\tilde{X}_p^i P_G \tilde{X}_p^j = {}^t\tilde{X}_p^i (Q \otimes D)\tilde{X}_p^j = \text{trace}({}^t X_p^i Q X_p^j D) = < X_p^i, X_p^j >_{Fps}.$$

La dernière égalité est en fait dépendante, du shéma de dualité découlant du diagramme tensoriel, abordé dans le III.2 :



où E = R^q, F = R^n, G = R^p. Ce schéma de dualité est visible sur la planche P1 déjà décrite.

Donc, la composante principale de l'ACP de X_G et le compromis obtenue par la méthode Pré-Statis sur les p tableaux, représentent le même tenseur : le vecteur composante principale étant le vectorialisé du compromis.

Mais la méthode Pré-Statis va plus loin car on effectue :

- l'ACP du compromis, pour analyser l'inter-sructure et,
- l'ACPVI orthogonale de chaque tableau par rapport au compromis pour analyser l'intra-structure.

Maintenant nous allons décrire des liens entre la méthode Pré-Statis et Statis. On peut remarquer par un calcul algébrique simple que l'opérateur de variance précédent est :

$$UP = {}^tX_G P_G X_G P = \sum_{j=1}^{q} q_j \, {}^tX_q D X_q P \text{ ,si Q est diagonale.}$$

Cet opérateur s'exprime donc comme un compromis des q opérateurs de variance de chaque tableau Xq avec comme pondération, les poids de Q.

Dans Statis on construit, dans une première étape, le meilleur compromis (au sens des moindres-carrés) des q opérateurs. Puis pour analyser l'inter-structure on diagonalise ce compromis.

A ce stade Statis est meilleur que Pré-Statis car le compromis est meilleur par rapport au critère des moindres-carrés.

Mais dans Pré-Statis, la diagonalisation de UP nous donne comme composante principale un tableau compromis des p tableaux. Alors la question de comparer l'opérateur de produits scalaires entre les lignes ou colonnes de ce tableau compromis et le compromis obtenu par Statis, se pose.

La réponse peut sembler triviale, du fait que l'on compare la première composante principale d'un compromis, au meilleur compromis. Mais on pourra remarquer que l'opérateur de produits scalaires fourni par Pré-Statis va tenir compte des liens entre tableaux différents alors que dans le compromis de Statis on ne tient compte que de l'importance qu'apportent les tableaux séparément. Suivant les données une méthode primera sur une autre.

On pourrait d'ailleurs pour tirer partie des deux méthodes, et utiliser l'axe principal fournit par la diagonalisation du compromis de Statis pour construire le compromis des tableaux dans Pré-Statis. On appellera cette

méthode le Statis-Pré-Statis ou SPS dont on pourrait étudier le caractère optimal.

Nous avons décrit précédemment (paragraphe III.3,c) l'ACP-3modes de Kroonenberg comme des modèles Pré-Statis sous contraintes alternés. On peut alors imaginer, dans la même optique, alterner des modèles Statis-Pré-Statis sous contraintes. La procédure sera alors optimale dans le sens où à chaque pas d'une itération l'on cherchera les axes principaux du meilleur compromis et non du compromis somme. On nomme cette méthode ACP-SPS-3modes dont on peut penser quelle fournira la même solution que l'ACP-3modes, mais plus rapidement.

Remarques :

-1 Nous avons fait ici quelques comparaisons formelles de méthodes par le biais de l'approche tensorielle, dans le but de mieux les comprendre. Ceci nous a permis d'en proposer d'autres.

Certains auteurs ont effectué des comparaisons de certaines méthodes citées ci-dessus : soit sur des données par le biais du résultat obtenu et de leur interprétation comme dans l'ouvrage :"Data analysis : the ins and outs of solving real problems" de Janssen et al.(1987), soit par comparaison des fonctions minimisées ou maximisées comme dans Kiers(1989).

-2 On remarquera aussi que deux des modèles Pré-Statis correspondent aux optiques univariée et multivariée qui ont été la base de la partie A de ce travail. Ils sont sur la planche P3 et font intervenir les matrices X_E et X_F.

III.5 Pré-Statis -Croisé :

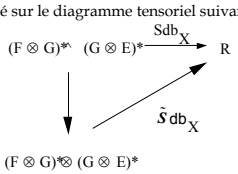
Nous avons décrit, dans le paragraphe III.2 et notamment sur les planches P1, P2 et P3, les différents schémas de dualité qui intervenaient dans la description tensorielle d'un tableau à trois entrées.

Un schéma de dualité faisant intervenir les produits tensoriels de deux des espaces avec le troisième peut être considéré et amener à décrire diverses méthodes portant le nom générique de Pré-Statis-Croisé.

Ce nom tient au fait que le modèle est construit sur deux modèles Pré-Statis, et que le troisième espace considéré intervient à droite et à gauche dans les produits tensoriels.

On distinguera et explicitera dans la suite le Pré-Statis-Croisé sans modèle, avec modèle et le Pré-Statis-Croisé-Fictif.

On a associé à deux schémas de dualité Pré-Statis pour donner un troisième schéma basé sur le diagramme tensoriel suivant :



Ce schéma de dualité fait intervenir l'opérateur ${}^tX_{FE}$ (planche P3) qui peut être représenté par une matrice de m(pq;np). Si l'on pense en terme de produit de Kronecker, le tableau devrait être sous la forme de blocs à q lignes et p colonnes, avec en ligne p blocs et en colonne n blocs. Mais on peut écrire le même tableau à une permutation des colonnes près, sous la forme de blocs à q lignes et n colonnes, répétés p fois en ligne et en colonne.

La construction la plus naturelle, **sans modèle**, consiste à affecter à chaque bloc la matrice X_p^k qui lui correspond, c'est à dire à l'application :

$$Sdb_X(f_i^* \quad g_k^* ; g_{k'}^* \quad e_j^*) = X_{ijk} \text{ si k=k' et 0 sinon.}$$

Mais on peut imaginer d'autres configurations avec par exemple :

$$Sdb_X(f_i^* \quad g_k^* ; g_{k'}^* \quad e_j^*) = (X_{ijk'} + X_{ijk})/2,$$

qui donnera une matrice ${}^tX_{FE}$ bloc-symétrique. Ces blocs extra-diagonaux peuvent être interprétés comme des prédictions du futur (si k<k') ou du passé (si k>k'). Ici on effectue une **modélisation**. On peut alors envisager d'autres modèles, et obtenir par exemple une matrice non-bloc-symétrique avec :

$$Sdb_X(f_i^* \quad g_k^* ; g_{k'}^* \quad e_j^*) = \left(P_{X_p^k} X_p^{k'}\right)_{ji}.$$

Pour chaque individu à un temps donné on regarde son impact (en tant que prédiction linéaire) dans le passé et dans le futur. C'est à dire, comment ses observations actuelles expliquent son futur, et ce qu'elles ont expliqué dans le passé. Pour les variables, à un temps donné, on observe leurs prédictions dans le passé et futur, où plutôt comment elles sont expliquées.

On peut envisager aussi :

$$Sdb_X(f_i^* \quad g_k^* ; g_{k'}^* \quad e_j^*) = \left(P_{X_p^k} X_p^k\right)_{ji}$$

qui a la même interprétation que précédemment mais en intervertissant les rôles de variables et des individus.

Dans le premier cas la matrice réordonnée représentant ${}^tX_{FE}$ est :

$$\begin{array}{c} \\ q \\ q \\ . \\ . \\ q \end{array} \begin{array}{c} \begin{array}{cccc} n & n & . & n \end{array} \\ \begin{pmatrix} X_1 & P_{X2}X_1 & . & . \\ P_{X1}X_2 & X_2 & . & . \\ . & . & . & . \\ . & . & . & . \\ . & . & . & Xp \end{pmatrix} \end{array}$$

où on a noté X_Z l'une des p matrices à q lignes et n colonnes. Dans le deuxième cas l'on a :

$$\begin{array}{c} \\ q \\ q \\ . \\ . \\ q \end{array} \begin{array}{c} \begin{array}{cccc} n & n & . & n \end{array} \\ \begin{pmatrix} X_1 & P_{X1}X_2 & . & . \\ P_{X2}X_1 & X_2 & . & . \\ . & . & . & . \\ . & . & . & . \\ . & . & . & Xp \end{pmatrix} \end{array} .$$

Le cas sans modèles serait la bloc-diagonale de l'une de ces matrices.

Chacune des ACP de ces matrices nous renseignera donc sur les trajectoires des individus ou des variables.

Elles correspondent aussi à une optique Pré-Statis sur un cube de données par exemple à n lignes qp colonnes et p fuyantes.

Dans une optique analogue, on peut associer les composantes principales issues des deux Pré-Statis qui définissent la base du modèle Pré-Statis-Croisé (i.e. avec X_F et X_E). On construit, en sommant leurs produits tensoriel pondérés, un tableau **fictif** X_{FE}. Ce tableau a une interprétation de prédiction à posteriori comme précédemment. Deux possibilités s'offrent pour l'association des composantes :

a) soit des associations, pour le même rang dans la suite des valeurs propres.

Alors la décomposition en valeurs singulières de X_{FE} redonne ces composantes. On peut noter que l'opérateur de produits scalaires sur les lignes de X_{FE} est comparable avec l'opérateur de produits scalaires sur les lignes de X_E (de même pour les colonnes avec X_F). Les deux opérateurs ont les mêmes vecteurs propres et par exemple si l'on choisit, comme poids pour

construire X_{FE}, les valeurs singulières de X_E et P=Id, alors il y a égalité de ces opérateurs.

b) soit des associations, de chaque composante du premier Pré-Statis avec toutes celles du deuxième.

Alors la décomposition en valeurs singulières ne redonnera pas les composantes des Pré-Statis.

Remarques :

-1 Un choix à préconiser pour les poids dans la construction de X_{FE} serait la moyenne géométrique ou arithmétique des valeurs singulières de X_E et X_F correspondantes.

-2 Il peut être intéressant de comparer l'opérateur de produit scalaire obtenu dans l'un des trois Pré-Statis-Croisé (surtout le Fictif qui peut être envisagé comme un compromis), avec le produit tensoriel des compromis d'opérateurs de produits scalaires, obtenus par la méthode Statis.

-3 Le Pré-Statis-Croisé fictif suggère que l'on puisse effectuer, des interprétations et lectures simultanées, de deux modèles Pré-Statis associés.

-4 Notons que l'approche sans modèle correspond à une remarque faite dans le paragraphe IV.2 du chapitre I de la partie A. On avait alors mis en évidence que chaque vecteur propre issu de l'ACP de cette matrice est une juxtaposition d'un vecteur propre issu d'une ACP d'une x_i et de vecteurs nuls ailleurs.

On peut essayer de résumer toutes ces méthodes sur cube de données, décrites ici, par le tableau suivant :

Méthodes	**Critères intervenants**	**Liens avec**	**Meilleur que...**	**Généralisable à k-modes**
Tucker1	carré scalaire du tenseur contracté	ACP		oui
ACG	corrélation des transformations de chaque tableau	ACP, Tucker1, métriques de Mahalanobis		non
Pré-Statis	carré scalaire d'une transformation des tableaux, puis carré scalaire d'une transformation du compromis	ACP et , Tucker1		oui

Statis	corrélation des WD, puis variance d'une transformation d'un tableau associé au compromis	SVD, ACG, Pré-Statis	ACG, si mêmes métriques. Pré-Statis, mais optiques différentes	non
Statis-Pré-Statis	carré scalaire d'une transformation du compromis optimisé	Statis, Pré-Statis	Pré-Statis. Statis, mais optique sdifférentes	non
Pré-Statis-Croisé-Modèle	"carré scalaire du passé, présent et futur" des tableaux	Pré-Statis, ACG, et		?
Pré-Statis-Croisé-Fictif	"double optimisation Pré-Statis"	Pré-Statis, et		?
ACP-3modes	moindres-carrés alternés et moindres-carrés pour le tenseur	Tucker1, Pré-Statis , ACPVI. et ^{k},	Tucker1	oui
ACP-SPS-3modes	moindres-carrés alternés pondérés,et moindres-carrés pour le tenseur	Statis-Pré-Statis, ACP-3modes	Tucker1. ACP-3modes pour la rapidité et cas pathologiques	non
ATP-3modes	valeurs singulières et moindres-carrés pour le tenseur	ACP-3modes, DVS, et	Tucker1, ACP-3modes	oui

IV . Contraintes dans les modèles à 3 modes (à k modes)

IV.1 ACPVI-3modes, ATPVI-kmodes :

Nous avons vu au § II que les expressions de l'ACPVI et de l'ACPVI double contraintes sous formes d'ACP particulières, pouvaient en terme de contraintes de sous-espaces se retrouver aisément par la définition de la recherche des valeurs singulières.

Il suffit donc de rechercher les valeurs singulières du tenseur, image du tenseur X, par le produit tensoriel des projecteurs sur chacun des sous-espaces.

Cette procédure se généralise naturellement à trois où k modes pour donner l'ATPVI-kmodes qui sera donc l'ATP-kmodes du tenseur :

$$P_{S_k} \otimes P_{S_{k-1}} \otimes \ldots \otimes P_{S_2} \otimes P_{S_1}(X) .$$

Pour l'ACPVI-3modes de Kroonenberg et De Leeuw(1980) et l'ACPVI-3modes sur critères de variances on effectue la même transformation sur X avant d'opérer l'analyse spécifique.

On peut essayer de résoudre le problème en terme de métriques ou semi-métriques et d'approximations d'opérateurs de produits scalaires sur chacun des modes. Mais on se heurte comme nous l'avons déjà souligné, au fait que de résumer l'information par la donnée d'un des opérateurs n'est plus comme dans le cas de deux modes suffisant. Seule la recherche de valeurs singulières garde le même genre d'information avec n'importe quel nombre de modes.

Remarques :

-1 En présence de plusieurs facteurs, la décomposition de la variance du tenseur X se fera de manière analogue aux modèles décrits par Sabatier(1987). Mais le fait d'avoir des contraintes sur plusieurs modes va entraîner une inflation des "effets" i.e. des projecteurs à considérer. Par exemple pour un tenseur centré à trois modes avec deux facteurs orthogonaux sur un mode et un chacun des deux autres on aura 12 effets à "mesurer" pour décomposer l'inertie de X.

-2 Les transformations initiales sur le tenseur, c'est à dire les centrages et réductions pour certains modes et pas d'autres, entraînent une multitude d'analyses, dont les interprétations ne seront pas toujours aisées. Une grande prudence sera à observer. Beffy(1992) sur la base de l'ACP-3modes de Kroonenberg a explicité ces difficultés et a proposé des solutions par rapport à des problématiques spécifiques à l'écologie.

IV.2 Pré-StatisVI, StatisVI, Pré-Statis-CroiséVI :

Pour le modèle Pré-Statis on peut rajouter, par rapport à ce qui a été dit précédemment que l'on va se retrouver avec une ACPVI classique ou une ACPVI double contrainte, mais la structure sur les lignes pourra avoir une forme tensorielle.

Notons que pour imposer une contrainte de sous-espaces dans les modèles Pré-Statis ou Statis on peut à priori le faire à deux endroits.

a. En effet, pour Pré-Statis, on peut choisir de définir un compromis des tableaux, puis sur l'analyse du compromis introduire les contraintes qui alors ne pourront être que sur deux modes.

b. On peut aussi le faire dés le départ, en cherchant un compromis des tableaux sous des contraintes de sous-espace.

a et b. On peut enfin combiner les deux en cherchant un compromis avec une contrainte sur le mode qui va être "résumé" puis sur l'analyse de ce compromis imposer d'autres contraintes sur les deux modes restants.

Suivant la problématique statistique l'un de ces trois choix pourra être envisagé. Leur comparaison ne pourra se faire qu'à travers une interprétation du mode opératoire. Si l'on met d'emblée toutes les contraintes, on va privilégier les interactions des facteurs sinon on emboîte les effets. Les choses pourront se compliquer à loisir dans le cas de plusieurs facteurs sur un même mode, et où l'on choisira de contraindre dès le départ pour certains et pas pour d'autres...

Dans l'analyse Pré-Statis-Croisé on peut faire les mêmes constatations. Mais suivant le Pré-Statis-Croisé choisi, diverses approches apparaîtront. Le Pré-Statis-Croisé-Fictif dépendant directement de deux modèles Pré-Statis, la façon de contraindre en découlera : soit des contraintes dans les modèles Pré-Statis pour chercher les tableaux compromis en terme d'ACPVI, soit on fera

l'ACPVI du tableau fictif. Éventuellement une combinaison des deux pourra être faite.

Remarque :

-1 Encore une fois, la prudence, quant aux choix des façons de contraindre les analyses sera de rigueur et pourra se justifier par la problématique statistique.

V . AFC-kmodes, AFC3 Multiple, AFCk' Multiple

V.1 Indépendance de k variables :

Une méthode qui découle directement de l'ATP-kmodes est l'AFC de k variables ou AFC-kmodes. Nous avons rappelé, au chapitre BII, quel était le but de l'AFC de deux variables, et comment sa généralisation à plus de deux variables avait été envisagée.

Escoufier(1985) a défini l'AFC comme une ACP particulière, celle du triplet :

$$(D_I^{-1} P D_J^{-1} - 1_{I,J}, D_J, D_I).$$

Cette analyse peut s'écrire sous forme tensorielle :

$$((D_J^{-1} \otimes D_I^{-1}) .. P - 1_{IJ}, D_J, D_I).$$

Cette dernière forme se généralise, lorsque l'on a k variables, au (k+1) uplet :

$$((D_{I_k}^{-1} \otimes \dots \otimes D_{I_2}^{-1} \otimes D_{I_1}^{-1}) .. P - 1_{I_k \dots I_2 I_1}, D_{I_k}, \dots, D_{I_2}, D_{I_1})$$

dont on va décomposer l'inertie i.e. l'écart à l'indépendance entre les k variables par l'ATP-kmodes. Cette analyse est bien évidemment différente d'une AFCM dans laquelle on modélise l'écart à l'indépendance des variables deux à deux alors qu'ici on le fait entre toutes les variables.

Dans cette généralisation, il faut distinguer métrique sur un des espaces et métrique sur les objets que l'on manipule. En effet sur deux modes la dualité consiste à dire que la métrique pour calculer les distances entre les individus est la métrique sur E (i.e la métrique sur l'espace déterminé par les variables), et inversement. Pour l'ATP-kmodes la métrique effective sur les individus dépend de toutes les autres métriques définis sur les autres espaces.

Ainsi pour l'AFC-kmodes la métrique qui intervient lorsqu'on veut calculer les distances par exemple entre les éléments de la variable I_l, est le produit tensoriel des métriques des autres variables :

$$D_{I_k} \otimes \dots D_{I_{l+1}} \otimes D_{I_{l-1}, \dots} \otimes D_{I_2} \otimes D_{I_1}.$$

Les aides à l'interprétation de l'AFC se généralisent de façon naturelle. Pour chaque variable on pourra comparer, les contributions des modalités à

la construction de chaque axe (contribution absolue), par la part de variance (norme au carré) de la modalité dans la variance de l'axe (le carré de la valeur singulière). Les reconstitutions des modalités sur chaque axe seront vues par la part de variance sur l'axe dans la variance de la modalité (la contribution relative ou cosinus carré).

Comme c'est le cas pour deux modes, il faut noter que, $1_{I_k \dots I_1} = 1_{I_k} \otimes \dots \otimes 1_{I_1}$ est la première solution de l'ATP-kmodes de $((D_{I_k}^{-1} \otimes \dots \otimes D_{I_2}^{-1} \otimes D_{I_1}^{-1}) .. P, D_{I_k}, \dots, D_{I_2}, D_{I_1})$. Ainsi le tenseur analysé dans l'AFC-kmodes est donc la projection du tenseur $(D_{I_k}^{-1} \otimes \dots \otimes D_{I_2}^{-1} \otimes D_{I_1}^{-1}) .. P$ sur l'orthogonal de cette solution triviale. On pourra donc choisir de faire une AFC-kmodes centrée (présentation faite) ou non centrée (ATP-kmodes avec la solution triviale). Elles n'auront pas les mêmes solutions, a contrario de deux modes, car l'orthogonal tensoriel d'un tenseur est différent de l'orthogonal du tenseur.

D'autres modélisations d'indépendance pourraient être prise en compte comme, par exemple, celles décrites par D'Ambra(1988) par exemple. Ceci nous permettrait d'être moins exigent sur l'absence de liens entres les variables car, plus le nombre de variable va augmenter plus l'écart à l'indépendance entre les k variables sera confus.

Une autre approche est celle des dépendances partielles, que l'on peut mettre à profit pour réduire les modes. Pour trois modes par exemple on a la décomposition suivante introduite par Lancaster(1951) :

$$\frac{p_{ijk}}{p_{i..}p_{.j.}p_{..k}} - 1 = \frac{p_{ij.} - p_{i..}p_{.j.}}{p_{i..}p_{.j.}} + \frac{p_{i.k} - p_{i..}p_{..k}}{p_{i..}p_{..k}} + \frac{p_{.jk} - p_{.j.}p_{..k}}{p_{.j.}p_{..k}} + \frac{p_{ijk} - \alpha_{ijk}}{p_{i..}p_{.j.}p_{..k}}$$

où le α_{ijk} assure l ' égalité des deux menbres .

On peut alors écrire :

$$\chi^2 / I_1 I_2 I_3 = \sum_{ijk} p_{i..}p_{.j.}p_{..k}\left(\frac{p_{ijk}}{p_{i..}p_{.j.}p_{..k}} - 1\right)^2$$

$$= \sum_{ij} p_{i..}p_{.j.}\left(\frac{p_{ij.} - p_{i..}p_{.j.}}{p_{i..}p_{.j.}}\right)^2 + \sum_{ik} p_{i..}p_{..k}\left(\frac{p_{i.k} - p_{i..}p_{..k}}{p_{i..}p_{..k}}\right)^2 + \sum_{jk} p_{.j.}p_{..k}\left(\frac{p_{.jk} - p_{.j.}p_{..k}}{p_{.j.}p_{..k}}\right)^2$$

$$+ \sum_{ijk} p_{i..}p_{.j.}p_{..k}\left(\frac{p_{ijk} - \alpha_{ijk}}{p_{i..}p_{.j.}p_{..k}}\right)^2.$$

Les trois premiers chi-deux correspondent à des AFC sur chacune des marges à deux modes du "cube" P (i.e. somme des tableaux sur un des

modes). Le dernier chi deux correspond à un AFCVI-3modes orthogonale par rapport aux précédents modèles marginaux.

On peut imaginer ce genre de décomposition pour plus de trois variables. On est amené alors à sommer sur plus de deux modes et la décomposition du chi-deux global en chi-deux partiels peut être assez longue. Par exemple pour 4 variables on pourra avoir une décomposition de l'AFC-4modes en 4 AFC-3modes et une AFCVI-4modes ou 12 AFC-2modes (AFC classique) et 4 AFCVI-3modes et une AFCVI-4modes.

Remarques :

-1 Carlier et Kroonenberg(1993) proposent les décompositions d'un tableau de contingences à trois entrées, en utilisant la décomposition de Lancaster et modélisent l'écart à l'indépendance globale en utilisant soit le modèle Tucker3 présenté par Kroonenberg(1983), soit le modèle Parafac.

V.2 Indépendance de paquets de variables :

Si l'on veut, comme c'est le cas dans les mesures répétées dans le temps, particulariser une variable par rapport aux autres, on peut effectuer une AFC de 3 variables (AFC-3modes) dans le sens d'une AFCM pour deux des modes.

C'est à dire que l'on disjonctivera les (k-1) variables étudiées à chaque date et on sera en présence d'un tenseur à trois modes : un mode individu, un mode variable-modalité, et un mode répétition.

Ceci n'est possible que pour l'évolution de tables de contingences. Dans le même esprit que pour l'AFCM, on peut montrer dans ce cas que l'AFC-3modes des trois variables est liée à l'AFC-3modes où les deux variables sont disjonctivées pour chaque modalité de la répétition.
L'avantage d'une telle analyse que nous appellerons AFC3M est, que l'on observe les "associations" des variables deux à deux et avec le temps. On aura aussi une représentation des individus.

On peut imaginer une AFC-k'modes de k variables où l'on regroupe des variables de manière à avoir k' paquets de variables. On analyse alors l'analogue des sous-tableaux de Burt. On appellera ce type d'analyse une AFC-k'modes Multiple de Burt : AFCk'B. On peut encore disjonctiver les

paquets de variables de façon croisée et effectuer ensuite une AFC-(k'+1)modes. On realise alors ce que l'on appellera une AFCk'M.

De telles analyses pourraient être intéressantes lorsque plusieurs variables représentent quelques domaines qui sont distincts à priori. C'est le cas par exemple, dans les questionnaires de qualité de vie où l'on mesure divers aspects tels que la mobilité, l'isolement social, la douleur, le sommeil, le tonus etc....par plusieurs variables qualitatives dans chaque aspects, ou encore en écologie dans les relevés faunistiques où la classification des espèces peut intervenir etc....

En effet, on souhaite alors observer l'écart à l'indépendance surtout entre les domaines. C'est à dire qu'à l'intérieur d'un domaine l'on aura que des dépendances deux à deux, et entre les domaines l'on aura une dépendance des k' domaines.

Remarques :

-1 Si toutes les variables ont leurs modalités définies de la même façon (i.e. même nombre et même sens) comme par exemple des variables ordinales graduées avec la même échelle (ex : léger, modéré, fort, très fort). On pourra identifier, une AFCM à une AFC par le modèle Pré-Statis. Mais alors on pourra également effectuer une AFC-3modes : les variables , les individus et les modalités.

-2 Kroonenberg(1989) a proposé des AFC de 3 variables sur la base de l'ACP-3modes, en se basant sur les résidus de modèles log-linéaires.

-3 Pour les mesures répétées, un avantage supplémentaire d'effectuer des AFC3M est que l'AFCVI proposée par Sabatier(1987) peut être menée. Donc la prise en compte de facteurs structurels d'une population ou plutôt d'une cohorte dans l'évolution de l'association de variables qualitatives est possible. On peut même décomposer, dans l'esprit de l'analyse de variance, ces évolutions par un modèle faisant intervenir ces facteurs.

VI . Conclusions et Remarques pratiques

L'approche tensorielle pour la description des données nous a permis, par le biais de la recherche des valeurs singulières d'un tenseur, de retrouver aisément les résultats de l'ACP liés à la DVS, ainsi que la prise en compte de contraintes comme en ACPVI et double ACPVI.
La généralisation naturelle de la recherche des valeurs singulières, grâce à cette approche, à plus de deux modes, nous a conduit à définir l'ATP-kmodes comme extension de l'ACP à k modes. On a pu auparavant examiner sur trois modes les éventualités d'analyses et décrire les schémas de dualité correspondants. Ces schémas nous ont permis de retrouver l'analyse Pré-Statis et de faire le lien avec la méthode Statis ce qui a conduit à deux nouvelles analyses le Statis-Pré-Statis et le Pré-Statis-Croisé.

L'ATP-kmodes fournit une généralisation du théorème d'Eckart et Young pour un tenseur d'ordre k. La comparaison, pour 3 modes, a prouvé l'optimalité de l'ATP-3modes par rapport à l'ACP-3modes de Tucker méthode 1 et l'ACP-3modes de Kroonenberg et De Leeuw.

L'introduction de métriques n'a posé aucune difficulté, ce qui nous a permis de définir l'AFC de k variables que nous avons appelé AFC-kmodes. Sur cette base on a définit l'AFC3M comme une AFC-3modes où le troisième mode est le temps le deuxième est l'ensemble des variables disjonctivées par le premier mode (le mode individu) ainsi que des AFCk'B comme des AFC-k'modes de k variables ($k'<k$). La prise en compte de contraintes de sous-espaces s'est trouvé être une généralisation simple de l'ACPVI : on peut avoir des contraintes sur les k modes.

Sur le plan programmation nous fournissons en annexes les programmes permettant d'effectuer les ATP-kmodes et l'ATP-3modes complète (i.e. avec les ACP pour les solutions associées). Nous avons écrit ces programmes en SAS/IML. Ce langage offre la simplicité d'un langage matriciel, mais malheureusement pas la capacité d'analyser de grandes quantités de données (du moins sur PC). En effet nous sommes limités sur SAS/IML à des vecteurs de 4095 éléments, ce qui sera pour nous (avec notre écriture) la taille maximale du tenseur à analyser.

L'algorithme de recherche d'une valeur singulière d'un tenseur est simple au niveau calcul, mais il y a constamment des réécritures d'un tenseur

sous une forme matricielle pour préparer la contraction sur un mode. En ce sens nous pensons, qu'une programmation dans un langage symbolique gagnerait en temps de calcul et peut être en capacité.

Pour l'analyse de mesures répétées, il est évident qu'un des modes sera le temps. On pourra à loisir imposer une contrainte sur le temps dans l'optique de modélisation ou bien mettre une métrique de voisinages temporels, comme dans les analyses en composantes de voisinages (ACV) de Meot et al.(1992).

L'évolution d'un facteur groupe sur les individus sera pris en compte, soit en imposant la contrainte de sous-espace au mode individu, soit dans le cas de variables qualitatives en rajoutant le mode facteur groupe. On pourra également pour des variables qualitatives effectuer une AFC3MVI en ayant le mode temps, le mode individu avec la contrainte du facteur groupe et le mode variables disjonctivées etc... Devant la multitude des choix possibles pour combiner les variables le temps et les facteurs groupes, un long dialogue entre le statisticien et la personne voulant traiter ses données sera indispensable. Rappelons-le encore une fois car si des méthodes plus simples de l'analyse des données sont devenus à tort un peu presse-bouton, avec des résultats "magiques", les erreurs d'interprétations étaient rares. Ici la prudence sera de rigueur.

VII . Références bibliographiques

CAILLIEZ, F. and PAGES, J.P. (1976) Introduction à l'Analyse des Données. SMASH, Paris.

CARLIER, A and KROONENBERG,P.M. (1993) Biplots and decompositions in two-way and three-way correspondence analysis. Publication du Laboratoire de Statistiques et Probabilités, Université Paul Sabatier Toulouse, 01-93.

CAROLL, J.D. and CHANG, J.J. (1970) Analysis of individual differences in multidimensional scaling via n-way generalization of 'Eckart et Young' decomposition. Psychometrika, 283-319.

CHAMBADAL, L. and OVAERT, J.L. (1968) Algèbre linéaire et algèbre tensorielle. DUNOD, Paris.

CHARLES, B. and ALLOUCH, D. (1984) Algèbre générale. PUF, Paris.

