

STATISTIQUE ET ANALYSE DES DONNÉES

FRANÇOIS LECHEVALIER

L'analyse en composantes conjointes d'une famille de triplets indexés

Statistique et analyse des données, tome 15, n° 2 (1990), p. 35-75

http://www.numdam.org/item?id=SAD_1990__15_2_35_0

© Association pour la statistique et ses utilisations, 1990, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

*Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques*

<http://www.numdam.org/>

**L'Analyse en Composantes conJointes
d'une famille de triplets indexés**

par

François Lechevalier

Laboratoire de Statistique et Probabilités (M 2)

Université des Sciences et Techniques de Lille Flandres Artois

59655 - Villeneuve d'Ascq Cedex (France)

Résumé :

Pour une famille de tableaux quantitatifs X_k , de type $I \times J_k$, $k \in K$ nous posons le problème du positionnement simultané des nuages d'individus, et nous proposons une méthode d'analyse des données qui généralise l'Analyse en Composantes Principales (ACP) d'un tableau au cas d'une famille de tableaux.

Nous gardons la possibilité de construire, pour chaque nuage, une image euclidienne équivalente qui contient le maximum d'inertie lorsqu'on se limite aux premiers axes. Nous demandons, de plus, une justification du positionnement simultané des différents nuages.

Abstract :

Let $\{X_k, k \in K\}$ be a set of data matrices. Each X_k is an n rows (objects) and p_k columns (variables) matrix. We are looking for a simultaneous representation of the objects in a common space.

We present a method which, applied to a set of data matrices, allows a generalization of P.C.A results. For each set of objects we define an exact mapping, and the simultaneous representation takes into account and respects the properties of duality.

Mots clés : Analyse conjointe de tableaux – Compromis – Statis – Intrastructure – Indscal – Liaisons entre groupes de variables – Analyse Factorielle Multiple – Analyse Canonique – Analyse Procuste – Repère Equisecteur – Composantes conJointes.

A.M.S. Subject classification : 62H99.

Manuscrit reçu le 3 mai 1990, révisé le 6 août 1990

INTRODUCTION : Les objectifs de la méthode.

- a) Rappel des propriétés de l'ACP.
- b) Objectifs pour l'intrastructure.

I - LES METHODES D'INTRASTRUCTURE.

- 1.1) Les intrastructures inexactes.
 - a) L'AFJ1 (ou l'AFM).
 - b) L'AFJ2 (ou Statis).
- 1.2) Les intrastructures exactes.
 - a) La "superposition".
 - b) L'Analyse Procuste.
- 1.3) L'AFJ et les autres méthodes.
 - a) Les méthodes d'intrastructure.
 - b) Les méthodes de reconstitution des données.

II - LES TABLEAUX A COMPOSANTES PROPORTIONNELLES.

- 2.1) Le Repère Equisecteur.
 - a) Les systèmes équisecteurs.
 - b) Signification statistique pour les données cubiques.
 - c) Le Repère Equisecteur pour les données non-cubiques.
- 2.2) Introduction à l'AFJ.
 - a) Les tableaux proportionnels.
 - b) Les tableaux à composantes principales proportionnelles.
 - c) Le Repère Déformant.

III - L'AFJ D'UNE FAMILLE DE TABLEAUX QUELCONQUES.

- 3.1) Caractérisation des tableaux à composantes proportionnelles.
 - a) Propriétés du Repère Compromis.
 - b) Propriétés du Repère Déformant.
- 3.2) L'Analyse Factorielle conJointe (AFJ).
 - a) Les composantes conjointes.
 - b) Nombre minimal admissible de composantes.
 - c) L'algorithme de l'AFJ.
- 3.3) Les résultats de l'AFJ.
 - a) Interprétation de la méthode.
 - b) Applications.
 - c) L'Analyse Procuste.

CONCLUSION.**BIBLIOGRAPHIE.**

0 - INTRODUCTION

Les objectifs de la méthode

Soit une famille de triplets indexés : (X_k, M_k, D) , $k \in K$. Dans chaque triplet, le tableau X_k dimensionné $(n \times p_k)$ contient les observations d'un paquet J_k de variables quantitatives sur un ensemble I d'individus, et q_k est le rang de X_k . Les entiers n , p_k , m désignent le cardinal respectivement de I , J_k , K ; et on note indifféremment $k \in K$ ou $k \in [m]$. On parle de données cubiques quand tous les paquets J_k sont identiques ($J_k = J$) et de données non cubiques dans le cas contraire que nous étudions plus particulièrement. Nous cherchons un positionnement simultané des nuages de points N_k définis dans les espaces Euclidiens $E_k = (\mathbf{R}_{(e^k)}^{p_k}, M_k)$, par les colonnes des tableaux X'_k (transposés de X_k) supposés centrés pour la métrique des poids D . Ce positionnement simultané – nommé **intrastructure** – peut contenir certains transformés des nuages N_k .

L'intrastructure sert à représenter simultanément les nuages dans des sous-espaces de petite dimension, exactement comme c'est le cas pour un seul nuage dans l'ACP d'un tableau. Il paraît donc judicieux, au plan méthodologique, de définir les critères que doit satisfaire une intrastructure, en s'inspirant des résultats de l'ACP.

Nous rappelons ceux-ci, puis nous précisons les deux critères exigibles pour une intrastructure.

a) Rappel des propriétés de l'ACP.

Dans l'ACP d'un triplet (X, M, D) on construit deux triplets équivalents au premier : (C, I, D) et (Z, M, I) .

- Le nuage des individus $\overline{[X']}$, défini par les colonnes du tableau $X'(p \times n)$, est positionné dans l'espace Euclidien $E = (\mathbf{R}_{(e)}^p, M)$. Avec (C, I, D) on définit, par exemple dans Cailliez et Pagès (1976), une image euclidienne $(\overline{[C]}, \mathbf{R}_{(u_j)}^q, I)$ équivalente à l'image euclidienne $(\overline{[X]}, \mathbf{R}_{(e)}^p, M)$. Si la décomposition en valeur singulière (s.v.d.) de X s'écrit $X = \Gamma \Delta U'$, on a : $C = \Gamma \Delta$ de dimension $(n \times q)$ avec $q = \text{rang } X$, et $Z = \Delta U'$.

• Avec (Z, M, I) , on obtient pour le nuage des variables, une image euclidienne $(\boxed{Z}, \mathbf{R}_{(\gamma^j)}^q, I)$ équivalente à $(\boxed{X}, \mathbf{R}_{(f_i)}^n, D)$. Ce résultat permet de positionner les variables de façon comparable à celle retenue pour les individus. Il permet aussi de définir très simplement le biplot : sur les axes de l'espace $E^q = (\mathbf{R}_{(u_j)}^q, I)$, les coordonnées du vecteur e_j portant un axe initial (et donnant la direction de la variable initiale j) forment la colonne j de : $U'M = \Delta^{-1}ZM$. Cette conception du biplot n'est pas exactement celle existant dans la langue anglaise (Jolliffe 1985).

On note X_i la ligne i de X et X^j sa colonne j . Alors $x_i = (X_i)'$ définit l'individu i comme vecteur colonne de E (indice de x en bas), et $x^j = X^j$ définit la variable j comme vecteur colonne de F (indice de x en haut) ; de même pour $u_j = U^j$ et $c^j = C^j$. L'indice k des situations échappe à ces règles et figure indifféremment en haut ou en bas. Il ne repère pas une rangée mais spécifie que l'objet qu'il indice est pris en situation k . $\|X\|_{\varphi_{DM}}$ désigne la norme de matrice induite par le produit scalaire $\langle X_1, X_2 \rangle_{\varphi_{DM}} = \text{trace}(X_1'DX_2M)$ (Glaçon 1981).

Si on complète le théorème de décomposition, par les théorèmes d'approximation (Eckart-Young 1936, Kristof 1970, Le Calvé 1976) on retrouve la propriété d'optimalité suivante de l'ACP :

Le tableau $(n \times s)$ des s premières composantes principales $(c^j)_{j \in [s]}$, noté C_s , définit la projection du nuage $\boxed{C^j}$ sur un sous-espace de dimension s , ayant la plus grande inertie. De manière équivalente $\|C - (C_s \ ; \ 0)\|_{\varphi_{DI}}^2$ est minimale pour le tableau C_s . En terme de reconstitution des données : $\|X - B\|_{\varphi_{DM}}^2$, avec $B(n \times p)$ de rang s , est minimale si $B = \sum_{j=1}^s c^j(u_j)'$.

Pour résumer, l'ACP présente les propriétés suivantes :

- obtention d'une image euclidienne équivalente (C, I, D) .
- le tableau C_s est le tableau de rang s qui restitue la plus grande part de l'inertie du nuage $\boxed{C^j}$.
- la description des variables initiales est optimale au sens de l'inertie d'un nuage non centré (pour une métrique M diagonale).
- le biplot permet d'indiquer la direction des variables initiales dans les plans principaux retenus pour représenter le nuage des individus, et d'orienter ces plans et les axes principaux.

b) Objectifs pour l'intrastructure.

Pour une représentation simultanée de plusieurs nuages il est évidemment intéressant de retrouver ces 4 propriétés et, dans un premier temps, au moins les deux premières, en remarquant que la seconde n'a de signification que si la première est vérifiée.

En se limitant aux méthodes d'étude des données non-cubiques on constate que cette première propriété n'est pas vérifiée par Statis ou l'AFM, tandis qu'elle peut l'être dans l'Analyse Procuste ou la méthode dite de "superposition des axes principaux d'inertie", ce qui sera montré ultérieurement.

Ces deux dernières méthodes peuvent fournir deux infrastructures totalement différentes. Cela impose un deuxième objectif pour la recherche d'une infrastructure qui est de justifier le positionnement simultané réalisé.

Nous demandons donc à l'infrastructure d'une famille de tableaux indexés de satisfaire deux conditions : être exacte et être significative.

- Elle est **exacte** si pour tout $k \in K$, le triplet (Y_k, I, D) représenté dans l'infrastructure est équivalent au triplet initial (X_k, M_k, D) .
- Elle est **significative** si le positionnement simultané réalisé dans l'infrastructure respecte les règles de la dualité. Nous dégagerons progressivement l'importance d'un tel critère.

En schématisant les choses nous montrerons que les méthodes françaises, et les méthodes de type INDSCAL, donnent une infrastructure inexacte mais significative, que la méthode procuste et la méthode de "superposition" offrent une infrastructure éventuellement exacte mais peu significative. L'Analyse Factorielle conJointe (AFJ) que nous proposons peut alors présentée comme une synthèse de ces méthodes donnant, sous certaines conditions, une infrastructure exacte et significative. Nous verrons qu'elle réalise une Analyse en Composantes conJointes d'une famille de triplets indexés.

Nous présentons ici les principaux éléments de la méthode dont un exposé plus complet figure dans notre thèse (Lechevalier 1990). Nous l'avons nommée AFJ en raison du sens très large pris par le terme analyse factorielle. On pourrait aussi l'appeler ACJ (Analyse en Composantes conJointes) et nous laissons ce choix ouvert.

I - LES METHODES D'INTRASTRUCTURE

Avant de traiter les données non-cubiques, rappelons que l'ACP permet d'obtenir facilement deux infrastructures, exactes ou non, pour des données cubiques $X_{IJK} = \{X_k(n \times \bar{p}), k \in [m]\}$.

- L'ACP du Tableau Conjoint Empilé (multicentré) $X_e = \begin{pmatrix} X_1 \\ \dots \\ \vdots \\ \dots \\ X_m \end{pmatrix}$ présente un positionnement simultané des m nuages définis par les tableaux (centrés) X_k , optimal –au sens de l'inertie– pour la réunion des nuages. L'infrastructure obtenue est exacte.
- L'ACP du Tableau Moyen $\bar{X} = \sum_k \rho_k X_k$, ($\rho_k > 0$ et $\sum_k \rho_k = 1$) avec projection des tableaux X_k en supplémentaires donne une autre infrastructure, non nécessairement exacte ni optimale.

L'intérêt de cette seconde méthode est d'être adaptable au cas non-cubique des deux façons présentées ci-dessous, dans lesquelles, les poids ρ_k peuvent être choisis pour retrouver les résultats des méthodes classiques.

1.1.- Les infrastructures inexactes.

a) L'AFJ1 (ou l'AFM)

Pour tout $k \in [m]$, considérons :

$\tilde{X}_k = (0 \quad \vdots \quad \dots \quad \vdots \quad 0 \quad \vdots \quad X_k \quad \vdots \quad 0 \quad \vdots \quad \dots \quad \vdots \quad 0)$ tableau $(n \times p)$ avec $p =$

$\sum_k p_k$ et $\mathcal{M} = \begin{pmatrix} M_1 & & 0 \\ & \ddots & \\ 0 & & M_m \end{pmatrix}$. Alors les triplets (X_k, M_k, D) et

$(\tilde{X}_k, \mathcal{M}, D)$ sont équivalents. Ce dernier triplet donne un positionnement du nuage $N_k = \boxed{X'_k}$ dans l'espace $E^p = \bigoplus E_k$ et on le note : $\tilde{N}_k = \boxed{\tilde{X}'_k}$.

Les triplets $(\tilde{X}_k, \mathcal{M}, D)$, $k \in [m]$ constituent alors un cube de données que l'on peut traiter par la méthode du Tableau Moyen. Il s'écrit :

$$\tilde{X} = \sum_k \rho_k \tilde{X}_k = (\rho_1 X_1 \quad \vdots \quad \dots \quad \vdots \quad \rho_m X_m) = X_\rho.$$

Dans l'ACP du tableau moyen on diagonalise $W_0D = X_\rho \mathcal{M} X'_\rho D = \sum_k \rho_k^2 W_k D$, avec : $W_k = X_k M_k X'_k$, pour trouver les éléments propres (C_0, Λ_0) .

C_0 est un tableau $(n \times r)$ ($r = \text{rang } X_\rho$). Ses lignes sont les coordonnées des n points du nuage moyen dans la base $(u_j)_{j \in [r]}$ du sous-espace **Compromis** E_C de E^p que ce nuage moyen engendre. La \sqrt{m} -projection (projection multipliée par \sqrt{m}) des tableaux \tilde{X}_k en supplémentaires sur ce sous-espace E_C définit l'intrastructure $\hat{X}_k = \sqrt{m} \rho_k W_k D C_0 \Lambda_0^{-1}$, $k \in [m]$.