CLARET, A. (1987) Contribution au problème de l'approximation factorielle d'un tableau de données. Thèse 3ème cycle Mathématiques, USTL Montpellier.

COPPI, R. and BOLASCO, S. Eds (1989) Multiway Data Analysis. Elsevier SciencePublishers B.V, North-Holland.

COURANT, R. (1920) Ueber die eigenverte lei den differentialgleichungen der mathematischen physik. Math. Z., 7, 1-57.

D'AUBIGNY, G. and POLIT, E. (1989) Some optimality properties of the generalization of the Tucker method to the analysis of N-way tables with specified metrics. In: (R.Coppi and S.Bolasco). Multiway Data Analysis, 39-52

D'AUBIGNY, G. (1989) L'Analyse Multidimensionnelle des données de dissimilarité. Thèse d'état, Grenoble.

DENIS, J.B. and DHORNE, T. (1989) Orthogonal tensor decomposition of 3-way tables. In: (R.Coppi and S.Bolasco). Multiway Data Analysis. 31-38.

ECKART, C. and YOUNG, G. (1936) The approximation of one matrix by another of lower rank. Psychometrika, 211-218.

ESCOUFIER, Y. (1985) L'Analyse des Correspondances : ses propriétés et ses extensions. Proceeding of 45th Session ISI,1985,28.2.1-28.2.16.

ESCOUFIER, Y. (1988) Cours de DEA : Analyse des Données. MontpellierII USTL.

FISHER, E. (1905) Ueber quadratische formen mit reellen koeffizienten. Monastsh.Math.Phys. 16, 234-249.

FRANÇ, A. (1989) Multiway matrices : some algebraic remarks. In: (R.Coppi and S.Bolasco). Multiway Data Analysis.19-30.

FRANÇ, A. (1992) Etude Algébrique des Multitableaux : Apports de l'Algèbre Tensorielle. Thèse de Doctorat, spécialité Statistiques, Montpellier II.

HARSHMAN, R.A. (1970) Foundations of the PARAFAC procedure : models and conditions for 'an explanatory' multi-mode factor analysis.UCLA working Papers in Phonetics, 16, 1-84

HENDERSON, H.V. and SEARLE, S.R. (1979) Vec and Vech operators for matrices with some uses in jacobian and multivariate statistics. J.Can.Stat. 65-81.

HILL, M.O. (1973) Reciprocal averaging: an eigenvector method of ordination. J.Ecol. 237-249.

HILL, M.O. (1974) Correspondence analysis: a neglected multivariate method. Appl.Stat. 340-354.

JAFFRENOU, P.A. (1978) Sur l'analyse des familles finies de variables vectorielles. Bases algébriques et application à la description statistique. Thèse 3e cycle Lyon I

JANSSEN, J. MARCOTORCHINO, F. and PROTH, J.M. Eds (1987) Data Analysis : The ins and outs of solving real problems. Plenum Press, New York.

KAPTEIN, A. NEUDECKER, H. and WANSBEEK, T. (1986) An approach to n-mode component analysis. Psychometrika, 269-275.

KIERS, H.L. (1988) Comparison of Anglo-saxon and french three-mode methods. S.A.D. 14-32.

KROONENBERG, P.M. (1983) Three mode principal component analysis. DSWO Press, Leiden.

KROONENBERG, P.M. (1989) Singular value decompositions of interactions in three-way contingency tables. In: (R.Coppi and S.Bolasco). Multiway Data Analysis. N-H. 169-184.

KROONENBERG, P.M.(1989) The analysis of multiple tables in factorial ecology. III three-mode principal component analysis:" Analyse triadique complète ". Ac.Oecol.[Oecol.Gen.], 245-256.

KROONENBERG, P.M. and De LEEUW, J. (1980) Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika, 69-97.

KRUSKAL, J.B. (1977) Three way arrays : rank and uniqueness of trilinear decomposition, with applications to arithmetic complexity and statistics. Lin.Al.Appl. 95-118.

KRUSKAL, J.B. (1984) Multilinear Methods. In: (H.G.Law). Research methods for multimode data analysis, Praeger.

KRUSKAL, J.B. (1989) Rank decomposition and uniqueness for 3-way and N-way arrays. In: (R.Coppi and S.Bolasco). In : Coppi an Bolasco(1989). 7-18.

LANCASTER, H.O. (1951) Complex contingency tables treated by the partition of the chi-square. J.R.S.S. B, 242-249.

LEIBOVICI, D. (1993) Décomposition en Valeurs Singulières d'un Tableau à k entrées : ATP-kmodes, AFC de k variables. In: . ASU XXVe Journées de Statistiques, Vannes 24-28 mai 1993.

OKAMOTO, M. (1969) Optimality of Principal Component. In: (P.R.Krishnaiah , Ed). Multivariate Analysis II, Academic Press, New-York.

RAO, C.R. (1964) The use and interpretation of principal component analysis in applied research. Sankhya A, 329-359.

RAO, C.R. and MITRA, S.K. (1971) Generalized Inverse of Matrices and its Applications. John Wiley and Sons,Inc. New York.

RAO, C.R. (1979) Separation theorems for singular values of matrices and their applications in multivariate analysis. J.Mult.An. 362-377.

ROBERT, P. and ESCOUFIER, Y. (1976) A unifying tool for linear-multivariate statistical methods : the rv-coefficient. Appl.Stat. 257-265.

SABATIER, R. (1987) Méthodes factorielles en analyse des données : approximations et prise en compte de variables concomitantes. Thèse d'état, USTL Montpellier.

SABATIER, R. (1991) Critères et contraintes pour l'ordination simultanée de k tableaux. In: (Lebreton, J-D et Asselain, B (eds)) Biométrie et Environnement. Masson, Paris, sous presse.

SABATIER, R. LEBRETON, J. and CHESSEL, D.. (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. In: (R.Coppi and S.Bolasco). Multiway Data Analysis, 341-352.

SCHWARTZ, L. (1975) Les Tenseurs. Herman, Paris.

TUCKER, L.R. (1966) Some mathematical notes on three-mode factor analysis. Psychometrika, 279-311.

WANSBEEK, T. and VERHEES, J. (1989) Models for multidimensional matrices in econometrics and psychometrics. In: (R.Coppi and S.Bolasco). Multiway Data Analysis, 543-551.

YOSHISAWA, T. (1987) Singular value decomposition of multiway-array data and its applications. Departement of Computer Science, Takeda, Kofu 400 Japan.

CI
Description de l'enquête SEROCO et Buts de notre étude

L'enquête SEROCO est l'une des 5 cohortes de l'action coordonnée n°7 menée par l'ANRS. Cette enquête épidémiologique prospective multicentrique, menée sur 1500 individus infectées par le VIH dans 18 sites cliniques, a débuté en janvier 1988 et le recrutement s'est terminé en décembre 1990, sauf pour les personnes dont la date de contamination est connue à moins de trois mois de l'entrée.

La population étudiée est constituée des personnes de plus de 18 ans dont la date de première sérologie positive au VIH est connue depuis moins d'un an, n'ayant pas reçu de thérapeutique anti-VIH avant l'inclusion, et n'étant pas encore classés dans le groupe IV de la classification internationale du CDC.

Il faut y ajouter les personnes dont la date de contamination est connue à au plus trois mois, et qui constitue une sous-cohorte encore en recrutement.

L'inclusion dans la cohorte s'est faite par consentement d'être suivi pendant au moins trois ans par l'un des 18 sites cliniques. Ont été exclus de la cohorte les hémophiles, une enquête spéciale, HEMOCO, leur était destinée.

Les personnes sont suivies tous les six mois ou tous les trois mois s'il y a aggravation de l'état des patients, selon certains critères cliniques et biologiques. A chaque visite, un bilan clinique et biologique est réalisé ainsi que le recueil d'un questionnaire sur l'utilisation de préservatifs dans les pratiques sexuelles depuis la dernière visite.

On dispose donc de deux types d'information, des mesures socio-démographiques ou socio-épidémiologiques qui ne varient pas au cours du suivi, et des mesures répétées.

Un état descriptif de la cohorte peut être consulté dans l'article de Bucquet et al. (1992), et nous en donnons quelques éléments dans l'introduction du chapitre CII.

L'enquête SEROCO poursuit quatre objectifs principaux :

- **évaluer l'incidence annuelle du SIDA dans la population suivie,**
- **étudier les facteurs pronostiques de cette incidence,**
- **étudier la cinétique de paramètres biologiques de manière à les établir comme critères de jugement dans les essais thérapeutiques,**
- **créer une sérothèque et une cytothèque**.

Pour notre part, nous nous sommes intéressés surtout aux thèmes des facteurs pronostiques, et de la cinétique ou plutôt de la dynamique de quelques paramètres immunobiologiques. Une approche de la cinétique des lymphocytes T CD4 communément appelés les T4 a été faite mais ne sera pas exposée ici.

Une des originalités de la cohorte SEROCO réside dans le questionnaire sur l'utilisation de préservatifs. Une des ambitions étant de chiffrer la réexposition au virus par voie sexuelle nous avons donc dans le chapitre CII A) construit et analysé des indices de pratiques sexuelles non-protégées. Au chapitre CII B) on a tenté de décrire l'évolution de l'utilisation de préservatifs en tenant compte de divers paramètres. Les méthodes multitableaux décrites dans la partie BIII ainsi que des méthodes plus classiques comme la régression logistique nous ont permis de réaliser ces objectifs.

Ayant établi un facteur pronostique lié à la réexposition sexuelle, on a souhaité expliquer la dynamique de certains paramètres, que nous appellerons immuno-biologiques. Ceci nous a amené à étudier dans le chapitre CIII divers facteurs socio-épidémiologiques et cliniques pouvant expliquer l'évolution vers la dépression immunitaire. Les modèles décrits dans la partie A et plus particulièrement dans le chapitre AI ont été appliqué ici, avec les ACP correspondantes.

Si les chapitres de cette partie font références aux parties A et B de ce travail, ils sont toutefois rédigés de façon autonome. De la sorte, ils peuvent être lues indépendamment des parties A et B, notamment par des épidémiologistes qui seraient d'abord intéressés par les applications. Toutefois on ne s'est pas préocupé des problèmes de troncatures et de censures des données qui sont difficilement conciliables avec les méthodes factorielles.

D'autres résultats prenant en compte ces phénomènes ont été menés, et sont menés à partir des données de l'enquête SEROCO, en continuité ou en parallèle avec ceux présentés ici, par des chercheurs de ou rattachés à l'unité U292 de l'INSERM. Nous en donnons quelques références. Mais, soulignons, que le comité exécutif de SEROCO souhaite par appels d'offres susciter l'utilisation, des données collectées, par des chercheurs pour leurs thèmes propres.

Références :

BUCQUET, D. LEIBOVICI, D. and MAYAUX, M.J. (1991) Is HIV sexual re-exposure a co-factor to the onset of a major form of HIV infection ? Modifications of sexual behaviours during follow-up in the SEROCO french cohort HIV+ population according to HIV serostatus of partners. In: . VII International Conference on AIDS. Florence1991.

BUCQUET, D. (1991) Histoire naturelle - Les études de cohorte - Florence 1991. Jo.Sida (31-32), 25-28.

BUCQUET, D. DEVEAU, C. BELANGER, F. and LEIBOVICI,D (1992) Présentation de la cohorte SEROCO. Bulletin de l'ANRS, n° 8, 24-27.

CARRÉ, N. (1991) Facteurs Pronostiques du SIDA chez les sujets VIH+ dont la contamination est datée. Mémoire de DEA Statistiques et Santé (Pr. J.Lellouch), U.F.R. Médicale de Kremlin-Bicêtre, Université Paris-Sud.

LEIBOVICI,D, BUCQUET.D, and the SEROCO commitee (1992) "Variation of immunologic descriptors among SEROCO HIV seropositives according to an indicator of Hiv sexual reexposure." In :VIII International Conference on AIDS. Amsterdam 1992.

MORELLO, R (1992) Etude de l'influence pronostique de cofacteurs fixes sur la survenue d'un SIDA chez les sujets VIH+ dont la date de contamination est connue. Mémoire de DEA Statistiques et Santé (Pr. J.Lellouch), U.F.R. Médicale de Kremlin-Bicêtre, Université Paris-Sud.

CII
Ré-exposition sexuelle au VIH :

A) construction d'un indicateur

I . Introduction et matériels

I. 1 aperçu de la sexualité et de la contamination par le VIH :

Les études faites sur la sexualité et le VIH ont porté jusqu'à présent, sur la contamination et l'influence de certaines pratiques à risques et/ou sur l'utilisation de préservatifs et le risque de contamination (observation de la séroconvertion). Plusieurs articles, Detels et al. (1990), Stevens et al. (1990) dans le livre édité par l'institut Kinsey ("AIDS and SEX") abordent ces questions.

Il apparaît donc maintenant établi, que certaines pratiques sexuelles sont plus à risque quant à la probabilité d'être contaminé et que celle-ci est augmentée lorsque le nombre de partenaires augmente (Detels et al. (1990), Messiah et al. (1991)). Ces pratiques sexuelles ne sont pas plus l'apanage d'une minorité que celui d'une orientation sexuelle déterminée (homosexuels, hétérosexuels, bisexuels) (Bolling et al. (1987), Reinish et al. (1990)). Il semble que le nombre de partenaires ait une grande importance quant à la contamination, mais que son influence dépende des pratiques sexuelles (Stevens et al. (1990). On a pu tout de même constater sur un suivi de 18 mois dans l'étude SEROCO que les homosexuels ont plus de partenaires que les bisexuels et que ces derniers en ont plus que les hétérosexuels (cf.Tableaux 1, 2).

I.2 la ré-exposition au virus par voie sexuelle :

Nous nous sommes intéressés à la question suivante : **"La ré-exposition sexuelle au virus VIH est-elle liée à l'apparition d'une forme majeure de l'infection ? "**.

A notre connaissance, nous n'avons pas encore vu abordé une telle question dans la littérature.

Pour cela nous avons décidé de construire un indicateur de ré-exposition sexuelle fondée sur les comportements sexuels ainsi que leurs modifications durant le suivi de la cohorte SEROCO. Cette construction nécessite, compte tenu des différents facteurs intervenants dans le comportement sexuel des individus séropositifs au VIH, une réflexion biométrique que nous mènerons dans la partie I de ce travail. Les choix de construction étants faits nous présenterons dans la partie II la méthode de calcul utilisée qui fait appel à des techniques multidimensionnelles. Enfin nous répondrons à notre question dans la partie III en utilisant des modèles inférentiels du type logistique, mesures répétées, tests non-paramétriques.

I.3 la cohorte SEROCO et notre sélection pour l'étude :

La cohorte SEROCO est une étude multicentrique menée en France depuis janvier 1988. Elle comporte actuellement (janvier 1991) 1500 séropositifs dont 53% d'hétérosexuels, 32% d'homosexuels et 15% de bisexuels. De cette cohorte, est recueillie dans des centres cliniques, une batterie de paramètres biologiques, cliniques et de comportements par un suivi tout les six mois. Ce suivi est maintenant de 30 mois pour 150 individus, 24 mois pour 400 individus, 18 mois pour 700 individus et 12 mois pour 800 individus.

En sélectionnant les individus dont la séropositivité a été dépistée durant l'année précédent leur entrée dans la cohorte, nous avons pu conserver 437 individus ayant un recul de 18 mois. Au terme des 18 mois 41 (soit environ 9,5%) sont passés en stade sida. On note les mêmes prévalences d'hétérosexuels, d'homosexuels et bisexuels que pour l'ensemble de la cohorte avec pour les femmes uniquement des hétérosexuelles. En mars 1992 on a également fait une sélection et l'on a obtenu 631 individus. Nous présenterons aussi les résultats par rapport à cette nouvelle cohorte (l'ancienne et 194 individus supplémentaires), mais en remarquant surtout les changements dans les résultats.

tableaux 1, 2 .

Le questionnaire des comportements sexuels est fondé sur les déclarations, faites par les individus, de contacts avec les types de partenaires donnés en précisant pour chaque type de partenaire une fréquence d'utilisation de préservatifs et le nombre de partenaires correspondant. Ce questionnaire porte sur les six derniers mois précédents chacun des quatre bilans (espacés de six mois) .

Les types de partenaires sont déterminés par :

. le sexe du partenaire, **m**, **f** (**m**asculin, **f**éminin),
. un qualificatif de **r**égulier, **o**ccasionnel, **p**rostitué, **r**, **o**, **p**,
. son statut sérologique, **-**, **+**, **s**, **i** (négatif, positif, atteint du SIDA, inconnu),
. la pénétration **a**nale ou **v**aginale **a**, **v** pour un partenaire féminin.

Ces différentiations nous ont donc donnés 27 types de partenaires, mais en raison des effectifs faibles (ou nuls pour certains) notamment pour les pénétrations féminines anales, nous en avons retenu que 11 : **mr-**, **mr+**, **mrs**, **mri**, **mo-**, **mo+**, **moi**, **fr-**, **fr+**, **fri**, et **foi**.

Les fréquences d'utilisation de préservatifs sont qualitatives ordinales : **J**amais (4), **R**arement (3),**S** ouvent (2), **T**oujours (1).

Le nombre de partenaires est une variable numérique et pourra donc éventuellement être recodée en classes.

II . Aspects biométriques de la construction de l'indicateur de ré-exposition

Il semble que pour mesurer une ré-exposition au VIH par voie sexuelle il faille tenir compte de deux principaux aspects :

. le "risque" de ré-exposition qui dépend du type de partenaire. Le statut sérologique du partenaire (+, s, i,-) y joue un grand rôle, mais il est difficilement quantifiable sur ce seul aspect. On mesure alors une intensité globale, dépendante de la cohorte.

. l'intensité de ré-exposition qui dépend du nombre de partenaires, du nombres de contacts, des pratiques sexuelles. On mesure alors une intensité individuelle.

Donc pour chaque type de partenaire j (j=1..11) on peut lui associer un nombre noté Re(j), l'intensité globale ou risque global. Pour chaque individu i, suivant le type de partenaire j, on lui associe le nombre $I_j(i)$, une intensité individuelle. Alors on peut calculer, pour chaque individu une mesure de la ré- exposition par le nombre :

$$\mathbf{exp(i)= \sum_{j=1}^{11} I_j(i)\ Re(j)}\ ,$$

et ceci pour chaque bilan.

Dans la mesure où l'on s'intéresse à l'évolution de chaque individu, plus qu'à l'évolution de la cohorte, on souhaitera que les nombres Re(j) soient indépendant des bilans alors que $I_j(i)$ dépendra directement du bilan considéré.

On peut constater que la protection sexuelle contrôle les deux aspects : risque global et intensité individuelle. En effet le risque global diminuera si la protection (l'utilisation de préservatifs) est importante et/ou fréquente; pour l'intensité c'est évident.

On peut donc choisir de calculer les valeurs des I et Re en utilisant le questionnaire de pratiques sexuelles. On pourra appeler l'indicateur

construit, un **I**ndicateur de **P**ratiques **S**exuelles **N**on-**P**rotégées (**IPSNP**). Il faut alors le distinguer d'un indice de re-contamination ou de ré-infection bien qu'un lien semble évident.

Chaque Re(j) relatera la non-protection observée, dépendante des fréquences déclarées au cours des bilans, pour la cohorte. Nous présenterons la méthode de calcul de ces coefficients dans la partie suivante.

Les intensités individuelles pourront être mesurées directement par le codage de la fréquence déclarée (**J**amais, **R**arement, **S**ouvent, **T**oujours). Le codage peut être 4, 3, 2, 1 ou 3, 2, 1, 0 ou 16, 9, 4, 1 (pour une gravité quadratique)...Cette intensité peut aussi être mesurée par le codage de la fréquence déclarée multipliée par le nombre de partenaires du type correspondant (ou un codage de celui-ci).

On peut remarquer que pour un individu donné le fait d'avoir déclaré plusieurs types de partenaires va dans l'idée d'augmentation de l'intensité individuelle, ainsi on pourra aussi ne considérer que le risque global, l'intensité individuelle étant intrinsèque à la formule de calcul de l'indicateur. On a donc différents indices suivant les choix faits :

(1) $\mathbf{FRexp(i)= \sum_{j=1}^{11} Fr_j(i)\ Re(j)}$,

(2) $\mathbf{NFRexp(i)= \sum_{j=1}^{11} N_j(i)\ Fr_j(i)\ Re(j)}$,

(3) $\mathbf{Rexp(i)= \sum_{j=1}^{11} Re(j)}$,

avec $Fr_j(i)$ codage de la fréquence, $N_j(i)$ codage du nombre de partenaires.

De plus, pour tenir compte dans la non-protection d'une différence de risque suivant le statut sérologique du partenaire, on décomposera l'indicateur en trois indices :

exp = exp+s + expi + exp-.

Chaque indice est calculé sur un groupe de types de partenaires (sur les partenaires **mr+**, **mrs**, **mo+**, **fr+** pour exp+s; sur les partenaires **mri**, **moi**, **fri**, et **foi** pour expi, et sur les partenaires **mr-**, **mo-**, **fr-** pour exp).

Remarque :

-1 On peut d'après la décomposition ci-dessus, écrire un autre indice plus simple, qui consiste pour un individu donné à ne sommer que ses intensités individuelles à l'intérieur d'un même groupe de partenaire (par exemple +s). Le risque global devient intrinsèque à la décomposition en trois groupes. Nous le noterons exS+s, exSi, exS- et exS leur somme et l'on précisera le choix d'intensité si besoin est.

On va maintenant expliciter dans la partie suivante la méthode utilisée pour calculer les pondérations Re(j). On peut noter qu'il sera difficile de valider les valeurs ainsi obtenues. En effet, on ne pourrait que valider les choix préconisés. La fiabilité des valeurs elles dépendent directement de la fiabilité du questionnaire.

Pour le calcul, de l'intensité individuelle intervenant dans les indices, on peut se demander si le fait de demander une fréquence d'utilisation de préservatifs tient compte indirectement du nombre de partenaires. N'y a t-il pas aussi un "biais" dû au fait que les multipartenaires sont peut-être plus sensibles à la protection, et que l'impact des campagnes de prévention a été peut-être plus important pour eux ?

Il y a un dernier problème à s'affranchir, c'est celui de l'effet cohorte qui va se répercuter dans le calcul des "risques globaux" (les Re(j)). En effet la seule sélection envisagée a été la date de séroconversion de manière a avoir une population assez homogène quant à la réponse (stade de la maladie). On n'a pas par exemple sélectionné les individus contaminés uniquement par voie sexuelle. Suivant nos remarques dans l'introduction de ce travail ce choix parait sans valeurs réelles. Par contre le choix de covariables dans les

modèles logistiques par exemple pourront nous sauvegarder contre, par exemple la prépondérance d'un autre effet que la non-protection sexuelle.

III . Calcul des pondérations de l'IPSNP

III.1 tableaux de contingences utilisés :

Pour calculer les coefficients Re(j), on peut organiser les données de manière à avoir 4 (un par bilan) tableaux de contingence des nombres de déclarations. Chaque tableau croise la variable ordinale fréquence d'utilisation à 4 modalités : **J**, **R**, **S**, **T**, et la variable qualitative identifiant le type de partenaire à 11 modalités : **mr-**, **mr+**, **mri**, **mrs**, **mo-**, **mo+**, **moi**, **fr-**, **fr+**, **fri**, **foi**.

Une case du multitableau Z_{ijk} correspond alors au nombre de déclarations obtenues pour le partenaire i, avec la fréquence d'utilisation de préservatifs j, au bilan k.

Nous avons présenté, dans le tableau suivant les 4 tableaux concaténés des observations faites pour les 437 individus sélectionnés de la cohorte SEROCO pour les 4 bilans retenus. On a les nombres de déclarations par partenaires et fréquences d'utilisation d'utilisation de préservatifs pour chacun des 4 bilans. Nous avons aussi rapporté, pour chaque type de partenaires, la moyenne observée du nombre de partenaires de ce type.

On remarquera les marges lignes et colonnes, déjà très instructives. Elles expriment une prépondérance de déclarations de la fréquence **J** (jamais) au premier bilan, et de la fréquence **T** (toujours) aux bilans suivants. Le nombre de patients déclarants des contacts diminue de 92% à 72% sur ce suivi. Les déclarations pour des types de partenaires inconnus diminuent plus sensiblement que pour les autres. Le nombre moyen de partenaires le plus élevé est déclaré pour le type **moi** (masculin occasionnel inconnu) et le reste au cours du suivi.

tableau 3 .

Repeated cross-sectional information on utilization of condoms
By partner type (sex, regular/casual, HIV serostatus)
Number of notifications

BILAN 0

	NEVER J	SELDOM R	OFTEN S	ALWAYS T	LINE MARGIN	AVERAGE NB OF PARTNERS	
mr-	**47**	5	10	13	75	1.173	Male Regular -
mr+	53	8	11	11	83	1.023	Male Regular +
mri	**61**	11	17	15	**104**	2.750	Male Regular ?
mrs	9	1	2	2	14	1.642	Male Regular Aids
mo-	4	2	0	0	6	3.000	Male Casual -
mo+	3	3	0	2	8	1.375	Male Casual +
moi	**50**	22	29	21	**122**	**9.180**	Male Casual Unknown
fr-	40	2	5	14	61	1.032	Female Regular -
fr+	19	4	4	4	31	1.000	Female Regular +
fri	22	2	2	4	30	1.266	Female Regular ?
foi	27	5	4	7	43	**4.295**	Female Casual ?
COL. MARGIN	335	65	84	93	577		

NB OF PATIENTS REPORTING SEXUAL PRACTICES 405/437

BILAN 6

	NEVER J	SELDOM R	OFTEN S	ALWAYS T	LINE MARGIN	AVERAGE NB OF PARTNERS	
mr-	10	0	4	**68**	**82**	1.036	Male Regular -
mr+	8	9	8	**54**	**79**	1.025	Male Regular +
mri	6	2	4	37	49	1.448	Male Regular ?
mrs	0	2	0	3	5	1.000	Male Regular Aids
mo-	0	1	0	4	5	1.600	Male Casual -
mo+	1	0	0	2	3	1.000	Male Casual +
moi	2	2	9	**54**	67	**12.80**	Male Casual Unknown
fr-	2	3	3	46	54	1.092	Female Regular -
fr+	7	2	3	20	32	1.000	Female Regular +
fri	1	0	0	11	12	1.750	Female Regular ?
foi	3	0	0	16	19	**2.631**	Female Casual ?
COL. MARGIN	40	21	31	315	407		

NB OF PATIENTS REPORTING SEXUAL PRACTICES 342/437

BILAN 12

	NEVER J	SELDOM R	OFTEN S	ALWAYS T	LINE MARGIN	AVERAGE NB OF PARTNERS	
mr-	9	5	5	**65**	84	1.023	Male Regular -
mr+	10	8	6	59	83	1.000	Male Regular +
mri	6	2	2	32	42	1.309	Male Regular ?
mrs	0	1	0	1	2	1.000	Male Regular Aids
mo-	1	1	1	7	10	2.600	Male Casual -
mo+	2	0	0	1	3	1.333	Male Casual +
moi	9	3	5	**50**	67	**7.865**	Male Casual Unknown
fr-	1	1	6	45	53	1.037	Female Regular -
fr+	5	2	2	21	30	1.000	Female Regular +
fri	1	0	0	13	14	1.428	Female Regular ?
foi	1	0	2	17	20	**2.900**	Female Casual ?
COL. MARGIN	45	23	29	311	408		

NB OF PATIENTS REPORTING SEXUAL PRACTICES 342/437

BILAN 18

	NEVER J	SELDOM R	OFTEN S	ALWAYS T	LINE MARGIN	AVERAGE NB OF PARTNERS	
mr-	7	3	3	**65**	**78**	1.051	Male Regular -
mr+	9	5	9	**58**	**81**	1.012	Male Regular +
mri	4	0	2	29	35	1.342	Male Regular ?
mrs	0	0	0	0	0	0	Male Regular Aids
mo-	0	0	0	3	3	**4.333**	Male Casual -
mo+	0	0	1	1	2	**5.500**	Male Casual +
moi	5	0	4	**52**	61	**7.114**	Male Casual Unknown
fr-	1	0	5	47	53	1.132	Female Regular -
fr+	7	1	2	17	27	1.000	Female Regular +
fri	0	0	1	11	12	1.416	Female Regular ?
foi	0	0	0	12	12	2.750	Female Casual ?
COL. MARGIN	33	9	27	295	364		

NB OF PATIENTS REPORTING SEXUAL PRACTICES 318/347

Bucquet D, Leibovici D, Mayaux MJ & the SEROCO Committe (Florence 1991)

III.2 synthèse multidimensionnelle par l'ACP-3modes :

On peut considérer ces tableaux sous la forme d'un tableau à trois entrées. Nous allons, pour traiter ce "cube" de données, utiliser une technique appelée par Kroonenberg l'ACP-3modes, introduite par Tucker, (Kroonenberg(1983), Tucker(1963, 1966)).

Cette méthode peut se présenter comme une extension de l'Analyse en Composantes Principales, décomposition d'une forme bilinéaire, au cas de trois indices, décomposition d'une forme tri-linéaire, d'où son nom, mais certaines propriétés d'optimalité de l'ACP ne sont pas vérifiées, (Kroonenberg(1983), Franc(1989), D'Aubigny et Polit(1989)).

Nous avons utilisé, pour calculer les coordonnées du modèle, le programme écrit par Kroonenberg (Tuckals3) qui réalise l'estimation par moindres carrés alternés, (Kroonenberg et De Leeuw (1980)). Le lecteur intéressé pourra consulter cet article ainsi que le livre de Kroonenberg pour une meilleure compréhension du modèle, de ces formes particulières ainsi que ces divers champs d'application.

Pour avoir une idée de la généralisation effectuée, on peut constater l'extension entre l'ACP et l'ACP 3-modes en écrivant les modèles respectifs :

$$\hat{Z}_{ij} = \sum_{p=1}^{r} C_p x_{ip} y_{jp} ,$$

ajustement de la case (i,j) d'un tableau Z par l'ACP,

$$\hat{Z}_{ijk} = \sum_{p1=1}^{r} \sum_{p2=1}^{s} \sum_{p3=1}^{t} c_{p1p2p3} x_{ip1} y_{jp2} z_{kp3} ,$$

ajustement de la case (i,j,k) d'un tableau à trois entrées Z par l'ACP-3modes. Ce modèle est en fait l'extension du modèle :

$$\hat{Z}_{ij} = \sum_{p1=1}^{r} \sum_{p2=1}^{s} c_{p1p2} x_{ip1} y_{jp2} ,$$

dont la solution, donne en fait la même écriture que pour l'ACP (si l'on choisit r=s).