On vérifie que : $\hat{X} = \sum_k \rho_k \hat{X}_k$ est égal à $\sqrt{m} C_0 = C$ (les propriétés barycentriques sont conservées dans une projection, un changement de base et une homothétie). $\hat{X} = C$ définit un positionnement du **nuage compromis** homothétique du nuage moyen pour l'homothétie de centre O_{E^p} et de rapport \sqrt{m} . Le nuage compromis est le nuage moyen des \sqrt{m} -projetés des nuages \tilde{N}_k sur E_C et C est le **tableau compromis**. L'utilisation d'une \sqrt{m} -projection sera justifiée ultérieurement. Cette méthode est appelée AFJ1. Elle est proche de l'Analyse Factorielle Multiple (AFM - Escofier et Pagès : 1983 et 1988) qui impose des poids ρ_k proportionnels à $1/\sqrt{\lambda_1^k}$ (ou une prémultiplication des tableaux par ce facteur), avec λ_1^k première valeur propre de $W_k D$.

b) L'AFJ2 (ou Statis)

L'opérateur $W_k D$ est caractéristique du triplet (X_k, M_k, D) . Il en résulte que l'ACP du triplet (W_k, D, D) redonne le positionnement des individus obtenu dans l'ACP de (X_k, M_k, D) . Pour ce dernier triplet, $W_k D$ admet les éléments propres (C_k, Λ_k) avec $C'_k D C_k = \Lambda_k$. Pour (W_k, D, D) on diagonalise $(W_k D)^2$ dont les éléments propres (B_k, A_k) avec $B'_k D B_k = A_k$, vérifient $A_k = \Lambda_k^2$ et donc $B_k = C_k \Lambda_k^{1/2}$, ou encore $\Lambda_k = A_k^{1/2}$ et $C_k = B_k \Lambda_k^{-1/2}$.

L'AFJ2 consiste à faire l'ACP du Tableau Moyen sur le cube de données défini par les triplets (W_k, D, D) , $k \in [m]$. On trouve pour triplet moyen : $(W = \sum_k \rho_k W_k, D, D)$. La diagonalisation de $(WD)^2$ donne (B, A) , puis $\Lambda = A^{1/2}$ et $C = B \Lambda^{-1/2}$, ou encore, la diagonalisation de WD donne directement (C, Λ) .

Il reste à projeter les W_k en supplémentaires (dans l'AFJ2 on considère des projections et non des \sqrt{m} -projections), pour obtenir $\hat{W}_k = W_k D C \Lambda^{-1/2}$. Peut-on en déduire un positionnement simultané \hat{X}_k , $k \in K$, des individus des différents tableaux ? Si on prend le positionnement simultané \hat{W}_k , $k \in K$ et qu'on lui applique la règle de passage de B (qui définit le positionnement \hat{W}

de $\bar{W} = W$) à $C = B\Lambda^{-1/2}$ (qui définit le positionnement \hat{X} d'un certain \bar{X}), on obtient : $\hat{X}_k = W_k D C \Lambda^{-1/2} \Lambda^{-1/2} = W_k D C \Lambda^{-1}$, $k \in [m]$, et le compromis $\hat{X} = \sum_k \rho_k \hat{X}_k = C$. Alors C désigne le tableau des composantes principales de $X_{\sqrt{\rho}} = (\sqrt{\rho_1} X_1 \quad \dots \quad \sqrt{\rho_m} X_m)$ considéré comme un tableau moyen \bar{X} .

On peut présenter ainsi l'intrastructure de la méthode Statis qui utilise des "poids" particuliers vérifiant $\sum_k \rho_k^2 = 1$ (Lavit : 1985 *in* SAD, et 1988).

Remarques :

- 1) Ces deux intrastructures sont inexactes sauf dans de rares cas particuliers que nous verrons plus loin.
- 2) Une isométrie transformant n'importe quel tableau X_k laisse ces intrastructures inchangées puisque $W_k = X_k M_k X_k'$ est invariant dans une isométrie.
- 3) Les appellations AFJ1 et AFJ2 peuvent surprendre puisque ces deux méthodes proposent des intrastructures très proches de celles de l'AFM ou de Statis. Nous les utilisons pour les raisons suivantes :

- L'AFM ou Statis ne se réduisent pas à la recherche d'une intrastructure.
- La pondération de l'AFM et celle de Statis sont fixées alors qu'elle est libre (à ce stade) dans l'AFJ1 et l'AFJ2.
- Les intrastructures de l'AFJ1 et de l'AFJ2 sont parfaitement identiques pour des poids égaux (d'où la ressemblance des noms). Comme nous utilisons l'AFJ1 pour initialiser l'algorithme de l'AFJ et qu'elle fournit un critère permettant d'apprécier sa faisabilité, nous considérons l'AFJ1 comme la première étape de l'AFJ.

1.2.- Les intrastructures exactes.

a) La "superposition des axes principaux d'inertie".

Notons $q = \sup_k q_k$ et considérons les tableaux $C_k^c(n \times q)$ obtenus en complétant chaque tableau $C_k(n \times q_k)$ par $q - q_k$ colonnes de 0. Pour tout k , les triplets (C_k^c, I_q, D) et (X_k, M_k, D) sont équivalents. Donc les C_k^c définissent une intrastructure exacte dans l'espace (\mathbf{R}^q, I_q) obtenue par superposition des axes principaux d'inertie de même rang. Si on remplace les C_k^c par les tableaux C_k^s formés des s premières composantes, cette intrastructure limitée à s composantes

conserve le maximum d'inertie. Cette propriété mise à part, il n'y a aucune raison pour superposer les axes de même rang des différents nuages. Ainsi on peut imaginer deux nuages du plan tels que le critère "Procrustes" (ou Procuste) conduise à superposer les axes 1 et 2 et non les axes 1 comme le veut le critère de "superposition". En fait ces deux infrastructures ne sont pas nécessairement significatives du point de vue du critère basé sur la dualité que nous construirons.

b) L'Analyse Procuste

Dans Ten Bergé et Knol (1984) le problème (orthogonal) procuste est ainsi posé : ayant $n \geq p_1 \geq p_2 \geq \dots \geq p_m$, on cherche les matrices T_k de dimensions $(p_k \times p_m)$ réalisant $\text{Min } f(T_1, T_2, \dots, T_m) = \sum_{k < \ell} \|X_k T_k - X_\ell T_\ell\|_{\varphi_{D_I}}^2$, avec $T_k' T_k = I_{p_m}$, $k \in [m]$. Ce problème dit "asymétrique" diffère du problème "symétrique", pour lequel $p_k = \bar{p}$, $k \in [m]$ (données cubiques par exemple), par le fait essentiel que les matrices T_k ne définissent pas des isométries de (\mathbf{R}^{p_k}, I) dans (\mathbf{R}^{p_m}, I) . Ainsi on trouve une infrastructure exacte pour l'analyse orthogonale procuste généralisée symétrique, mais inexacte dans le cas asymétrique. En revanche, si on choisit des matrices $T_k (p_k \times p_1)$ définissant des isométries de \mathbf{R}^{p_k} dans \mathbf{R}^{p_1} , on a $T_k T_k' = I_{p_1}$ et le cas asymétrique équivaut au cas symétrique obtenu en complétant chaque X_k par $p_1 - p_k$ colonnes de 0.

Il existe donc bien deux infrastructures exactes pour des données non-cubiques. L'une s'obtient par la méthode de "superposition" et l'autre par l'Analyse Procuste "symétrisée" présentée ci-dessus. Nous verrons qu'elles ne sont significatives que dans de rares cas particuliers.

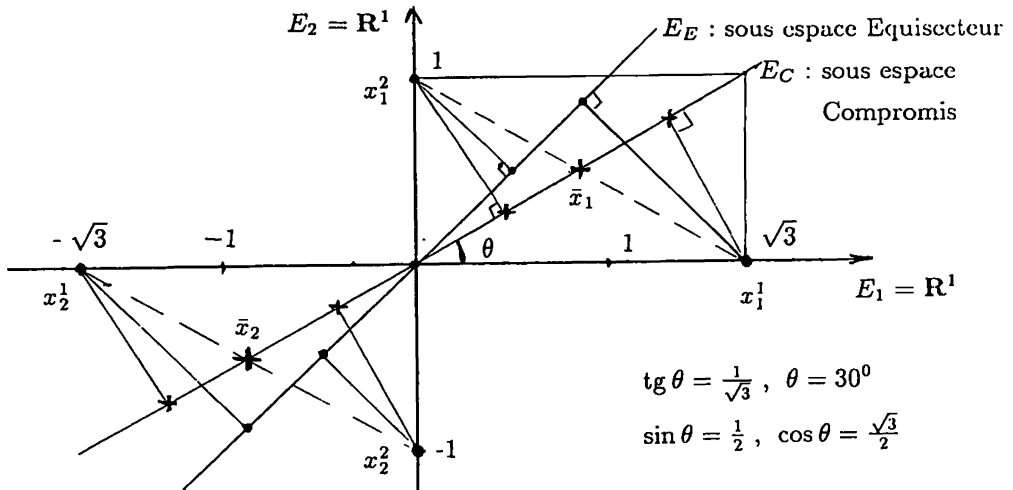
1.3.- L'AFJ et les autres méthodes.

a) Les méthodes d'infrastructure.

Dans les deux parties suivantes, nous allons développer l'Analyse Factorielle conJointe – mais aussi l'ensemble des méthodes d'infrastructure – à partir de considérations géométriques qui généralisent la méthode dite "AFJ1". Celle-ci se résume très simplement : les nuages N_k sont positionnés dans l'espace (E^p, \mathcal{M}) somme directe orthogonale des espaces (E_k, M_k) . Dans l'AFJ1 l'infrastructure s'obtient en projetant les nuages \tilde{N}_k sur le sous-espace Compromis E_C et elle est nommée Infrastructure Compromis (I.C.). De même, l'AFJ est une méthode qui obtient une infrastructure exacte en projetant sur un sous-espace particulier de

E^P appelé sous-espace Equisecteur et donnant l'Intrastructure Equisecteur (I.E.). Nous précisons cette première définition de l'AFJ, puis nous définissons dans E^P le sous-espace Déformant donnant l'Intrastructure Déformante (I.D.), et nous évoquons les sous-espaces procuste (cf 3.3.c), discriminant, de superposition, donnant les intrastructures correspondantes. Cette présentation établit donc un lien très simple entre toutes ces méthodes d'intrastructure.

Le petit exemple suivant peut être représenté dans le plan ; il aide à bien comprendre l'exposé du problème dans le cas général. Deux paquets de une variable ont été observés sur deux individus, et on a : $X_1 = \begin{pmatrix} \sqrt{3} \\ -\sqrt{3} \end{pmatrix}$ et $X_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. L'AFJ1 avec des poids égaux donne le tableau moyen $X_p = \frac{1}{2} \begin{pmatrix} \sqrt{3} & 1 \\ -\sqrt{3} & -1 \end{pmatrix}$, et les individus moyens engendrent le sous espace Compromis E_C que l'on représente dans $\mathbf{R}^2 = \mathbf{R} \oplus \mathbf{R}$ comme suit :



Les projections des nuages \tilde{N}_k sur E_C sont définies respectivement par : $\frac{\sqrt{3}}{2}X_1$ et $\frac{1}{2}X_2$, donc les \sqrt{m} -projections sont : $\hat{X}_1 = \frac{\sqrt{6}}{2}X_1$ et $\hat{X}_2 = \frac{\sqrt{2}}{2}X_2$. Cette Intrastructure Compromis est inexacte et aucune homothétie ne peut la rendre exacte.

Les projections des nuages \tilde{N}_k sur le sous espace Equisecteur E_E (première bissectrice) s'écrivent : $\frac{1}{\sqrt{2}}X_k, k \in K$, donc les \sqrt{m} -projections sont : $\hat{X}_k = X_k$. Cette Intrastructure Equisecteur est exacte.

b) Les méthodes de reconstitution des données.

Nous établirons cette deuxième définition : l'AFJ est une méthode qui recherche des triplets $(Y_k (n \times q), I, D)$ équivalents à (X_k, M_k, D) , $k \in K$, des matrices diagonales $A_k (q \times q)$ et un tableau Y de dimensions $(n \times q)$ tels que : $\Phi(Y_k, A_k, Y) = \sum_k \|Y_k - Y A_k\|_{\varphi_{DI}}^2$ soit minimale. Cette formulation est voisine de celle des problèmes INDSCAL et éventuellement Candecomp (Carroll et Chang 1970) ou Parafac (Harshman 1970, 1984). Avec Kiers (1987, 1988), remarquons que Parafac et Candecomp cherchent une reconstitution des données sous la forme $\hat{X}_k = Y A_k T'$, $k \in K$, telle que $\sum_k \|X_k - \hat{X}_k\|_{\varphi_{DM}}^2$ soit minimale, la dimension $(r \times r)$ de A_k diagonale étant choisie. Pour sa part, INDSCAL recherche $\hat{W}_k = Y A_k^2 Y'$, $k \in K$, tels que $\sum_k \|W_k - \hat{W}_k\|_{\varphi_{DD}}^2$ soit minimale.

L'originalité de l'AFJ, qui a notre sens justifie qu'on lui donne un nom, ne tient pas tant au problème d'optimisation posé, qu'à la manière géométrique de l'aborder et de lui donner une solution :

- 1) Nous ne cherchons pas la meilleure reconstitution des données (avec \hat{X}_k ou \hat{W}_k) mais une reconstitution exacte des données (avec Y_k) permettant de plus, d'avoir une infrastructure significative.
- 2) Nous établissons des liaisons simples entre toutes les méthodes précédemment citées.
- 3) Nous contruisons un algorithme qui prouve l'existence de minima pour Φ , avec une initialisation qui donnera, en général, une convergence rapide vers un éventuel optimum global.

En proposant une double définition de l'AFJ, l'une géométrique, l'autre analytique, nous alourdissons nécessairement notre exposé. L'utilisation de la seule définition analytique permettrait de réduire l'exposé de la méthode à quelques pages (la proposition 11 du 3.2.b. et l'algorithme qui en découle en 3.2.c.). Il nous faut donc motiver l'autre définition et les considérations géométriques parfois délicates qui sont développées en deuxième partie.

Leur principale raison d'être, d'ordre "historique", tient au fait que l'égalité (1) de la proposition 11 a été imaginée dans ce contexte géométrique. Cette raison historique, qui a un intérêt momentané, se double d'une raison "géographique" plus permanente : la géométrisation du problème permet de comparer simplement les

différentes méthodes (françaises ou anglo-saxonnes) et de choisir une initialisation très performante de l'algorithme.