L'ACP-3modes ne nous permet pas de lecture simultanée des réductions sur chaque mode. Par le "joint plot", Kroonenberg(1983) fournit le moyen d'y remédier. Cette technique utilise par exemple les coordonnées des deux premiers modes, et une tranche de la matrice noyau C (i.e. pour p_3 fixé) qui exprime les liens entre les composantes.

III.3 pré-traitement, le double centrage par bilan :

Dans les options de centrage du programme de Kroonenberg nous avons choisi, un bicentrage pour chaque bilan (fuyante du "cube") avec une standardisation par bilan.

En effet, on peut voir ces données sous la forme d'une seule variable identifiant pour chaque bilan un nombre de déclarations faites avec deux facteurs contrôlés : le type de partenaire à 11 modalités et le type de protection, à 4 modalités . On rejoint alors l'idée de l'ANOVA ou plutôt ici de FANOVA (Factor ANalysis Of VAriance, Gollob(1968)) : décomposition

factorielle de l'interaction. Kroonenberg dans son livre rend compte des différentes discussions de ce modèle.

Le bicentrage par tableau correspond à la transformation suivante :

$$Z_{ijk} - Z_{..k} - (Z_{i.k} - Z_{..k}) - (Z_{.jk} - Z_{..k}) = \hat{Z}_{ijk} .$$

C'est à dire que l'on enlève l'effet ligne (effet partenaire) et l'effet colonne (effet exposition) et l'effet moyen tout ceci pour chaque bilan k. Donc par l'ACP-3modes on modélise l'interaction ligne-colonne (partenaire exposition) croisée avec la répétition de k bilans.

Remarques:

-1 Dans une optique totalement univariable les trois facteurs (partenaires, expositions, bilans) sont en fait hiérarchisés avec une observation par cellule. L'écriture d'un modèle anova sur ce "design" peut s'avérer complexe.

-2 Nous recherchons un codage ordinal de la fréquence d'utilisation de préservatifs pour induire un ordre de réexposition suivant le type de partenaire. On aurait pu alors imposer une contrainte ordinale (voir BI) sur le mode fréquence (J> R>S>T) mais fort heureusement cet ordre s'est trouvé dans les résultats de l'ACP-3modes.
Nous avons préféré utiliser une méthode déjà bien connue des psychométriciens plutôt que de mettre en oeuvre l'ATP-kmodes ici (qui d'ailleurs n'existait pas encore au moment où nous avons construit l'IPSNP).

-3 D'autres modèles liés à l'ACP-3modes auraient pu être étudiés tels qu'une généralisation à trois modes de l'Analyse des Correspondances proposée par Kroonenberg dans "Multiway Data Analysis"(1989), ou encore l'AFC-3modes du chapitre BIII. Mais il nous a semblé que la variable Bilan jouait, par rapport à l'objectif cherché, un rôle secondaire et plutôt de répétition par rapport aux deux autres.

On aurait pu encore modéliser, une association ou une indépendance, uniquement sur les deux premières variables. Les marges étants très différentes (autant en ligne quand colonne) on a choisi, pour une

interprétation plus simple, l'interaction qui est tout de même assez proche de la notion d'association.

-4 On pourra aussi consulter Okamoto (1972) ou Cazes et al. (1988) sur le bi-centrage en ACP.

III.4 échelle des poids obtenue :

L'ACP-3modes a été réalisée avec un choix de 2 axes sur chaque modes, ce qui nous a permis de modéliser 98 % de la variabilité. Les graphiques sur les axes standardisées font apparaître les différences entre les modalités de chaque mode ainsi que les liens entre le mode 1 et le mode 2 par les "joint-plots" (voir fig.1).

On peut observer l'opposition du premier bilan aux trois autres, ainsi que nous avions déjà pu constater par le simple examen des tableaux. On a un étalement des modalités de protection sur le premier axe avec une nette opposition de **T**oujours avec les autres. Le lien entre les différents partenaires et cet étalement se voit sur le premier "joint-plot".

Les poids cherchés sont les coordonnées des partenaires sur le premier axe du premier "joint-plot" translatées au zéro choisi. En effet nous voulons des valeurs positives. On peut alors choisir le zéro pour le partenaire situé le plus à gauche du graphique, ou bien encore fixer le zéro au niveau de la modalité **T**oujours. C'est ce dernier choix que nous avons fait. L'échelle de poids obtenue pour chaque sélection est donnée dans le tableau 4.

figure.1.

tableau 4. Echelle des coefficients des onze types de partenaires

	sélection mars 1991			sélection janvier 1992	
1	mo+	3.8382	2	mrs	3.8350
2	mrs	3.8236	1	mo+	3.7735
3	mo-	3.6276	3	mo-	3.5524
4	fri	3.1528	4	fri	3.2061
5	fr+	2.8711	7	foi	3.0401
6	mri	1.9667	5	fr+	2.8751
7	foi	1.5923	6	mri	2.0364
8	fr-	1.1186	8	fr-	0.9160
9	moi	0.8817	9	moi	0.7918
10	mr+	0.7659	10	mr+	0.7125
11	mr-	0.0300	11	mr-	0.2167

Chaque coefficient est l'abscisse du type de partenaire sur le graphique "joint plot component 1 of mode 3" translaté sur T (l'axe des ordonnées passe par T) .

Cette échelle nous renseigne donc :

- sur un ordre des partenaires suivant la non-protection,
- ainsi qu'une distance entre ces partenaires.

Remarques :

-1 On notera que les modifications d'ordre des poids ne changent pas fondamentalement suivant la sélection de janvier 91 et mars 92. On ne peut comparer directement les poids (à moins d'une transformation de standardisation). Les changements d'ordre des poids peuvent s'expliquer d'abord, par un changement de cohorte (on ne travaille pas sur des échantillons stricto sensu). De plus, les nouveaux individus sont entrés dans l'étude plus récemment, donc ont peut-être été mieux informés et sensibilisés sur la maladie et les risques.

IV . Résultats

IV.1 calcul des indices mensualisés :

Après avoir calculé les indices de non-protection sexuelle pour chaque individu, à chaque bilan, on a étudié plusieurs liens éventuels des indices avec :

- le passage au stade 4 sida (CDC 4, B, C1, D),
- le mode de contamination,
- des variables biologiques pertinentes (voir annexe),
- l'évolution au cours du temps.

Pour ceci, on a calculé des indices mensuels, pour chaque individu, en considérant la période, jusqu'au passage au stade 4 sida ou jusqu'au dernier bilan choisi. Pour cela, on a cumulé, pour chaque individu, l'exposition en sommant la valeur de l'indice obtenue à chaque bilan, jusqu'au bilan de passage au stade 4 ou jusqu'au dernier bilan. Cette valeur a été rapporté au nombre de mois d'observation.

Suivant les choix décrit dans la partie II (Aspects biométriques...) on peut calculer trois types d'indicateurs mensuels ((1) , (2) et (3)) et chacun d'eux est alors décomposé en trois indices mensuels suivant les groupes de partenaires :

mexp = mexp+s + mexpi + mexp- .

Dans la partie suivante, sauf précision contraire, on ne considérera que l'indicateur (1) et sa décomposition. Nous ne discuterons des résultats des deux autres indicateurs qu'à la fin (partie V) car les conclusions obtenues, en examinant les différents choix de calcul de l'intensité individuelle, sont apparus relativement similaires à ceux discutés ci-après.

Le choix de mensualiser les indices est surtout utile pour les modèles logistiques (voir ***IV.3***) et pour faire une typologie des individus de manière simple.

IV.2 description des indices par des facteurs socio-épidémiologiques :

IV.2 a. analyse de variance sur mesures répétées

En utilisant des modèles d'analyse de variance pour des mesures répétées (4 mesures) on a testé l'influence de variables telles que le sexe, l'orientation sexuelle et le mode de contamination sur les indices (voir quelques sorties en annexe). Ici les indices n'étant pas mensualisés, le codage de la fréquence d'utilisation de préservatif $Fr_j(i)$ pour le calcul de l'intensité individuelle est important. On peut noter des résultats variables suivant celui-ci.

Le premier modèle fait intervenir le facteur sexe et le facteur orientation sexuelle.

• Il est apparu que l'indice **exp+s** était lié au sexe et à l'orientation sexuelle indépendamment de l'évolution dans le temps. Les hommes ont un indice supérieur aux femmes. En moyenne les hétérosexuels ont un indice supérieur à celui des homosexuels qui est à son tour supérieur à celui des bisexuels. Mais on n'a pas observé de différence dans l'évolution qui est surtout linéaire décroissante (existence aussi d'un effet quadratique et cubique).

Si on utilise l'indicateur (3) (i.e. sans intensité individuelle explicite, donc sans influence de codage) on observe une évolution différente suivant l'orientation sexuelle. Cette évolution est marquée entre les deux derniers bilans, avec un accroissement nul pour les homosexuels et un accroissement négatif pour les hétérosexuels et bisexuels. On peut noter, pour la sélection de mars 1992, que si la différence, pour l'orientation sexuelle, indépendamment du temps subsiste, alors qu'il n'y en a plus pour le sexe, on n'a pas observé d'évolution ni de différences d'évolution. Ceci est

certainement dû à des variances très élevées par rapport à la moyenne : i.e. une grande hétérogénéité.

• Pour l'indice **expi** on a constaté un effet indépendamment du temps du sexe et l'orientation sexuelle, mais pour ce dernier uniquement pour l'indicateur (1), avec des évolutions linéaires différentes. Les Hommes ont en moyenne un indice supérieur aux femmes mais avec une décroissance plus forte. Les bisexuels ont en moyenne un indice supérieur aux deux autres orientations sexuelles et avec une décroissance moins forte et un effet cubique (oscillations).
Pour la sélection de mars 1992, l'évolution générale de l'indice est remise en question pour l'indicateur (3) et demeure problématique pour l'indicateur (1) (pas d'effet manova, un effet anova mais la validité du F est douteuse : voir la partie A I); l'évolution différente suivant le sexe n'est pas observée.

• Pour l'indice **exp-** les deux facteurs précédents ont un effet global (indépendant du temps). Une évolution moyenne linéaire et cubique et différente suivant les modalités du facteur orientation sexuelle. est constaté.

Les hommes ont en moyenne un indice trois fois supérieur à celui des femmes. Les bisexuels ont en moyenne leur indice **exp-** supérieur à celui des hétérosexuels qui l'ont supérieur à celui des homosexuels. L'indice des bisexuels décroît alors que pour les deux autres on a peu de variations. Avec la sélection 1992 et pour l'indicateur (3) on n'observe plus que l'effet orientation sexuelle indépendamment du temps tandis que pour l'indicateur (1) on conserve l'effet d'évolution différente suivant l'orientation sexuelle, mais on n'a plus d'évolution globale.

Le deuxième modèle fait intervenir le mode de contamination et le passage en stade 4 sida pendant le suivi (CDC 4b, c1, d).
Nous avons seulement pu mettre en évidence, pour l'indice **expi,** une évolution différente suivant le mode contamination et suivant le mode contamination croisé avec le passage en stade sida, pour la sélection mars 1992 avec l'indicateur (1). En moyenne la décroissance linéaire est plus forte pour les individus toxicomaniaques et beaucoup plus forte pour ceux d'entres eux passés en stade sida.

IV.2 b. statistiques non-paramétriques sur les indices mensualisés

On a testé l'absence de différence suivant l'orientation sexuelle puis suivant le mode de contamination, par la statistique de somme de rangs de Wilcoxon, la statistique de Van der Waerden (que sous-entend un modèle normal sur les rangs) et celle de Savage (que sous-entend un modèle exponentiel sur les rangs) avec les indices mensualisés venant du choix (1).

On a pu constater que si l'orientation sexuelle n'influençait pas l'indice total **mexp**, l'indice **mexp+s** était en moyenne double pour les hétérosexuels (test de Van der Waerden et test de Savage). Les indices **mexpi** et **mexp-** sont en moyenne légèrement supérieurs pour les homosexuels (test de Van der Waerden, test de Savage et test de Wilcoxon).

Pour le mode de contamination, on a un effet sur l'indice **mexp-** ou les contaminés par toxicomanie ont en moyenne un indice deux fois supérieur à celui des contaminés par transfusion et trois fois supérieur à celui despar voie sexuelle (test de Van der Waerden, test de Savage et test de Wilcoxon). On constate un effet sur l'indice **mexp+s** par un autre test : le test de la médiane, où pour le groupe des contaminés par voie sexuelle on a 15 % d'individus au dessus de la médiane de plus que prévus.

On peut noter que pour la sélection de mars 92 (631 individus) on a eu en plus, l'effet du mode de contamination (pour tous les tests) et de l'orientation sexuelle (test de Savage uniquement) sur l'indice total **mexp**.

IV.3 influence de la ré-exposition sexuelle sur la survenue d'une forme majeure de l'infection par le vih :

La ré-exposition sexuelle est mesurée par les indices mensualisés de non-protection sexuelle pour l'indicateur (1). La survenue d'une forme majeure de l'infection est le passage au stade 4 sida (classification CDC 4, B, C1, D).

IV.3 a. modèles sans covariables

Les tests non-paramétriques déjà mis en oeuvre précédemment pour les facteurs socio-épidémiologiques ont été effectués par rapport à la variable dichotomique de passage en stade 4 sida au cours des 4 bilans considérés.

Les individus atteints du sida ont, en moyenne, un indice mensuel total (**mexp**) presque double (test de Van der Waerden, test de Savage et test de Wilcoxon). Si l'on regarde les sous-indices, on a le même résultat pour **mexpi** et **mexp+s** avec les mêmes tests (sauf pour **mexp+s** ou le test de Wilcoxon et celui de Van der Waerden ne sont pas significatifs). On a aussi un effet sur **mexp-**, avec un indice en moyenne deux fois plus petit pour les individus atteint du sida.

Les tests de Savage nous font penser a des effets de modèles logistiques que nous avons d'ailleurs testés ensuite. Ainsi on constate que chaque indice (le total et les sous-indices) ont des β ("log odd ratio") significatifs et inférieur à -2 pour **mexp-**, ce qui fait penser à un effet protecteur, et supérieurs à 0,5 pour les autres, ce qui marque donc un risque grandissant avec la non-protection sexuelle.

IV.3 b. modèles avec covariables

Pour confronter et ajuster, l'effet observé d'une association de la non-protection sexuelle sur la survenue d'une forme majeure de l'infection par le vih, nous avons introduit dans le modèle logistique d'autres variables susceptibles de modifier nos conclusions.

-covariables socio-épidémiologiques.

Tout d'abord, nous avons testé des modèles avec le sexe l'orientation sexuelle et le mode contamination comme covariables des indices
Il s'est avéré que seul la variable sexe entrait dans les modèles avec les indices, mais toujours après l'indice, dans la recherche pas à pas du modèle sauf pour l'indice **mexp-**. Le coefficient relate toujours d'un risque plus grand pour les hommes. On peut noter que pour la sélection de janvier 1991 la variable sexe entre dans le modèle uniquement avec **mexp-**, alors qu'elle est présente dans tous les modèles pour la sélection de mars 1992.

-covariables biologiques.

Nous avons choisi quelques variables biologiques décrivant le comportement immunitaire, pour confronter le risque de ré-exposition sexuelle au vih avec la dépression immunitaire caractéristique du sida.
Ces variables sont :

- le nombre de T4,
- l'hématocrite,
- le rapport entre le nombre de T4 et le nombre de T8,
- le dosage en IgM,
- le dosage en IgG,
- l'antigénémie au vih.

La question que l'on se pose est : **" est-ce qu'indépendemment du "profil immunitaire" observé à l'entrée dans l'enquête, le comportement présent et futur de non-protection sexuelle va accélérer l'apparition du sida ? "**.

On peut noter tout d'abord que le modèle pas à pas avec uniquement les variables biologiques rend compte de l'effet connu du nombre de T4 sur l'apparition du sida et ensuite ici du dosage en IgG (non-observé pour la sélection de 1992). Les autres variables ne rentrent pas dans le modèle.

L'ordre d'entrée dans le modèle logistique pas à pas est le nombre de T4 puis **mexp** puis le dosage en IgG. Pour les indices **mexp+s** et **mexpi** la variable de dosage en IgG passe devant l'indice. L'indice **mexp-** ne rentre plus dans le modèle.
Donc la ré-exposition sexuelle (mesurée ici par l'indicateur de non-protection sexuelle) augmente le risque dû au nombre de T4 et on peut remarquer que l'association avec le dosage en IgG augmente encore le risque. L'effet "protecteur" dû à **mexp-** devient négligeable.

-covariables biologiques et de comportement de protection sexuelle

Aux modèles précédents nous avons ajouté des variables simples (dichotomiques) de comportement de protection sexuelle (observée au premier bilan). Elles seront en concurrence avec notre indicateur qui lui est plus élaboré.

Tout d'abord, pour enregistrer le comportement d'utilisation de préservatifs, nous avons construit les variables MT MS MR et MJ construites comme :

. avoir déclaré, pour chaque type de partenaire, soit aucun contact, soit une fréquence d'utilisation Toujours, avec au moins pour un des partenaires une déclaration Toujours (MT),

. pour (MS) avoir déclaré pour chaque type de partenaire, soit aucun contact, soit une fréquence d'utilisation Toujours, soit une fréquence Souvent, avec au moins pour un des partenaires une déclaration Souvent

. mêmes constructions pour Rarement (MR) et Jamais (MJ).

Un deuxième type de variable êst construit avec les mêmes variables que précédemment mais non plus sur les onze partenaires possibles mais sur ceux porteurs du virus (+ ou s) pour M1T, M1S, M1R, M1J, sur ceux dont le statut sérologique est inconnu (i) pour M2T,..., et sur ceux déclarés non-porteurs du virus (-) pour M3T,...

.Enfin nous avons aussi considéré comme troisième façon, les mêmes variables pour la valeur 1, mais pour la valeur 0 avoir déclaré au moins un contact avec une fréquence inférieure pour l'ordre J<R<S<T (MaT,...,Ma1T, etc...) .

Ces dernières variables relatent donc d'une moins bonne fréquence, au sens de l'ordre J<R<S<T et de l'ensemble des partenaires, donnée contre une moins bonne fréquence que celle-là.

Elles sont plus pertinentes que les premières et l'on peut remarquer que l'on opère une sélection qui consiste à exclure de l'analyse des individus qui se sont moins exposés qu'un niveau donné, lorsqu' on introduit une de ces variables dans un modèle.

Par exemple : Ma1S vaudra 1, pour un individu qui a pour un des quatres types de partenaires mr+, mrs, mo+, fr+, déclaré une protection Souvent, et Souvent, Toujours ou aucune pour les trois autres. Si il a 0 pour cette variable il aura au moins déclaré une protection Rarement ou Jamais pour un des types de partenaires. Si il a une valeur manquante c'est qu'il a déclaré pour tous ces partenaires soit aucun contact soit une protection Toujours.

Nous avons donc calculé plusieurs régressions logistiques faisant intervenir l'exposition sexuelle avec les variables biologiques , et chacune des variables dichotomiques définies ci-avant.

avec MaT :

La sélection diminue la population de 45 individus (n=392 i.e 89%). On a plus que les individus ayant déclaré des contacts avec au moins l'un des onze partenaires.

On obtient le même résultat qu'avec les covariables biologiques seules mais avec des probabilités améliorées et des coefficients supérieurs. Ceci peut nous faire penser que l'on constate mieux un effet de l'exposition sexuelle sur ceux qui s'exposent effectivement.

avec MaS :

La sélection diminue la population de 93 individus (n=344 i.e 78%). On a les individus ayant eu des contacts avec au moins l'un des onze partenaires, mais avec comme meilleure protection déclarée Souvent, Rarement ou Jamais. Sont exclus de l'analyse ceux qui n'ont, pour chaque partenaire, rien déclaré ou déclaré Toujours).

On obtient les mêmes résultats mais ici l'indice **mexp** entre en premier dans le modèle devant les T4 et le dosage en IgG.

Ici on peut dire que la sélection étudiée comporte uniquement des individus dont l'exposition est réelle.

avec Ma1T :

La sélection diminue la population de 311 individus (n=126 i.e 29%) et donc l'on a plus que les individus ayant déclaré des contacts avec au moins l'un des quatres partenaires séropositifs ou atteints du SIDA.

Pour **mexp** l'ordre d'entrée est **mexp** et l'hématocrite. Pour **mxp+s** l'hématocrite passe devant dans le modèle. Pour **mexpi** l'ordre d'entrée est hématocrite, T4/T8, **mexpi** avec un "log odd ratio" supérieur à 2.

avec Ma1S :

La sélection diminue la population de 328 individus (n=109 i.e 27%) et donc l'on a les individus ayant eu des contacts avec au moins l'un des quatres partenaires, mais avec comme meilleure protection déclarée Souvent, Rarement ou Jamais.

Le changement par rapport à Ma1T est que la variable Ma1s entre en premier pour les modèles avec l'indice **mexp+s** et **mexp**,avec un "log odd ratio" de 1,94. L'indice entre avant le dosage en IgG et avant l'hématocrite pour **mexp+s**. Par contre dans le modèle avec **mexpi** l'ordre d'entrée est **mexpi**, T4/T8 puis hématocrite. Le "log odd ratio" de **mexpi** est 2,72.

avec Ma2T,Ma2S :

Pour Ma2T la population observée est n=26 i.e 51% et pour Ma2S n=185 i.e 43%

Pour **mexp+s** on a comme ordre d'entrée **mexp+s** ,T4, dosage en IgG, avec un "log odd ratio" pour **mexp+s** de 1,53. Alors qu'avec **mexp** et Ma2T l'ordre est Antigénémie au vih , T4, igG puis **mexp**. Pour Ma2S **mexp** passe devant puis on a les variables T4, IgG, avec un"log odd ratio" pour **mexp** supérieur à 1. L'indice **mexpi** ne rentre pas dans le modèle. Pour la sélection de mars 1992 celui-ci rentre dans le modèle précédé de l'Antigénémie et le nombre de T4. On peut noter d'ailleurs que pour cette sélection l'indice **mexp+s** ne rentre pas.

V . Conclusions et discussions

Sur 237 individus (qui n'avaient aucune données manquantes dans la répétition des mesures biologiques) on a exprimé le lien de l'indicateur de pratiques sexuelles non-protégées (IPSNP) et les stades 4 A, B, C1, C2, D (tableau .5). Une partition de ces 237 individus, obtenue après une classification ascendante hiérarchique avec les moments d'ordre 2, sur la base des indices mensuels, a donné quatre classes notées :

- IPSNP+s, très liée à l'indice mexp+s, n= 22,
- IPSNPi, très liée à l'indice mexpi, et pas à l'indice mexp+s, n= 70,
- IPSNPii, liée à l'indice mexpi et un peu à mexp+s, n= 12,
- IPSNP-, très liée à l'indice mexp- et pas aux autres, n= 133.

tableau 5: Associations entre la ré-exposition sexuelle au VIH et les stades CDC

Stades typologie	A N	A O		B N	B O		C1 N	C1 O		C2 N	C2 O		D N	D O	
IPSNP+s	7%	19%	**	9%	14%	ns	8%	30%	**	6%	19%	**	9%	14%	ns
IPSNPi	28%	38%	ns	30%	14%	ns	29%	23%	ns	29%	29%	ns	28%	57%	*
IPSNPii	5%	3%	ns	5%	0%	ns	5%	8%	ns	5%	4%	ns	5%	14%	
IPSNP-	58%	38%	**	55%	71%	ns	57%	38%	ns	58%	47%	ns	57%	14%	**

*sur 237 individus ***: p<0,001, **:p<0,05, *: p<0,1, ns :p>0,1 pour un* χ^2

Ce tableau confirme et détaille les résultats précédents. On note aussi que le stade B se distingue des autres stades Sida (C1, D).

La conclusion de ce travail est donc la réponse affirmative à la question que l'on s'était posé en introduction : " La ré-exposition sexuelle au virus VIH est-elle liée à l'apparition d'une forme majeure de l'infection ? ". La construction d'un indicateur de ré-exposition au virus par voie sexuelle, appelé, Indicateur de Pratiques Sexuelles Non-Protégées, nous a permis de répondre à la question et d'observer les liens de cette mesure avec des facteurs socio-épidémiologiques.

Ainsi des liens avec le sexe et l'orientation sexuelle ont été décrits mais avec tout de même une variation suivant l'indicateur choisi et la cohorte observée (notamment pour le sexe ou l'effet est observé pour la sélection de 1991 mais pas pour celle de 1992).

On a pu voir dans les modèles logistiques l'association de l'indice surtout mexpi avec les variables biologiques nombre de T4 et dosage en IgG. Pour les individus réellement exposés au virus le risque due à la réexposition sexuelle est indépendant des autres variables biologiques et notamment des T4.

Il semble que la non-protection sexuelle pour des partenaires inconnus est un aspect important de la non-protection sexuelle car les "odds-ratios" sont les plus élevés. Notons d'ailleurs (cf.tableau.3.) les marges et nombres moyens de partenaires élevées pour les partenaires de statut sérologique inconnu.

La construction de l'indicateur aurait certainement pu se faire avec d'autres choix, et d'ailleurs nous en avons décrit quelques uns. On peut noter que le fait de mensualiser les indices reduit les imperfections (valeurs aberrantes). On peut noter aussi que la décomposition de l'indice aurait pu se faire en deux indices i.e. par rapport aux partennaires séronégatifs, et par rapport aux partenaires séropositifs, sida ou inconnus.

VI . Annexes

1

Réferences

BUCQUET, D. LEIBOVICI, D. and MAYAUX, M.J. (1991) Is HIV sexual re-exposure a co-factor to the onset of a major form of HIV infection ? Modifications of sexual behaviours during follow-up in the SEROCO french cohort HIV+ population according to HIV serostatus of partners. In: . VII International Conference on AIDS. Book of Abstracts,

BOLLING ,D.R. and VOELLER, B. (1987) Anal intercourse among heterosexuals. J.A.M.A. 458-474.

CAZES, P. CHESSEL, D. and DOLEDEC, S. (1988) L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie. R.S.A. 39-54.

D'AUBIGNY, G. and POLIT, E.(1989) Some optimality properties of the generalization of the tucker method to the analysis of n-way tables with specified metrics. In: (R.Coppi and S.Bolasco). Multiway Data Analysis, North-Holland, Amsterdam, 39-52.

DETELS , R. VISSCHER,B.R. JACOBSON, L.P. KINGSLEY,L.A. CHMIEL, J.S. ELDRED, L.J. ENGLISH,P. and GINZBURG,H. (1990) .Sexual activity, Condom use, HIV-1 Seroconversion. In: AIDS and SEX, Voeller.B Reinish.J.M and Gottlib.M (editors), Kinsey Institute Series. Oxford University Press, New-York.

FRANC, A. (1989) Multiway matrices : some algebraic remarks. In: (R.Coppi and S.Bolasco). Multiway Data Analysis, North-Holland, Amsterdam, 19-30.

GOLLOB, H.F. (1968) Confouding sources of variation in factor analysis and analysis of variance. Psychometrika, 73-115.

KROONENBERG, P.M. and De LEEUW, J. (1980) Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika, 69-97.

KROONENBERG P, M. (1983) Three mode principal component analysis.
DSWO Press, Leiden.

KROONENBERG, P.M. (1989) Singular value decompositions of interactions in three-way contingency tables. In: (R.Coppi and S.Bolasco). Multiway Data Analysis, North-Holland, Amsterdam, 169-184.

MESSIAH, A. BUCQUET, D. METTETAL, J.F. LARROQUE, B. ROUZIOUX, C. and BRUGEAT GROUP. (1992) Factors correlated with homosexually acquired HIV infection in the era of 'Safer Sex' : was the prevention message well understood ?. Sexually Transmitted Diseases. (Accepté).

OKAMOTO, M. (1972) Four techniques of Principals Components Analysis. Jap.Stat.Soc. 63-69.

REINISH, J.M. ZIEMBA-DAVIS, M. and SANDERS,S.A. (1990) Sexual behavior and AIDS : lessons from art and sex research. In: AIDS and SEX, Voeller.B Reinish.J.M and Gottlib.M (editors), Kinsey Institute Series. Oxford University Press, New-York.

STEVENS, C.E. TAYLOR,P.E. RODRIGUEZ de CORDOBA,S. ZANG,E.A. and RUBINSTEIN, P. (1990) Sexual activity and Human Immunodeficiency Virus Type 1 infection in a cohort of homosexual men in New-York city. In: AIDS and SEX, Voeller.B Reinish.J.M and Gottlib.M (editors), Kinsey Institute Series. Oxford University Press, New-York.

TUCKER, L.R. (1966) some mathematical nots on three-mode factor analysis. Psychometrika, 279-311.

TUCKER, L.R. (1963) Implications of factor analysis of three-way matrices for measurement of change. In: . Problems in measuring change, Madison University of Wisconsin Press, Harris, C W.(ed).