II - LES TABLEAUX A COMPOSANTES PROPORTIONNELLES

Précisons le contenu de la deuxième partie :

- En 2.1 on définit les concepts principaux de l'AFJ. Une application aux données cubiques est proposée. Pour les données non-cubiques, une infrastructure exacte et parfaitement significative n'existe que sous l'hypothèse de tableaux à composantes proportionnelles (elle donne son titre à cette partie).

- Sous cette hypothèse, on compare en 2.2, les infrastructures de différentes méthodes. Pour obtenir des résultats précis, nous partons d'hypothèses encore plus sévères qui sont progressivement relâchées. Il s'agit d'une introduction à l'AFJ car :

- 1) on montre que seule son infrastructure est exacte et significative.
- 2) le Repère Déformant introduit dans cette démonstration, est un concept de base pour la construction de la méthode dans le cas général, faite en troisième partie.

2.1.- Le Repère Equisecteur.

a) Les systèmes équisecteurs.

Définition 1.- Soit $(e^k)_{k \in [m]}$ une base orthonormale de l'espace Euclidien $E^m = (\mathbf{R}^m, I_m)$. On appelle vecteur équisecteur de la base (e^k) le vecteur unitaire de E^m défini par : $e = \frac{1}{\sqrt{m}} \sum_{k \in [m]} e^k$.

De même Δe est la direction équisectrice des directions Δe^k .

Remarques

- Cette définition généralise à \mathbf{R}^n la notion de bissectrice.
- Dans l'espace des variables F , si $\mathbf{1}_n$ désigne le vecteurs de coordonnées égales à 1, l'axe $\Delta \mathbf{1}_n$ est la direction équisectrice des Δf^i et la D -projection d'une variable sur $\Delta \mathbf{1}_n$ définit sa moyenne.
- Les propriétés de ces objets vont être exploitées dans la généralisation suivante de la définition 1.

Définition 2.- Soit $E = (\mathbf{R}^{mq}, I_{mq})$ l'espace Euclidien somme directe orthogonale d'une famille de m espaces Euclidiens de même dimension $E_k = (\mathbf{R}^q, I_q)$, $k \in [m]$. Une base orthonormale $(e_j^k)_{j \in [q]}$ de l'espace E_k plongé dans E , sera notée $(\tilde{e}_j^k)_{j \in [q]}$. On appelle système équisecteur de E le système des vecteurs e_j , $j \in [q]$, équisecteurs des $(\tilde{e}_j^k)_{k \in [m]}$; soit : $\{e_j = \frac{1}{\sqrt{m}} \sum_{k \in [m]} \tilde{e}_j^k, j \in [q]\}$.

Remarques :

- Le système équisecteur $(e_j)_{j \in [q]}$ engendre un sous-espace de E noté E_e .
- Le système de vecteurs $(\tilde{e}_j^k)_{j \in [q], k \in [m]}$ constitue une base orthonormale de E .
- Il faut bien noter que chaque Δe_j est direction équisectrice d'axes $\Delta \tilde{e}_j^k$ situés dans des espaces E_k différents (ils sont supplémentaires dans E). Il n'y a donc pas d'analogie avec les propriétés de $\Delta \mathbf{1}_n$ visibles dans F .

Nous pouvons alors énoncer la proposition suivante :

Proposition 1.-

- 1) $(e_j)_{j \in [q]}$ est une base orthonormale de E_e .
 - 2) Un vecteur quelconque x de E_k est caractérisé par un q -uplet de coordonnées. \tilde{x} désigne ce vecteur plongé dans E et \hat{x} sa \sqrt{m} -projection sur le sous-espace E_e . Dans la base (e_j) de E_e on a $\hat{x} \approx x$ (où \approx signifie que les deux vecteurs ont les mêmes coordonnées).
 - 3) Le système équisecteur, ainsi que le sous-espace E_e dépendent de l'ordre dans lequel sont rangés les vecteurs de chacune des bases $(e_j^k)_{j \in [q]}$, $k \in [m]$.
- 1) E étant muni de la métrique $I_{mq} = I$, on a d'après la remarque précédente :

$$\forall (i, j) \in [q]^2: e_i^l I e_j = \frac{1}{m} \left(\sum_k \tilde{e}_i^k \right)' I \left(\sum_\ell \tilde{e}_j^\ell \right) = \frac{1}{m} \sum_k (\tilde{e}_i^k)' I \tilde{e}_j^k = \delta_{ij}$$

- 2) $\forall k \in [m], \forall x \in E_k$, on a : $x = \sum_{j=1}^q x^j e_j^k$, et $\tilde{x} \in \tilde{E}_k$ s'écrit $\tilde{x} = \sum_j x^j \tilde{e}_j^k$. Sa \sqrt{m} -projection sur Δe_i s'écrit : $\hat{x}^i = \sqrt{m} e_i^l I (\sum_j x^j \tilde{e}_j^k)$, $i \in [q]$, soit :

$$\hat{x}^i = \sqrt{m} \frac{1}{\sqrt{m}} \left(\sum_\ell \tilde{e}_i^\ell \right)' I \sum_j x^j \tilde{e}_j^k = \sum_\ell \sum_j ((\tilde{e}_i^\ell)' I \tilde{e}_j^k) x^j$$

$$\hat{x}^i = x^i, i \in [q] \quad \text{et} \quad \hat{x} \approx x$$

\hat{x} est la \sqrt{m} -projection de \tilde{x} sur le sous-espace E_e .

3) L'ordre dans lequel sont rangés les vecteurs d'une base est arbitraire et on peut construire un système équisecteur et un sous-espace E_e pour chacune des $(q!)^{m-1}$ façons de choisir, sans ordre, q m -uplets de vecteurs du type : $(e_{j_1}^1, e_{j_2}^2, \dots, e_{j_m}^m)$, chacun donnant un vecteur e_j . Ainsi dans $\mathbf{R}^4 = \mathbf{R}^2 \oplus \mathbf{R}^2$, on a la base :

$$\tilde{e}_1^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tilde{e}_2^1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \tilde{e}_1^2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \tilde{e}_2^2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

qui donne deux systèmes équisecteurs :

$$\begin{aligned} \bullet e_1 &= \frac{1}{\sqrt{2}}(\tilde{e}_1^1 + \tilde{e}_1^2) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{et} \quad e_2 = \frac{1}{\sqrt{2}}(\tilde{e}_2^1 + \tilde{e}_2^2) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \\ \bullet f_1 &= \frac{1}{\sqrt{2}}(\tilde{e}_1^1 + \tilde{e}_2^2) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad \text{et} \quad f_2 = \frac{1}{\sqrt{2}}(\tilde{e}_2^1 + \tilde{e}_1^2) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \end{aligned}$$

Ils engendrent deux sous-espaces E_e différents.

Le lecteur pourra vérifier qu'un changement des bases modifie le sous espace E_e sauf si les matrices de changement de base sont identiques dans tous les espaces E_k . ■

b) Signification statistique pour les données cubiques.

Soit un cube de données X_{IJK} . Dans chaque espace $E_k = (\mathbf{R}_{(e_j^k)}^{\bar{p}}, I_{\bar{p}})$, X_{IJK} définit un nuage N_k , noté \tilde{N}_k lorsqu'on le plonge dans $E = \bigoplus E_k = (\mathbf{R}^{m\bar{p}}, I_{m\bar{p}})$ rapporté à la base $(\tilde{e}_j^k)_{j \in [\bar{p}], k \in [m]}$ (pour ce cas particulier, la dimension q des espaces E_k du cas général, devient \bar{p} qui est le nombre de variables de chaque tableau). Considérons un système équisecteur quelconque : $\{e_j = \frac{1}{\sqrt{m}} \sum_k \tilde{e}_j^k, j \in [\bar{p}]\}$.

Corollaire

- 1) les \sqrt{m} -projections \hat{N}_k des nuages \tilde{N}_k sur un sous-espace E_e quelconque définissent une inrastructure exacte. Celle-ci coïncide avec celle qu'on obtiendrait dans $\mathbf{R}^{\bar{p}}$ par superposition des axes $\Delta e_j^k, k \in K$, en les associant comme les \tilde{e}_j^k le sont dans le calcul de e_j
- 2) Il y a $(\bar{p})^{m-1}$ sous-espaces E_e différents et autant d'intrastructures exactes.

■ 1) Chaque infrastructure est exacte car les images euclidiennes suivantes sont équivalentes : $(N_k, E_k, I_{\bar{p}}) \iff (\tilde{N}_k, E, I_{m\bar{p}}) \iff (\hat{N}_k, E_e, I_{\bar{p}})$. La première équivalence a été vue avec l'AFJ1 et la seconde résulte de la proposition précédente.

2) Le dénombrement $(\bar{p}!)^{m-1}$ déjà calculé est aussi le nombre de permutations des indices $j \in J$, à la puissance m car il y a m ensembles J ; divisé par $\bar{p}!$ car les permutations des e_j de E_e définissent le même sous-espace. ■

Proposition 2.-

1) *Il existe un système équisecteur qui possède un sens statistique : c'est celui qui, pour tout $j \in J$, associe dans le calcul des e_j , les axes Δe_j^k , $k \in K$, liés par dualité à la même variable j (indice j attaché à une seule variable).*

2) *L'infrastructure obtenue coïncide avec le positionnement simultané des nuages N_k réalisé dans $\mathbf{R}^{\bar{p}}$ par le Tableau Conjoint Empilé X_e .*

■ 1) Le corollaire précédent précise que l'infrastructure obtenue par \sqrt{m} -projection sur E_e s'obtient également en superposant les axes Δe_j^k qui sont associés dans le calcul des e_j . Une infrastructure obtenue par \sqrt{m} -projection sur un sous-espace E_e quelconque a une existence géométrique mais, en général, aucun sens statistique puisqu'elle superpose sur un seul axe, des axes relatifs à des variables différentes (sauf cas très particuliers). En effet, la dualité stipule qu'un axe quelconque de l'espace des individus est lié à une variable **déterminée**. Seul le cas mentionné ci-dessus permet une interprétation statistique puisqu'il associe dans le calcul de chaque e_j , et donc superpose, des axes liés à la même variable statistique $j \in J$.

2) Le Tableau Conjoint Empilé $X_e = \begin{pmatrix} X_1 \\ \dots \\ \vdots \\ \dots \\ X_m \end{pmatrix}$ définit le nuage $N = \bigcup_{k \in K} N_k$

qui réalise un positionnement simultané dans $\mathbf{R}^{\bar{p}}$ des nuages N_k . Celui-ci s'obtient également en superposant les axes Δe_j^k associés pour tout k , $k \in K$, à une même variable j , $j \in J$. ■

Définition 3.- *Ce système équisecteur qui possède un sens statistique sera nommé Repère Equisecteur de E , et le sous-espace qu'il engendre s'appellera sous-espace Equisecteur de E , notée E_E .*

La dualité nous permet aussi de définir une instructure de données non-cubiques.

c) Le Repère Equisecteur pour les données non-cubiques.

Définition 4.- Dans l'espace choisi pour représenter simultanément les nuages d'individus, l'instructure qui respecte les propriétés de dualité s'obtient en superposant les axes des différents nuages liés (par dualité) à une même variable ou plus généralement à une même direction de l'espace des variables. Une telle instructure est dite **significative**.

Dans cette définition, le premier cas envisagé (axes associés à une même variable j) correspond à l'instructure des données cubiques. Celle-ci sera maintenant considérée comme l'instructure de la méthode du Tableau Conjoint Empilé. Le second cas définit l'instructure de données non-cubiques dans le cas de composantes (combinaisons linéaires de variables initiales) proportionnelles. En effet des composantes proportionnelles occupent une même direction de l'espace des variables $F = (\mathbf{R}^n, D)$. Nous allons donc chercher des triplets (Y_k, I, D) équivalents aux triplets initiaux (X_k, M_k, D) vérifiant, si possible, cette propriété. Soit $q_k = \text{rang} X_k$ avec $X_k(n \times p_k)$. Si $q_k < p_k$, il est préférable de partir des triplets équivalents (C_k, I, D) avec $C_k(n \times q_k)$ définissant un positionnement des individus dans l'espace $E_k^{q_k} = (\mathbf{R}_{(u_k^t)}^{q_k}, I_{q_k})$, $k \in K$.

Définition 5.- La famille de triplets (X_k, M_k, D) est à composantes proportionnelles si pour tout k , $k \in K$, il existe une isométrie H_k de $E_k^{q_k}$ dans $E_k^q = (\mathbf{R}^q, I_q)$ avec $q = \sup_k(q_k)$, telle que $Y_k = C_k H_k^t$ vérifie : $Y_k = Y A_k$ avec A_k diagonale ($q \times q$) et $Y(n \times q)$.

Nous dirons plus simplement : tableaux à composantes proportionnelles. De tels tableaux vérifient INDSCAL avec des semi-métriques car $W_k = Y A_k^2 Y'$.

Remarque : Si $p_k = \dim E_k \leq q$, on peut compléter l'espace E_k par $q - p_k$ directions telles que l'espace obtenu E_k^c soit égal à E_k^q . Sinon on pose : $E_k^c = E_k$. On a ainsi : $E_k \subset E_k^q = E_k^c$ ou $E_k^q \subset E_k = E_k^c$, $k \in K$.

Proposition 3.-

Pour une famille de tableaux à composantes proportionnelles :

- 1) Il existe un Repère Equisecteur dans $E^{p^c} = \bigoplus E_k^c$.

2) En considérant les \sqrt{m} -projections des nuages \tilde{N}_k sur le sous-espace Equisecteur E_E de E^{p^c} , l'AFJ définit une infrastructure exacte et significative appelée **Intrastructure Equisecteur**. Elle s'écrit : $\hat{Y}_k = Y_k$, $k \in K$, et pour le nuage compromis : $\hat{Y} = \sum_k \rho_k Y_k = Y \sum_k \rho_k A_k$.

■ 1) Il suffit de remarquer l'interprétation géométrique suivante, particulièrement simple si $M_k = I_{p_k}$:

Le tableau $X_k(n \times p_k)$ positionne le nuage N_k dans $E_k = (\mathbf{R}_{(e^k)}^{p_k}, I_{p_k})$,

$C_k(n \times q_k)$ définit le même nuage par ses coordonnées dans le sous-espace $E_k^{q_k}$ de $E_k \subset E_k^c$.

$Y_k(n \times q)$ définit le même nuage par ses coordonnées dans le sous-espace E_k^q de E_k^c avec $E_k^q = (\mathbf{R}_{(v^k)}^q, I_q)$.

Ce sont des changements de base orthonormale dans E_k^c avec variation du nombre des composantes nécessaires à restituer N_k dans E_k^c . Alors l'espace $E^{mq} = \bigoplus_{\perp} E_k^q$ est un sous-espace de $E^{p^c} = \bigoplus_{\perp} E_k^c$ et le Repère Equisecteur de E^{mq} est aussi le Repère Equisecteur de E^{p^c} .

2) D'après les définitions et propositions précédentes, les \sqrt{m} -projections des nuages \tilde{N}_k positionnés dans E^{mq} , sur le sous-espace Equisecteur E_E définissent une infrastructure exacte et significative. $E^{mq} \subset E^{p^c}$ donc l'AFJ est une méthode de \sqrt{m} -projection des nuages \tilde{N}_k sur le sous-espace Equisecteur E_E de E^{p^c} . ■

Remarques :

1) Si un espace E_k a été complété à la dimension q , on distingue $E^{p^c} = \bigoplus_{\perp} E_k^c$ et $E^p = \bigoplus_{\perp} E_k$ pour constater que :

- le sous-espace Compromis E_C vérifie : $E_C \subset E^p \subset E^{p^c}$,
- le sous-espace Equisecteur E_E vérifie : $E_E \subset E^{mq} \subset E^{p^c}$ mais pas : $E_E \subset E^p$.
Donc la comparaison de méthodes qui projettent les nuages \tilde{N}_k sur E_C (AFJ1) ou sur E_E (AFJ) peut se faire dans E^{p^c} et non dans E^p . (L'hypothèse $p_k \leq q$ ou rang $X_k = p_k$ entraîne $E^{p^c} = E^{mq}$).
- l'exposé de l'AFJ seule se fait dans E^{mq} . C'est la comparaison des méthodes qui impose d'introduire E^p et E^{p^c} .

2) Le sous-espace E_E de E^{p^c} dépend des tableaux X_k , en particulier E^{p^c} dépend des dimensions p_k et des rangs q_k de ces tableaux.

- 3) L'AFJ déterminera l'Intrastructure Equisecteur (I.E.) sans préciser la position de E_E sauf dans les rares cas particuliers où il coïncide avec E_C .

2.2.- Introduction à l'AFJ.

Avant d'exposer en troisième partie, la façon dont l'AFJ détermine l'I.E. pour les tableaux à composantes proportionnelles (ou même dans le cas général) nous allons vérifier que les méthodes usuelles ne donnent pas des intrastructures exactes et significatives. En partant d'hypothèses très strictes, qui s'élargiront progressivement nous montrerons les lacunes de ces intrastructures.

a) Les tableaux proportionnels.

Définition 6.- La famille de triplets (X_k, M_k, D) constitue des tableaux de composantes proportionnels (ou tableaux proportionnels) si les tableaux des composantes principales vérifient : $C_k = a_k C$ avec $a_k \in \mathbf{R}_+^*$, $k \in K$ et C ($n \times q$).

Sous cette hypothèse (opérateurs proportionnels) les nuages sont homothétiques donc très simples à comparer. Pourtant, nous pouvons montrer que certaines méthodes proposent des intrastructures inexactes !

Proposition 4.-

Pour les tableaux proportionnels qui vérifient $C_k = a_k C$, $k \in K$:

- 1) l'AFJ1 à poids égaux et Statis définissent la même intrastructure inexacte.
- 2) l'AFJ1 avec les poids $\rho_k = \frac{1}{m a_k}$ (AFM à un facteur près), l'analyse procuste et la superposition donnent l'Intrastructure Equisecteur $\hat{C}_k = C_k$, $k \in K$.

■ Explicitons les calculs relatifs à l'AFJ1 puis l'AFJ2

• Pour l'AFJ1 on a : $W_0 = \sum_k \rho_k^2 W_k = \sum_k \rho_k^2 a_k^2 W$ avec $W = CIC'$.

Adoptons une normalisation des a_k et de C telle que $\sum_k \rho_k^2 a_k^2 = \frac{1}{m}$.

(C, Λ) sont les éléments propres de WD , avec $C'DC = \Lambda$, et (C_0, Λ_0) ceux de W_0D , avec $C_0'DC_0 = \Lambda_0$. Comme $W_0D = \frac{1}{m}WD$, C et C_0 sont proportionnels, donc : $\frac{1}{m}WDC = C\Lambda_0 = \frac{1}{m}C\Lambda$, et : $\Lambda = m\Lambda_0$, $C = \sqrt{m}C_0$.

L'introduction du coefficient multiplicateur \sqrt{m} dans l'Intrastructure Compromis (I.C.) de l'AFJ1 est ainsi doublement justifiée :

- L'I.E. retient des \sqrt{m} -projections.

- En considérant les \sqrt{m} -projections sur le sous-espace compromis E_C , si $C_k = a_k C$ avec $\sum_k \rho_k^2 a_k^2 = \frac{1}{m}$, le tableau moyen C_0 obtenu par diagonalisation de $W_0 D$ est remplacé par le tableau compromis : $\hat{X} = C = \sqrt{m} C_0$.

Pour des tableaux proportionnels l'I.C. s'écrit donc :

$$\begin{aligned} \hat{X}_k &= \sqrt{m} \rho_k W_k D C_0 \Lambda_0^{-1} \\ &= \sqrt{m} \rho_k a_k^2 W_k \left(\frac{C}{\sqrt{m}} \right) m \Lambda^{-1} = m \rho_k a_k^2 C, \quad k \in K \end{aligned}$$

et $\hat{X} = C = \sqrt{m} C_0$.

• Pour l'AFJ2 on a $W_{II} = \sum_k \rho_k W_k$, et les "poids" ρ_k de Statis sont proportionnels aux coefficients a_k^2 : $\rho_k = \frac{a_k^2}{\beta}$, $k \in K$. Ils vérifient $\sum_k \rho_k^2 = 1$, donc $\sum_k a_k^4 = \beta^2$. Comme $W_k = a_k^2 W$, les éléments propres (C_{II}, Λ_{II}) de $W_{II} D = \sum_k \frac{a_k^2}{\beta} a_k^2 W D = \beta W D$, vérifient : $\Lambda_{II} = \beta \Lambda$ et $C_{II} = \sqrt{\beta} C$.

Statis normalise W_{II} pour que : $\|W_{II}^*\| = \frac{1}{m} \sum_k \|W_k\|$, $\|W_{II}^*\| = \frac{1}{m} \sum_k a_k^2 \|W\|$ et la normalisation $\sum_k a_k^2 = m$ conduit à : $W_{II}^* = W$ (et non $W_{II} = \beta W$). L'intrastructure s'écrit : $\hat{X} = C$ et $\hat{X}_k = W_k D C \Lambda^{-1} = a_k^2 C$, $k \in K$.

On a donc les résultats suivants :

1) Si $C_k = a_k C$, $k \in K$ avec $\sum_k a_k^2 = m$, l'intrastructure de Statis s'écrit : $\hat{X} = C$ et $\hat{X}_k = a_k C$, $k \in K$ (cas de l'interstructure non normée).

Dans l'AFJ1 à poids égaux, $\rho_k = \frac{1}{m}$ et $\sum_k \rho_k^2 a_k^2 = \frac{1}{m}$ entraînent $\sum_k a_k^2 = m$ d'où l'Intrastructure Compromis : $\hat{X} = C$ et $\hat{X}_k = a_k C$, $k \in K$.

Ces deux intrastructures identiques sont inexactes car elles grossissent les gros nuages ($a_k > 1$) et diminuent les petits. Mais elles sont significatives car elles superposent des axes (principaux) de l'espace des individus liés par dualité à des variables proportionnelles.

2) Pour les poids $\rho_k = \frac{1}{m a_k}$ qui correspondent à la normalisation : $\sum_k \frac{1}{a_k} = m$, on a l'I.C. : $\hat{X} = C$ et $\hat{X}_k = a_k C = C_k$, $k \in K$. Elle coïncide avec l'I.E. qui s'écrit $\hat{C} = C$ et $\hat{C}_k = C_k$, $k \in K$ (Pour ne pas alourdir les notations on a gardé le symbole C quoique ce tableau ne soit pas égal à celui du 1)).

Il est évident que la superposition des axes principaux donne la même intrastructure. Pour l'analyse procuste, la propriété a été démontrée sous l'hypothèse plus large qui va suivre. ■

b) Les tableaux à composantes principales proportionnelles.

Définition 6 bis.- Les tableaux à composantes principales proportionnelles sont tels que : $C_k^q = CA_k$, avec A_k diagonale ($A_k = \text{diag}(a_{kj})$, $j \in [q]$), $k \in K$. Le tableau C_k^q ($n \times q$) est le tableau C_k ($n \times q_k$) complété par $q - q_k$ colonnes de 0, dont les colonnes sont convenablement réordonnées.

En effet la composante c_k^2 de C_k peut être proportionnelle, par exemple, à c^1 . Elle passe alors en première colonne de C_k^q . Si $q_k < q$, il existe $q - q_k$ indices j , $j \in [q]$, tels que $a_{kj} = 0$ et $(C_k^q)^j = \mathbf{0}_{\mathbf{R}^n}$.

Les composantes principales ne sont plus globalement proportionnelles. Elles le sont individuellement. Donc les nuages ne sont plus globalement homothétiques mais ils le sont "axe par axe".

On adopte la normalisation : $\sum_k \rho_k^2 A_k^2 = \frac{1}{m} I_q$, et on suppose que $C = (c^1 \quad \vdots \quad c^2 \quad \vdots \quad \dots \quad \vdots \quad c^q)$ vérifie : $\|c^1\| \geq \|c^2\| \geq \dots \geq \|c^q\|$.

Ici presque toutes les méthodes offrent une intrastucture inexacte !

Proposition 5.-

Pour des tableaux à composantes principales proportionnelles :

- 1) *l'Intrastucture Compromis (I.C.) de l'AFJ1 : $\hat{X} = C$ et $\hat{X}_k = m\rho_k C_k^q A_k$, $k \in K$ est inexacte (elle déforme les nuages par des homothéties opérant dans les directions des axes principaux). Avec la normalisation adoptée, C est bien le tableau compromis du 1.2.a.*
- 2) *S'il y a proportionnalité de la première composante des différents nuages (si $\forall k \in K, \forall j \in [q]$, on a : $\|a_{k_1} c^1\| \geq \|a_{k_j} c^j\|$), alors les poids $\rho_k = \frac{1}{m a_{k_1}}$ font coïncider le premier axe du nuage compromis de l'AFJ1 avec l'axe 1 du Repère Equisecteur (R.E.) de l'AFJ. Les coordonnées sur l'axe 1 de l'Intrastucture Compromis sont exactes. C'est l'AFM à un facteur près, elle n'est donc exacte que pour la première composante de chaque nuage et sous les hypothèses considérées.*
- 3) *S'il y a proportionnalité des composantes principales de même rang dans les différents nuages ($\forall k \in K : \|a_{k_1} c^1\| \geq \|a_{k_2} c^2\| \geq \dots \geq \|a_{k_q} c^q\|$) alors l'intrastucture "de superposition" coïncide avec l'Intrastucture Equisecteur*

(I.E.) de l'AFJ. Sinon, elle est exacte mais totalement dépourvue de signification.

4) L'Intrastructure Procuste coïncide avec l'Intrastructure Equisecteur de l'AFJ qui s'écrit : $\hat{C}_k = C_k^q = CA_k$, $k \in K$ et $\hat{C} = \sum_k \rho_k C_k^q$.

5) La part d'inertie de chaque nuage conservée dans l'Intrastructure Equisecteur limitée à s axes ($s \leq q$) vaut : $p_{sk} = (\sum_{j=1}^s a_{kj}^2 \lambda_j) / (\sum_{j=1}^q a_{kj}^2 \lambda_j)$, $k \in K$. Pour l'ensemble des nuages c'est aussi la part de l'inertie conservée pour le nuage compromis : $p_s = (\sum_{j=1}^s \lambda_j) / (\sum_{j=1}^q \lambda_j)$.

■ • Certaines propositions découlent de façon évidente de la proposition 1) qui sera seule démontrée. La proposition 4) n'est pas immédiate, sa démonstration pourra être trouvée ailleurs (Lechevalier 1990).

• On a : $C_k^q = CA_k$, $k \in K$ avec $\sum_k \rho_k^2 a_{kj}^2 = \frac{1}{m}$, $j \in [q]$. Soient (C, Λ) les éléments propres de $WD = CC'D$, (C_k^q, Λ_k^q) ceux de $W_k D$, $k \in K$.

$C_k^q = CA_k \Rightarrow \lambda_j^k = \|c_k^j\|^2 = a_{kj}^2 \lambda_j$, $j \in [q]$ (ou $\Lambda_k^q = \Lambda A_k^2$), $k \in K$.

Comme C ($n \times q$) est de rang q : $a_{kj} = 0 \iff \lambda_j^k = 0$, et

$W_k DC_k^q = C_k^q \Lambda_k^q \iff W_k DCA_k = CA_k \Lambda_k^q \iff W_k DC = C \Lambda_k^q$.

Alors $W_0 DC = \sum_k \rho_k^2 W_k DC = C(\sum_k \rho_k^2 \Lambda_k^q) = C \sum_k \rho_k^2 A_k^2 \Lambda = C \Lambda \frac{1}{m} = C \Lambda_0$.

Les éléments propres (C_0, Λ_0) de $W_0 D$ vérifient donc $\Lambda_0 = \Lambda/m$ et $C_0 = C/\sqrt{m}$.

• L'I.C. de tableaux à composantes principales proportionnelles s'écrit alors :

$$\hat{X} = \sqrt{m} C_0 = C$$

et $\hat{X}_k = \sqrt{m} \rho_k W_k DC_0 \Lambda_0^{-1} = \sqrt{m} \rho_k W_k D \frac{C}{\sqrt{m}} m \Lambda^{-1}$

soit $\hat{X}_k = m \rho_k CA_k^2 = m \rho_k C_k^q A_k$, $k \in K$

• L'I.C. avec des poids égaux s'écrit : $\hat{X} = C$ et $\hat{X}_k = C_k^q A_k$, $k \in K$ et les déformations subies par les nuages sont évidentes. ■

Concluons en soulignant que sous l'hypothèse de composantes principales proportionnelles seules l'AFJ et l'analyse procuste proposent une intrastructure exacte et significative. Nous verrons que pour des composantes (quelconques) proportionnelles l'intrastructure procuste n'est plus parfaitement significative (voir proposition 12). Examinons donc le cas des composantes proportionnelles.

c) Le Repère Déformant.