2

modèles logistiques sans covariables

```
The LOGISTIC Procedure  de SAS
  Data Set: WORK.TRAV
 Response Variable: SID
 Response Levels: 2
 Number of Observations: 437
 Link Function: Logit
      Response Profile
  Ordered
    Value     SID      Count
        1       1         41
        2       2        396
             Simple Statistics for Explanatory Variables
                                    Standard
Variable             Mean         Deviation         Minimum         Maximum
 mexp            0.526266          0.542357               0         3.59469
                      Intercept
                Analysis of Maximum Likelihood Estimates
              Parameter     Standard        Wald              Pr >       Standardized
Variable      Estimate       Error     Chi-Square      Chi-Square      Estimate
 INTERCPT       -2.8191      0.2363       142.3631          0.0001             .
 mexp            0.8588      0.2214        15.0507          0.0001        0.256788
The LOGISTIC Procedure
  Data Set: WORK.TRAV
 Response Variable: SID
 Response Levels: 2
 Number of Observations: 437
 Link Function: Logit
      Response Profile
  Ordered
    Value     SID      Count
        1       1         41
        2       2        396
             Simple Statistics for Explanatory Variables
                                    Standard
Variable             Mean         Deviation         Minimum         Maximum
 mexp+s          0.201641          0.448587               0         3.39876
                      Intercept
                   Intercept        and
Criterion             Only       Covariates    Chi-Square for Covariates
 AIC                 274.069        269.370          .
SC                   278.149        277.530          .
-2 LOG L             272.069        265.370        6.699 with 1 DF (p=0.0096)
                Analysis of Maximum Likelihood Estimates
              Parameter     Standard        Wald              Pr >       Standardized
Variable      Estimate       Error     Chi-Square      Chi-Square      Estimate
  INTERCPT      -2.4593      0.1866       173.6539          0.0001             .
  mexp+s         0.7112      0.2537         7.8560          0.0051        0.175897
   The LOGISTIC Procedure
  Data Set: WORK.TRAV
 Response Variable: SID
 Response Levels: 2
 Number of Observations: 437
 Link Function: Logit
      Response Profile
  Ordered
    Value     SID      Count
        1       1         41
        2       2        396
             Simple Statistics for Explanatory Variables
                                    Standard
Variable             Mean         Deviation         Minimum         Maximum
 mexpi           0.259724          0.336704               0         2.69500
                Analysis of Maximum Likelihood Estimates
              Parameter     Standard        Wald              Pr >       Standardized
Variable      Estimate       Error     Chi-Square      Chi-Square      Estimate
 INTERCPT       -2.6775      0.2221       145.2991          0.0001             .
  mexpi          1.2666      0.3824        10.9715          0.0009        0.235123
   The LOGISTIC Procedure
  Data Set: WORK.TRAV
 Response Variable: SID
 Response Levels: 2
 Number of Observations: 437
 Link Function: Logit
      Response Profile
  Ordered
    Value     SID      Count
        1       1         41
        2       2        396
             Simple Statistics for Explanatory Variables
                                    Standard
Variable             Mean         Deviation         Minimum         Maximum
 mexp-           0.064902          0.165891               0         2.00823
                      Intercept
                Analysis of Maximum Likelihood Estimates
              Parameter     Standard        Wald              Pr >       Standardized
Variable      Estimate       Error     Chi-Square      Chi-Square      Estimate
 INTERCPT       -2.1189      0.1753       146.1523          0.0001             .
  mexp-         -3.5619      2.1714         2.6908          0.1009       -0.325772
```

3

4

Un modèle de mesures répétées pour exp+s

```
General Linear Models Procedure  SAS
Class Level Information
Class    Levels    Values

sexe          2    1 2

MdeC          3    1 2 3
General Linear Models Procedure
Repeated Measures Analysis of Variance
Repeated Measures Level Information

Dependent Variable     exp+s0   exp+s6  exp+s12  exp+s18

     Level of TIME          0        6       12       18

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no TIME Effect
H = Type III SS&CP Matrix for TIME   E = Error SS&CP Matrix

                                S=1    M=0.5    N=214.5

Statistic                      Value          F      Num DF   Den DF  Pr > F

Wilks' Lambda             0.92406780    11.8053           3      431  0.0001
Pillai's Trace            0.07593220    11.8053           3      431  0.0001
Hotelling-Lawley Trace    0.08217168    11.8053           3      431  0.0001
Roy's Greatest Root       0.08217168    11.8053           3      431  0.0001

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no TIME*sexe Effect
H = Type III SS&CP Matrix for TIME*sexe    E = Error SS&CP Matrix

                                S=1    M=0.5    N=214.5

Statistic                      Value          F      Num DF   Den DF  Pr > F

Wilks' Lambda             0.99671852     0.4730           3      431  0.7013
Pillai's Trace            0.00328148     0.4730           3      431  0.7013
Hotelling-Lawley Trace    0.00329228     0.4730           3      431  0.7013
Roy's Greatest Root       0.00329228     0.4730           3      431  0.7013

Manova Test Criteria and F Approximations for
the Hypothesis of no TIME*sexe2 Effect
H = Type III SS&CP Matrix for TIME*sexe2   E = Error SS&CP Matrix

                                S=2    M=0    N=214.5

Statistic                      Value          F      Num DF   Den DF  Pr > F

Wilks' Lambda             0.99034316     0.6987           6      862  0.6507
Pillai's Trace            0.00965823     0.6988           6      864  0.6507
Hotelling-Lawley Trace    0.00974959     0.6987           6      860  0.6507
Roy's Greatest Root       0.00960301     1.3828           3      432  0.2474

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

General Linear Models Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source                   DF     Type III SS   Mean Square  F Value    Pr > F

sexe                      1      85.332900     85.332900     6.32    0.0123
MdeC                      2     177.232698     88.616349     6.57    0.0015
sexe*MdeC                 0       0.000000            .        .         .

Error                   433    5842.822071     13.493815

General Linear Models Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

Source: TIME

                                                                 Adj  Pr > F
      DF     Type III SS    Mean Square   F Value   Pr > F    G - G    H - F
       3   273.21244576   91.07081525     24.31   0.0001   0.0001   0.0001

Source: TIME*sexe

                                                                 Adj  Pr > F
      DF     Type III SS    Mean Square   F Value   Pr > F    G - G    H - F
       3    3.49340602     1.16446867      0.31   0.8176   0.6865   0.6887

Source: TIME*MdeC

                                                                 Adj  Pr > F
      DF     Type III SS    Mean Square   F Value   Pr > F    G - G    H - F
       6   10.26418315     1.71069719      0.46   0.8406   0.7277   0.7296

Source: TIME*sexe*MdeC

                                                                 Adj  Pr > F
      DF     Type III SS    Mean Square   F Value   Pr > F    G - G    H - F
       0            .              .         .        .        .        .

Source: Error(TIME)

     DF     Type III SS    Mean Square
   1299   4866.35858189     3.74623447

Greenhouse-Geisser Epsilon = 0.5404
      Huynh-Feldt Epsilon = 0.5459
```



```
General Linear Models Procedure
Repeated Measures Analysis of Variance
Analysis of Variance of Contrast Variables

TIME.N represents the nth degree polynomial contrast for TIME

Contrast Variable: TIME.1

Source                   DF     Type III SS   Mean Square  F Value    Pr > F

MEAN                      1    192.7867416   192.7867416    32.48    0.0001
sexe                      1      1.4028106     1.4028106     0.24    0.6271
MdeC                      2      4.8169353     2.4084677     0.41    0.6667
sexe*MdeC                 0      0.0000000            .        .         .

Error                   433   2569.7587904     5.9347778

Contrast Variable: TIME.2

Source                   DF     Type III SS   Mean Square  F Value    Pr > F

MEAN                      1    58.08979128   58.08979128    15.70    0.0001
sexe                      1     0.09813160    0.09813160     0.03    0.8707
MdeC                      2     2.84909742    1.42299871     0.38    0.6810
sexe*MdeC                 0     0.00000000            .        .         .

Error                   433   1602.1490377     3.70011325

Contrast Variable: TIME.3

Source                   DF     Type III SS   Mean Square  F Value    Pr > F

MEAN                      1    22.33591287   22.33591287    13.93    0.0002
sexe                      1     1.99246384    1.99246384     1.24    0.2656
MdeC                      2     2.60125041    1.30062521     0.81    0.4451
sexe*MdeC                 0     0.00000000            .        .         .

Error                   433   694.45075373     1.60381236


Level of          -----------exp+s0-----------    -----------exp+s6----------
sexe                N       Mean          SD          Mean          SD

1                 314    1.68240255    3.79841880     0.79508229    2.21125347
2                 123    1.83089027    3.56201122     0.64106967    1.63127067

Level of          -----------exp+s12----------    -----------exp+s18----------
sexe                N       Mean          SD          Mean          SD

1                 314    0.72510618    2.16758291     0.65848790    1.97967670
2                 123    0.58517163    1.31108549     0.41719211    0.76117994

Level of          -----------exp+s0-----------    -----------exp+s6----------
MdeC          N       Mean          SD          Mean          SD

1                 233    2.03856785    3.93293308     0.97917927    2.29637659
2                 143    1.41349902    3.54560281     0.48231399    1.55225623
3                  61    1.25177607    3.27031530     0.51455311    2.18464435

Level of          -----------exp+s12----------    -----------exp+s18----------
MdeC          N       Mean          SD          Mean          SD

1                 233    0.83877000    1.99187005     0.76241644    1.96563450
2                 143    0.39653853    1.48176305     0.35902713    0.96235227
3                  61    0.77903328    2.68201410     0.47698230    2.07882197
```

5

```
The LOGISTIC Procedure
  Data Set: WORK.TRAV
  Response Variable: SID
  Response Levels: 2
  Number of Observations: 423
  Link Function: Logit
            Response Profile
    Ordered
    Value      SID     Count
        1        1        38
        2        2       385
    WARNING: 14 observation(s) were deleted due to missing values for the response or explanatory variables.
                    Simple Statistics for Explanatory Variables
                                          Standard                              Variable
Variable            Mean       Deviation         Minimum         Maximum       Label
  V524        42.848700       4.514935          25.000          59.00       hematocrite
V551         520.860520     279.396441           2.000         1552.00    nbre t4
T4_8           0.612280       0.349904           0.010            2.04
V555        2067.118203     673.075748         590.000         5050.00    igG
V557         236.799054     142.064616          30.000         1400.00    igM
V577         106.328605     503.410500           0.000         9999.00    ag vih1 do
 mexp          0.527373       0.546797           0.000            3.59
    Stepwise Selection Procedure
                                Summary of Stepwise Procedure
                  Variable      Number          Score         Wald                Pr >       Variable
Step    Entered   Removed         In    Chi-Square    Chi-Square          Chi-Square       Label
    1   V551                       1      27.8026             .                0.0001    nbre t4
    2   mexp                       2       7.1805             .                0.0074
    3   V555                       3       5.2100             .                0.0225    igG
                        Analysis of Maximum Likelihood Estimates
                Parameter     Standard        Wald          Pr >       Standardized    Variable
Variable        Estimate      Error      Chi-Square    Chi-Square        Estimate        Label
  INTERCPT         -1.9698     0.7341          7.1994        0.0073               .    Intercept
V551            -0.00502    0.00189         21.2711        0.0001         -0.772982    nbre t4
V555          0.000572    0.000255          5.0359        0.0248          0.212134    igG
 mexp             0.6226      0.2382          6.8301        0.0090          0.187677
    Association of Predicted Probabilities and Observed Responses
  Concordant = 82.4%          Somers' D = 0.652
 Discordant = 17.2%           Gamma     = 0.655
  Tied       =  0.4%          Tau-a     = 0.107
 (14630 pairs)                c         = 0.826
  The LOGISTIC Procedure
    Data Set: WORK.TRAV
    Response Variable: SID
    Response Levels: 2
    Number of Observations: 423
    Link Function: Logit
              Response Profile
      Ordered
      Value      SID     Count
          1        1        38
          2        2       385
      WARNING: 14 observation(s) were deleted due to missing values for the response or explanatory variables.
                      Simple Statistics for Explanatory Variables
                                            Standard                              Variable
Variable            Mean       Deviation         Minimum         Maximum       Label
  V524        42.848700       4.514935          25.000          59.00       hematocrite
V551         520.860520     279.396441           2.000         1552.00    nbre t4
T4_8           0.612280       0.349904           0.010            2.04
V555        2067.118203     673.075748         590.000         5050.00    igG
V557         236.799054     142.064616          30.000         1400.00    igM
V577         106.328605     503.410500           0.000         9999.00    ag vih1 do
 mexp+s          0.202431       0.452934           0.000            3.40
    Stepwise Selection Procedure
                                Summary of Stepwise Procedure
                  Variable      Number          Score         Wald                Pr >       Variable
Step    Entered   Removed         In    Chi-Square    Chi-Square          Chi-Square       Label
    1   V551                       1      27.8026             .                0.0001    nbre t4
    2   V555                       2       5.1452             .                0.0233    igG
    3   mexp+s                     3       4.4929             .                0.0340
                        Analysis of Maximum Likelihood Estimates
                Parameter     Standard        Wald          Pr >       Standardized    Variable
Variable        Estimate      Error      Chi-Square    Chi-Square        Estimate        Label
  INTERCPT         -1.6194     0.6905          5.4897        0.0190               .    Intercept
V551            -0.00526    0.00189         23.4833        0.0001         -0.810030    nbre t4
V555          0.000566    0.000249          5.1739        0.0229          0.210001    igG
 mexp+s           0.5328      0.2647          4.0522        0.0441          0.133036
    Association of Predicted Probabilities and Observed Responses
  Concordant = 81.9%          Somers' D = 0.643
 Discordant = 17.5%           Gamma     = 0.647
  Tied       =  0.6%          Tau-a     = 0.105
 (14630 pairs)                c         = 0.822
  The LOGISTIC Procedure
    Data Set: WORK.TRAV
    Response Variable: SID
    Response Levels: 2
    Number of Observations: 423
    Link Function: Logit
              Response Profile
      Ordered
      Value      SID     Count
          1        1        38
          2        2       385
      WARNING: 14 observation(s) were deleted due to missing values for the response or explanatory variables.
                      Simple Statistics for Explanatory Variables
                                            Standard                              Variable
Variable            Mean       Deviation         Minimum         Maximum       Label
  V524        42.848700       4.514935          25.000          59.00       hematocrite
V551         520.860520     279.396441           2.000         1552.00    nbre t4
T4_8           0.612280       0.349904           0.010            2.04
V555        2067.118203     673.075748         590.000         5050.00    igG
V557         236.799054     142.064616          30.000         1400.00    igM
V577         106.328605     503.410500           0.000         9999.00    ag vih1 do
 mexpi           0.259962       0.336764           0.000            2.70
```

```
     Stepwise Selection Procedure
                                  Summary of Stepwise Procedure
                  Variable          Number       Score         Wald            Pr >       Variable
 Step     Entered     Removed          In    Chi-Square    Chi-Square      Chi-Square      Label
     1    VS51                          1       27.8026             .          0.0001    nbre t4
     2    V555                          2        5.1452             .          0.0233    igG
     3    mexpi                         3        4.6725             .          0.0306
     4    V524                          4        2.2796             .          0.1311    hematocrite
     5                 V524             3        1.5225        2.2525          0.1334    hematocrite
                                 Analysis of Maximum Likelihood Estimates
              Parameter     Standard      Wald          Pr >      Standardized    Variable
 Variable     Estimate        Error    Chi-Square    Chi-Square     Estimate       Label
 INTERCPT       -1.7398       0.7105        5.9953        0.0143            .     Intercept
 V551          -0.00511      0.00108       22.2424        0.0001      -0.787486   nbre t4
 V555          0.000539     0.000249        4.6784        0.0305       0.199949   igG
 mexpi           0.9196       0.4394        4.3795        0.0364       0.176738
     Association of Predicted Probabilities and Observed Responses
   Concordant = 82.7%          Somers' D = 0.658
  Discordant = 16.8%           Gamma     = 0.662
  Tied       =  0.5%           Tau-a     = 0.198
  (14630 pairs)                c         = 0.829
 The LOGISTIC Procedure
   Data Set: WORK.TRAV
 Response Variable: SID
 Response Levels: 2
 Number of Observations: 423
 Link Function: Logit
          Response Profile
   Ordered
   Value     SID     Count
       1       1        38
       2       2       385
   WARNING: 14 observation(s) were deleted due to missing values for the response or explanatory variables.
                          Simple Statistics for Explanatory Variables
                                        Standard                              Variable
 Variable            Mean       Deviation         Minimum         Maximum      Label
   V524         42.848700        4.514935       25.000          59.00    hematocrite
 V551          520.860520      279.396441        2.000        1552.00    nbre t4
   T4_8          0.612280        0.349904        0.010           2.04
 V555         2067.118203      673.075748      590.000        5050.00    igG
 V557          236.799054      142.064616       30.000        1400.00    igM
 V577          106.328605      503.410500        0.000        9999.00    ag vih1 do
  mexp-          0.064980        0.167528        0.000           2.00
     Stepwise Selection Procedure
                                  Summary of Stepwise Procedure
                  Variable          Number       Score         Wald            Pr >       Variable
 Step     Entered     Removed          In    Chi-Square    Chi-Square      Chi-Square      Label
     1    VS51                          1       27.8026             .          0.0001    nbre t4
     2    V555                          2        5.1452             .          0.0233    igG
     3    mexp-                         2        2.1376             .          0.1437
     4                mexp-             2             .        2.2325          0.1351
     SAS                                                                                 11:02 Tuesday, April 28, 1992
165
    The LOGISTIC Procedure
                                 Analysis of Maximum Likelihood Estimates
              Parameter     Standard      Wald          Pr >      Standardized    Variable
 Variable     Estimate        Error    Chi-Square    Chi-Square     Estimate       Label
 INTERCPT       -1.3201       0.6629        3.9659        0.0464            .     Intercept
 V551          -0.00549     0.00108       25.9564        0.0001      -0.845542   nbre t4
 V555          0.000539     0.000243        4.9185        0.0266       0.199967   igG
     Association of Predicted Probabilities and Observed Responses
   Concordant = 81.7%          Somers' D = 0.640
  Discordant = 17.7%           Gamma     = 0.644
  Tied       =  0.6%           Tau-a     = 0.105
  (14630 pairs)                c         = 0.820
The LOGISTIC Procedure

Data Set: WORK.TRAV
Response Variable: SID
Response Levels: 2
Number of Observations: 344
Link Function: Logit


     Response Profile

Ordered
  Value     SID     Count

      1       1        32
      2       2       312

WARNING: 93 observation(s) were deleted due to missing values for the response or explanatory variables.

                       Simple Statistics for Explanatory Variables
                                        Standard                              Variable
Variable            Mean       Deviation         Minimum         Maximum      Label
V518           9.857558        0.545376        6.000           10.00    indice karnofsky
V551         524.674419      277.005339        2.000        1552.00    nbre t4
T4_8           0.621097        0.355559        0.010           2.04
V555        2062.180233      672.799904      590.000        5050.00    igG
V557         233.636628      139.358898       40.000        1400.00    igM
V577         108.261628      547.867763        0.000        9999.00    ag vih1 do
MAS            0.125000        0.331201        0.000           1.00
MXTOT          0.594562        0.562740        0.005           3.59


Stepwise Selection Procedure
                                  Summary of Stepwise Procedure
                  Variable          Number       Score         Wald            Pr >       Variable
 Step     Entered     Removed          In    Chi-Square    Chi-Square      Chi-Square      Label

     1    MXTOT                         1       24.8711             .          0.0001
     2    V551                          2       12.0563             .          0.0005    nbre t4
```

```
     3    V555                          3        7.2980             .          0.0069    igG
     4    MAS                           4        2.5792             .          0.1083
     5                 MAS              3             .        2.4492          0.1176
                                 Analysis of Maximum Likelihood Estimates
              Parameter     Standard      Wald          Pr >      Standardized    Variable
 Variable     Estimate        Error    Chi-Square    Chi-Square     Estimate       Label
 INTERCPT       -2.9129       0.8684       11.2520        0.0008            .     Intercept
 V551          -0.00363     0.00114       10.0969        0.0015       -0.554026   nbre t4
 V555          0.000716     0.000272        6.9265        0.0085        0.265542   igG
 MXTOT           0.7889       0.2603        9.1897        0.0024        0.244773
 Association of Predicted Probabilities and Observed Responses

  Concordant = 80.6%          Somers' D = 0.616
  Discordant = 19.0%          Gamma     = 0.619
  Tied       =  0.4%          Tau-a     = 0.104
  (9984 pairs)                c         = 0.808
 The LOGISTIC Procedure

 Data Set: WORK.TRAV
 Response Variable: SID
 Response Levels: 2
 Number of Observations: 344
 Link Function: Logit
      Response Profile

 Ordered
   Value     SID     Count

       1       1        32
       2       2       312

 WARNING: 93 observation(s) were deleted due to missing values for the response or explanatory variables.
                        Simple Statistics for Explanatory Variables
                                        Standard                              Variable
 Variable            Mean       Deviation         Minimum         Maximum      Label
 V518           9.857558        0.545376        6.000           10.00    indice karnofsky
 V551         524.674419      277.005339        2.000        1552.00    nbre t4
 T4_8           0.621097        0.355559        0.010           2.04
 V555        2062.180233      672.799904      590.000        5050.00    igG
 V557         233.636628      139.358898       40.000        1400.00    igM
 V577         108.261628      547.867763        0.000        9999.00    ag vih1 do
 MAS            0.125000        0.331201        0.000           1.00
 MX4SPS         0.229610        0.481380        0.000           3.40
  Stepwise Selection Procedure
                                  Summary of Stepwise Procedure
                  Variable          Number       Score         Wald            Pr >       Variable
 Step     Entered     Removed          In    Chi-Square    Chi-Square      Chi-Square      Label

     1    VS51                          1       18.1279             .          0.0001    nbre t4
     2    V555                          2        6.7934             .          0.0091    igG
     3    MX4SPS                        3        5.6906             .          0.0171
     4    V518                          4        2.0817             .          0.1491    indice karnofsky
     5                 V518             3             .        1.9913          0.1582    indice karnofsky
                                 Analysis of Maximum Likelihood Estimates
              Parameter     Standard      Wald          Pr >      Standardized    Variable
 Variable     Estimate        Error    Chi-Square    Chi-Square     Estimate       Label
 INTERCPT       -2.2571       0.7793        8.3891        0.0038            .     Intercept
 V551          -0.00413     0.00113       13.2941        0.0003       -0.638352   nbre t4
 V555          0.000685     0.000262        6.8876        0.0091        0.253981   igG
 MX4SPS          0.6115       0.2738        4.9901        0.0255        0.162303
  Association of Predicted Probabilities and Observed Responses

  Concordant = 79.8%          Somers' D = 0.600
  Discordant = 19.8%          Gamma     = 0.602
  Tied       =  0.4%          Tau-a     = 0.101
  (9984 pairs)                c         = 0.800
 The LOGISTIC Procedure

 Data Set: WORK.TRAV
 Response Variable: SID
 Response Levels: 2
 Number of Observations: 344
 Link Function: Logit

      Response Profile

 Ordered
   Value     SID     Count

       1       1        32
       2       2       312

 WARNING: 93 observation(s) were deleted due to missing values for the response or explanatory variables.

                        Simple Statistics for Explanatory Variables
                                        Standard                              Variable
 Variable            Mean       Deviation         Minimum         Maximum      Label
 V518           9.857558        0.545376        6.000           10.00    indice karnofsky
 V551         524.674419      277.005339        2.000        1552.00    nbre t4
 T4_8           0.621097        0.355559        0.010           2.04
 V555        2062.180233      672.799904      590.000        5050.00    igG
 V557         233.636628      139.358898       40.000        1400.00    igM
 V577         108.261628      547.867763        0.000        9999.00    ag vih1 do
 MAS            0.125000        0.331201        0.000           1.00
 MX4III         0.292922        0.358366        0.000           2.70
  Stepwise Selection Procedure
```



```
                                  Summary of Stepwise Procedure
                  Variable          Number       Score         Wald            Pr >       Variable
 Step     Entered     Removed          In    Chi-Square    Chi-Square      Chi-Square      Label

     1    VS51                          1       18.1279             .          0.0001    nbre t4
     2    V555                          2        6.7934             .          0.0091    igG
     3    MX4III                        3        6.1554             .          0.0131
                                 Analysis of Maximum Likelihood Estimates
              Parameter     Standard      Wald          Pr >      Standardized    Variable
 Variable     Estimate        Error    Chi-Square    Chi-Square     Estimate       Label

 INTERCPT       -2.4804       0.8189        9.1752        0.0025            .     Intercept
 V551          -0.00391     0.00114       11.8902        0.0006       -0.597716   nbre t4
 V555          0.000665     0.000264        6.3710        0.0116        0.246744   igG
 MX4III          1.0737       0.4527        5.6243        0.0177        0.212134
  SAS                                                                                 11:02 Tuesday, April 28, 1992 133

 The LOGISTIC Procedure

 Data Set: WORK.TRAV
 Response Variable: SID
 Response Levels: 2
 Number of Observations: 344
 Link Function: Logit
       Response Profile

 Ordered
   Value     SID     Count

       1       1        32
       2       2       312

 WARNING: 93 observation(s) were deleted due to missing values for the response or explanatory variables.
                        Simple Statistics for Explanatory Variables
                                        Standard                              Variable
 Variable            Mean       Deviation         Minimum         Maximum      Label

 V518           9.857558        0.545376        6.000           10.00    indice karnofsky
 V551         524.674419      277.005339        2.000        1552.00    nbre t4
 T4_8           0.621097        0.355559        0.010           2.04
 V555        2062.180233      672.799904      590.000        5050.00    igG
 V557         233.636628      139.358898       40.000        1400.00    igM
 V577         108.261628      547.867763        0.000        9999.00    ag vih1 do
 MAS            0.125000        0.331201        0.000           1.00
 MX4MMM         0.072829        0.180737        0.000           2.00
  Stepwise Selection Procedure
                  Variable          Number       Score         Wald            Pr >       Variable
 Step     Entered     Removed          In    Chi-Square    Chi-Square      Chi-Square      Label

     1    VS51                          1       18.1279             .          0.0001    nbre t4
     2    V555                          2        6.7934             .          0.0091    igG
     3    MX4MMM                        3        2.3673             .          0.1239
     4                 MX4MMM           2             .        2.5242          0.1121
                                 Analysis of Maximum Likelihood Estimates
              Parameter     Standard      Wald          Pr >      Standardized    Variable
 Variable     Estimate        Error    Chi-Square    Chi-Square     Estimate       Label

 INTERCPT       -1.8104       0.7354        6.0599        0.0138            .     Intercept
 VS51          -0.00451     0.00112       16.1313        0.0001       -0.689526   nbre t4
 V555          0.000643     0.000255        6.3583        0.0117        0.238592   igG


 Association of Predicted Probabilities and Observed Responses

  Concordant = 79.3%          Somers' D = 0.590
  Discordant = 20.3%          Gamma     = 0.593
  Tied       =  0.4%          Tau-a     = 0.100
  (9984 pairs)                c         = 0.795
  SAS                                                                                 11:02 Tuesday, April 28, 1992 180


 The LOGISTIC Procedure
   Data Set: WORK.TRAV
 Response Variable: SID
 Response Levels: 2
 Number of Observations: 109
 Link Function: Logit
       Response Profile
  Ordered
  Value     SID     Count
      1       1        16
      2       2        93
  WARNING: 328 observation(s) were deleted due to missing values for the response or explanatory variables.
                          Simple Statistics for Explanatory Variables
                                        Standard                              Variable
 Variable            Mean       Deviation         Minimum         Maximum      Label
   V524         41.211009        4.832516       26.00           53.00    hematocrite
 V551          521.715596      287.163107        2.00         1307.00    nbre t4
   T4_8          0.650713        0.377657        0.01            1.89
 V555         2110.422018      700.352931     1070.00         5050.00    igG
 V557          250.394495      186.008065       40.00         1400.00    igM
 V577          194.394495      963.328393        0.00         9999.00    ag vih1 do
 MA1S          0.137615        0.346086        0.00            1.00
  mexpi        0.093129        0.189842        0.00            1.13
   Stepwise Selection Procedure
                  Variable          Number       Score         Wald            Pr >       Variable
 Step     Entered     Removed          In    Chi-Square    Chi-Square      Chi-Square      Label
     1    mexpi                         1        4.9770             .          0.0257
     2    T4_8                          2        4.6191             .          0.0316
     3    V524                          3        4.7251             .          0.0297    hematocrite
     4    MA1S                          4        2.6793             .          0.1017
     5                 MA1S             3             .        2.5425          0.1108
                                 Analysis of Maximum Likelihood Estimates
              Parameter     Standard      Wald          Pr >      Standardized    Variable
 Variable     Estimate        Error    Chi-Square    Chi-Square     Estimate       Label

 INTERCPT       4.2353       2.5108        2.8454        0.0916            .     Intercept
 V524          -0.1264       0.0602        4.4086        0.0358       -0.336859   hematocrite
 T4_8          -2.1409       0.9429        5.1549        0.0232       -0.445753
  mexpi         2.7230       1.2954        4.4186        0.0355        0.284998
  Association of Predicted Probabilities and Observed Responses
  Concordant = 77.0%          Somers' D = 0.554
 Discordant = 22.2%           Gamma     = 0.556
 Tied       =  0.2%           Tau-a     = 0.140
 (1488 pairs)                 c         = 0.777
```

CIII
Évolution du profil biologique en fonction :

de facteurs pronostiques et socio-épidémiologiques

I . Introduction et premiers résultats

Ayant établi les rapports de la ré-exposition sexuelle au VIH, dans le chapitre CII avec le passage en stade 4 SIDA (CDC : B, C1 ou D), nous nous sommes intéressés à l'impact de cette ré-exposition sur le profil immunitaire des individus.

En effet, si l'évolution du profil immunitaire vers la déficience est caractéristique de l'évolution vers le sida, on peut s'interroger sur l'effet accélérateur ou accentuateur d'un facteur pronostique du sida sur cette dépression immunitaire, dans ses caractéristiques et ses formes.

Les variables retenues pour décrire l'état immuno-biologique des patients sont :

-le nombre de lymphocytes T4, notée **T4**,
-le nombre total de lymphocytes, notée **Ly**,
-le rapport nombre de T4 sur nombre de T8, notée **T4-8**,
-l'hématocrite notée **HEM**,
-le dosage en immunoglobuline M, notée **IgM**,
-le dosage en immunoglobuline G, notée **IgG**,
-l'antigénémie au VIH, **AvD**.

Pour mesurer les influences de notre indicateur de pratiques sexuelles non-protégées (IPSNP) sur l'évolution de ces variables immuno-biologiques, nous avons mis en oeuvre les techniques statistiques de modèles de mesures répétées associées à l'analyse factorielle que nous avons exposés dans le chapitre AI de ce travail.

On a pu aussi évaluer de la même manière les influences d'autres facteurs. Nous ne présenterons ici à titre d'exemple que les résultats pour le facteur de stade 4, et un facteur issu de l'IPSNP.

On a tout d'abord mis en oeuvre les modèles univariables en optique univariée et multivariée pour chaque variable sur les 237 individus satisfaisant la sélection décrite au chapitre CII et n'ayant pas de données manquantes parmi les variables étudiées. La procédure de SAS/GLM pour des mesures répétées a été utilisée. Cette procédure donne les résultats d'un modèle Split-Plot et d'un modèle multivarié pour chaque variable.

Les résultats univariables peuvent être résumés de la façon suivante :

- les individus passés en stade sida ont globalement des niveaux de T4, Lymphocytes (total), Hématocrites et rapport T4 sur T8 plus faibles et des niveaux de IgG, IgM, et antigénémie plus élevés,
- la décroissance linéaire (parfois quadratique) du premier groupe de variables est générale mais avec décroissance plus forte pour ceux qui passent en stade sida,
- les niveaux des T4 et Lymphocytes sont globalement ordonnés croissants suivant les individus avec un indice mexp+s élevé, puis ceux avec un indice mexpi élevé, puis enfin ceux liés à l'indice mexp-,
- des différences ont été significatives pour l'évolution des variables rapport T4 sur T8 et l'antigénémie pour le classement des individus selon les indices mensualisés.