Définition 7.- Soient les tableaux X_k à composantes proportionnelles avec les poids ρ_k et $(v_j^k)_{j \in [q]}$ la base de E_k^q telle que $Y_k = Y A_k$, $k \in K$. On appelle **Repère Déformant (R.D.)** de E^{p_c} le système de vecteurs défini par : $\{v_j = \sqrt{m} \sum_{k \in K} \rho_k a_{kj} \tilde{v}_j^k, j \in [q]\}$, et l'hypothèse de normalisation $\sum_k \rho_k^2 A_k^2 = \frac{1}{m} I_q$. Le R.D. engendre le sous-espace Déformant E_D et les \sqrt{m} -projections des nuages \tilde{N}_k sur celui-ci définissent l'**Intrastructure Déformante**.

Le Repère Déformant est un Repère Equisecteur pondéré. Il nous donne une expression simplifiée de l'**Intrastructure Compromis** qui éclaire la façon dont elle est inexacte quand les tableaux sont à composantes (quelconques) proportionnelles. Dans le cas général, l'I.D. constitue un trait d'union entre l'I.C. qui est connue, et l'I.E. que l'on recherche.

Proposition 6.-

Pour une famille de tableaux X_k à composantes proportionnelles :

- 1) $(v_j)_{j \in [q]}$ est base orthonormale de E_D et un vecteur quelconque $\tilde{y} = (y^1, \dots, y^q)$ de $E_{k(\tilde{v}_j^k)}^q$ a pour projection sur $E_{D(v_i)}$ le vecteur $\hat{y} = \sqrt{m}(\rho_k a_{k1} y^1, \dots, \rho_k a_{kq} y^q)$.
- 2) L'**Intrastructure Déformante** s'écrit : $\hat{Y}_k = m \rho_k Y A_k^2$, $k \in K$, et le compromis $\hat{Y} = \sum_k \rho_k \hat{Y}_k$ vaut Y .
- 3) Les sous-espaces compromis E_C et Déformant E_D sont confondus. Donc les **Intrastructures Compromis (I.C.)** et **Déformante (I.D.)** obtenues par \sqrt{m} -projection sur E_C et E_D sont identiques. Leurs expressions diffèrent car elles ne sont pas rapportées à la même base (sauf si on a les composantes principales proportionnelles).

■ 1) • $\forall (i, j) \in [q]^2 : v_j' I v_i = \sqrt{m} (\sum_k \rho_k a_{kj} \tilde{v}_j^k)' I \sqrt{m} (\sum_\ell \rho_\ell a_{\ell i} \tilde{v}_i^\ell)$

avec $(\tilde{v}_j^k)' I \tilde{v}_i^\ell = \begin{cases} 1 & \text{si } k = \ell \text{ et } j = i \\ 0 & \text{sinon} \end{cases}$

d'où : $v_j' I v_i = \begin{cases} 0 & \text{si } j \neq i \\ m \sum_k \rho_k^2 a_{kj}^2 = 1 & \text{si } j = i \end{cases}$

• $\tilde{y}' I v_i = (\sum_{j \in [q]} y^j \tilde{v}_j^k)' I \sqrt{m} \sum_{\ell \in K} \rho_\ell a_{\ell i} \tilde{v}_i^\ell$

d'où $\hat{y}^i = \tilde{y}' I v_i = \sqrt{m} \rho_k a_{ki} y^i$

et $\hat{y} = \sqrt{m}(\rho_k a_{k1} y^1, \dots, \rho_k a_{kq} y^q)$

Pour des poids égaux, la \sqrt{m} -projection de \tilde{y} sur E_D s'écrit :

$$\hat{y} = (a_{k1}y^1, a_{k2}y^2, \dots, a_{kq}y^q).$$

2) L'I.D. s'obtient par \sqrt{m} -projection des nuages \tilde{N}_k sur E_D et donc les formules d'intrastructure découlent immédiatement de ce qui précède :

$\hat{Y}_k = m\rho_k Y_k A_k = m\rho_k Y A_k^2$, $k \in K$, et $\hat{Y} = \sum_k \rho_k \hat{Y}_k = Y$, du fait de la normalisation : $\sum_k \rho_k^2 a_{kj}^2 = \frac{1}{m}$, $j \in [q]$.

3) Montrons d'abord $E_C \subset E_D$, ou encore que tous les individus du nuage compromis (qui constituent un système de générateurs de E_C) sont dans E_D . On raisonne dans E^{mq} rapporté à la base $(\tilde{v}_j^k)_{j \in [q], k \in K}$. Dans E_k^q on a $Y_k = Y A_k$, $k \in K$ d'où les coordonnées de l'individu \tilde{y}_i^k dans la base $(\tilde{v}_j^k)_{j \in [q], k \in K}$:

$$\tilde{y}_i^k = \sum_{j \in [q]} y_i^{jk} \tilde{v}_j^k + 0, \text{ avec } y_i^{jk} = a_{kj} y_i^j.$$

Le i -ème individu compromis \tilde{y}_i , $i \in I$, est le \sqrt{m} -barycentre des \tilde{y}_i^k , $k \in K$, et :

$$\tilde{y}_i = \sqrt{m} \sum_k \rho_k \tilde{y}_i^k = \sqrt{m} \sum_k \rho_k \sum_j a_{kj} y_i^j \tilde{v}_j^k = \sum_j y_i^j \sqrt{m} \sum_k \rho_k a_{kj} \tilde{v}_j^k = \sum_j y_i^j v_j.$$

Tous les individus compromis sont dans E_D . Comme $\text{Dim} E_C = \text{Dim} E_D = q$, on a bien $E_C = E_D$. L'I.C. et l'I.D. sont donc confondues mais l'expression de l'I.C. : $\hat{X}_k = m\rho_k W_k D C \Lambda^{-1}$ est relative à la base $(u_j)_{j \in [q]}$ portant les axes principaux du compromis (elle sera nommée Repère Compromis dans la définition 9), et l'I.D. à la base $(v_j)_{j \in [q]}$ (Repère Déformant) de $E_D = E_C$. Soit $Q = (u_j)_{(v_j)}$ matrice des coordonnées des u_j dans la base v_j . Alors $Y = C Q'$ et $\hat{X}_k = \hat{Y}_k Q$, $k \in K$. ■

Définition 8.- *L'Intrastructure Déformante exprimée dans la base liée aux axes principaux du nuage moyen sera nommée Intrastructure Déformante Principale (I.D.P.) et cette base le Repère Déformant Principal (R.D.P.) ; sous l'hypothèse considérée l'Intrastructure Compromis est identique à l'I.D.P. de même que le Repère Compromis (R.C.) coïncide avec le R.D.P. (u_j)*

Conclusion.- L'introduction du Repère Déformant nous a permis de montrer plusieurs propriétés de l'Intrastructure Compromis de tableaux à composantes proportionnelles :

- Elle est déformante, avec, si $\rho_k = \frac{1}{m}$: $\hat{Y}_k = Y_k A_k$, $k \in K$ et $\hat{Y} = Y$. Les déformations sont le fait d'homothéties de rapport a_{kj} et de centre 0 dans chaque direction Δv_j .

- Elle est significative puisqu'on superpose les axes Δv_j^k des différents nuages, liés par dualité à la même direction Δy^j de l'espace des variables, mais elle est inexacte.

L'Intrastructure Equisecteur de l'AFJ : $\hat{Y}_k = Y_k$, $k \in K$ et $\bar{Y} = \sum_k \rho_k \hat{Y}_k$ est exacte et significative car elle superpose les mêmes axes Δv_j^k . La différence est claire et bien maîtrisée : l'I.C. s'obtient en \sqrt{m} -projetant les nuages sur le sous-espace E_D engendré par le R.D., alors que l'I.E. est une \sqrt{m} -projection sur E_E engendré par le R.E.

L'appellation "Repère Déformant" ne doit pas être mal interprétée. Tous les repères (considérés comme bases de sous-espaces particuliers) sont déformants, y compris le R.E. qui multiplie les nuages par $1/\sqrt{m}$. Ce nom précise que les déformations subies par les différents nuages sont encore très simples. Pour des tableaux quelconques, les déformations peuvent être, selon les méthodes (et donc selon les sous-espaces de projection), beaucoup plus complexes.

Dans cette partie géométrique, nous avons montré la possibilité de définir une intrastructure exacte et significative de tableaux à composantes proportionnelles. Par ailleurs, nous avons établi que les méthodes classiques donnent une intrastructure inexacte ou peu significative sauf dans des conditions très restrictives. Cette démarche permet maintenant de proposer une caractérisation des tableaux à composantes proportionnelles dont la mise en œuvre algorithmique donne une solution au problème de l'AFJ, y compris dans le cas général.

III - L'AFJ D'UNE FAMILLE DE TABLEAUX QUELCONQUES

3.1.- Caractérisation des tableaux à composantes proportionnelles.

Rappelons que l'hypothèse des tableaux à composantes proportionnelles (définition 5) permet la vérification du modèle INDSCAL avec des semi-métriques. Ce modèle recherche pour une famille de tableaux $W_k(n \times n)$, un tableau $Y(n \times r)$ et des (semi)-métriques diagonales D_k tels que $W_k = YD_kY'$, $k \in K$. En général, le modèle INDSCAL ne se vérifie pas exactement. Les méthodes existantes recherchent alors $Y(n \times r)$ et D_k , $k \in K$, tels que $\sum_k \|W_k - YD_kY'\|_{\varphi_{D,D}}^2$ soit minimale. On peut le voir dans Carroll et Chang (1970), Tucker (1972), De Leeuw et Pruzanski (1978) ou Lafaye *in* SAD (1985). Ici, nous minimiserons la

fonction $\sum_k \|Y_k - YA_k\|_{\varphi_D}^2$, beaucoup plus significative pour la recherche d'une infrastructure.

En fait, nous allons travailler sur une autre fonction, qui lui est égale à une constante près, mais beaucoup plus simple à minimiser. Cette autre fonction a été trouvée en recherchant une nouvelle caractérisation des tableaux à composantes proportionnelles qui complète les caractérisations des tableaux vérifiant IDIOSCAL et INDSCAL énoncées par Glaçon (1981) ou de Leeuw et Heiser (1982). Cette caractérisation traduit certaines propriétés géométriques des Repères Compromis et Déformant.

a) Propriétés du Repère-Compromis.

Considérons les espaces : $E^p = \bigoplus E_k$, $E^{mq} = \bigoplus E_k^q$ et $E^{pc} = \bigoplus E_k^c$ définis dans la proposition 3, et reprenons dans E^p , l'AFJ1 qui définit l'I.C., sous une autre forme que celle exposée en 1.2. a).

- Dans celle-ci, on considérait les triplets (X_k, M_k, D) et $E_{(\tilde{e}_j^k)}^p$ c'est-à-dire E^p rapporté à la base $(\tilde{e}_j^k)_{k \in K, j \in [p_k]}$.
- On prend maintenant les triplets équivalents (C_k, I_{q_k}, D) et $E_{(\tilde{u}_j^k)}^{\Sigma q_k} \subset E^p$ avec les conventions suivantes :

- $(u_j^k)_{j \in [q_k]}$ est dite la base principale du sous-espace $E_k^{q_k}$ de E_k car elle contient les vecteurs M_k -orthonormés qui portent les axes principaux du nuage N_k .
- elle peut être complétée pour donner $(u_j^k)_{j \in [p_k]}$ base de E_k ou $(\tilde{u}_j^k)_{j \in [q]}$ base de E_k^q . Le tableau C_k ($n \times q_k$) devient alors $C_k^0 = (C_k \vdots 0)$ ($n \times p_k$) ou $C_k^c = (C_k \vdots 0)$ ($n \times q$). Ainsi les propriétés obtenues dans $E^{\Sigma q_k}$ s'étendront à E^p et E^{mq} .

Définition 9.- On appelle Repère Compromis (R.C.) la base $(u_j)_{j \in [r]}$ du sous-espace compromis E_C de $E^{\Sigma q_k}$ formée des vecteurs orthonormés qui portent les axes principaux d'inertie du nuage compromis N défini par le tableau $\sqrt{m}C_\rho = \sqrt{m} \sum_k \rho_k \tilde{C}_k$, de rang r .

Cette définition crée une similitude des notations pour les sous-espaces E_E , E_D , E_C et leurs bases appelées R.E., R.D., R.C..

Rappelons que $C_0(n, r)$ désigne le tableau des composantes principales de C_ρ ou de X_ρ , et que $C = \sqrt{m}C_0$ est le tableau compromis.

Proposition 7.-

Supposons que $\rho_k = \frac{1}{m}$, $k \in K$.

- 1) Les \sqrt{m} -projections des bases principales sur le Repère Compromis sont les lignes du tableau $P_k = C'_k D C \Lambda^{-1}$, $k \in K$.
- 2) $C_k = C P'_k$, $k \in K$: chaque nuage N_k est le \sqrt{m} -projeté du nuage compromis N sur le sous-espace E_k .

■ 1) Les triplets équivalents (X_k, M_k, D) et (C_k, I, D) définissent le même tableau W_k , $k \in K$, donc le même "compromis" $W_0 = \sum_k \rho_k^2 W_k$ et le même tableau $C(n \times r)$ des composantes du nuage compromis avec $C = \sqrt{m} \cdot C_0$.

(C_0, Λ_0) sont les éléments propres de $W_0 D$ et on a, en ACP, les coordonnées des vecteurs $(u_j)_{j \in [r]}$ avec : $U = C'_\rho D C_0 \Lambda_0^{-1}$

$$U = \frac{1}{m} \begin{pmatrix} C'_1 \\ \dots \\ C'_2 \\ \dots \\ \vdots \\ \dots \\ C'_m \end{pmatrix} D \frac{C}{\sqrt{m}} m \Lambda^{-1} = \frac{1}{\sqrt{m}} \begin{pmatrix} P_1 \\ \dots \\ P_2 \\ \dots \\ \vdots \\ \dots \\ P_m \end{pmatrix} \text{ avec } P_k = C'_k D C \Lambda^{-1}, k \in K.$$

Le tableau $U(\sum_k q_k \times r)$ contient en élément $(\sum_{l=1}^{k-1} q_l + i, j)$, la coordonnée du vecteur u_j sur l'axe $\Delta \tilde{u}_i^k$ qui vaut : $u'_j I \tilde{u}_i^k = (\tilde{u}_i^k)' I u_j$. C'est aussi la coordonnée du vecteur \tilde{u}_i^k sur l'axe Δu_j . On a : $P_k = \sqrt{m} \tilde{U}'_k I U$, $k \in K$. Ses lignes sont les \sqrt{m} -projections des \tilde{u}_i^k sur la base (u_j) .