Après ces premiers résultats, pour mettre en oeuvre les analyses de façon multivariable, nous avons utilisé les méthodes du chapitre AI.

On a été obligé de réduire le nombre d'individus à cause d'un problème de capacité mémoire sur notre PC. En effet l'utilisation de nos propres programmes, à cause du langage SAS/IML, nécessite une grande place mémoire. Ceci nous a permis toutefois d'équilibrer les groupes, issus du classement réalisé au chapitre CII sur l'IPSNP, avec 12 individus tirés au hasard dans chacun d'eux.

Nous avons dans les paragraphes II et III les résultats des différents modèles respectivement pour le facteur stade 4, et le facteur IPSNP.

Enfin la mise en évidence des liens, existant entre divers facteurs socio-épidémioloques et/ou pronostiques, dans l'évolution des variables immuno-biologiques, a été réalisée dans le paragraphe IV. Les méthodes d'analyses à trois modes sous contraintes décrites dans le chapitre BIII ont été appliquées. On a pu reconsidérer, ici, à nouveau tous les individus dans ces analyses puisque les contraintes nous ramènent à des dimensions des tableaux assez faibles.

II . Facteur issu des stades 4 CDC

A partir de la classification des stades du Sida établie par le Center for Disease Control d'Atlanta, nous avons construit trois groupes évalués au terme de deux ans de suivi (4 bilans espacés de six mois) :

- le groupe **s0** des individus n'étant pas en stade 4 ou étant en stade 4 A,
- le groupe **sC** des individus étant en stade 4 C2,
- le groupe **sS** des individus étant en stade Sida i.e. stade 4 B, C1 ou D.

Ce facteur est tautologiquement prédictif du Sida puisqu'il le définit lui-même. Il est toutefois intéressant de constater quelle est la différence, sur le plan immunobiologique entre les stades.

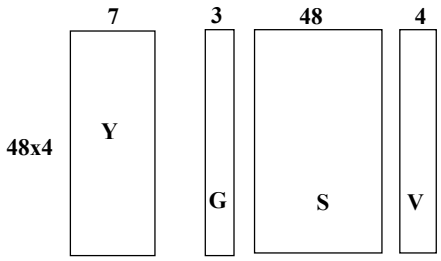
Remarque :

Il est à noter qu'au cours du suivi, les patients peuvent "naviguer" d'un stade à un autre et donc il pourrait être pertinent d'utiliser alors un modèle S.U.R. (chapitre AI) sur le facteur défini par tous les stades.

II.1 modèle Split-Plot :

On a 48 individus (12 dans chaque groupe de l'IPSNP) repartis en 35 individus dans le groupe **s0** (dont 1 était en stade A), 5 individus dans le groupe **sC** et 8 individus dans le groupe Sida, **sS**.

La vision univariée Split-Plot multivariable revient à considérer les n=48 individus mesurés sur q=7 variables en p=4 bilans de la façon suivante :



où le tableau Y est le tableau des 7 variables centrées réduites observées sur les 192 unités statistiques, mesures des 4 bilans sur les 48 patients. G est le tableau des indicatries des groupes, S est le tableau des indicatrices des patients (i.e. des sujets) et V est le tableau des indicatrices des bilans (i.e. des visites). Les données sont dans l'annexe générale.

Rappelons que le modèle Split-plot revient à écrire pour une variable l, mesurée sur l'individu i, du groupe j, à la k ème visite :

$$y^l_{ijk} = \alpha^l_j + \beta^l_{i(j)} + \gamma^l_k + \delta^l_{jk} + rsd^l_{ijk} .$$

On décompose la variable mesurée en somme d'un effet groupe, d'un effet sujet (hiérarchisé par rapport au facteur groupe), un effet visite, un effet intéraction goupe visite, et le résidu.

Cette écriture, écrite traditionellement en terme de variance, devient, lorsque l'on considère toutes les variables, sous forme d'inertie :

$$I_{totale} = I_G + I_{S(G)} + I_{V/(S+G)} + I_{(G.V)/S} + I_{rsd} .$$

L'inertie totale est décomposée en inerties dues à chacun des effets décrits plus haut.

Chaque inertie est décomposée à son tour par l'ACPVI correspondante. Par exemple, l'inertie due à l'effet Groupe, I_G,, est décomposée par l'ACPVI de (Y,Q,D) par rapport à G (i.e. l'ACP de (P_GY,Q,D)) où Q et D sont les métriques sur les lignes et colones de Y choisies a priori .

Remarques pratiques :

-1 P_G est le projecteur sur l'espace engendré par les colonnes de G. Pour le calcul de projecteurs plus complexes, tel que le projecteur sur V/(S+G), on peut se reporter à Sabatier(1987), Yoccoz(1988) ou encore Pontier et al.(1990). Mais notons que si les groupes sont équilibrés on a orthogonalité des sous-espaces en jeu et $P_{V/(S+G)}=P_V$. Si tel n'est pas le cas on calcule V/(S+G) par analyse canonique complête (ACC), méthode que nous avons rappelé dans lechapitre BII à propos de la méthode LONGI.

-2 On analyse alors l'écart à l'hypothèse $\mathbf{H_{01}}$, l'intéraction Groupe .visite par l'ACP de ($P_{G.V}$,Q,D), ou celle de ($P_{(G.V)/S}$,Q,D) si l'on veut une décomposition additive orthogonale de l'inertie totale. Pour le test $\mathbf{H_{02}}$ de l'effet visite on réalise l'ACP de (P_V,Q,D) ou celle de ($P_{V/S}$,Q,D). Celle qui correspond au test $\mathbf{H_{03}}$ est l'ACP de(P_G,Q,D).

-3 Q et D sont en général l'identité d'ordre q et la diagonale des poids uniformes des individus (i.e. 1/np). Si l'on veut se rapprocher des test F, classiquement réalisés pour chaque effet lorsque l'on a une seule variable, on pourra pour chaque analyse changer Q. Par exemple pour le test $\mathbf{H_{03}}$ on pourra prendre $Q = ({}^tYP_{S(G)}Y)^{-1}$. Mais dans un tel cas on n'a plus de décomposition additive de l'inertie totale.

Interprétations des analyses :

On a représenté, sur les deux pages suivantes les résultats des ACPVI décomposant le modèle Split-Plot.L'inertie totale se décompose comme suit :

$$I_{totale} = I_G + I_{S(G)} + I_{V/(S+G)} + I_{(G.V)/S} + I_{rsd} ,$$

100% = 8 % + 71% + 2% + 2% + 17%.

On peut y voir l'ACP totale, c'est à dire celle de (Id_q, Id_{np}), avec les trajectoires des individus.

Pour l'effet sujet, on constate un premier axe opposant les T4 et T4/T8 à l'antigénémie AvD. Cet aspect traduit la dépression immunitaire globalement sur le suivi associée à la presence plus importante de virus. Le deuxième axe, expliqué par les dosages en IgG et IgM, traduit une reaction immunitaire a priori à une infection.

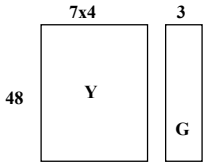
Notons que, s'il semble y avoir plus d'individus des groupes sS ou sC du coté gauche de l'axe 1, aucun groupe net n'apparait. L'individu du groupe sS tout seul en haut à gauche est certainement très "contaminé" et en prise de multiples infections.

L'effet groupe confirme ces constatations et ordonne les groupes par rapport à la dépression immunitaire et l'infection, du groupe s0 au groupe sS en "passant" (ce n'est pas forcément l'histoire des individus) par le groupe sC.

Sur l'intéraction groupe.visite on voit la grande mouvance du groupe sS, la quasi stabilité du groupe s0 et le parcours intermédiaire du groupe sC. Ce dernier a une histoire immunitaire assez étrange puisqu'il a une forte infection et une réaction immunitaire au départ. Ensuite il évolue vers une certaine stabilité, avec un niveau de défense assez faible, mais aussi une moindre présence virale et de réaction immunitaire.

II. 2 modèle "Doubly Multivariate Model" :

L'approche multivariée multivariable revient à considérer les n=48 individus mesurés sur q=7 variables en p=4 instants sous la forme :



On réalise alors le modèle linéaire décrit dans le chapitre AI :

$$Y = G\beta + \varepsilon,$$

et l'on décompose la statistique de Lawley-Hotteling trace(HE^{-1}), qui correspond à chaque test souhaité, par les ACP correspondantes (chapitre AI .III.3).

Pour le test $\mathbf{H_{01}}$ on effectue l'ACP de $(PYM,({}^tM{}^tYP_G^{\perp}YM)^-,D)$; pour le test $\mathbf{H_{02}}$ l'ACP de $(P_GY,M({}^tM{}^tYP_G^{\perp}YM)^{-}{}^tM,D)$ et pour le test $\mathbf{H_{03}}$ l'ACP de $(PY,({}^tYP_G^{\perp}Y)^-,D)$, avec les choix adéquats de M et L pour chaque test. Notons que dans le cas multivariable M a une forme tensorielle qui ici est $M' \ Id_7$.

Interprétations des analyses :

On a présenté sur les pages suivantes l'ACP totale, pour pouvoir comparer en quelque sorte les deux approches univariées et multivariées, ainsi que les décompositions des tests $\mathbf{H_{01}}$ et $\mathbf{H_{03}}$. Les inerties, trace(HE^{-1}), et les significations statistiques (***: $p<10^{-3}$, **: $p<5.10^{-2}$, * : $p< 10^{-1}$) y sont rapportées pour une meilleure lecture simultanée des résultats inférentiels et descriptifs.

Pour le test $\mathbf{H_{02}}$ on a seulement donné la valeur de la statistique car le test $\mathbf{H_{01}}$ étant rejeté on doit utiliser (voir chapitre AI .III.3) un contraste L qui ne permet pas d'ACP.

Remarque :

-1 Appliquer le projecteur P_X revient à faire une analyse des moyennes par groupe, et appliquer le projecteur P revient à faire une analyse des contrastes entre chaque groupe et les autres comme nous l'avons illustré dans un cas hypothétique.

Projecteurs sur X et sur $X({}^tXX)^-{}^tL$:

$$X = \begin{pmatrix} 0&0&1\\ 0&0&1\\ 0&1&0\\ 0&1&0\\ 0&1&0\\ 1&0&0\\ 1&0&0\\ 1&0&0 \end{pmatrix} \qquad L = \begin{pmatrix} 1&0&-1\\ 0&1&-1 \end{pmatrix}$$

$$P_X = X({}^tXX)^-{}^tX = \begin{pmatrix} 0.5&0.5&0&0&0&0&0&0\\ 0.5&0.5&0&0&0&0&0&0\\ 0&0&0.33&0.33&0.33&0&0&0\\ 0&0&0.33&0.33&0.33&0&0&0\\ 0&0&0.33&0.33&0.33&0&0&0\\ 0&0&0&0&0&0.33&0.33&0.33\\ 0&0&0&0&0&0.33&0.33&0.33\\ 0&0&0&0&0&0.33&0.33&0.33 \end{pmatrix}$$

$$X({}^tXX)^-{}^tL = \begin{pmatrix} 0&0.5\\ 0&0.5\\ 0.33&0\\ 0.33&0\\ 0.33&0\\ -0.33&-0.33\\ -0.33&-0.33\\ -0.33&-0.33 \end{pmatrix} \quad P_{X({}^tXX)^-{}^tL} = \begin{pmatrix} 0.375&0.375&-0.125&-0.125&-0.125&-0.125&-0.125&-0.125\\ 0.375&0.375&-0.125&-0.125&-0.125&-0.125&-0.125&-0.125\\ -0.125&-0.125&0.2083&0.2083&0.2083&-0.125&-0.125&-0.125\\ -0.125&-0.125&0.2083&0.2083&0.2083&-0.125&-0.125&-0.125\\ -0.125&-0.125&0.2083&0.2083&0.2083&-0.125&-0.125&-0.125\\ -0.125&-0.125&-0.125&-0.125&-0.125&0.2083&0.2083&0.2083\\ -0.125&-0.125&-0.125&-0.125&-0.125&0.2083&0.2083&0.2083\\ -0.125&-0.125&-0.125&-0.125&-0.125&0.2083&0.2083&0.2083 \end{pmatrix}$$

.

On peut décrire les graphiques de la façon suivante :

- l'intération groupe.visite (ACP du test H_{01}) montre que la différence des contrastes se fait pour l'évolution de la variable AvD, liée au groupe sC au début du suivi et liée au groupe sS en fin de suivi,

- les T4, T4/T8 et Hem traduisent la différence constante sur le suivi entre le groupe s0 et les autres : avec le groupe sS sur le premier axe, avec sC sur le deuxième axe, (ACP du test H_{01} etH_{03})
- les tests valident et confirment les résultats,
- les résultats amenés par l' ACP du test H_{03} confirme tout ceci, ainsi que les tests sur l'effet visite ; à noter le graphique sur les niveaux moyens des variables, très lisible.

Remarque :

-1 L'association de lecture des tests et des graphiques est très pratique et permet à notre avis une vision pertinente du phénomène.

-2 On pourra faire la comparaison entre le test H_{01} mené avec un contraste linéaire au lieu des différences successives, et le test H_{01} mené avec une modélisation linéaire dans le modèle de courbes de croissances.

III . Facteur issue de l'IPSNP

III.1 classement suivant l'IPSNP :

L'IPSNP se présente sous la forme de trois indices mensuels *mexp+s, mexpi, mexp-* qui chiffrent l'exposition mensuelle jusqu'à la fin du suivi (de quatre bilans) ou jusqu'au passage en stade Sida (par rapport à des partenaires séropositifs au VIH ou en stade sida (+s), par rapport à des partenaires de statut sérologique inconnu (i), et par rapport à des partenaires séronégatifs au VIH (-)).

Une typologie des 237 individus sélectionnés précédemment a été réalisée sur la base de ces indices par classification hiérarchique, sur moments centré d'ordre deux. Nous avons extrait une partition en 4 classes :

- une notée +s trés liée à l'indice mexp+s
- une notée ii liée à mexpi et peu à mexp+s
- une notée i trés liée à mexpi et pas du tout à mexp+s
- une notée - liée à mexp- et pas aux deux autres

On peut noter le caractère relativement exclusif de comportement de protection sexuelle de la cohorte, qui rappelons-le s'exprime à travers les déclarations des individus.

La répartition des 237 individus est :

- classe +s 22 individus,
- classe ii 12 individus,
- classe i 70 individus,
- classe - 133 individus.

Nous n'avons donc pour les analyses du paragraphe II et du paragraphe III considéré que 12 individus dans chaque classe. La sélection a été aléatoire dans les groupes.

Nous n'avons pas pour ce facteur présenté les résultats du modèle Split-Plot, préférant à celui-ci le modèle multivarié qui nous permet une lecture commune des tests et des graphiques.

III. 2 modèle DMM et courbes de croissances :

Ici le facteur groupe G a quatre modalité et nous procédons comme pour le facteur stade 4. On a aussi modélisé l'évolution de façon linéaire, linéaire et quadratique, en utilisant le modèle de croissance décrit au chapitre AI .III.4.

Rappelons son écriture et l'ACP qui lui est associée :

$$Y = G\beta\Gamma + \varepsilon$$

où Γ est la modélisation. On effectue l'ACP de $(PY^{t}P_{t\Gamma^{\perp}}, ({}^{t}YP_{G}{}^{\perp}Y)^{-}, Id)$ pour tester le test H_{01}.

- on constate une opposition des +s et i aux - et ii avec une très grande variation de l'antigènémie par rapport aux autres variables. C'est à dire une augmentation plus forte d'antigènes chez les - et ii au cours du suivi que chez les +s et i. Pour les variables immunitaires (T4 etc...) on a un lien de ces deux derniers groupes en opposition aux - et ii qui s'expliquent ici par une décroissance plus forte chez les +s et i de ces variables (il faudrait aussi regarder le test H02) ,
- on a aussi une opposition entre le groupe - et le groupe ii par rapport aux T4 et AvD avec une augmentation plus faible d'antigène (aspect quadratique) et pour les T4 une chute moins forte (deuxième axe),
- et enfin un lien avec les variables IgG et IgM pour les i et ii.

Remarque :

-1 On notera les valeurs des statistiques assez fortes, avec des significativités inférieures à 0.001.

IV . modèles multitableaux à 3 modes pour plusieurs facteurs

IV. 1 facteurs étudiés :

Dans la problématique précédente l'idée a été de faire appel à des techniques inférentielles et de leurs associer des méthodes factorielles. Mais, pour observer l'évolution des variables immuno-biologiques par rapport à plusieurs facteurs, on peut aussi utiliser des méthodes factorielles adaptées. Ce sont les méthodes multitableaux comme par exemple l'ACP-3modes ou encore l'ATP-3modes que nous avons explicités au chapitre BIII. On peut alors effectuer une ACPVI-3modes ou une ATPVI-3modes par rapport à la structure sur les individus qui sera la reunion des structures définies par chaque facteur groupe.

Ce sera d'ailleurs ici l'occasion d'avoir une comparaison pratique, sur un exemple, des deux méthodes (ACP-3modes et ATP-3modes).

Le choix des facteurs groupes et leurs répartitions sur les 237 individus est :

le classement précédent sur l'indicateur IPSNP,

- IPSNP+s	..	22 individus
- IPSNPi	..	70 individus
- IPSNP-	..	133 individus
- IPSNPii	..	12 individus

le classement suivant le sexe,

- Hommes	H	172 individus
- Femmes	F	65 individus

le classement suivant l'orientation sexuelle,

- Hétérosexuels Hetro		121 individus
- Homosexuels	Homo	71 individus
- bisexuels hétéro	BiHe	11 individus
- bisexuels homo	BiHo	24 individus

selon le mode de contamination,

- transfusion	Transf	2 individus
- toxicomanie	Tox	51 individus
- sexe	sex	175 individus
- autres autr		9 individus

et selon le stade au cour du suivi (voir explication ci-après),

- pas groupe 4	sta0	177 individus
- stade A	staA	6 individus
- stade C2	staC2	29 individus
- stade B	staB	6 individus
- stade C1	staC1	12 individus

- stade D staD 7 individus.

Pour avoir un classement fixe du facteur stade, on a considéré au cour du suivi, le plus mauvais stade suivant l'ordre donnée par la liste ci-dessus, par lequel était passé l'individu.

Cet ordre, voulant signifier la gravité de l'état de santé, peut être considéré comme arbitraire, mais la classification obtenue est assez proche du facteur établi à la fin du suivi. Notons tout de même qu'il s'est inspiré des résultats des analyses précédentes. On a voulu distinguer, les individus passants au stade 4 au bilan 18 et au stade A, des individus étants passés en stade 4 avant le bilan 18 (par exemple au stade C2) et étant revenus au stade A au bilan 18.

On peut par cette méthode avoir une vision dynamique de l'evolution des variables.

IV. 2 Résultats par l'ACPVI-3modes :

On a donc ici utilisé le modèle de Tucker décrit par Kroonenberg et De Leeuw(1980), Kroonenberg(1983). Le programme utilisé est l'Algorithme TUCKALS3 écrit par Kroonenberg et Brouwer(1985). Nous renvoyons au chapitre BIII pour une description du modèle. Nous avons demandé lors de chaque analyse 2 axes par mode.

On travaille sur le "cube" 20∞7∞4 des moyennes des 7 variables sur les 20 classes de chacun des groupes à chacun des quatre bilans étudiés.

On se sert des représentations conjointes pour décrire l'analyse obtenue. La façon de les obtenir est décrite dans Kroonenberg et Brouwer(1985). On peut dire qu'elles sont basées sur les composantes de deux modes et la matrice noyau obtenue pour une composante du troisième mode.

De multiples opérations de centrages et réductions préalables au traitement du "cube" de données peuvent être faites par le programme

TUCKALS3. Nous en donnons une illustration ici par rapport à l'objectif d'analyse suivi. On obtient ainsi :

-une analyse des différences des groupes dans l'évolution en centrant les tableaux des variables par bilan et en standardisant ces tableaux,

-une analyse des différences des groupes indépendamment de l'évolution en centrant les tableaux des bilans par variables et en standardisant ces tableaux,

-une analyse des différences des groupes indépendamment de l'évolution et des variables en bi-centrant les tableaux des variables par classe et en standardisant ces tableaux,

Nous donnons sur la page suivante les trois graphiques correspondants aux objectifs précédents.

On retrouve certains résultats déjà décrits pour le facteur issu des stades 4CDC (paragraphe II) et le facteur issu de l'IPSNP qui sont en général responsables de la variabilité. Ils sont plus affinés et il ya des résultats complémentaires.

- Sur la différence des groupes dans l'évolution, on retrouve l'idée du stade C2 intermédiaire au stade SIDA (B, C1, D) et au groupe 0, avec pour le stade SIDA une grande distinction entre le stade B et les deux autres due à l'évolution par rapport à la variable IgG.

On constate une association, déjà vérifiée au chapitre CII, entre IPSNP+s ou IPSNPii avec les stades avancés ou avec les stades SIDA avec ici plus de nuances.

Les homosexuels sont associés à une transmission sexuelle du virus, tandis que les hétérosexuels sont plus associés aux transmissions par toxicomanie, transfusion ou autre. Les contaminés par transfusion sont plus associés au stade B avec les réactions immunitaires traduites par les variables IgG et IgM (ceci se retrouve sur les autres graphiques).

- Sur les différences, indépendamment de l'évolution, on retrouve un peu la même analyse. On constate aussi un profil immuno-biologique identique entre les individus "entrant" en stade 4 au stadeA et ceux qui ont des formes avancées de sida, les stades C1 et D.

Sur les différences, indépendamment de l'évolution et des variables, on a en fait l'interaction variables et bilans. Le premier axe décrit la présence d'IgG plutôt pour les individus de stades 0, B, C2 et les IPSNPi, ii et -. Le deuxième axe décrit la présence d'Antigènes (AvD) plutôt chez les individus des stades C1, D ou 0et les IPSNPi, ii, en opposition au niveau de T4 chez les individus du stade A et IPSNP-, IPSNP+s certainement lié au bilan 0.

On remarquera la particularité des individus classés dans le stade A qui ont pour la plupart quitté les groupes d'infections primaires (voir classification CDC dans l'annexe générale). Le fort lien avec la variable AvD, visible sur le graphique d'ACP-3modes des différences des groupes dans l'évolution, est due à une trés forte valeur de cette variable au bilan 0.

IV. 2 Résultats par l'ATPVI-3modes :

On a mis en oeuvre ici, en guise de comparaison, un modèle à trois modes visant les mêmes objectifs que pour le paragraphe de l'ACP-3modes mais en ayant la meilleure décomposition orthogonale du tenseur étudié, et en ayant donc des solutions emboitées.

On réalise, de façon analogue au cas de deux modes, une décomposition en valeurs singulières du tenseur d'ordre trois.

Pour plus de précisions on se reportera au chapitre BIII III.3. Nous avons utilisé les programmes que nous avons écrit en SAS/IML (donné dans les annexes générales). De la même façon qu'au paragraphe précédent, on travaille sur le même "cube" 20∞7∞4 et l'on effectue les mêmes opérations préalables.

Les résultats sont donnés par les graphiques de la page suivante. Nous donnons, pour chaque analyse, une partie du listing qui retrace la décomposition obtenue (résultats du programme d'ATP-3modes).

Une vsiii est la valeur singulière de la i ème solution k-modes. Une asXiii est une valeur singulière d'une solution associée au x de la i ème solution k-modes (de même pour asYiii et asZiii). Le tenseur solution correspondant à cette valeur singulière est obtenu par SVD classique du tenseur contracté par la solution x de vsiii.

Pour la première analyse on a :

```
ATP-3modes Différences des groupes dans
          l'évolution
        ---------------------------------
                 ---------
  Décomposition après la solution 333
   reconstruction de    99.9994  %
 ------------------------------------
          -----
       Valeurs Singulières
            Pourcent
vs111     15.930159              45.316
             %
 asX111 15.930159                    .
 asX111 1.1846981                 00.251
 asX111 0.4410898                 00.035
 asX111 0.1834004                 00.006
 asY111 15.930159                    .
 asY111  8.289769                 12.271
 asY111 2.4887554                 01.106
 asY111 1.6420372                 00.481
```

```
asZ111 15.930159                    .
asZ111 9.3773327                 15.703
asZ111 4.4169064                 03.484
asZ111 3.0189544                 01.628
asZ111 0.1309897                 00.003
asZ111  0.062945                 00.001
asZ111 0.0023196                 00.000
vs222     9.9780126              17.779
 asX222 9.9780126                    .
 asX222 1.2339213                 00.272
 asX222  0.04513                  00.000
 asX222 7.270E-17                 00.000
 asY222 9.9780126                    .
 asY222 1.2302799                 00.270
 asY222 0.4352337                 00.034
 asY222 4.547E-16                 00.000
 asZ222 9.9780126                    .
 asZ222 0.7942221                 00.113
 asZ222 0.3674464                 00.024
 asZ222  0.031683                 00.000
 asZ222  0.022497                 00.000
 asZ222 0.0009272                 00.000
```

```
  asZ222 3.425E-16          00.000          asY333 1.285E-16          00.000
vs333    2.1565818          00.831          asZ333 2.1565818             .
  asX333 2.1565818             .            asZ333 0.6803893          00.083
  asX333 0.0949193          00.002          asZ333 0.1372313          00.003
  asX333 1.781E-17          00.000          asZ333 0.0273456          00.000
  asX333 1.622E-18          00.000          asZ333 0.0010201          00.000
  asY333 2.1565818             .            asZ333 5.611E-16          00.000
  asY333 0.7215433          00.093          asZ333 8.209E-17          00.000
  asY333 8.757E-16          00.000        vs444    1.0896534          00.212
```

On peut écrire la décomposition du tenseur X centrée, par variable et standardisé par bilan :

$$X = vs111(x_1 \otimes y_1 \otimes z_1) + as111^{x2}(x_1 \otimes y_1^{x2} \otimes z_1^{x2}) + as111^{x3}(x_1 \otimes y_1^{x3} \otimes z_1^{x3})$$
$$+as111^{x4}(x_1 \otimes y_1^{x4} \otimes z_1^{x4}) + \mathrm{L} + as111^{z2}(x_1^{z2} \otimes y_1^{z2} \otimes z_1) + \mathrm{L} + as111^{z7}(x_1^{z7} \otimes y_1^{z7} \otimes z_1)$$
$$+$$
$$vs222(x_2 \otimes y_2 \otimes z_2) + as222^{x2}(x_2 \otimes y_2^{x2} \otimes z_2^{x2}) + as222^{x3}(x_2 \otimes y_2^{x3} \otimes z_2^{x3})$$
$$+as222^{x4}(x_2 \otimes y_2^{x4} \otimes z_2^{x4}) + \mathrm{L} + as222^{z2}(x_2^{z2} \otimes y_2^{z2} \otimes z_2) + \mathrm{L} + as222^{z7}(x_2^{z7} \otimes y_2^{z7} \otimes z_2)$$
$$+$$
$$vs333(x_3 \otimes y_3 \otimes z_3) + as333^{x2}(x_3 \otimes y_3^{x2} \otimes z_3^{x2}) + as333^{x3}(x_3 \otimes y_3^{x3} \otimes z_3^{x3})$$
$$+as333^{x4}(x_3 \otimes y_3^{x4} \otimes z_3^{x4}) + \mathrm{L} + as333^{z2}(x_3^{z2} \otimes y_3^{z2} \otimes z_3) + \mathrm{L}\ as333^{z7}(x_3^{z7} \otimes y_3^{z7} \otimes z_3)$$
$$+vs444(x_4 \otimes y_4 \otimes z_4).$$

Les x sont des coordonnées du mode classements en groupes, les y celles du mode variable et les z celles du mode bilan.

Une approximation satisfaisante à 98 % sera :

$$\hat{X} = vs111(x_1 \otimes y_1 \otimes z_1) + as111^{y2}(x_1^{y2} \otimes y_1 \otimes z_1^{y2}) + as111^{y3}(x_1^{y3} \otimes y_1 \otimes z_1^{y3}) +$$
$$as111^{z2}(x_1^{z2} \otimes y_1^{z2} \otimes z_1) + as111^{z3}(x_1^{z3} \otimes y_1^{z3} \otimes z_1) + as111^{z4}(x_1^{z4} \otimes y_1^{z4} \otimes z_1)$$
$$+$$
$$vs222(x_2 \otimes y_2 \otimes z_2)$$
$$+$$
$$vs333(x_3 \otimes y_3 \otimes z_3).$$

La reconstitution, uniquement avec les trois solutions k-modes, est de 64 %.