Ce résultat établi dans $E^{\sum q_k}$ s'étend à E^p et E^{mq} en prenant C_k^0 ou C_k^c .

On l'obtient très simplement dans $E^p = (\mathbf{R}^p, \mathcal{M})$ en rappelant que

$U = X'_\rho D C_0 \Lambda_0^{-1}$ et $\tilde{U}'_k = (0 \vdots \dots \vdots \Lambda_k^{-1} C'_k D X_k \vdots \dots \vdots 0)$, et en remarquant que :

$$\sqrt{m} \tilde{U}'_k \mathcal{M} U = \sqrt{m} \Lambda_k^{-1} C'_k D X_k M_k \frac{1}{m} X'_k D \frac{C}{\sqrt{m}} m \Lambda^{-1} = C'_k D C \Lambda^{-1} = P_k, k \in K.$$

2) Dans notre première démonstration du 1) on peut écrire la formule de reconstitution des données suivante : $C_\rho = (C_0 \Lambda_0^{-1/2}) \Lambda_0^{1/2} U' = \frac{1}{\sqrt{m}} C U'$,

soit $\frac{1}{m} (C_1 \vdots C_2 \vdots \dots \vdots C_m) = \frac{1}{m} C (P'_1 \vdots \dots \vdots P'_m)$.

Donc $C_k = C P'_k$, avec $P_k = C'_k D C \Lambda^{-1}$, $k \in K$.

Soit \hat{x} la \sqrt{m} -projection d'un x quelconque de E_C sur $E_k^{q_k}$; on a $\hat{x} = P_k x$.
 En effet, $\forall x \in E_C : x = \sum_{j=1}^r x^j u_j$ et sa \sqrt{m} -projection sur $\Delta \tilde{u}_i^k$ s'écrit :
 $\hat{x}^i = \sqrt{m}(\tilde{u}_i^k)' I x = \sqrt{m}(\tilde{u}_i^k)' I \sum_j x^j u_j = (P_k)_i x$, et $\hat{x} = P_k x$.
 P_k est un \sqrt{m} -projecteur de E_C sur $E_k^{q_k}$. ■

b) Propriétés du Repère Déformant.

Le Repère Déformant nous permet de trouver ici une caractérisation des tableaux à composantes proportionnelles qui est à la base de notre méthode.

Supposons que les tableaux X_k soient à composantes proportionnelles et considérons le Repère Déformant de $E^{mq} = \bigoplus_k E_k^q$. Dans chaque espace E_k^q on dispose des bases orthonormales (\hat{u}_j^k) et (v_j^k) et on note $Q_k = (\hat{u}_j^k)_{(v_j^k)}$ le tableau des coordonnées des (\hat{u}_j^k) dans la base des (v_j^k) , d'où $Y_k' = Q_k (C_k^c)'$, $k \in K$. De même dans le sous-espace Déformant ou Compromis ($E_C = E_D$) on a : $Q = (u_j)_{(v_j)}$, et $Y' = QC' \iff Y = CQ' \iff C = YQ$.

Proposition 8.-

Soient, pour $k \in K$, les triplets (X_k, M_k, D) équivalents à (C_k, I_{q_k}, D) , $C(n \times r)$ et les tableaux $P_k^c = (C_k^c)' D C A^{-1}$ de dimensions $(q \times r)$.

- 1) IDIOSCAL admet une solution exacte à la Condition Nécessaire et Suffisante que : $r = q = q_k$, (ou $r = q = \sup_k q_k$ si on accepte des semi-métriques).
- 2) INDSCAL admet une solution exacte à la Condition Nécessaire et Suffisante que : $r = q = q_k$ (ou $r = q = \sup_k q_k$ pour des semi-métriques) et $P_k^c = Q_k' A_k Q$ avec Q et Q_k orthogonales $(q \times q)$, A_k diagonale (s.)d.p., $k \in K$. Le cas $(r = q = \sup_k q_k)$ et $P_k^c = Q_k' A_k Q$ caractérise les tableaux à composantes proportionnelles.

■ 1) On utilise la proposition 7-3) :

$C_k = C P_k^c \Rightarrow W_k = C P_k^c P_k C' = C O_k C'$ avec $O_k (r \times r)$ s.d.p., $k \in K$. Donc IDIOSCAL admet toujours une solution exacte avec des semi-métriques. On peut exiger, en notant $q = q_0 = \sup_k q_k$, qu'au moins une matrice O_{k_0} soit une métrique. Rang $C = r \geq q_k = \text{rang } W_k = \text{rang } O_k$ entraîne la CNS indiquée.

2) CN : Pour des tableaux à composantes proportionnelles, $r = q = \sup_k q_k$, et l'I.C. coïncide avec l'I.D.P.. On a donc : $\hat{X}_k = W_k D C A^{-1} = Y_k' A_k Q = C_k^c Q_k' A_k Q$, $k \in K$. Remplaçons C_k^c par I_q pour \sqrt{m} -projeter les vecteurs (\hat{u}_j^k) et non les nuages

\tilde{N}_k sur le sous-espace compromis, on obtient : $P_k^c = Q_k' A_k Q$, $k \in K$. La condition A_k s.d.p. revient à choisir convenablement l'orientation des axes Δv_j^k de chaque nuage.

CS : $W_k = C P_k' P_k C'$ s'écrit sous les conditions retenues :
 $W_k = C Q' A_k Q_k Q_k' A_k Q C' = Y A_k^2 Y'$ et,
 (X_k, M_k, D) est équivalent à $(Y A_k, I, D)$, $k \in K$. ■

Cette caractérisation des tableaux à composantes proportionnelles (proposition 8-2) équivaut à la caractérisation suivante établie par Glaçon (1981) :

INDSCAL admet une solution exacte si $r = q = \sup_k q_k$, et si les $O_k(q \times q)$ solutions de $W_k = C O_k C'$ admettent un même système orthonormal de vecteurs propres.

En effet : $\forall k \in K$, $O_k = P_k' P_k = Q' A_k Q_k Q_k' A_k Q = Q' A_k^2 Q$.

Dans un contexte plus large, de Leeuw et Heiser (1982) proposent une autre caractérisation équivalente (avec comme deuxième condition : $W_l W W_k = W_k W W_l$, pour tout k, l , de K), et la proposition suivante :

Proposition 9.- (unicité)

Pour des tableaux à composantes proportionnelles, une normalisation et une orientation des axes données, l'écriture $Y_k = Y A_k$, $k \in K$ est unique à la Condition Nécessaire et Suffisante que pour tout i et j de $[q]$, il existe $k \in K$ tel que $i \neq j$ entraîne $a_{ki} \neq a_{kj}$.

■ Pour tout $i \neq j$ de $[q]$, il existe $k \in K$ tel que les valeurs propres a_{ki}^2 et a_{kj}^2 du tableau $O_k = Q' A_k^2 Q$ soient distinctes. Donc le système Q' des vecteurs propres orthonormés est unique (au signe près). L'écriture : $W_k = C O_k C' = C Q' A_k^2 Q C' = Y A_k^2 Y'$, $k \in K$ se vérifie pour des tableaux $Y_k = Y A_k$, $k \in K$ uniques (car A_k s.d.p.). Et réciproquement. ■

Notre caractérisation est cependant plus utilisable car elle permet de vérifier l'hypothèse ou de voir jusqu'à quel point on s'en écarte. Il suffit pour cela de calculer la distance $d = \sum_k \|P_k^c - Q_k' A_k Q\|^2$. Le choix de la norme induite par φ_{IA} donnera un résultat particulièrement intéressant.

La condition nécessaire $r = q = \sup_k q_k$ mérite un commentaire : elle signifie que les variables des tableaux X_k engendrent dans \mathbf{R}^n des sous-espaces inclus

dans l'un deux. Pour l'Analyse Canonique Généralisée ce cas est trivial et dans l'Intrastructure Discriminante associée (cf. Lechevalier 1990) tous les nuages sont confondus. Si en revanche on s'écarte trop de cette situation, c'est la notion d'Intrastructure Equisecteur (et d'AFJ) qui perd tout sens et l'ACG permet de voir pourquoi. Ces deux méthodes sont donc complémentaires au sens où elles concernent des familles de tableaux aux caractéristiques différentes.

3.2.- L'Analyse Factorielle conJointe.

Nous abordons maintenant le cas général avec une famille de triplets n'admettant pas des composantes exactement proportionnelles. Les résultats précédents conduisent rapidement à l'algorithme recherché.

a) Les composantes conjointes

Avec des triplets (Y_k, I, D) équivalents à (X_k, M_k, D) , on peut écrire $Y_k = YA_k + \varepsilon_k$ avec A_k diagonale s.d.p., $\varepsilon_k(n \times q)$ et $Y(n \times q)$.

Les tableaux ε_k peuvent s'interpréter comme les erreurs dans l'approximation de Y_k par YA_k , $k \in K$; et s'ils sont nuls, les tableaux sont à composantes proportionnelles ce qui se vérifie très rarement.

Définition 10.- On appelle composantes conjointes d'une famille de triplets (X_k, M_k, D) équivalents à (Y_k, I, D) , les q ensembles de m colonnes de même indice j dans les tableaux Y_k , tels que :

$$\Phi(Y_k, A_k, Y) = \sum_k \|\varepsilon_k\|_{\varphi_{DI}}^2 = \sum_k \|Y_k - YA_k\|_{\varphi_{DI}}^2$$

soit minimale. Les composantes conjointes sont donc les ensembles $\{y_k^j, k \in K\}_{j \in [q]}$ qu'on désignera simplement par composantes (des tableaux) Y_k , $k \in K$. L'AFJ est une méthode de détermination de ces composantes qui sont les plus proches de la proportionalité au sens des moindres carrés.

A ces composantes sont associés un Repère Equisecteur et un Repère Déformant et les intrastructures correspondantes. L'I.E. calculée par l'AFJ est d'autant plus significative que Φ est voisin de 0, ce que nous apprécierons à l'aide du rapport $e = (\sum_k \|\varepsilon_k\|^2)/m\|Y\|^2$. L'unicité des composantes conjointes se vérifie sauf dans le cas où la fonction Φ présente plusieurs optimums locaux

de même valeur auxquels correspondent plusieurs infrastructures dont aucune ne peut être considérée comme la plus significative. La multiplicité des solutions paraît d'autant plus improbable que e est petit (cf. proposition 9). Dans le cas contraire, l'I.E. étant moins significative, elle devra s'interpréter avec d'autant plus de prudence qu'elle est susceptible d'être multiple, avec des optimums locaux égaux qui seraient obtenus avec des initialisations différentes de l'algorithme que nous allons présenter.

Proposition 10.-

1) Pour des couples (Y_k, A_k) , $k \in K$ donnés, avec $\sum_k A_k^2 = mI_q$, $\Phi(Y_k, A_k, Y)$ est minimale si : $Y = \frac{1}{m} \sum_k Y_k A_k$.

2) Alors, si $D = \frac{1}{n} I_n$, les tableaux ε_k vérifient : $\sum_k \varepsilon_k A_k = 0$.

■ 1) $\Phi = \sum_k \|Y_k - Y A_k\|_{\varphi_{D,I}}^2 = \sum_k \text{tr}(Y_k' D Y_k + Y' D Y A_k^2 - 2Y_k' D Y A_k)$. Prenons $D = \frac{1}{n} I_n$ pour simplifier les écritures. La minimisation de Φ équivaut à la minimisation de

$$\Phi_1(Y) = \sum_{i=1}^n \sum_{j=1}^q (y_i^j)^2 \sum_k a_{kj}^2 - 2 \sum_k \sum_i \sum_j y_i^j y_{ki}^j a_{kj}.$$

Ce minimum classique s'obtient lorsque : $\frac{\partial \Phi_1}{\partial y_i^j} = 0$, $i \in I$, $j \in [q]$. Avec les hypothèses de normalisation et de poids égaux on a $\sum_k a_{kj}^2 = m$, $j \in [q]$, et les solutions vérifient $2m y_i^j = 2 \sum_k y_{ki}^j a_{kj}$, $i \in I$, $j \in [q]$; soit $Y = \frac{1}{m} \sum_k Y_k A_k$.

2) L'Infrastructure Déformante s'écrit : $\hat{Y}_k = Y_k A_k = Y A_k^2 + \varepsilon_k A_k$, $k \in K$;

et le tableau moyen : $\bar{Y} = \frac{1}{m} \sum_k \hat{Y}_k = Y \frac{1}{m} \sum_k A_k^2 + \frac{1}{m} \sum_k \varepsilon_k A_k$.

On a aussi : $\bar{Y} = \frac{1}{m} \sum_k Y_k A_k = Y$ et $\frac{1}{m} \sum_k A_k^2 = I$ d'où $\sum_k \varepsilon_k A_k = 0$.

A l'optimum, on a donc : $Y = \frac{1}{m} \sum_k Y_k A_k$ et $\sum_k \varepsilon_k A_k = 0$. ■

b) Nombre minimal admissible de composantes.

L'optimum obtenu pour Φ définit une Infrastructure (I.E.) significative sous réserve que le critère e précédent soit suffisamment petit, par exemple inférieur à $e_0 = 10$ ou 20%. On dit alors que les tableaux sont à **composantes quasi-proportionnelles**.

Il serait souhaitable d'avoir un critère permettant d'apprécier, au début des calculs, la valeur de e . On a des composantes proportionnelles à la condition

nécessaire que r (le rang de C), soit égal à $q = \sup_k(q_k)$, c'est-à-dire, pour les valeurs propres de WD : $\lambda_{q+1} = \lambda_{q+2} = \dots = \lambda_n = 0$.

On admettra donc la condition nécessaire : $p_q = (\sum_{i=1}^q \lambda_i) / (\sum_{i=1}^n \lambda_i) \geq p_0$ avec $p_0 = 80$ à 90% pour qu'il existe des composantes quasi-proportionnelles.

Règle n° 1 : Soit $q = \sup_k q_k$ (éventuellement $q = \sup_k p_k$), si la part d'inertie p_q portée par les q premiers axes d'inertie du nuage compromis dépasse le seuil p_0 on peut chercher une Intrastructure Equisecteur. Sinon, cette recherche est douteuse.

Nous avons traité des cas pour lequel $p_q = 99\%$ et on peut alors se limiter à un nombre s de composantes ($s < q$) tel que $p_s = (\sum_{i=1}^s \lambda_i) / (\sum_{i=1}^n \lambda_i) \geq p_0^s$ avec $p_0^s = 95\%$ à 99%, à condition que dans chaque nuage on vérifie également $p_s^k \geq p_0^s$.

Règle n° 2 : Si $p_s \geq p_0^s$ et $p_s^k \geq p_0^s$, $k \in K$, on peut remplacer les tableaux $C(n \times q)$ et $C_k(n \times q)$ par les tableaux C_s et C_k^s obtenus en ne gardant que les s premières colonnes de ces tableaux. Les tableaux $Y_k(n \times s)$ définissant l'Intrastructure Equisecteur contiennent la même inertie que C_k^s et donc perdent moins de 5% de l'inertie des nuages N_k .

Les seuils p_0 et p_0^s sont donnés ici à titre indicatif. Alors que p_0 repère l'adéquation des données à un modèle hypothétique ($Y_k = Y A_k$), le seuil p_0^s exprime un abandon de précision dans le positionnement des nuages et on doit être beaucoup plus strict sur ce point.

Proposition 11.-

Soient une famille de triplets (X_k, M_k, D) , les tableaux $C_k(n \times s)$ des s premières composantes principales de chaque nuage ($s \leq q$), un tableau $C(n \times s)$ vérifiant $C'DC = \Lambda$ (Λ diagonale définie positive), et les tableaux $P_k = C_k'DC\Lambda^{-1}$, $k \in K$. Pour des matrices orthogonales Q et Q_k telles que : $Y = CQ'$ et $Y_k = C_kQ_k'$, $k \in K$, et des A_k diagonales s.d.p., on a :

$$1) \sum_k \|P_k - Q_k' A_k Q\|_{\varphi_{I\Lambda}}^2 + \delta_s = \sum_k \|Y_k - Y A_k\|_{\varphi_{DI}}^2 \quad (1)$$

2) C étant le tableau des composantes principales du nuage compromis, on a $\delta_s = 0$ si $s = q = r = \text{rang} X_p$.

■ 1) Rappelons que $\langle A, B \rangle_{\varphi_{M_1 M_2}} = \text{tr}(A'M_1 B M_2) = \text{tr}(B M_2 A' M_1)$,

• $\varphi_k(Q'_K, A_k, Q) = \|P_k - Q'_k A_k Q\|_{\varphi_{r\Lambda}}^2 = \|C'_k D C \Lambda^{-1} \Delta - Q'_k A_k Q \Delta\|_{\varphi_{r\Lambda}}^2$,
avec $\Delta^2 = \Lambda$.

$$\begin{aligned}\varphi_k(Q'_K, A_k, Q) &= \text{tr}(C'_k D [C \Lambda^{-1} C' D] C_k + Q'_k A_k Q \Lambda Q' A_k Q_k - 2C'_k D C Q' A_k Q_k) \\ &= \text{tr}(C'_k D (P_c C_k) + \Lambda Q' A_k^2 Q - 2C'_k D C Q' A_k Q_k)\end{aligned}$$

$P_c = C \Lambda^{-1} C' D$ définit le projecteur d'une variable sur le sous-espace F_s engendré par les variables du tableau C .

Si C est le tableau compromis et si $s = r$ (alors $s = q = r$), on a $c'_k \in F_s$, donc $P_c C_k = C_k$ et $\text{tr}(C'_k D P_c C_k) = \text{tr}(\Lambda_k) = \sum_{i=1}^{q_k} \lambda_i^k = I_k$

Sinon $\text{tr}(C'_k D P_c C_k) \leq I_k^s = \sum_{i=1}^s \lambda_i^k$, qui est l'inertie du nuage k pour $C_k (n \times s)$.

• $\Phi_k(Y_k, A_k, Y) = \|Y_k - Y A_k\|_{\varphi_{DI}}^2$

$$\begin{aligned}\Phi_k(Y_k, A_k, Y) &= \text{tr}(C_k Q'_k Q_k C'_k D + C Q' A_k^2 Q C' D - 2Q_k C'_k D C Q' A_k) \\ &= \text{tr}(C'_k D C_k + \Lambda Q' A_k^2 Q - 2C'_k D C Q' A_k Q_k)\end{aligned}$$

$$\Phi_k(Y_k, A_k, Y) = \varphi_k(Q'_K, A_k, Q) + d_k, \text{ en posant } I_k^s - \text{tr}(C'_k D P_c C_k) = d_k \geq 0.$$

En sommant selon k , on obtient (1) avec $\delta_s = \sum_{k=1}^m d_k \geq 0$. On notera $\Phi_s = \varphi_s + \delta_s$, où l'indice s rappelle que les calculs sont faits avec des tableaux C et C_k de dimensions $(n \times s)$.

2) C étant le tableau compromis, $s = q = r$ entraîne $d_k = 0$ et $\delta_s = 0$. On retrouve que l'égalité $P_k = Q'_k A_k Q$ est bien une propriété caractéristique des tableaux à composantes proportionnelles. ■

Cette remarquable propriété permet surtout de passer d'une caractérisation des tableaux à composantes proportionnelles, à la détermination des composantes conjointes de l'AFJ. Il suffit d'emboîter deux algorithmes simples. Le premier, dit interne (exposé en c1, c2, c3), travaille sur des tableaux $(s \times s)$ pour déterminer (Q_k, A_k) , $k \in K$. Le second, dit externe (exposé en c4), traite des tableaux $(n \times s)$ pour déterminer Y et donc C .

c) L'algorithme de l'AFJ.

c1) Principe et initialisation de l'algorithme interne.

Les tableaux C et C_k , $k \in K$ étant donnés, l'algorithme est fondé sur la possibilité de modifier un tableau Q_k , A_k ou Q de façon à faire décroître la

fonction φ_s , et donc aussi Φ_s . De même, une modification du tableau C pourra faire décroître Φ_s .

Comme le note Ten Berge (1986, 1988), cela montre qu'il existe des solutions, puisqu'on construit une suite de valeurs décroissantes pour Φ_s qui est minorée par 0, et fournit une solution dans le même temps. Celle-ci peut être un minimum local ou global, d'où l'idée de faire plusieurs initialisations différentes pour comparer les résultats.

Pour des composantes quasi-proportionnelles l'initialisation de C par le tableau compromis de l'AFJ1 nous met suffisamment près de l'optimum global pour éviter les minima locaux liés à l'algorithme externe.

On peut alors fixer presque arbitrairement $Q_k, A_k, k \in K$ pour chercher Q . Il semble que le choix $Q_k = I_s$ et $A_k = \text{diag}(P_k)$, $k \in K$, avec renormalisation des A_k telle que $\sum_k A_k^2 = mI$, soit simple et efficace. Dans le cas où une composante c_k^j serait proportionnelle à c^i (avec $j \neq i$) on pourrait trouver une initialisation meilleure, mais plus complexe. Cela se remarque à la forme particulière du tableau P_k . On peut aussi envisager plusieurs initialisations de l'algorithme interne pour vérifier qu'on atteint bien un minimum global.

c2) Calcul de Q minimisant φ_s .

Pour (Q_k, A_k) fixés, $k \in K$ avec $\sum A_k^2 = mI$. On a :

$$\Phi_s = \sum_{k=1}^m \Lambda_k^2 + m\Lambda - 2 \text{tr} \left(\sum_{k=1}^m C_k' DC Q' A_k Q_k \right) = \varphi_s + \delta_s$$

et avec $C_k' DC = P_k \Lambda$ on a :

$$\text{Min}_Q \varphi_s \iff \text{Max}_Q \text{tr} \left(Q' \sum_k (A_k Q_k P_k \Lambda) \right)$$

Avec la décomposition en valeurs singulières on écrit : $\sum_k (A_k Q_k P_k \Lambda) = V \Delta U'$. En appliquant le théorème de Kristof (Kristof 1970, Le Calvé 1976, Ten Berge 1983), on trouve que l'extremum recherché s'obtient avec $Q = VU'$.

c3) Calcul de A_k et Q_k minimisant φ_s .

On détermine A_ℓ , puis Q_ℓ , pour tous les éléments successifs ℓ de K en supposant fixé $Q, (A_k, Q_k), k \in K - \{\ell\}$

- $\text{Min } \tau(A_\ell) = \text{Min}[\text{tr}(A_\ell^2 Q \Lambda Q') - 2 \text{tr}(A_\ell Q_\ell P_\ell \Lambda Q')] \quad (Q_\ell \text{ fixé}).$ Ce minimum s'obtient en annulant toutes les dérivées :

$$\frac{\partial \tau}{\partial a_{\ell j}} = 2a_{\ell j}(Q \Lambda Q')_j^j - 2(Q_\ell P_\ell \Lambda Q')_j^j, \quad j \in [s]$$

d'où :

$$A_\ell = \text{diag}(Q_\ell P_\ell \Lambda Q') [\text{diag}(Q \Lambda Q')]^{-1}, \quad \ell \in K$$

- $\text{Min } \nu(Q_\ell) \iff \text{Max tr}(Q_\ell P_\ell \Lambda Q' A_\ell) \quad (A_\ell \text{ fixé}).$ Avec la *s.v.d.* on a : $P_\ell \Lambda Q' A_\ell = V_\ell \Delta_\ell U_\ell'$. On obtient :

$$Q'_\ell = V_\ell U'_\ell \quad \text{ou} \quad Q_\ell = U_\ell V'_\ell, \quad \ell \in K$$

On réitère le calcul de Q et $(A_k, Q_k), k \in K$ tant que φ_s décroît de façon significative. La fonction $\Phi_s = \sum_k \|Y_k - Y A_k\|^2$ a évidemment subi la même décroissance et on la calcule uniquement lors de cette dernière étape.

c4) L'algorithme externe : calcul de C .

Pour Y_k donc Q_k et A_k fixés, $k \in K$, on calcule une nouvelle valeur de Y faisant décroître Φ_s . C'est : $Y = \frac{1}{m} \sum_k Y_k A_k$ en vertu de la proposition 10.

On en déduit, au moyen de la *s.v.d.* par exemple, un nouveau tableau $C(n, s)$ qui contient les composantes principales du triplet (Y, I, D) . La *s.v.d.* de Y s'écrit $Y = \Gamma \Delta U'$. On prendra $C = \Gamma \Delta$ et la matrice Q telle que $Y = C Q'$ est $Q = U$.

Avec le nouveau tableau C , on réitère l'algorithme interne pour s'arrêter lorsque la fonction $\Phi_s = \sum_k \|Y_k - Y A_k\|^2$ ne décroît plus de façon significative. Notons qu'après cette première itération les tableaux $P_k = C'_k D C \Lambda^{-1}$ ne représentent plus les \sqrt{m} -projections $\sqrt{m} \tilde{U}'_k M U$ puisque C et U ne sont plus relatifs au nuage compromis de l'I.C.

3.3.- Les résultats de l'AFJ.

a) Interprétation de la méthode

- A l'optimum on connaît $Y, C, Q, (Q'_k, A_k, C_k), k \in K$.

On connaît donc l'I.E. : $\hat{Y}_k = Y_k = C_k Q'_k, k \in K$ et $\bar{Y} = \frac{1}{m} \sum_k Y_k$;

on connaît aussi l'I.D. : $\hat{X}_k = Y_k A_k = C_k Q'_k A_k, k \in K$ et $\bar{X} = \frac{1}{m} \sum_k \hat{X}_k = Y$.

- Si les tableaux sont à composantes proportionnelles, l'I.D. est confondue avec l'I.C. et le tableau compromis C vérifie $C = YQ$. L'AFJ se résoud sans itérer l'algorithme externe. Cela nous montre que la méthode est fondée sur la recherche de l'**Intrastructure Déformante** (dont le tableau Y) et le fait que celle-ci détermine immédiatement l'**Intrastructure Equisecteur**.

- Si les tableaux sont à composantes quasi-proportionnelles, on montre une très grande proximité entre les sous-espaces E_C et E_D , avec des particularités décrites par ailleurs (Lechevalier 1990). C'est pourquoi l'initialisation de l'algorithme avec le tableau compromis nous met très près de l'optimum. L'algorithme externe se réitère d'autant moins que les composantes sont proches de la proportionalité.

- Par analogie à l'I.C. ou l'I.D.P. (cf Définition 8) nous présentons l'I.E. dans la base de E_E portant les axes principaux du nuage moyen, nommée Repère Equisecteur Principal (R.E.P.). L'expression de l'I.E. Principale est alors : $\hat{C}_k = Y_k Q_E = C_k Q'_k Q_E$, $k \in K$ et $\bar{C} = \bar{Y} Q_E$. Dans cette I.E.P. nous indiquons les directions des axes Δv_j qui définissent les composantes conjointes $Y_k = Y A_k + \varepsilon_k$ avec $\sum_k \|\varepsilon_k\|^2$ minimum.

b) Applications

Toutes les données traitées par INDSCAL, Statis ou l'AFM, en particulier les tableaux quantitatifs $X_k(n \times p_k)$ ou $(n \times p)$ auront avantage à être traités par l'AFJ.

- Le critère minimisé par l'AFJ est voisin de celui de INDSCAL qui s'écrit : $\text{Min} \sum_k \|W_k - \hat{W}_k\|^2$. Nous obtenons au plan théorique la démonstration de l'existence de l'optimum et un algorithme simple à convergence très rapide. Au plan pratique nous reconstituons $\hat{W}_k = Y A_k^2 Y'$, $k \in K$ avec en plus une Intrastructure Exacte : $\hat{Y}_k = Y_k = Y A_k + \varepsilon_k$ et la position des axes de plus grande proportionalité relativement à ceux de plus grande inertie.