Pour la deuxième analyse, centrage par bilan et standardisation par variable, la suite des solutions jusqu'à la 4 ème solution k-modes est :

```
ATP-3modes diférences des groupes indépendamment de
                  l'évolution
 ------------------------------------
                 ------
 Décomposition après la solution 333
   reconstruction de  99.612752  %
 ------------------------------------
                 ------

      Valeurs Singulières
          POURCENT

vs111    13.065322          30.483
                  %             .
  asX111 13.065322             .
  asX111 2.9380031          01.541
  asX111 2.5782767          01.187
  asX111 1.2402687          00.275
  asY111 13.065322             .
  asY111 3.8542006          02.653
  asY111 3.2265253          01.859
  asY111 0.8786849          00.138
  asZ111 13.065322             .
  asZ111 12.206459          26.607
  asZ111 7.7221322          10.648
  asZ111 4.4579759          03.549
```



```
  asZ111 3.2061547          01.836          asZ222 3.578E-16          00.000
  asZ111 2.3698723          01.003        vs333    3.3198328          01.968
  asZ111 1.4459859          00.373          asX333 3.3198328             .
vs222    7.011082           08.778          asX333 1.3954947          00.348
  asX222  7.011082             .             asX333 2.546E-16          00.000
  asX222 2.0372733          00.741          asX333 1.042E-16          00.000
  asX222  0.341089          00.021          asY333 3.3198328             .
  asX222 2.821E-18          00.000          asY333 1.1703967          00.245
  asY222  7.011082             .            asY333 2.505E-16          00.000
  asY222 0.7356257          00.097          asY333 2.282E-16          00.000
  asY222 0.6410856          00.073          asZ333 3.3198328             .
  asY222 4.547E-16          00.000          asZ333 2.6698442          01.273
  asZ222  7.011082             .            asZ333 2.0035114          00.717
  asZ222 2.7919876          01.392          asZ333 0.6580236          00.077
  asZ222 1.5993321          00.457          asZ333 0.6038358          00.065
  asZ222 1.3596334          00.330          asZ333 3.832E-16          00.000
  asZ222 0.9036723          00.146          asZ333 1.159E-16          00.000
  asZ222 0.5580535          00.056        vs444    1.9499364          00.679
```

Pour la troisième analyse, bi-centrage par bilan et variable et standardisation par classement, la suite des solutions jusqu'à la 4 ème solution k-modes est :

```
ATP-3modes diférences des groupes
indépendamment de l'évolution et des variables
 ------------------------------------
                 ------
 Décomposition après la solution 333

 reconstruction de  100.0000  %
 ------------------------------------
                 ------

Valeurs Singulières              POURCENT
   vs111     14.933992          39.826  %
                  %             .
    asX111 14.933992             .
    asX111 6.2402146          06.954
    asX111 0.9811672          00.172
    asX111 4.926E-15          00.000
     asY111 14.933992            .
    asY111 9.5017933          16.122
    asY111  6.370939          07.248
    asY111 2.977E-14          00.000
     asZ111 14.933992            .
    asZ111 8.1214031          11.778
    asZ111 4.6939093          03.934
    asZ111 2.0433808          00.746
    asZ111 0.1846008          00.006
    asZ111   0.05113          00.000
    asZ111 6.515E-16          00.000
  vs222    7.0988743          08.999
                  %             .
    asX222 7.0988743             .
    asX222 1.0290031          00.189
    asX222 9.276E-16          00.000
    asX222 3.398E-16          00.000
     asY222 7.0988743            .
    asY222 2.7204494          01.322
    asY222 1.958E-14          00.000
    asY222 3.929E-16          00.000
     asZ222 7.0988743            .
    asZ222 2.5333632          01.146
    asZ222 1.3851579          00.343
    asZ222 0.2008562          00.007
    asZ222 0.0594677          00.001
    asZ222  9.72E-16          00.000
    asZ222 8.795E-16          00.000
  vs333    2.2405073          00.896
                  %             .
    asX333 2.2405073             .
    asX333 9.944E-16          00.000
    asX333 2.798E-17          00.000
    asX333 2.652E-18          00.000
     asY333 2.2405073            .
    asY333 6.009E-15          00.000
    asY333 3.158E-16          00.000
    asY333 2.382E-16          00.000
     asZ333 2.2405073            .
    asZ333 1.3109952          00.307
    asZ333 0.1550124          00.004
    asZ333 0.0399778          00.000
    asZ333 1.072E-15          00.000
    asZ333 6.642E-16          00.000
  vs444    9.101E-15          00.000
```


On remarque immédiatement qu'il y a des valeurs singulières dont le pourcentage d'inertie expliqué est très proche de celles obtenus pour l'ACP-3modes avec les "Joint plots" (représentations conjointes). Les graphiques correspondants sont alors relativement semblables (à un changement de signe près des axes).

Rappelons que :

- l'ACP-3modes fournit d'abord les composantes pour chaque mode et construit ensuite les représentations conjointes de deux modes associées à une composante du troisième,
- les composantes sont influencées par le choix du nombre de composantes par mode.

Notons que les graphiques de l'ATP-3modes semblables aux "joint plots" sont justement ceux associés à une même composante du mode bilan (voir deuxième page de graphiques) qui se trouve alors représenté sur la première diagonale.

Rappelons aussi un résultat (chapitre BIII III.3) qui assure que si il n'y a que des solutions k-modes alors l'ATP-3modes est équivalente à l'ACP-3modes.

Enfin remarquons pour la première analyse (où le meilleur premier plan est obtenu pour les deux premières solutions 3-modes et où il y a aussi une solution associée à z1 proche de la deuxième valeur singulière) que si l'on retrouve à peu prés les mêmes graphiques entre l'ACP-3modes et l'ATP-3modes, il n'y a pas concordance dans la variabilité expliquée.

V . Conclusions et discussions

Nous avons pu décrire ici certains aspects de la dépression immunitaire, chez les patients séropositifs au VIH, en relation avec quelques facteurs socio-épidémiologiques, dont certains ont un aspect pronostique du sida.

On a pu préciser tout d'abord l'évolution de la réponse immunitaire par rapport aux stades cliniques marqués (facteur issu des stades CDC) et par rapport à un facteur pronostique lié à la ré-exposition sexuelle au virus (facteur issu de l'IPSNP). Ceci a été fait en illustrant les statistiques inférentielles classiques des modèles à mesures répétées, mais aussi en la présence d'autres facteurs, en utilisant des méthodes multidimensionnelles appropriées à l'analyse simultanée de plusieurs tableaux. Ceci nous a permis d'avoir une vision dynamique de cette évolution.

Miedema et al.(1990) ont décrit des aspects dynamiques de la relation entre le virus et le système immunitaire expliquant la variabilité entre les stades.

Benzecri(1990a b) ainsi que Tekaïa et al.(1990) ont présenté des résultats issus d'analyse multidimensionnelle du profil immunitaire en accord avec ceux décrits ici. Les précédents auteurs avaient utilisé toutefois une classification introduisant des subdivisions des stades CDC d'Atlanta engendrées par des variables biologiques, et effectué une analyse transversale. Il semble donc que l'aspect longitudinal et transversal du profil immunitaire des patients vih+ peut avoir des concordances.

La prise en compte de toutes ces informations peut fournir un guide pour le suivi clinique des patients.

Une analyse des transitions des stades cliniques serait intéressante mais nous avons pas ici assez d'individus pour la réaliser. On pourrait aussi (si assez d'individus dans les classes) modéliser la dynamique des variables immuno-biologiques avec ce facteur évolutif par un modèle S.U.R. décrit au chapitre AI.

VI . Références

BENZECRI, J.P. (1990) État du système immunitaire et histoire clinique chez les patients infectés par le virus VIH. Cahiers de l'Analyse des Données. vol XV n°3, 279-284.
BENZECRI, J.P. (1990) Analyse des données biologiques et pathologie clinique. Cahiers de l'Analyse des Données. vol XV n°3, 285-304.
PONTIER, J. DUFOUR, A.B. and NORMAND, M. (1990) Le modèle euclidien en analyse des données. Editions Ellipses, Paris.
SABATIER, R. (1987) Méthodes factorielles en analyse des données : approximations et prise en compte de variables concomitantes. Thèse doct.ès-sciences, U.S.T.L., Montpellier
YOCCOZ,N.G. (1988) Le rôle du modèle euclidien d'analyse des données en biologie évolutive. Thèse de doctorat, Université Lyon 1.
KROONENBERG, P.M. and De LEEUW, J. (1980) Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika 45(1), 69-97.
KROONENBERG, P.M. (1983) Three mode principal component analysis. DSWO Press
KROONENBERG,P.M. and BROUWER, P. (1985) User's guide to TUCKALS3 version 4.0. Section W.E.P., WR 85-09-RP, Departement of education , University of Leiden.
LEIBOVICI, D. (1992) Modèles Linéaires et Analyses Factorielles pour l'Analyse de Facteurs à Mesures Répétées. Communication aux XXIVe Journées de Staistique A.S.U Bruxelles.
MIEDEMA,F. TERSMETTE, M. and Van LIER, R.A.W. (1990) AIDS pathogenesis : a dynamic interaction between HIV and the immune system. Immunology Today, vol XI, n°8,
TEKAÏA,F. SANSONETTI,P. and CLAVERIE, J.M. (1990) Estimation du stade de l'infection par le VIH chez les sujets séropositifs . Cahiers de l'Analyse des Données. vol XV n°3, 264-278.

ANNEXES GÉNÉRALES

I . Classification des Stades du SIDA CDC Atlanta :

avec n° des variables dans l'enquête SEROCO

Les stades CDC (Atlanta) du SIDA

- ***Groupe I*** Infection aigue (primo-infection)
- ***Groupe II*** Infection asymptomatique
- ***Groupe III*** Lymphadenopathies chroniques
- ***Groupe IV*** stades A, B, C1, C2, D, E : SIDA =(B ou C1 ou D).

stade A (dp4a)
- v289 Amaigrissement
- v292 Diarrhée
- v297 Fièvre

stade B (dp4b)
- v405 Démence
- v409 Myélopathies
- v413 Neuropathies périfériques

stade C1 (dp4c1)
- v422 Pneumonie
- v426 Cryptosporidiose chronique
- v430 Toxoplasmose
- v434 Strongyloïdose extra-intestinale
- v438 Isosporose
- v442 Candidose oesophagienne ou bronco-pulmonaire
- v446 Cryptococcose
- v450 Histoplasmose
- v454 Infections à mycobactéries atypiques
- v458 Infections à CMW
- v462 Infections chroniques à virus herpès simplex
- v466 Leuco-encéphalopathie multi-focale progressive

stade C2 (dp4c2)
- v470 Leucophasie chevelue de la langue
- v474 Zona
- v478 Salmonellose récurrente
- v482 Nocardiose
- v486 Tuberculose
- v490 Muguet

stade D (dp4d)
- v495 Sarcome de Kaposi
- v499 Lymphomes non hodgkiniens
- v503 Lymphomes primitifs cérébraux
- v507 Maladie de Hodgkin

stade E (dp4e)
- v511 autres pathologies

II . Programmes Informatiques en SAS/IML :

Les programmes sont écrits en SAS/IML(version 6.03) et en Langage MACRO.

Ceux utilisés dans le chapitre CIII pour les analyses à mesures répétées décrites au chapitre AI ne sont pas exibés ici. Il s'agit d'ACP de triplets particuliers et il faut simplement calculer les éléments du triplet i.e. le tableau et les métriques puis avoir un programme d'ACP avec métriques quelconques.

Nous donnons les programmes d'ATP-kmodes que nous avons décrit dans le chapitre BIII. Pour 3-modes l'écriture est spécifique; l'exécution est plus rapide que le progrmme k-modes avec k=3. Nous donnons d'ailleurs au I.2 le programme qui ne fournit que les solutions k-modes et au I.3 un programme d'ATP-kmodes qui fournit toutes les solutions. Une macro calculant le k+1 uplet pour effectuer une AFC de k variables y est insérée, ce qui en fait un programme d'AFC-kmodes.

Ces programmes ont fonctionné sur un PC 486 avec SAS-PC et à partir de k=4 n'ont pas été trés performant : 40 mn pour l'AFC des 4 variables du programme. Pour l'ATP-3modes il faut environ 5 mn.

La manipulation des tenseurs explique ces contre-performances. L'Optimisation proprement dite est trés rapide, mais la préparation des tenseurs sous forme de tableaux est trés longue. Ceci suggère de reprogrammer chaque macro avec un langage formel ou symbolique, un langage objet. Nous pensons réaliser ce travail prochainement...

II.1 ATP-3modes

```
/* PROCEDURE DVS3 avec métriques*/
/* Didier Leibovici nov 1992*/
/*****************************************************************
********/
/* Realisation de l'acp3mode compromis de
variance et rang   */
/*  identique à la recherche de valeurs singulières
succésives  */
/*    i.e relations de transitions entre (ou dans) les
solutions */
/*****************************************************************
********/
 data es;
 do i=1 to 176;
 x=ranuni(552881962)*1210;

 output;
 end;
    run;

 options linesize=80 pagesize=500 nomprint
nomtrace nosymbolgen;
proc iml worksize=220;

remove _all_;

reset  nolog;

/*****************************************************************
********/

/*********************MACROS************************/

 %global n p q  sol lig col fuy;

 %macro configu(x=X);
/*************************************************/
/*    configuration en tableaux                  */
/*************************************************/
   %if %upcase(&ex) ne X and %upcase(&ex) ne NWX
%then load &ex ;
        ;
 %let nq=%eval(&n*(&q-1));

   tXw=repeat(0,&n,1);
   tXv=repeat(0,1,&q);
   tXu=repeat(0,(&nq +&n),1);

 %if %upcase(&diago)=O %then %do;
Xvr=x;x=x#(sqrt(P)#sqrt(Q)#sqrt(D));%end;;;
 ;
 %if %upcase(&diago)=N %then %do;
Xvr=x;x=(root(P)#root(Q)#root(D))*x;%end;;;
;

  %do t=1 %to &p;
   %let r=%eval(1+(&t-1)*&n*&q);
   TX&t=&X[&r:(&r+&n-1)];

     %do tt=&n %to &nq %by &n;
       TX&t=TX&t | |&X[(&r-1+(&tt+1)):(&r-
1+((&tt/&n)+1)*&n)];

  %end;

  %if &sol=111 %then %do; colvs=["
"];vsing=0;pct=0; %end;
       ;
  colvs=colvs | |["vs&sol"];
  VSing=vSing//vs&sol;
  PCT=PCT//PCTOT;
  k=nrow(vsing);
  Valeurs=vsing[2:k,];
  Pourcent=pct[2:k,];
  colvs2=colvs[,2:k];

    print " ------------------------------------------ ";
    print   valeurs[rowname=colvs2 ]  "  " Pourcent
[format=z6.3] " % " ;
    print " ------------------------------------------ ";

  %if %upcase(&imp)=O %then
  print " Solution &sol" sx &lig sy &col sz &fuy;
                  ;
  %if %upcase(&stoK)=O %then %do;
       sx&sol= sx ; sy&sol= sy ; sz&sol= sz ;
       store sx&sol sy&sol sz&sol;

                  %end;
%mend soluc;
/*****************************************************************
/**********/

 %macro apsolu;
/*************************************************************/

      call svd(szsx&sol,vssx&sol,sysx&sol,wX);
    print " ------------------------------------------ ";
    print "   solutions associées à sx&sol          ";
    print " ------------------------------------------ ";
    vpsx&sol=vssx&sol##2;
  qq=min(&p,&q)  ;
    som= (vpsx&sol[2:qq])[+];
  pctsx&sol=(som/tot)*100;
  pctx&sol=(vpsx&sol/tot)*100;
  pctx&sol[1]=. ;
   if som<>0 then do;
  Pc=(vpsx&sol/som)*100;
  Pc[1]=.;
              end;

  print "% tenseur initial reconstruit en plus =
"Pctsx&sol;
print "Val singulières "vssx&sol" Val propres
"vpsx&sol" " Pc[format=z6.3] "%";
     szsx_&sol=szsx&sol[,1:qq];
     sysx_&sol=sysx&sol[,1:qq];
   %if %upcase(&diago)=O %then  %do;
                  sysx_&sol=(sqrt(Q)##(-
1))#sysx_&sol;      szsx_&sol=(sqrt(P)##(-
1))#szsx_&sol;;
                  %end;
   ;
   %if %upcase(&diago)=N %then  %do;

sysx_&sol=ginv(root(Q))*sysx_&sol;

szsx_&sol=ginv(root(P))*szsx_&sol;;
                  %end;

   %end;

   tXw=tXw | |TX&t ;
   tXv=tXv//TX&t;
   tXu=tXu | |(shape((TX&t)`,1)`);
 %end;

  Xw=(tXw[,2:(&p*&q+1)])`;  free tXw ;
  Xv=tXv[2:(&n*&p +1),] ; free tXv;
  Xu=tXu[,2:(&p+1)]; free tXu;


  %mend configu;
/*************************************************************/
/********/


 %macro inituck1(imp=O);
/*************************************************************/
/*solution initiale Tucker1                    */
/*************************************************************/

        /* Xv  p.n  q */
call svd(scyT,lv,syT,Xv);
sy=syT[,1];
        /* Xw  p.q  n */
call svd(scxT,lw,sxT,Xw);
sx=sxT[,1];
        /* Xu  q.n  p */
call svd(sczT,lu,szT,Xu);
sz=szT[,1];
  free syT sxT szT;

%if %upcase(&imp)=O %then %do;
print "   ";
print  " ----------initialisation tucker1 ----------
--" ;
print "Valeurs singulières &sol de Xw Xv Xu
solution initiale &sol";
 print " ------------------------------------------- ";
print lw  lv  lu  sx &lig sy &col sz &fuy;
     free  lw lv lu;
           %end;
%mend inituck1;
/*************************************************************/
/********/

%macro solucc(X=x,imp=O,stok=N);
/*************************************************************/
/*recherche de la première solution par contractions
*/
/*succèssives                                  */
/*************************************************************/
/* ATTENTION LE X DOIT ETRE CELUI DU
DERNIER CONFIGU*/

   print " ------------------------------------------- ";
   print " +          solution &sol            + ";
   print " ------------------------------------------- ";

 %macro xy_z;
        /* x y -> z */

  cwX=Xw*sx ;     /* tableau p.q 1*/

    twX=cwX[1:&q];
     %do t=1 %to (&p-1);
      twX=twX| |cwX[(&t*&q+1):((&t+1)*&q)] ;
     %end;
     wX=twX`;    /* tableau p q  */


    sz=wX*sy ;   /* tableau p 1  */
     nz=sz`*sz ;;
      sz=sz/sqrt(nz);
 %mend xy_z;


 %macro zx_y ;
       /* z x -> y */

 cuX=Xu*sz ;     /* tableau q.n 1 */
    tuX=cuX[1:&n];
    %do t=1 %to (&q-1);
     tuX=tuX| |cuX[(&t*&n+1):((&t+1)*&n)] ;
    %end;
    uX=tuX`;    /* tableau q n  */
      free tuX ;

    sy=uX*sx;   /* tableau q 1  */
    ny=sy`*sy ;; ;
    sy=sy/sqrt(ny);
 %mend zx_y;

 %macro yz_x ;
       /* y z -> x */

 cvX=Xv*sy;      /* tableau p.n 1 */
     vX=cvX[1:&n];
     %do t=1 %to (&p-1);
      vX= vX| |cvX[(&t*&n+1):((&t+1)*&n)] ;
     %end;
             /* tableau n p  */

    sx=vX*sz;
     nx=sx`*sx ;;   ;
     sx=sx/sqrt(nx);
 %mend yz_x ;

iter=0;
  %if %upcase(&diago) ne  %then x=xvr ;;
                                ;
 do until (test<0.0000000001);

   iter=iter+1;

   s0x=sx; s0y=sy; s0z=sz;  ;


                        %xy_z;


                        %zx_y;


                        %yz_x;  ;
print " ------------------------------------------- ";
print "     Normes aux carrés des solutions      ";
print " ------------------------------------------- ";
print nz " " ny " " nx;

   vs&sol= sqrt(nx);
   testx=(s0x-sx)`*(s0x  sx);;
   testy=(s0y-sy)`*(s0y-sy) ;;
   testz=(s0z-sz)`*(s0z-sz);;
        ;
   test=testx +testy +testz;
print " ------------------------------------------- ";

print iter "    " test "     " vs&sol;
  if iter>200 then test=0.000000000001;
;
  end;

  /***fin boucle*****/
  %if %upcase(&diago)=O %then  %do;
                  sx=(sqrt(D)##(-1))#sx;
                  sy=(sqrt(Q)##(-1))#sy;
                  sz=(sqrt(P)##(-1))#sz;;
                   %end;
     ;;
  %if %upcase(&diago)=N %then  %do;
                  sx=ginv( root(D))*sx;;
                  sy=ginv( root(Q))*sy;;
                  sz=ginv( root(P))*sz;;
                   %end;
 ;

   estX=(sz@sy@sx)*vs&sol;
  %if %upcase(&diago)=  %then
    %do;
    resx=(&ex-estX)`*(&ex-estX);;
     xxx=&ex`*&ex;;
    xxes=estx`*estx;;
    %end;;
  %if %upcase(&diago)=O %then
      %do;
       mm=(P@Q@D);
      resx=(&ex-estX)`*(mm#(&ex-estX));
       xxx=&ex`*(mm#&ex);
      xxes=estx`*(mm#estx);
      %end; ;
  %if %upcase(&diago)=N %then
      %do;
       mm=(P@Q@D);;     /* npq<63 */
      resx=(&ex-estX)`*mm*(&ex-estX);;
       xxx=&ex`*mm*&ex;;
      xxes=estx`*mm*estx;;
      %end;;
    ;
 pctot=(xxes/tot)*100;
 pctes=(xxes/xxx)*100;
   print " ------------------------------------------- ";
   print " +   DVS-3modes-CC    solution &sol      + ";
   print " ------------------------------------------- ";
   print " +   totale  +  expliquée   + résiduelle + ";
   print " ------------------------------------------- ";
print     xxx "   " xxes "   "resx ;
print    " % expliqué      % tenseur initial reconstruit
";
print pctes[ format=z6.3] "              "pctot [
format=z6.3];
   print " ------------------------------------------- ";
```


Thèse de doctorat spécialité Biostatistiques

Didier LEIBOVICI - INSERM Montpellier


```
  %if &sol=111 %then %do; colvs=["
"];vsing=0;pct=0; %end;
       ;
  colvs=colvs | |["vs&sol"];
  VSing=vSing//vs&sol;
  PCT=PCT//PCTOT;
  k=nrow(vsing);
  Valeurs=vsing[2:k,];
  Pourcent=pct[2:k,];
  colvs2=colvs[,2:k];

    print " ------------------------------------------ ";
    print   valeurs[rowname=colvs2 ]  "  " Pourcent
[format=z6.3] " % " ;
    print " ------------------------------------------ ";

  %if %upcase(&imp)=O %then
  print " Solution &sol" sx &lig sy &col sz &fuy;
                  ;
  %if %upcase(&stoK)=O %then %do;
       sx&sol= sx ; sy&sol= sy ; sz&sol= sz ;
       store sx&sol sy&sol sz&sol;

                  %end;
%mend soluc;
/*****************************************************************
/**********/

 %macro apsolu;
/*************************************************************/

      call svd(szsx&sol,vssx&sol,sysx&sol,wX);
    print " ------------------------------------------ ";
    print "   solutions associées à sx&sol          ";
    print " ------------------------------------------ ";
    vpsx&sol=vssx&sol##2;
  qq=min(&p,&q)  ;
    som= (vpsx&sol[2:qq])[+];
  pctsx&sol=(som/tot)*100;
  pctx&sol=(vpsx&sol/tot)*100;
  pctx&sol[1]=. ;
   if som<>0 then do;
  Pc=(vpsx&sol/som)*100;
  Pc[1]=.;
              end;

  print "% tenseur initial reconstruit en plus =
"Pctsx&sol;
print "Val singulières "vssx&sol" Val propres
"vpsx&sol" " Pc[format=z6.3] "%";
     szsx_&sol=szsx&sol[,1:qq];
     sysx_&sol=sysx&sol[,1:qq];
   %if %upcase(&diago)=O %then  %do;
                  sysx_&sol=(sqrt(Q)##(-
1))#sysx_&sol;      szsx_&sol=(sqrt(P)##(-
1))#szsx_&sol;;
                  %end;
   ;
   %if %upcase(&diago)=N %then  %do;

sysx_&sol=ginv(root(Q))*sysx_&sol;

szsx_&sol=ginv(root(P))*szsx_&sol;;
                  %end;

   ;
  print szsx_&sol &fuy; free szsx_&sol szsx&sol;
  print sysx_&sol &col; free sysx_&sol sysx&sol;

   ass=repeat(" asX&sol",1,qq);
  colvs=colvs | |ass; free ass;
  PCT=PCT//(Pctx&sol[1:qq]);
  VSing=vSing//(vssx&sol[1:qq]);

      call svd(sxSY&sol,vssy&sol,szSY&sol,vX);
    print " ------------------------------------------ ";
    print "   solutions associées à sy&sol          ";
    print " ------------------------------------------ ";

   vpsy&sol=vssy&sol##2;
  qq=min(&p,&n);
    som= (vpsy&sol[2:qq])[+];
  pctsy&sol=(som/tot)*100;
  pcty&sol=(vpsy&sol/tot)*100;
  pcty&sol[1]=.  ;
     if som<>0 then do;
  Pc=(vpsy&sol/som)*100;
  Pc[1]=.;            end;
  print "% tenseur initial reconstruit en plus =
"Pctsy&sol;
print "Val singulières "vssy&sol" Val propres
"vpsy&sol" " Pc[format=z6.3] "%";
    szsy_&sol=szsy&sol[,1:qq];
    sxsy_&sol=sxsy&sol[,1:qq];
    %if %upcase(&diago)=O %then  %do;
                  sxsy_&sol=(sqrt(D)##(-
1))#sxsy_&sol;      szsy_&sol=(sqrt(P)##(-
1))#szsy_&sol;;
                  %end;
   ;
   %if %upcase(&diago)=N %then  %do;

sxsy_&sol=ginv(root(D))*sxsy_&sol;

szsy_&sol=ginv(root(P))*szsy_&sol;;
                  %end;
   ;
  print szSY_&sol &fuy; free szsy&sol szsy_&sol;
  print sxSY_&sol &lig; free  sxsy&sol sxsy_&sol;
   ass=repeat(" asY&sol",1,qq);
  colvs=colvs | |ass;  free ass;
  PCT=PCT//(Pcty&sol[1:qq]);
  VSing=vSing//(vssy&sol[1:qq]);

      call svd(sxsz&sol,vssz&sol,syszsol,(uX)`);
    print " ------------------------------------------ ";
    print "   solutions associées à sz&sol          ";
    print " ------------------------------------------ ";

   vpsz&sol=vssz&sol##2;
  qq=min(&q,&n);
    som= (vpsz&sol[2:qq])[+];
  pctsz&sol=(som/tot)*100;
  pctz&sol=(vpsz&sol/tot)*100;
  pctz&sol[1]=.   ;
      if som<>0 then do;
  Pc=(vpsz&sol/som)*100;
  Pc[1]=.;            end;
```


Thèse de doctorat spécialité Biostatistiques

Didier LEIBOVICI - INSERM Montpellier

```
  print "% tenseur initial reconstruit en plus =
"Pctsz&sol;
print "Val singulières "vssz&sol" Val propres
"vpsz&sol" " Pc[format=z6.3] "%";
   szsz_&sol=szsz&sol[,1:qq];
   sxsz_&sol=sxsz&sol[,1:qq];
   %if %upcase(&diago)=O %then  %do;
                  sysz_&sol=(sqrt(Q)##(-
1))#sysz_&sol;      sxsz_&sol=(sqrt(D)##(-
1))#sxsz_&sol;;
                  %end;
   ;
   %if %upcase(&diago)=N %then  %do;

sysz_&sol=ginv(root(Q))*sysz_&sol;

sxsz_&sol=ginv(root(D))*sxsz_&sol;;
                  %end;
   ;
  print szSZ_&sol &fuy; free szsz&sol szsz_&sol;
  print sxSZ_&sol &lig; free  sxsz&sol sxsz_&sol;
   ass=repeat(" asZ&sol",1,qq);
  colvs=colvs | |ass;  free ass;
  PCT=PCT//(Pctz&sol[1:qq]);
  VSing=vSing//(vssz&sol[1:qq]);
  k=nrow(vsing);
  Valeurs=vsing[2:k,];
  Pourcent=pct[2:k,];
  colvs2=colvs[,2:k];
  pto=pourcent[+];
    print " ------------------------------------------ ";
    print " Décomposition après la solution &sol      ";
    print " reconstruction de " pto   " %          " ;
    print " ------------------------------------------ ";
    print   valeurs[rowname=colvs2 ]  "  " Pourcent
[format=z6.3] " % " ;
    print " ------------------------------------------ ";



 %mend apsolu;
/*************************************************/
/********/

%macro ppo_x2(x=X,stok=N,sx=sx,sy=sy,sz=sz);
/*************************************************************
/
/*projection sur x orto(o/n) et y orto(o/n) et z
orto(o/n)*/
/*************************************************************
/
   %if %upcase(&ex)  ne X   %then load &ex ;;
   %if %upcase(&sx) ne SX %then load &sx ;;
   %if %upcase(&sy) ne SY %then load &sy ;;
   %if %upcase(&sz) ne SZ %then load &sz ;;

        %let b=&diago;
        %let diago= ;
   %if &b=O %then %do; %let qm=#(Q`)*;%let
dm=#(D`)*;%let pm=#(P`)*;%end;;;
   %if &b=N %then %do; %let qm=*Q*;%let
dm=*D*;%let pm=*P*;%end;;;

   %if &b= %then %do; %let qm= ; %let dm= ; %let
pm= ; %end;;;

 %configu(x=&ex);

        %if %upcase(&sy) ne %then %do;

vX2=Xv&qm&sy*&sy`;

     tv2=vX2[1:&n,1];
    %do tt=2 %to &q;
     tv2=tv2//vX2[1:&n,&tt];
    %end;
    %do t=1 %to (&p-1);
      %do tt=1 %to &q;
 tv2=tv2//vX2[(&t*&n+1):((&t+1)*&n),&tt] ;
      %end;
    %end;

 nwx=&ex-tv2;

 %configu(X=nwx);
                  %end;
        %if %upcase(&sx) ne %then %do;
wx2=Xw&dm&sx*&sx`;
    tw2=(wX2[1,])`;
   %do t=2 %to (&p*&q);
    tw2=tW2//(wX2[&t,])`;
   %end;
   %if %upcase(&sy) ne %then
                nwx=nwx-tw2;
               %else
                nwx=&ex -tw2;

 %configu(X=nwx);
                  %end;
        %if %upcase(&sz) ne %then %do;

uX2=Xu&pm&sz*&sz`;
    tu2=uX2[,1];
    %do t=2 %to &p;
    tu2=tu2//uX2[,&t];
    %end;
   %if (&sy= and &sx= ) %then nwx=&x-tu2;
                %else nwx=nwx -tu2;;
                  %end;

 %if %upcase(&stok)=O %then %do;
xa&sol=nwx;store xa&sol
                  ;free /x tot Nlig Ncol
                       Nfuy vsing pct colvs &met; ;
                   %end;
              %else free /x nwx tot Nlig Ncol Nfuy
vsing pct
                       colvs &met;
          %let diago=&b;
 %mend ppo_x2;
/*************************************************************/
/**********/
    %macro tot;
 %if %upcase(&diago)=    %then
    tot=x`*x;;;

 %if %upcase(&diago)=O %then
```

```
      %do;
       mm=(P@Q@D);
       tot=x`*(mm#x);
      %end;

 %if %upcase(&diago)=N %then
      %do;
       mm=(P@Q@D);     /* npq<63 */
       tot=x`*mm*x;
      %end;;
 print "-----------------------------------------------------------
-----";
 print "------Procédure SVD3COMP réalisée par
D.Leibovici nov1992------";
 print " Analyse par Décomposition en Valeurs
Singulières avec Métriques";
 print "              d'un Tableau à 3 entrées
";
 print "-----------------------------------------------------------
-----";
 print ,"    ", &tit;
  print "-----Inertie totale-----" ,tot;
 %mend tot;


          /*******FIN des MACROS*************/
/*************************************************************/
/*     macro configu(x=X)                                    */
/*  x tenseur: configuration des tableaux                    */
/*************************************************************/
/*     macro inituck1(imp=O)                                 */
/*  solution initiale Tucker1                                */
/* imprémssion si O numero de la solution dans sol
*/
/*************************************************************/
/*     macro solucc(imp=O,stok=N)                            */
/*  calcul de la solution par doubles  contractions
*/
/*              alternées                                    */
/* imprémssion si O , stockage si O, numéro dans sol
*/
/*************************************************************/
/*    macro ppo_x2(x=X,stok=N,sx=sx,sy=sy,sz=sz)
*/
/*projection sur sx orto(O/N) sy orto(O/N) sz
orto(O/N)*/
/*************************************************************/

/*****************************************************************
********/
/**********************     PROGRAM
**********************/
/*****************************************************************
********/
      /**** préparations des données ****/
 edit es ,read all var[x] into X;

 %let x=x;
 %let n=11;
 %let q=4;
 %let p=4;
  %let tit="Données aléatoires sur tableaux ...";

  Nlig={"mr-", "mr+","mri","mrs","mo-
","mo+","moi","fr-","fr+","fri",
        "foi"} ;
  Ncol={"J", "R","S","T"};
  Nfuy={"0", "6","12","18"};
  D={2,4,5,6,5,1,4,1,2,8,3};
  Q={1,2,3,4};
  P={1,2,3,4};

  %let diago=O;    /* O (vecteurs) ou  (pas de
métriques) ou N (matrices)*/
  %let met=D P Q;

  %let lig=[rowname=Nlig] ;
  %let col=[rowname=Ncol];
  %let fuy=[rowname=Nfuy];

          /***** prog************************/

          %tot;

          %let sol=111;
%configu(X=x);
%inituck1(imp=o);
%solucc(x=X,imp=o,stok=o);
          %apsolu;

          %let sol=222;
%ppo_X2(X=x,stok=o,sx=sx111,sy=sy111,sz=sz111);
%configu(X=Xa222);
%inituck1(imp=o);
%solucc(x=Xa222,imp=o,stok=o);
          %apsolu;

          %let sol=333;
%ppo_X2(X=Xa222,stok=o,sx=sx222,sy=sy222,sz=sz
222);

%configu(X=Xa333);
%inituck1(imp=o);
%solucc(x=Xa333,imp=o,stok=o);
          %apsolu;
          %let sol=444;
%ppo_X2(X=Xa333,stok=o,sx=sx333,sy=sy333,sz=sz
333);

%configu(X=Xa444);
%inituck1(imp=o);
%solucc(x=Xa444,imp=o,stok=o);


  remove _ALL_;
/*****************************************************************
********/
/**********************  FIN  de PROGRAM
**********************/
/*****************************************************************
********/
```


II.2 ATP-kmodes (solutions k-modes)

```
/* Procédure svdkm     ou DVSk-modes avec
métriques*/
  /* Singular Value Decomposition of k modes by
Contractions of the Tensor*/
  /* Didier Leibovici nov 1992*/

/*************************************************************
********/
  /* Realisation de l'acpkmode compromis de
variance et rang       */
  /*  identique à la recherche de valeurs singuliäres
successives   */
  /*     i.e relations de transitions  dans les solutions
*/

/*************************************************************
********/
  data es;
   do i=1 to 3000;
  x=ranuni(88199562)*60;
  nom="a";
   output;
  end;
     run;

  options linesize=80 pagesize=500  ;
  proc iml worksize=250;

  remove _all_;

  reset  nolog;


/*************************************************************
********/

/**********************MACROS**********************/

  %macro reord(A=X,l=1,tl=);
  /*************************************************************/
       %let clas= ;
       %let ind= ;
  %let fa=%eval(&k+1);
    vd=repeat(0,1,&fa);
    create vs from vd;
    close vs;
  %do y=1 %to &k;
        %let v=1 ;
        %if &y ne &l %then %let clas=col&y &clas  ;
     %if &y ne &tl %then %do;

    %do z=&k %to 1 %by -1;
          %if &z ne &tl %then %do;
            %if &y ne &z  %then
                %let v=&v @ j(&&n&tl,1,1);
                    %else
                %let v=&v @ (do(1,&&n&z,1))` ;
                      %end;
    %end;
          %if &tl ne %then col&y=&v @ j(&&n&tl,1,1);
              %else col&y=&v;
                      %end;
                  %else
                  col&tl=j((&nrowx/&&n&tl),1,1) @
(do(1,&&n&tl,1))`;;
       %let ind=&ind col&y;
   %end;

   col&fa=&A;

   %let vind=&ind col&fa;
   edit vs;
   append var{&vind};
  close vs;
   free &ind vd v;

  sort vs by &&clas;       /* classement en k @.
1+1@l-1 .. @1 (@l) */

  edit vs;
  read all var{col&fa} into o&l.&A;
  close vs;

  %mend reord;
  /*************************************************************/
  /***********/

  %macro configu(x=X,tot=O,s=);
  /*************************************************************/
  /*    configuration en tableaux  n-i,ni          */
  /*************************************************************/
     %if %upcase(&x) ne X  %then load &x ;
                ;
  %if %upcase(&diago)=O %then %do;
Xvr=&x;&x=&x#(&sqrtPQD);%end;;;
   ;
   %if %upcase(&diago)=N %then %do;
Xvr=&x;&x=(&rootPQD)*&x;%end;;;
   ;

     %if %upcase(&tot)=O %then %do;
       %do ttt=1 %to &k;
        %let fin=&nrowx/&&n&ttt -1;

        %if &ttt ne 1 %then
       %reord(A=&x,l=&ttt,tl=);
                %else o1&x=&x;
                ;
     t&X&ttt=o&ttt.&x[1:&&n&ttt];
         %do tt=1 %to &fin ;

t&X&ttt=t&X&ttt| |o&ttt.&x[(1+&ttt*&&n&ttt):(&tt*
&&n&ttt +&&n&ttt)];
         %end;
       X&ttt=(t&X&ttt)`;  free t&X&ttt o&ttt.&x;
        %end;


                        %end;
                    %else
                        %do;
     %let  fin=&nrowx/&&n&es -1;
     %reord(A=&x,l=&es,tl=);
     t&X&s=o&s.&x[1:&&n&es];
```



```
        %do tt=1 %to &fin ;

t&X&s=t&X&s| |o&s.&x[(1+&tt*&&n&es):(&tt*&&n
&es +&&n&es)];
         %end;
      X&s=(t&X&s)`; free t&X&s o&s.&x;

                        %end;
    %mend configu;
  /*************************************************************/
  /********/

   %macro inituck1(imp=O);
  /*************************************************************/
  /*solution initiale Tucker1                        */
  /*************************************************************/


    %do ttt=1 %to &k ;

  call svd(sl&ttt,l&ttt,sc&ttt,X&ttt);
  s&ttt=sc&ttt[,1];
      free  sc&ttt sl&ttt;

  %if %upcase(&imp)=O %then %do;
  print "     ";
  print  " ----------------initialisation tucker1 ------------
---" ;
  print "Valeurs singuliäres &sol de &X&ttt solution
initiale &sol";
   print " ------------------------------------------------------- ";

  print l&ttt s&ttt &&col&ttt;
        free  l&ttt;
                %end;
    %end;

   %mend inituck1;
  /*************************************************************/
  /********/

   %macro solucck(X=x,imp=O,stok=N);
  /*************************************************************/
  /*recherche de la première solution par
contractions  */
  /*succéssives                                      */
  /*************************************************************/

  iter=0;
    %macro calc(r);
       %let cc=1;
     %do z=&k %to 1 %by -1;
       %if &r ne &z %then
        %let cc=(&cc @ s&z);
     %end;
     s&r=(X&r)`*&cc;

   %mend calc;

         %if %upcase(&diago) ne %then &x=xvr;;;


      /*** debut boucle  ***/

  do until (test<0.0000000001);

    iter=iter+1;
   %do ttt=1 %to &k;
     s0&ttt=s&ttt;
   %end;

     %do r=1 %to  &k;
       %calc(&r);;
      %if &r=&k %then vs&sol=sqrt(s&r`*s&r);;
       s&r=s&r/sqrt(s&r`*s&r);
     %end;

    %let tes=(s0&k-s&k)`*(s0&k-s&k);

    %do r=(&k-1) %to  1 %by -1;
     %let tes=&tes + (s0&r-s&r)`*(s0&r-s&r);
    %end;

    test=&tes;
  print " ------------------------------------------- ";
  print iter "     " test "     " vs&sol;
  print " ------------------------------------------- ";
    if iter>200 then test=0.000000000001;

   end;
    /***fin boucle*****/
    %if %upcase(&diago)=O %then  %do;
                    %do r=1 %to &k;
                      s&r=(sqrt(D&r)##(-1))#s&r;
                    %end;
                      %end;
       ;;
    %if %upcase(&diago)=N %then  %do;
                    %do r=1 %to &k;
                      s&r=ginv( root(D&r))*s&r;;
                    %end;
                      %end;
    ;

     %let est=s&k ; free s0&k;
    %do r=(&k-1) %to  1 %by -1;
     %let est=&est @ s&r;
      free s0&r;
     %end;

     estX=&est*vs&sol;

    %if %upcase(&diago)=O %then %let pm=(&mm)
#;;
    %if %upcase(&diago)=N %then %let pm=(&mm)
*;;
    %if %upcase(&diago)=  %then %let pm= ;;
              ;
     resx=(&x-estX)`*(&pm(&x-estX));
      xxx=&x`*(&pm &x);
     xxes=estx`*(&pm estx); free estx ;

   pctot=(xxes/tot)*100;
   pctes=(xxes/xxx)*100;
   print " ------------------------------------------------------
------ ";
      print " +   SVD-&k.modes-   solution &sol     + ";
      print " ------------------------------------------- ";
      print " +   totale  +   expliquée   + résiduelle + ";
      print " ------------------------------------------- ";
```



```
  print     xxx "   " xxes "   "resx ;
  print    " % expliquC      % tenseur initial
reconstruit ";
  print pctes[ format=z4.2] "            "pctot [
format=z4.2];
  print " ------------------------------------------------------
------ ";
              ;
        %let sso=1;
      %do r=2 %to &k;
        %let sso=1&sso ;
      %end;
    %if &sol=&sso %then %do; colvs={"
"};vsing=0;pct=0; %end;;

   colvs=colvs| |{"vs&sol"};
   VSing=vSing//vs&sol;
   PCT=PCT//PCTOT;
   wk=nrow(vsing);
   Valeurs=vsing[2:wk,];
   Pourcent=pct[2:wk,];
   colvs2=colvs[,2:wk];

     print " ------------------------------------------- ";
     print   valeurs[rowname=colvs2 ]  "  " Pourcent
[format=z6.3] " % " ;
     print " ------------------------------------------- ";

   %if %upcase(&imp)=O %then  %do;
          %do ttt=1 %to &k;
    print ," Solution &sol" s&ttt &&col&ttt ,;
            %end;          ;
                %end;
    %if %upcase(&stoK)=O %then %do;
                %do ttt=1 %to &k;
        s&ttt._&sol= s&ttt ;
        store s&ttt._&sol ;  free  s&ttt._&sol;
                %end;
                    %end;
   %mend solucck;

  /*************************************************************/
   /**********/

   %macro ppo_x2(B=X,stok=N);

  /*************************************************************
  /
    /*projection sur s1 orto et s2 orto et ..sk orto
  */

  /*************************************************************
  /
       %if%upcase(&B) ne X %then load &B ;;

   %macro rx(A=x,ttl=1);
     rx=(&A[1,])`;
    %do t=2 %to (&nrowx/&&n&ttl);
      rx=rx//(&A[&t,])`;
    %end;
     %reord(A=rx,l=,tl=&ttl);
      free rx;
   %mend rx;

          %let bo=&diago;
          %let diago= ; ;
    %if &bo=O %then %do;
                %do r=1 %to &k;
                %let Dm&r=#(D&r`);;
                %end;
              ;%end;;;
    %if &bo=N %then %do;
                %do r=1 %to &k;
                %let Dm&r=*D&r;;
                %end;
              ;%end;;
    %if &bo=  %then %do;
                %do r=1 %to &k;
                %let Dm&r= ;;
                %end;
              ;%end;;;


          %do ttt=1 %to &k;

     %configu(x=&B,tot=N,s=&ttt);
   deb=X&ttt &&Dm&ttt*s&ttt*s&ttt` ;
        %rx(A=deb,ttl=&ttt);
   d=&B-orx; free orx s&ttt;
    %let B=d;

          %end;

    %if %upcase(&stok)=O %then %do;
xa&sol=&B;store xa&sol
                    ;free /x tot &col &met
                      vsing pct colvs; ;
                    %end;
                %else free /x d tot &col &met vsing
pct
                      colvs ;
   ;
    %let diago=&bo;
    %mend ppo_x2;

  /*************************************************************/
   /**********/

    %macro tot;

  /*************************************************************/
    %if %upcase(&diago)=     %then
      tot=x`*x;;;

    %if %upcase(&diago)=O %then
        %do;
          %global sqrtPQD mm;
          %let mm=1 ;
          %let sqrtPQD=1;;
        %do r=&k %to 1 %by -1;
          %let  mm=&mm @ D&r;
          %let sqrtPQD=&sqrtPQD @ sqrt(D&r);
        %end;;
        tot=x`*((&mm) #x);
        %end; ;

    %if %upcase(&diago)=N %then
        %do;          ;
          %global rootPQD mm;;
```

```
        %let mm=1;
        %let rootPQD=1;
       %do r=&k %to 1 %by -1;
        %let rootPQD=&rootPQD @root(D&r);
        %let mm=&mm @ D&r;     /* utilisable sur
PC si n1n2...nk<63  */
        %end;;
       tot=x`*(&mm)*x;;
        %end;;
  print "-------------------------------------------------------------
------";
  print "------Procédure DVS -&k. modes réalisée par
D.Leibovici nov1992------";
  print " Analyse par Décomposition en Valeurs
Singulières avec Métriques";
  print "              d'un Tableau à k entrées
";
  print "-------------------------------------------------------------
------";
  print ,"   ", &tit;;
  print "--Inertie totale-- =",tot;

  %mend tot;

/*************************************************************
 /**********/


        /******FIN des MACROS************/
 /*************************************************************/
 /*    macro configu(x=X, tot=O,s=,nrowx=)
*/
 /* x tenseur: configuration des tableaux            */
 /*************************************************************/
 /*     macro inituck1(imp=O)                      */
 /* solution initiale Tucker1                     */
 /* imprÇssion si O numero de la solution dans sol
*/
 /*************************************************************/
 /*     macro solucck(X=x,imp=O,stok=N)
*/
 /* calcul de la solution par   contractions          */
 /*            alternÇes                              */
 /* imprÇssion si O , stockage si O, numÇro dans
sol   */
 /*************************************************************/
 /*    macro ppo_x2(x=X,stok=N)                    */
 /*projection sur s1 orto et ... sk orto             */
 /*************************************************************/
 /*    macro tot                                  */
 /* calcul de l'inertie totale et prÇparation          */
 /*              de la mÇtrique tensorielle          */


/*************************************************************
********/
 /**********************     PROGRAMME
************************/

/*************************************************************
********/
          /*******prep   **********/
  edit es ;read all var{x} into X;

  %let x=x;
  %let k=5;
  %let nrowx=3000;
%let n1=50;
%let n2=3;
%let n3=2;
%let n4=5;
%let n5=2;
 edit es read all avr{nom} into Ncol1;
  Ncol2={"J", "R","S"};
  Ncol3={"H", "F"};
  Ncol4{"0", "6","12","18","24"};
  Ncol5={"S", "NS"};

    %let col=Ncol1 Ncol2 Ncol3 Ncol4 Ncol5;
%let col1={rowname=Ncol1} ;
%let col2={rowname=Ncol2};
%let col3={rowname=Ncol3};
%let col4={rowname=Ncol4};
%let col5={rowname=Ncol5};

  D1=repeat(1, 50,1);
  D2={3,2,1};
  D3={4,2};
  D4=do(1,5,1)`;
  D5={0.5,0.5};

  %let met=D1 D2 D3 D4 D5;

  %let diago=O;  /* (O) vecteurs ou() pas de
mÇtriques ou (N) matrices*/

  %let tit= "Essai sur donnÇes alÇatoires";
          /******Prog***** SOLUTIONS k-
modes*****/
    %tot;

        %let sol=11111;
%configu(X=x,tot=O,s=);
%inituck1(imp=o);
%solucck(x=X,imp=o,stok=o);

   %ppo_x2(B=x,stok=O);
        %let sol=22222;
%configu(X=xa111,tot=O,s=);
%inituck1(imp=o);
%solucck(x=Xa111,imp=o,stok=o);

   %ppo_x2(B=xa11111,stok=O);
        %let sol=33333;
%configu(X=xa22222,tot=O,s=);
%inituck1(imp=o);
%solucck(x=Xa11111,imp=o,stok=o);


/*************************************************************
********/
 /******************* FIN de   PROGRAMME
************************/

/*************************************************************
********/
```


II.3 AFC-kmodes par ATP-kmodes

```
/* Procédure DVS-kmodes avec métriques/
/* Singular Value Décomposition of k modes by
Contractions of the Tensor */
/* Didier Leibovici nov 1992*/
/*************************************************************
********/
/*                    ATP-kmodes
/
/* Réalisation de l'acpkmode compromis de
variance et rang      */
/* identique à la recherche de valeurs singulières
successives  */
/*    i.e relations de transitions  dans les solutions
*/
/*************************************************************
********/
/*************************************************************
********/
 data es;
 input x @@;
cards; /*]     R      S      T */
100   11    19    36    /* mr-   Un   0*/
107   25    28    28    /* mr+   Un   0*/
 99    9    18    18    /* mri   Un   0*/
 17    0     3     3    /* mrs   Un   0*/
  7    0     0     2    /* mo-   Un   0*/
  5    3     2     3    /* mo+   Un   0*/
 37    3     7     7    /* moi   Un   0*/
 87    5    10    34    /* fr-   Un   0*/
 39    5     5     7    /* fr+   Un   0*/
 35    3     5     8    /* fri   Un   0*/
 16    1     1     4    /* foi   Un   0*/
  7    3     4     4    /* mr-   Qq   0*/
  4    0     2     1    /* mr+   Qq   0*/
 27   10    11    16    /* mri   Qq   0*/
  0    0     0     0    /* mrs   Qq   0*/
  5    4     1     0    /* mo-   Qq   0*/
  1    4     1     1    /* mo+   Qq   0*/
 35   15    27    31    /* moi   Qq   0*/
  6    1     1     0    /* fr-   Qq   0*/
  1    0     0     0    /* fr+   Qq   0*/
  7    2     0     2    /* fri   Qq   0*/
 28    2     5     9    /* foi   Qq   0*/
  0    0     1     0    /* mr-   Be   0*/
  0    0     0     0    /* mr+   Be   0*/
  2    3     5     4    /* mri   Be   0*/
  0    1     0     0    /* mrs   Be   0*/
  1    1     1     1    /* mo-   Be   0*/
  0    0     0     0    /* mo+   Be   0*/
 38   27    44    30    /* moi   Be   0*/
  0    0     0     0    /* fr-   Be   0*/
  0    0     0     0    /* fr+   Be   0*/
  0    0     0     0    /* fri   Be   0*/
  7    5     0     2    /* foi   Be   0*/
      16          5          16         148
      22         19          17          98
      11          6           6          56
       1          2           0           3
       0          0           0           7
       1          0           1           4
       0          0           1           6
       5          2           3          20
       2          1           3          28
       1          0           0           0
       1          1           0           3
       0          0           0           2
       3          2           9          70
       1          1           3          22
       4          1           0          10
       1          0           0           4
       4          0           1          15
       0          0           0           1
       0          0           0           0
       0          0           1           0
       0          0           0           1
       0          0           0           0
       1          0          11          47
       0          0           0           1
       0          0           0           0
       0          0           0           2
       0          0           0           2
      20          7          13         160
      19         15          10          84
      14          5           8          39
       0          1           0           2
       1          0           0           8
       3          0           1           0
       4          0           0          17
       2          3           0          81
       4          2           2          18
       1          0           0          14
       0          0           0           8
       0          0           2           5
       8          1           5          33
       4          1           1          29
       0          0           0           0
       0          1           0           6
       0          0           0           2
       3          2           6          59
       1          1           1          23
       4          0           1          15
       1          1           0          10
       1          0           3          13
       0          0           0           1
       0          0           0           0
       0          0           0           4
       0          0           0           0
       0          0           0           1
       0          0           0           0
       3          1           9          51
       0          0           0           0
       0          0           0           0
       0          0           0           1
       0          0           0           2
      19          9          16         136
      19          8           9          75
      12          3           8          40
       0          0           1           3
       1          0           0           3
       1          0           1           3
       3          0           0          21
       2          0           8          74
       6          1           3          17
       0          0           1          11
       0          1           1           9
       0          1           1           8
       6          0           9          36
       2          0           2          27
       0          0           1           1
       0          0           0           2
       0          0           0           2
       4          2           3          44
       2          1           2          29
       5          0           0          12
       0          1           0           7
       0          0           0          12
```



```
        0         0         0         0
        0         0         0         0
        0         0         0         1
        0         0         0         0
        0         0         0         1
        0         0         1         0
        0         1         6        50
        0         0         0         0
        0         0         0         0
        0         0         0         0
        0         0         1         4
;
    r;

 options linesize=80 pagesize=500  ;
proc iml worksize=222;

remove _all_;

reset  nolog;

/*************************************************************
********/

/*********************MACROS**********************/

 %macro reord(A=X,l=1,tl=);
/*************************************************************/
    %let clas= ;
    %let ind= ;
 %let fa=%eval(&k+1);
  vd=repeat(0,1,&fa);
  create vs from vd;
  close vs;
 %do y=1 %to &k;
     %let v=1 ;
    %if &y ne &l %then %let clas=col&y &clas  ;
   %if &y ne &tl %then %do;

   %do z=&k %to 1 %by -1;
       %if &z ne &tl %then %do;
         %if &y ne &z  %then
             %let v=&v @ j(&&n&z,1,1);
                 %else
             %let v=&v @ (do(1,&&n&z,1))`;
                   %end;
   %end;
       %if &tl ne %then col&y=&v @ j(&&n&tl,1,1);
                %else col&y=&v;

                   %end;
              %else
              col&tl=j(((&nrowx/&&n&tl),1,1) @
(do(1,&&n&tl,1))`;;
    %let ind=&ind col&y;
 %end;

 col&fa=&A;

 %let vind=&ind col&fa;
 edit vs;
 append var{&vind};
close vs;
 free &ind vd v;

sort vs by &&clas;       /* classement en k @. l+1@l-
1 .. @1 (@l) */

edit vs;
read all var{col&fa} into o&l.&A;
close vs;

 %mend reord;
/*************************************************************/
/************/

 %macro configu(x=X,tot=O,s=);
/*************************************************************/
/*    configuration en tableaux  n-ni,ni          */
/*************************************************************/
   %if %upcase(&x) ne X  %then load &x ;
         ;
 %if %upcase(&diago)=O %then %do;
Xvr=&x;&x=&x#(&sqrtPQD);%end;;;
;
 %if %upcase(&diago)=N %then %do;
Xvr=&x;&x=(&rootPQD)*&x;%end;;;
;

  %if %upcase(&tot)=O %then %do;
    %do ttt=1 %to &k;
     %let fin=&nrowx/&&n&ttt -1;

     %if &ttt ne 1 %then
   %reord(A=&x,l=&ttt,tl=);
            %else o1&x=&x;
              ;
   t&X&ttt=o&ttt.&x[1:&&n&ttt];
        %do tt=1 %to &fin ;

t&X&ttt=t&X&ttt | |o&ttt.&x[(1+&tt*&&n&ttt):(&tt*
&&n&ttt +&&n&ttt)];
          %end;
    X&ttt=(t&X&ttt)`;  free t&X&ttt o&ttt.&x;
    %end;

                  %end;
               %else
                  %do;
   %let  fin=&nrowx/&&n&s -1;
   %reord(A=&x,l=&s,tl=);
   t&X&s=o&s.&x[1:&&n&s];
        %do tt=1 %to &fin ;

t&X&s=t&X&s | |o&s.&x[(1+&tt*&&n&s):(&tt*&&n
&s +&&n&s)];
          %end;
   X&s=(t&X&s)`; free t&X&s o&s.&x;

                  %end;
  %mend configu;
/*************************************************************/
/********/

 %macro inituck1(imp=N);
/*************************************************************/
/*solution initiale Tucker1                    */
/*************************************************************/


  %do ttt=1 %to &k ;
```



```
call svd(sl&ttt,l&ttt,sc&ttt,X&ttt);
s&ttt=sc&ttt[,1];
    free  sc&ttt sl&ttt;

print  " ------ initialisation tucker1   &sol --------" ;
%if %upcase(&imp)=O %then %do;
print "     ";
print "Valeurs singulières &sol de &X&ttt solution
initiale &sol";
 print " ------------------------------------------------------- ";

print l&ttt s&ttt &&col&ttt;
     free  l&ttt;
                %end;
  %end;

 %mend inituck1;
/*************************************************************/
/********/

 %macro solucck(X=x,imp=O,stok=N);
/*************************************************************/
/*recherche de la première solution par contractions
*/
/*successives       algorithme  RPVSCC
*/
/*************************************************************/

iter=0;
   %macro calc(r);
      %let cc=1;
    %do z=&k %to 1 %by -1;
      %if &r ne &z  %then
        %let cc=(&cc @ s&z);
    %end;
    s&r=(X&r)`*&cc;

   %mend calc;;

        %if %upcase(&diago) ne %then &x=xvr;;;


     /*** debut boucle  ***/

 do til (test<0.0000000001);

   iter=iter+1;
   %do ttt=1 %to &k;
     s0&ttt=s&ttt;
   %end;

     %do r=1 %to  &k;
       %calc(&r);;
     %if &r=&k  %then vs&sol=sqrt(s&r`*s&r);;
       s&r=s&r/sqrt(s&r`*s&r);
     %end;

     %let tes=(s0&k-s&k)`*(s0&k-s&k);

    %do r=(&k-1) %to  1 %by -1;
     %let tes=&tes + (s0&r-s&r)`*(s0&r-s&r);
    %end;

    test=&tes;
print " ------------------------------------- ";

print iter "     " test "     " vs&sol;
print " ------------------------------------- ";
  if iter>200 then test=0.000000000001;

  end;
   /***fin boucle*****/
   %if %upcase(&diago)=O %then  %do;
                %do r=1 %to &k;
                  s&r=(sqrt(&&D&r)##(-1))#s&r;
                %end;
                  %end;
     ;;
  %if %upcase(&diago)=N %then  %do;
                %do r=1 %to &k;
                  s&r=ginv( root(&&D&r))*s&r;;
                %end;
                  %end;
  ;

    %let est=s&k ; free s0&k;
    %do r=(&k-1) %to  1 %by -1;
    %let est=&est @ s&r;
     free s0&r;
    %end;

    estX=&est*vs&sol;

   %if %upcase(&diago)=O %then %let pm=(&mm)
#;;
   %if %upcase(&diago)=N %then %let pm=(&mm)
*;;
   %if %upcase(&diago)=  %then %let pm= ;;
         ;
    ress=(&x-estX)`*(&pm(&x-estX));
     xxx=&x`*(&pm &x);
    xxes=estx`*(&pm estx); free estx ;

 pctotx=(xxes/totx)*100;
 pctot=(xxes/tot)*100;
 pctes=(xxes/xxx)*100;
 print " ------------------------------------------------------
---- ";
   print " +  SVD-&k.modes-CC    solution &sol     +
";
   print " ------------------------------------ ";
   print " +   totale  +   expliquée   + résiduelle + ";
   print " ------------------------------------ ";
print     xxx "    " xxes "    "resx ;
print " % expliqué      % tenseur analysé reconstruit
% tenseur initial";
print pctes[ format=z4.2] "    " pctot [ format=z4.2] "
" pctotx [ ormat=z4.2];
print " ------------------------------------------------------
---- ";
         ;
  if totx=tot then  do;
     %let sso=1;
    %do r=2 %to &kx;
     %let sso=1&sso ;
    %end;
  %if &sol=&sso %then %do; colvs={"
"},vsing=0;pct={0 0}; pctx={0 0};%end;;
              end;
         ;
```

```
colvs=colvs||"&cont.vs&sol";
VSing=vSing//vs&sol;
temp=J0 0);
temp[,1]=PCTOI';
temp[,2]=PCTOI'x;
%if (1=&nk)  and (&k ne &kx) %then temp=[. .];;

PCT=PCT//temp;
wk=nrow(vsing);
Valeurs=vsing[2:wk,];
Pourcent=pct[2:wk,];
  do a=1 to (wk-2);
  if valeurs[a]=valeurs[(wk-1)] then pourcent[(wk-
1),]=[. .];
  end;
colvs2=colvs[,2:wk];

  print " ------------------------------------------ ";
  print  valeurs[rowname=colvs2 ]  "  " Pourcent
[format=z6.3] " % " ;
  print " ------------------------------------------ ";

%if %upcase(&imp)=O %then  %do;
       %do ttt=1 %to &k;
print ," Solution &sol" s&ttt &&col&ttt ,;
       %end;              ;
                  %end;
%if %upcase(&stoK)=O %then %do;
              %do ttt=1 %to &k;
     s&ttt._&sol= s&ttt ;
     store s&ttt._&sol ;  free  s&ttt._&sol;
              %end;
                  %end;
%mend solucck;
/*************************************************************/
/**********/

%macro apsolu3;
/*************************************************/
/* solution deux modes associées aux solutions 3
modes &k=3*/
   %do a=1 %to 3;

 free cox coy;

   Zd=X&a#(sqrt(&&D&a)`)*s&a;

%if &a=1 %then %do;%let ly=3;
              %let lx=2;
            %end;
%if &a=2 %then %do;%let ly=3;
              %let lx=1;
            %end;

%if &a=3 %then %do;%let ly=2;
              %let lx=1;
            %end;

     xZy=Zd[1:&&n&lx];
     %do t=1 %to (&&n&ly-1);

xZy=xZy||Zd[(&t*&&n&lx+1):((&t+1)*&&n&lx)] ;
     %end;

      call svd(cox,vs&a._&sol,coy,xZy);
  print " ------------------------------------------ ";
  print " solutions associées Ö s&a._x&sol  sous
&cont";
  print " ------------------------------------------ ";
  vp&a._&sol=vs&a._&sol##2;

  qq=min(nrow(xZy),ncol(xZy))  ;
   som= (vp&a._&sol[2:qq])[+];
   pcts&a._&sol= {0 0};
   pcts&a._&sol[,2]=(som/totx)*100;
   pcts&a._&sol[,1]=(som/tot)*100;
 pct&a._&sol=repeat(0 ,nrow(vp&a._&sol),2);
 pct&a._&sol[,1]=(vp&a._&sol/tot)*100;
 pct&a._&sol[,2]=(vp&a._&sol/totx)*100;
 pct&a._&sol[1,]=[. .];
  if som<>0 then do;
 Pc=(vp&a._&sol/som)*100;
 Pc[1]=.;
                end;

 print "% tenseur initial reconstruit en plus =
"Pcts&a._&sol;
print "Va singuliäres "vs&a._&sol"Va propres
"vp&a._&sol" " Pc[format=z6.3] %";
    sx_&a._&sol=cox[,1:qq];
    sy_&a._&sol=coy[,1:qq];
                sx_&a._&sol=(sqrt(&&D&lx)##(-
1))#sx_&a._&sol;
                sy_&a._&sol=(sqrt(&&D&ly)##(-
1))#sy_&a._&sol;;
 ;
 print ,sx_&a._&sol &&col&lx ,;
 print ,sy_&a._&sol &&col&ly ,;

  ass=repeat(" as_&a._&sol",1,qq);
 colvs=colvs||ass; free ass;
 PCT=PCT//(Pct&a._&sol[1:qq,]);
 VSing=vSing//(vs&a._&sol[1:qq]);

wk=nrow(vsing);
 Valeurs=vsing[2:wk,];
 Pourcent=pct[2:wk,];

   do a=1 to (wk-2);
   if valeurs[a]=valeurs[(wk-1)] then pourcent[(wk-
1),]=[. .];
   end;

 colvs2=colvs[,2:wk];
 pto=pourcent[+,];
 colvs2=colvs[,2:wk];
 pto=pourcent[+,];
  print " ------------------------------------------ ";
  print " Décomposition aprés la solution &sol      ";
  print " reconstruction de " pto    " %          " ;
  print " ------------------------------------------ ";
  print   valeurs[rowname=colvs2 ]  "  " Pourcent
[format=z6.3] " % " ;
  print " ------------------------------------------ ";

 %end;

%mend apsolu3;
/*************************************************/
```



```
/********/

 %macro ppo_x2(B=X,stok=N);
/*************************************************************
/
/*projection sur s1 orto et s2 orto et ..sk orto        */
/*************************************************************
/
  %if %upcase(&B) ne X %then load &B ;;

%macro rx(A=x,ttl=1);
  rx=(&A[1,])`;
 %do t=2 %to (&nrowx/&&n&ttl);
   rx=rx//(&A[&t,])`;
 %end;
  %reord(A=rx,l=,tl=&ttl);
   free rx;
%mend rx;

       %let bo=&diago;
       %let diago= ; ;
  %if &bo=O %then %do;
                %do r=1 %to &k;
                %let Dm&r=#(&&D&r`);;
                %end;
               ;%end;;;
  %if &bo=N %then %do;
                %do r=1 %to &k;
                %let Dm&r=*&&D&r;;
                %end;
               ;%end;;;
  %if &bo=  %then %do;
                %do r=1 %to &k;
                %let Dm&r= ;;
                %end;
               ;%end;;;



       %do ttt=1 %to &k;

   %configu(x=&B,tot=N,s=&ttt);
deb=X&ttt &&Dm&ttt*s&ttt*s&ttt`;
     %rx(A=deb,ttl=&ttt);
d=&B-rx; free orx s&ttt;
  %let B=d;

       %end;

 %if %upcase(&stok)=O %then %do;
xa&sol=&B;store xa&sol
                    ;free /x totx tot &cl  &metri
                         vsing pct colvs; ;
                   %end;
              %else free /x &B tot totx &cl &metri
vsing pct
                    colvs ;
 ;
  %let diago=&bo;
 %mend ppo_x2;
/*************************************************************/
/**********/

%macro tot(x=x);
/*************************************************************/
 load &x;

 %if %upcase(&diago)=    %then

   tot=&x`*&x;;;

 %if %upcase(&diago)=O %then
     %do;
      %global sqrtPQD mm;
      %let mm=1 ;
      %let sqrtPQD=1;;
      %do r=&k %to 1 %by -1;
       %let  mm=&mm @ &&D&r;
       %let sqrtPQD=&sqrtPQD @ sqrt(&&D&r);
      %end;;
      tot=&x`*((&mm) #&x);
     %end; ;

 %if %upcase(&diago)=N %then
     %do;         ;
      %global rootPQD mm;;
      %let mm=1;
      %let rootPQD=1;
      %do r=&k %to 1 %by -1;
       %let rootPQD=&rootPQD @root(&&D&r);
       %let mm=&mm @ &&D&r;     /* utilisable sur
PC si n1n2...nk<63  */
      %end;;
      tot=&x`*(&mm)*&x;;
     %end;;

 %if %upcase(&x)=X %then totx=tot;;


 print "-----------------------------------------------
-----";
 print "------ProcÉdure DVS&k.-modes réalisée par
D.Leibovici nov1992------";
 print " Analyse par Décomposition en Valeurs
Singulières avec Métriques";
 print "            d' Tableau à &k entrées
";
 print "-----------------------------------------------
-----";
 print ,"   ", &ttit,;
 print "--Inertie totale-- " ,tot;

 %mend tot;
/*************************************************************/
/**********/


        /******FIN des MACROS  DE
BASE**********/
/*************************************************************/
/*     macro configu(x=X, tot=O,s=,nrowx=)
*/
/*  x tenseur: configuration des tableaux            */
/*************************************************************/
/*     macro inituck1(imp=O)                        */
/*  solution initiale Tucker1                       */
/*  imprm      si O numero de la solution dans sol
*/
/*************************************************************/
/*     macro solucck(X=x,imp=O,stok=N)              */
/*  calcul de la solution par  contractions         */
```



```
/*            alternées                          */
/* imprÉssion si O , stockage si O, numÉro dans sol
*/
/*************************************************************/
/*   macro ppo_x2(x=X,stok=N)                     */
/*projection sur s1 orto et ... sk orto            */
/*************************************************************/
/*   macro tot                                    */
/* calcul de l'inertie totale et prÉparation      */
/*            de la mÉtrique tensorielle          */

/*************************************************************
********/
/**********************      PROGRAMME
************************/
/*************************************************************
********/
          /********prep   **********/
 edit es ;read all var[x] into X ;
    %let x=x;
    %let k=4;      %let kx=4;
    %let nrowx=528;
    %let vrowx=528;

 %let n1=4;
 %let n2=11;
 %let n3=3;
 %let n4=4;


%let vn1=4;
%let vn2=11;
%let vn3=3;
%let vn4=4;
    %let cont=;
  Ncol2={"mr-", "mr+","mri","mrs","mo-
","mo+","moi","fr-","fr+","fri",
       "foi"} ;
  Ncol1={"J", "R","S","T"};
  Ncol4={"0", "6","12","18"};
  Ncol3={"Un", "Qq","Be"};


    %let cl=Ncol1 Ncol2 Ncol3 Ncol4;
 %let col1=[rowname=Ncol1] ;
 %let col2=[rowname=Ncol2];
 %let col3=[rowname=Ncol3];
 %let col4=[rowname=Ncol4];

 %let vcol1=[rowname=Ncol1] ;
 %let vcol2=[rowname=Ncol2];
 %let vcol3=[rowname=Ncol3];
 %let vcol4=[rowname=Ncol4];

  /***************/
  %macro pr_afc;
  /***************/
    x=x/x[+];
   %let diago=;
   %configu(X=x, Tot=O,s=);
   %let met=1 ;

   %do z=&k %to 1 %by -1;
  D&z=(X&z[+,])`;
   %let met=&met @(D&z ##(-1));
    %end;

 x=x#( &met )-j( &nrowx ,1,1);
 print "-----------------------------------------------
-----";
 print "              AFC de &k variables
";
 print "      généralisation de l'AFC de 2 variables
";
 print "         e ATP-kmodes particulières
"
 print "-----------------------------------------------
-----";
   /***************/
  %mend pr_afc;
   /***************/

  %pr_afc;

  %let metri=D1 D2 D3 D4;

  %let D1=D1;
  %let D2=D2;
  %let D3=D3;
  %let D4=D4;

  %let vD1=D1;
  %let vD2=D2;
  %let vD3=D3;
  %let vD4=D4;

  %let metri=D1 D2 D3 D4;

  %let diago=O;  /* (O) vecteurs ou() pas de
mÉtriques ou (N) matrices*/

  %let tit="Déclarations utilisation de préservatifs :
freq x typart x nbpart  bil";



          /**************PROG*********/


/*************************************************************
********/
  %macro DVSk(x=x,k= ,nb=3);

/*************************************************************
********/


     /* recherche des solution kmodes*/
 print "-----------------------------------------------
-----";
 Print "              Solutions &k modes
";
 print "-----------------------------------------------
-----";
   %tot(x=&x);
   %local nk;
   %global nn;
       %let nk=1;
```



```
       %let nn=&nk;

   %let sol=;
   %do aa=1 %to &k;%let sol=&sol.1; %end;
        /* sol=1111  si k=4*/;
%configu(X=&x,tot=O,s=);
%inituck1(imp=o);
%solucck(x=&X,imp=o,stok=o);

  %if &k=3 %then %do; %apsolu3;
%ppo_x2(B=&x,stok=O);%end;
        %else
            %do; %ppo_x2(B=&x,stok=O);
        %do cont=1 %to  &k;
        %DVS_km12(x=&x,apsol=1,nbm=1);
        %end;
            %end;



  %do nk=2 %to &nb;

  print "-----------------------------------------------
-----";
  Print "              Solutions &k modes
";
  print "-----------------------------------------------
-----";

   %let asol=&sol;
    %tot(x=xa&asol);
    %let sol=;
    %do aa=1 %to &k;%let sol=&sol.&nk;%end;
            /* sol=2222  si k=4*/;
%configu(X=xa&asol,tot=O,s=);
%inituck1(imp=N);
%solucck(x=Xa&asol,imp=o,stok=o);

  %if &k=3 %then %do; %apsolu3;
%ppo_x2(B=xa&asol,stok=O);%end;
        %else
            %do; %ppo_x2(B=xa&asol,stok=O);
        %do cont=1 %to  &k;
        %DVS_km12(x=xa&asol,apsol=&nk,nbm=2);
        %end;
            %end;
  remove xa&asol;

  %end;


  %mend DVSk;

/*************************************************************
********/

  %macro DVS_km12(x=x,apsol=1,nbm=3);
  /*************************************************/
  /* Apres e DVSk solutions kmodes*/
   %local sol ;
/*DVS_km1 contracte et fait l'analyse . DVS_km12
est la variante qui contracte puis projette sur
l'orthogonal tensoriel du "reste" de la solution avant
de continuer l'analyse (l'avantage est que les vs
redondantes sont déjà éliminées*/
          %let aps=;
        %do t=1 %to &k; %let  aps=&aps&apsol;
%end;

  print "-----------------------------------------------
-----";
  print "-----------------------------------------------
-----";
  print " Solutions associées à s&cont de &aps ";
  print "-----------------------------------------------
-----";
  print "-----------------------------------------------
-----";

     load s&cont._&aps;
      %let vdiago=&diago;
      %let diago=;
     %configu(x=&x,tot=N,s=&cont);
      %let diago=&vdiago;
  /*
   %if &diago=O %then
     bx= x&cont#( (((&&D&cont))`)*s&cont._&aps;
          %else
     bx= x&cont*&&D&cont*s&cont._&aps;
  */
     bx= x&cont#( (((&&D&cont))`)*s&cont._&aps;

     %let   nvk= %eval((&k -1));
   %let nrowx=(&vrowx /&&n&cont);

      %do e=1 %to &k;load s&e._&aps;
s&e=s&e._&aps;%end;

   %if &cont ne &k %then %do;
    %do f=&cont %to  &nvk;
       %let w=%eval(&f+1);
      %let n&f=&&vn&w;
      %let D&f=&&vD&w;
      %let col&f=&&vcol&w;
      free s&f;
      s&f=s&w;
    %end;
                %end;

   %let kq=&k;
   %let k=&nvk;
    %ppo_x2(B=Bx,stok=N);

    free Bx; Bx=D; free D;
   %let k=&kq;

     %DVSk(x=bx,k=&nvk,nb=&nbm);

    free bx;

    %do t=1 %to &kx;
     %let D&t=&&vD&t;
     %let n&t=&&vn&t;
     %let col&t=&&vcol&t;
    %end;
   %let nrowx=&vrowx;

   %mend  DVS_km12;
```

```
/*************************************************************
********/

  %DVSK(x=x,k=4,nb=2);

/*************************************************************
********/
 /****************** FIN de   PROGRAMME
************************ /

/*************************************************************
********/
```