- Nous avons vu que les intrastructures de Statis ou de l'AFM étaient inexactes. Nous y avons ajouté la possibilité d'en déduire une intrastructure exacte avec une formulation précise de la géométrie le permettant, et les conditions nécessaires à l'obtention de ce résultat. Elles n'ont presque rien à voir avec le choix des poids ρ_k affectés aux tableaux. A priori, les meilleurs poids sont ceux

qui donnent aux tableaux $p_k X_k$ la même inertie mais, d'un autre côté, les poids égaux sont les plus simples. Pour des données non-cubiques, les tableaux sont centrés et réduits et une prémultiplication par $1/p_k$ assure les deux conditions. Avec les données cubiques, pour éviter la prémultiplication des tableaux on peut prendre des poids proportionnels à $1/\|X_k\|$.

• L'interstructure de Statis offre un critère suffisant pour que l'Intrastructure soit significative : il suffit de vérifier que le premier axe de l'interstructure (nuage des W_k) porte plus de 70% de l'inertie totale. La dispersion des W_k n'est donc pas gênante à condition qu'elle se situe essentiellement dans une seule direction. Nous avons développés ces deux derniers points par ailleurs (Lechevalier 1990), de même que les exemples évoqués ci-dessous.

Rappelons la condition nécessaire liée au compromis : $p_q \geq p_0$ ou $p_s \geq p_0^*$, et à titre indicatif, donnons pour 3 exemples, les valeurs de ces deux critères et l'erreur moyenne constatée une fois l'I.E. trouvée (en pourcentages) :

	Accidents	Cantons profils	Cantons bruts
inertie axe 1 interstructure	96, 36	91, 75	86, 68
inertie 3 axes compromis	97, 73	94, 58	94, 80
erreur moyenne e	2, 30	3, 90	4, 97

Nous voyons que les deux conditions sont largement vérifiées et que les composantes sont presque proportionnelles ($e < 5\%$).

c) L'Analyse Procuste.

Avec $T(\sum_k q_k \times q_m)$, $q_m = \text{Inf}_k q_k$, $T' = (T'_1 \quad \dots \quad T'_m)$ et $T'_k T_k = I_{q_m}$ le problème procuste asymétrique s'écrit aussi $\text{Min } \psi(T) = \sum_k \sum_\ell \|C_k T_k - C_\ell T_\ell\|^2$. L'interstructure associée n'est pas exacte et on peut maintenant lui substituer deux problèmes procustes symétriques. Le premier déjà signalé correspond à $T(mq \times q)$ avec $q = \sup_k q_k$ et $T'_k T_k = I_q$. Il porte sur les tableaux $C_k^c = (C_k \quad 0)$ ($n \times q$), $k \in K$. On peut aussi considérer les tableaux $C_k(n \times s)$ limités aux s premières composantes, à condition qu'ils contiennent une grande part de l'inertie des nuages N_k (plus de 95%), et faire une analyse procuste symétrique qui perd très peu d'information. On aura alors $T(ms \times s)$. Dans la suite nous ne distinguerons plus ces deux cas.

Proposition 12.- *La minimisation de la fonction $\psi(T)$ de l'Analyse Procuste Symétrique (A.P.S.) équivaut à la minimisation de la fonction : $\Phi(Y_k, I, Y) = \sum_k \|Y_k - YI\|_{\varphi_{DI}}^2$ qui est la fonction de l'AFJ contrainte par $A_k = I, k \in K$. L'APS est donc une AFJ contrainte.*

■ En introduisant des poids ρ_k égaux on montre que :

$$\sum_k \sum_{\ell} \|C_k T_k - C_{\ell} T_{\ell}\|^2 = 2m \sum_k \|C_k T_k - \frac{1}{m} \sum_{\ell} C_{\ell} T_{\ell}\|^2$$

Dans l'AFJ, on a : $Y_k = C_k Q'_k = C_k T_k$ avec $T_k = Q'_k, k \in K$ et $Y = \frac{1}{m} \sum_k Y_k A_k$. Si $\forall k \in K A_k = I$, on a : $Y = \frac{1}{m} \sum_k C_k T_k$. Et on peut écrire :
 $\Phi(Y_k, I, Y) = \sum_k \|Y_k - YI\|_{\varphi_{DI}}^2 = \sum_k \|Y_k Q - YQ\|_{\varphi_{DI}} = \sum_k \|Z_k - C\|_{\varphi_{DI}}^2$. ■

Définition 11.- *De même que les composantes conjointes $Y_k = Y A_k + \varepsilon_k$ sont les composantes les plus proches de la proportionalité, on appelle Composantes Procustes d'une famille de triplets (X_k, M_k, D) , équivalents à (Z_k, I, D) les composantes des tableaux $Z_k = C + \varepsilon_k$ minimisant $\sum_k \|\varepsilon_k\|^2$ avec $C = \frac{1}{m} \sum_k Z_k$. Ce sont les composantes les plus proches, au sens des moindres carrés, de l'égalité à des composantes principales.*

L'APS, en tant qu'AFJ contrainte par $A_k = I$, vérifie les mêmes propriétés : elle définit une intrastucture exacte par \sqrt{m} -projection sur le sous-espace Procuste E_P . L'introduction de la contrainte $A_k = I$ dans l'algorithme conduit à la solution, mais moins rapidement que dans l'AFJ classique car E_C est "moins proche" de E_P que de E_D . L'Intrastucture Procuste est moins significative que l'I.E. car bien évidemment $\sum_k \|Y_k - Y A_k\|^2 \leq \sum_k \|Z_k - C\|^2$. L'AFJ devrait donc être préférée à l'Analyse Procuste, et à défaut, l'algorithme de l'AFJ contrainte par $A_k = I$ donnera sans doute d'aussi bons résultats que les algorithmes classiques d'A.P. du fait qu'il travaille souvent sur des tableaux $(s \times s)$ et non $(n \times s)$. La recherche pourra être approfondie dans ce sens.

En conclusion :

Ce travail apporte les trois résultats suivants :

- 1) Avec l'AFJ on dispose d'une méthode donnant une intrastucture exacte et significative. "Exacte" indique la possibilité d'obtenir des images euclidiennes équivalentes. "Significative" précise que le positionnement simultané respecte

“le mieux possible” au sens des moindres carrés les propriétés de dualité. On réalise ainsi l'Analyse en Composantes conJointes d'une famille de triplets.

- 2) Plusieurs méthodes de l'analyse conjointe de tableaux de données ont été présentées comme techniques de \sqrt{m} -projection sur un certain sous-espace, des nuages d'individus \tilde{N}_k , $k \in K$, positionnés dans l'espace somme directe orthogonale des espaces initiaux. Ainsi voit-on un lien très simple entre les méthodes d'origine française ou anglo-saxonne.
- 3) Cette présentation a permis de dégager des propriétés duales sur les espaces des individus et sur l'espace des variables.
 - Pour que l'intrastructure exacte soit significative il faut que les sous-espaces F_k engendrés respectivement par les variables des tableaux X_k soient inclus dans l'un deux. Elle l'est d'autant moins qu'on s'écarte de cette situation.
 - l'Intrastructure Equisecteur correspond aux variables les plus proches de la proportionalité au sens des moindres carrés.
 - l'Intrastructure Procuste correspond aux variables les plus proches de l'égalité à des composantes principales.
 - l'Intrastructure Compromis (non exacte) correspond aux variables canoniques au sens de la liaison $L^2(y, X_k)$ (Escofier et Pagès 1983, 1988).
 - l'Intrastructure Discriminante (non exacte) correspond aux variables canoniques de l'ACG au sens de Carroll (Casin et Turlot 1986, Lechevalier 1990).

BIBLIOGRAPHIE.

- BOUROCHE J.M. (1975)** - Analyse des données ternaires : la double analyse en composantes principales. *Thèse de 9^{ème} cycle. Université de Paris VI.*
- BOUROCHE J.M. & SAPORTA G. (1980)** - L'Analyse des Données. *Que Sais-je ?*, PUF - PARIS.
- CAILLIEZ F. & PAGES J.P. (1976)** - Introduction à l'Analyse des Données. *SMASH - PARIS.*
- CARROLL J.D. (1968)** - A generalisation of canonical correlation analysis to three or more sets of variables. *Proc. 76th Conv. Amer. Psych. Assoc. p.227-228.*
- CARROLL J.D. & CHANG, J.J. (1970)** - Analysis of individual differences in multidimensional scaling via an n-way generalizations of "Eckart-Young" decomposition. *Psychometrika, 35, p. 283-319.*
- CASIN Ph. (1985)** - L'Analyse Canonique Généralisée et l'analyse des tableaux évolutifs. *Thèse de 9^{ème} cycle. Université L. Pasteur Strasbourg.*
- CASIN Ph. & TURLOT J.C. (1986)** - Une présentation de l'Analyse Canonique Généralisée dans l'espace des individus. *R.S.A. 1986 - Vol. XXXV n° 3, p. 65-75.*
- COPPI R. & BOLASCO S. (1989)** - Multiway data analysis. Volume originated from Multiway'88 international meeting on the multiway data matrices. CNR Rome Italie 28-30 Mars. *North Holland.*
- DE LEEUW J. & PRUZANSKY S. (1978)** - A new computational method to fit the weighted euclidean distance model. *Psychometrika, 43, n° 4, p. 479-490.*
- DE LEEUW J. & HEISER W. (1982)** - Theory of Multidimensionnal Scaling. in P.R. Krishnaiah & L.N. Kanal (Eds.) : *Handbook of Statistics, Vol. 2. North Holland, p. 285-316.*
- ECKART C. & YOUNG G. (1936)** - The approximation of one matrix by another of low rank. *Psychometrika, 1, p. 211-218.*
- ESCOFIER B. & PAGES J. (1983)** - L'Analyse Factorielle Multiple. Une méthode de comparaison de groupes de variables. in Diday E. et al. (Eds) : *Data Analysis and Informatics III. North Holland 1984.*
- ESCOFIER B. & PAGES J. (1988)** - Analyses Factorielles Simples et Multiples. *DUNOD - PARIS.*
- ESCOUFIER Y. et al. (1985)** - Numéro spécial sur les tableaux à 3 dimensions et données évolutives. *Statistique et Analyse des Données. Juin 1985. (SAD Vol. 10, n° 1).*

- GLACON F. (1981)** - Analyse conjointe de plusieurs matrices de données : comparaison de méthodes. *Thèse de 3^{ème} cycle - Université Médicale de Grenoble.*
- HARSHMAN R.A. (1970)** - Foundations of the PARAFAC procedure : models and conditions for an "explanatory" multi-mode analysis. *UCLA Working Papers in Phonetics, 16, p. 1-84.*
- HARSHMAN R.A. & LUNDY, M.E. (1984)** - The PARAFAC model for Three-Way Factor Analysis and Multidimensional Scaling. in H.G. Law, C.W. Snyder, J.A. Hattie, and R.P. McDonald (Eds.) : *Research methods for multimode data analysis (p. 122-215). New York : Praeger.*
- JOLLIFFE J.T. (1986)** - Principal Component Analysis. *New York : Springer-Verlag.*
- KIERS H.A.L. (1987)** - Comparaison of "anglo-saxon" and "French" three way methods and their application to the analysis of nominal variables. *présenté à la workshop "données évolutives" Data Analysis and Informatics V.*
- KIERS H.A.L. (1988)** - Hierarchical relations between three-way methods. *présenté aux "XX Journées de Statistique" Grenoble 2-6-1988.*
- KIERS H.A.L. (1989)** - Three ways methods for the analysis of qualitative and quantitative two ways data. *Leiden : DSWO-press.*
- KRISTOF W. (1970)** - A theorem on the trace of certain matrix products and some applications. *Journal of Mathematical psychology, Vol. 7, p.515-530.*
- KROONENBERG P.M. (1983)** - Three mode principal component analysis : Theory and applications. *Leiden : DSWO-press.*
- LAFAYE J.Y. (1985)** - Application du modèle INDSCAL. in Escoufier Y. et al. : *Vol. 10, n° 1 p. 89-102.*
- LAFOSSE R. (1985)** - Analyses procustéennes de deux tableaux. *Thèse de 3^{ème} cycle. Université P. Sabatier Toulouse.*
- LAVIT C. (1985)** - Application de la méthode Statis. in Escoufier Y. et al. : *Vol. 10, n° 1 p. 103-116.*
- LAVIT C. (1988)** - Analyse conjointe de tableaux quantitatifs. Méthode et programmes. *MASSON - PARIS.*
- LE CALVE G. (1976)** - Problèmes d'Analyse des données. *Thèse d'Etat. Université de Rennes I.*
- LECHEVALIER F. (1990)** - L'Analyse Factorielle conJointe d'une famille de triplets indexés. *Thèse . Université de Lille I.*
- L'HERMIER des plantes H. (1976)** - S.T.A.T.I.S. : théorie et application d'une méthode d'analyse conjointe. *Thèse de 3^{ème} cycle. Université des Sciences et Techniques du Languedoc.*

- MOUTTET F. (1982)** - Comparaison de deux tableaux par la méthode procruste. *Thèse de 3^{ème} cycle. Université de Paris VI.*
- POLIT M.O. E. (1986)** - Une n-ACP d'un hypercube de données. *Thèse de 3^{ème} cycle - Université de Grenoble II.*
- SAPORTA G. (1975)** - Liaison entre plusieurs ensembles de variables et codage de données qualitatives. *Thèse de 3^{ème} cycle. Université de Paris VI.*
- SAPORTA G. (1990)** - Probabilité, Analyse des données et Statistique. *TECHNIP - PARIS .*
- TEN BERGE J.M.F. (1977)** - Orthogonal Procrustes rotation for two or more matrices. *Psychometrika, 42, p. 267-276.*
- TEN BERGE J.M.F. (1983)** - A generalization of Kristof's theorem on the trace of certain matrix products. *Psychometrika, 48, p. 519-523.*
- TEN BERGE J.M.F. (1986)** - A general solution for the Maxbet problem. in J. de Leeuw, W.J. Heiser, J. Meulman & F. Critchley (Eds.) : *Multidimensional Data Analysis. Leiden : DSWO-press, p. 81-87.*
- TEN BERGE J.M.F. (1988)** - Generalized approaches to the Maxbet problem and the maxdiff problem, with applications to canonical correlations. *Psychometrika, 53, n° 4, p. 487-494.*
- TEN BERGE J.M.F. & KNOL D.L. (1984)** - Orthogonal rotations to maximal agreement for two or more matrices of different column orders. *Psychometrika, 49, p. 49-55.*
- TUCKER L.R. (1966)** - Some mathematical notes on three-mode factor analysis. *Psychometrika, 31 n° 3, p. 279-311.*
- TUCKER L.R. (1972)** - Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika, 37 n° 1, p. 3-27.*