III . Références Bibliographiques des trois parties :

CHESSEL, D. et DOLEDEC, S.(1992) A.D.E.software (version 4.3) Multivariate analysis and graphical display for environmental data. Chessel, D et Dolédec, S, URA CNRS 1451, Lyon.

BARCIKOWSKI, R.S. and ROBEY, R.R. (1983) ivariate and Multivariate repeated measures analysis through PROC's REG, GLM and MATRIX. In: Proceeding of the SUGI conference, SAS Institute, Cary, NC.

BARCIKOWSKI, R.S. and ROBEY, R.R. (1984) Decisions in single group repeated measures analysis: Statistical tests and three computer packages. Am. Stat. 148-150.

BERENBLUT, I.I. and WEEB, G.I. (1974) Experimental design in the presence of autocorrelated errors. Biometrika, 427-437.

BERK, K. (1985) Computing for balanced repeated measures experiments. In: Proceeding of the SUGI conference, SAS Institute Cry, NC.

BESSE, P. and RAMSAY, J.O. (1986) Principal components analysis of sampled fctions. Psychometrika, 285 311.

BESSE, P. (1987) Choix de la métrique pour l'ACP de series d'évènements discrets. S.A.D. 1-16.

BOIK, R.J. (1981) A priori tests in repeated measures designs : effects of nonsphericity. Psychometrika, 241-255.

BOIK, R.J. (1988) The mixed model for multivariate repeated measures: validity conditions and an approximate test. Psychometrika, 489-486.

BRILLINGER, D.R. (1984) Analysis of variance and problems der time series models. In: Handbooks of Statistics. Krishnaiah, P.R. (ed) North-Holland Publishing Company, Amsterdam, New-York, vol.1.

BURMAN, P. (1991) Regression fction estimation from dependent observations. J.Mult.An. 263-279.

BYRNE, P.J. and ARNOLD, S.F. (1983) Inference about multivariate means for a nonstationary autoregressive model. Journal of the American Statistical Association, 850-855.

CAUSSINUS, H. and FERRE, L. (1989) Analyse en Composantes Principales d'individus définis par modèle. S.A.D. 19-28.

CHAKRAVORTI, S.R. (1974) On some tests of growth curve model der Behrens-Fisher situation. J.Mult.An. 31-51.

CHURCH, A.Jr. (1966) Analysis of data when the response is curve . Technometrics, 229-246.

CLEROUX, R. and DUCHARME, G. (1986) Vector correlation for elliptical distributions. Rapport n° 586, Département Informatique et Recherche Opérationnelle, iversité de Montreal.

CORNELIUS, P.L. (1979) Curve fitting by regression on smoothed singular vectors. Biometrics, 849-859.

DEVILLE, J.C. (1974) Méthodes statistiques et numérique de l'analyse harmonique. Ann. INSEE , 3-101.

D'AUBIGNY, G. (1989) L'analyse multidimensionnelle des tableaux de dissimilarité. , Thèse de doctorat Grenoble.

DCAN, G.M. (1983) Estimation and inference for heteroscedastic systems ofequations. Int.Eco.Rev. 559-566.

EL FAOUZI, N. and ESCOUFIER, Y. (1991) Modélisation de courbes de croissance par les I-splines. R.S.A. XXXIX, 51-64.

ESCOUFIER, Y. (1973) Le traitement de variables vectorielles. Biometrics (29), 4, 751-760.

ESTEVE, J. and SHIFFLERS, E. (1976) Discussion et illustration de quelques méthodes d'analyse longitudinale. In : Proceeding of the 9th International biometric conference, biometric society, Boston 1976, 463-358.

GABRIEL, K.R. (1961) The model of ante-dependence for data of biological growth. Bull. I.S.I. 253-264.

GABRIEL, K.R. (1962) Ante-dependence analysis of an ordered set of variables. Ann. Math. Stat. 201-212.

GOLDSTEIN, H. and McDONALD, P. (1988) A general method for the analysis of multilevel data. Psychometrika, 445 467.

GRIEVE, A.P. (1984) Tests of shericity of normal distributions and the analysis of repeated measures designs. Psychometrika, 257-267.

GRIZZLE, J.E. and ALLEN, D.M. (1969) Analysis of growth and dose response curves. Biometrics, 357-381.

HAND, D. J. and TAYLOR, C.C. (1987) Multivariate analysis of variance and repeated measures. Chapman and Hall (eds), London, New-York.

HANNAN, E.J. (1967) Canonical correlation and multiple equation systems in economics. Econometrica (35), 123-138.

HOULLIER, F. (1987) Comparaison de courbes et de modèles de croissance choix d'e distance entre individus. S.A.D. 17-36.

HUYNH, H. and FELDT, L.S. (1970) Conditions der which mean square ratios in repeated measurements designs have exact F-distributions. J.A.S.A. 1585-1589.

JORESKROG, K.G. (1970) A general method for analysis of covariance structures. Biometrika, 239-251,

JOERESKOG ,K G. (1986) Analysis of longitudinal data with LISREL. In: Statistical Software (3rd Conference on the use of statistical software) , Lehmacher, W et Hoermann, A (eds), New-York.

KANG, G. and BATES, D.M. (1990) Approximate inferences in multiresponse regression analysis. Biometrika, 231-331.

KENWARD, M.G. (1987) A method for comparing profiles of repeated measurements. Appl. Stat. 296-308.

KESELMAN, H.J. and KESELMAN, J. C. (1984) The analysis of repeated measures designs in medical research. Stat.in Med. 185-195.

KHATRI, C.G. (1966) A note on MANOVA model applied to problems in growth curves. Stat.Math. 75-86.

KHATRI, C.G. (1973) Testing some covariance structure der a growth curve model. J.Mult.An. 102-116.

KOCH, G.G. AMARA, I.A. STOKES, M.E. and GILLINGS, D.B. (1980) Some views on parametric and non-parametric analysis for repeated measurements and selected bibliography. Int.Eco.Rev. 249-265.

LAVIT, C. (1988) Analyse conjointe de tableaux quantitatifs.Masson, Paris.

LEBRETON, J.D. ROUX, M. BANCO,G. et BACOU, A.M. (1992) logiciel BIOMECO (version 4.1) Analyse statistique et modélisation des processus écologiques. CEFE-CNRS, Montpellier.

LECOUTRE, B(1991) A Correction for ε (Huynh et Feldt) approximate test in repeated measures designs with two or more independent groups . J.Edu.Stat. 371-372.

LECOUTRE, B. and ROUANET, H. (1981) Deux structures statistiques fondamentales en analyse de la variance ivariée et multivariée. Math.Sci.Hum. 71-82.

LEE, J.C. (1988) Prediction and estimation of growth curves with special covariance structures. Journal of the American Statistical Association 83(402), 432-440.

LEE, Y.K. (1974) A note on Rao's reduction of Potthoff & Roy'sgeneralized linear model. Biometrika, 349-352.

LEIBOVICI, D. BUCQUET, D. SABATIER, R. CURTIS, S. and COLVEZ, A. (1990) Data analysis with dependent structure applied to epidemiological survey data in gerontology. In: 11th Meeting of the International Society for Clinical Biostatistics, Volume d'abstracts, 18-21 Sept Nîmes.

LEIBOVICI, D. (1992) Modèles Linéaires et Analyses Factorielles pour l'Analyse de Facteurs à Mesures Répétées. Commication aux XXIVe Journées de Staistique A.S.U Bruxelles.

LOONEY, S.W. and STANLEY, W.B. (1989) Exploratory repeated measures analysis for tow or more groups, review and update. Am. Stat. 220-225.

MAUCHLY, J.W. (1940) Significannce test for sphericity of a n-variate distribution. An.Math.Stat. 204-209.

MENDOZA, J.L. (1980) A significant test for multisample sphericity. Psychometrika, 495-498.

MONLEZ, C. and BLOUIN, D. (1988) A general nested split-plot analysis of covariance. J.A.S.A. 818-823.

MORRISON, D.F. (1970) The optimal spacing measure. Biometrics, 281-290.

MORRISON, D.F. (1972) The analysis of single sample of repeated measurements. Biometrics, 55-71.

PATEL, H.I. ROY, T. and HUQUE, M.F. (1985) Some preliminary tests in repeated measures design. In: . Proceeding of the SUGI conference, SAS Institute, Cary, NC.

PATEL, H.I. (1986) Analysis of repeated measures designs with changing covariates in clinical trials. Biometrika, 707-715.

PONTIER, J. DUFOUR, A.B. and NORMAND, M. (1990) Le modèle euclidien en analyse des données. Editions Ellipses, Paris.

POTTHOF, R.F. and ROY, S.N. (1964) A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 313-626.

RAO, C.R. (1964) The use and interpretation of principal component analysis in applied research. Sankhya A, 329-359.

RAO, C.R. (1965) The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. Biometrika, 447-458.

RAO, C.R. (1967) Least square theory using an estimated dispersion matrix and its application to measurement of signals. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics, iversi Press, B, 355 372.
RAO, C.R. (1972) Recent trends of research work in multivariate analysis. Biometrics, 3-22.
RAO, C.R. and YANAI, H. (1979) General definition and decomposition of projectors and some applications to statistical problems. J.Stat.Plan.Inf., 3, 1-17.
RAO, C.R. (1987) Prediction of future observations in growth curve models. Stat.Sci. 434-471.
RAZ, J. (1989) Analysis of repeated measurements using nonparametric smoothers and randomization tests. Biometrics, 851-871.
REINSEL, G. (1982) Multivariate repeated-measurement or growth curve model with multivariate random-effects covariance structure. J.A.S.A. 190-195.
REVANKAR, N.S. (1974) Some finite sample results in the context of two seemingly related regression equations. J.A.S.A. 187-190.
ROBERTS, J.S. and LAUGHLIN, J.E. (1983) Analyzing anlysis of covariance models in ivariante repeated measures designs using SAS. In: . Proceeding of the SUGI conference, SAS Institute, Cary, NC.
ROUANET, H. and LEPINE, D. (1970) Comparison between treatments in a repeated-measurement design : anova and multivariate methods. Brit.Jo.Math.Stat.Psyc. 147-163.
SABATIER, R. (1987) Méthodes factorielles en analyse des données : approximations et prise en compte de variables concomitantes. Thèse doct.ès-sciences, U.S.T.L., Montpellier
SABATIER, R. LEBRETON, J. and CHESSEL, D. (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. In: Multiway data analysis, Coppi, R et Bolasco, S (eds), North-Holland, Amsterdam, 341-352.
SCHATZOFF, M. (1966) Sensitivity comparisons among tests of general linear hypothesis. J.A.S.A. (part 1), 415-437.
SINGH, S. and ULLAH, A. (1974) Estimation of seemingly related regressions with random coefficients. J.A.S.A. 191-195.
SNEE, R.D. (1972) On the analysis of response curve data. Technometrics 14 (1) ,47-62.
SNEE, R.D. , ACUFF, S.K. and GIBSON, J.R. (1979) A useful method for tha analysis of growth studies. biometrics, 835-848.
SPECTOR, P. (1985) Repeated Measures Analysis in PROC GLM. In: . Proceeding of the SUGI conference, SAS Institute, Cary, NC
SRIVASTAVA, V.K. and DWIVEDI, T.D. (1979) Estimation of seemingly related regression equations. Jo.Eco. 15-32.
SRIVASTAVA, M.S. and CARTER, E.M. (1983) An introduction to applied multivariate statistics. Elsevier Science Pub Co, New York.
STANEK, E.J. and KOCH, G.J. (1985) The equivalence of parameter estimates from growth curve models and seemingly related regression models. Am. Stat. 149-152.
STANEK, E.J. and DIEHL, S.R. (1988) Growth curve models of repeated response. Biometrics, 973-983.
TANDON, P.K. and MOESCHBERGER, M.L. (1983) The SAS macros for nonparametric analysis of repeated measures designs. In: Proceeding of the SUGI conference, SAS Institute, Cary, NC
THOMAS , D.R. (1983) ivariate repeated measures techniques applied to multivariate data. Psychometrika, 454 464.
TIMM, N H. (1984) Multivariate analysis of variance of repeated measurements.In: Handbooks of Statistics Krishnaiah, P R (ed) North-Holland Publishing Company, Amsterdam, New-York, vol.1, 41-87.
VERBYLA, A.P. (1988) Analysis of repeated measures designs with changing covariates. Biometrika, 172-174.
VERBYLA, A.P. and VENABLES, W.N. (1988) An extension of the growth curve model. Biometrika, 129-138.
WARE, J.H. (1985) Linear models for the analysis of longitudinal studies. Am. Stat. 95-101.
WINER, B.J. (1971) Statistical principle in experimental design. Mc Graw-Hill Book Company, New-York, Toronto.
WISHART, J. (1938) Growth rate deteermination in nutrition studies with the baccon pig and their analysis. Biometrika, 16-28.
WRIGHT, P.S. (1982) Repeated measures ANOVA : The multivariate approach. In: Proceeding of the SUGI conference, SAS Institute, Cary, NC.
YATES, F. (1982) Regression models for repeated measurement. Biometrics, 850-853.

YOSHIMASA, U. KAZUO, N. and ETSUO, M. (1992) UMP invariants tests for generalized linear model. J.Mult.An. 1-12.
ZELLNER, A. (1962) An efficient method of estimating seemingly related regressions and tests for aggregation bias. J.A.S.A. 348-368.
ZERBE, G.O. and WALKER, S.H. (1977) A randomization test for comparison of groups of growth curves with different polynomius design curve. Biometrics, 653-657.
ZERBE, G.O. and JONES R, H. (1980) On application of growth curve techniques to time series. J.A.S.A. 507-509.
ZUPKIS, R.V. and SHARP, M. (1982) SAS macro "REP2" for weighted means analysis of repeated measures. In: Proceeding of the SUGI conference, SAS Institute, Cary, NC.
AGRESTI, A. (1988) Logit models for repeated ordered categorical response data. Sas Sugi 997-1005.
BISHOP, Y.M. FIENBERG, S.E. and HOLLAND, P.W. (1975) Discrete multivariate Analysis : Theory and Practice. M.I.T. Press, Cambridge, Massachussets.
ANDREWS, D.M. and DAVID, H.A. (1990) Nonparametric analysis of balanced paired-comparison or ranked data. J.A.S.A. 1140-1146.
BECKER, M.P and AGRESTI, A (1992) Log-linear modelling of pairwise interobserver agreement on a categorical scale. Stat.Med. 11, 101-114.
BHAPKAR, V.P. and KOCH, G.G. (1968) On the Hypotheses of 'NO Interaction' in Contingency Tables. Biometrics, 567-594.
BONETT, D.G. WOODWARD, J.A. and BENTLER, P.M. (1985) Some extension of a linear model for categorical variables. Biometrics, 745-750.
CLAYTON, D.G. (1974) Some odds ratio statistics for the analysis of ordered categorical data. Biometrika, 525-531.
COX , D.R. (1972) Regression Models and Life Tables. J. R. Stat. Soc. B.34, 187-220.
DAUDIN, J.J and TRECOURT,P. (1980) Analyse factorielle des correspondances et modèle log-linéaire : comparaison des deux méthodes sur exemple. R.S.A. XXVIII, 5-24.
ESCOFIER, B. (1984) Analyse factorielle en référence à modèle : application à l'analyse de tableaux d'échanges. Rapport Technique n°337. INRIA, Rocquencourt.
FISHER, G.H. (1989) An IRT-Based Model for Dichotomous Longitudinal data. Psychometrika, 599-624.
GRIZZLE, J.E. STARMER, C.F. and KOCH, G.G (1969) Analysisof categorical data by linear models. Biometrics, 489-504.
GRIZZLE, J.E. and WILLIAMS, D.O (1972) Log linear models and tests of independance for contingency tables. Biometrics, 137-156.
GOODMAN, L.A. (1986) Some useful extensions of the usual correpondence analysis approach an the usual log-linear models approach in the analysis of contingency tables. In.Stat.Rev. 243-309.
IMREY, P.B. KOCH, G.G. and STOKES, M.E. et al. (1981) Categorical data analysis : Some reflections on the log-linear model and logistic regression. Part I :Historical and Methodoligical overview. Int.Stat.Rev. 265-283.
IMREY, P.B. KOCH, G.G. and STOKES, M.E. et al.(1982) Categorical data analysis : Some reflections on the log-linear model and logistic regression. Part II : Data analysis. Int.Stat.Rev. 35-63.
JÖRESKOG, K.G. (1970) A general method for analysis of covariance structures. Biometrika, 239-251.
KOCH, G.G. LANDIS, J.R. FREEMAN, J.L. FREEMAN, D.H. AND LEHNEN, R.G. (1977) A general methodology for analysis of experiments with repeated measurement of categorical data. Biometrics, 133-158.
KOCH, G.G. and REINFURT, D.W. (1971) The analysis of categorical data from mixed models. Biometrics, 157-173.
KOCH, G.G. STOKES, M.E. and BROCK, D. (1980) Applications of weighted least squares methods for fitting variational models to health survey data. In : Proceedings of the American Statistical Association.Section on Survey Research Methods. 218-223.
LANDIS, J.R. and KOCH, G.G. (1977) The measurement of observer Agreement for categorical data. Biometrics, 159-174.
LEIBOVICI, D. BUCQUET, D. SABATIER, R. CURTIS, S. and COLVEZ, A. (1990)D ata analysis with dependent structure applied to epidemiological survey data in gerontology. In: 11th Meeting of the International Society for Clinical Biostatistics, Volume d'abstracts, 18-21 Sept.
MCLEAN, R.A. SANDERS, W.L. and STROUP, W.W. (1991) A ified approach to mixed linear models. Am. Stat. 55-64.

SABATIER, R. and LEBRETON,J.D. (1990) Comparaisons d'Analysesfactorielles sous-contraintes linéaires et desmodèles log-linéaires pour l'analyse de tables de contingenes. In: XXII journnées de satistique . ASU. Tours mai 1990.
VAN Der HEIJDEN, P.M. De FALGUEROLLES, A. and De LEEUW, J. (1989) A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. Appl. Stat. 249-292.
WARE, J.H. LIPSITZ, S. and SPEIZER, F.E. (1988) Issues in the analysis of repeated categorical outcomes. Stat.Med. 95-107.