Structure de l'ensemble des analyses multivariées des tableaux de données à trois entrées : Eléments théoriques et appliqués

Mohamed HANAFI

Mots clés :

Three mode data, Three way data, Linear Multivariate Analysis, Canonical Analysis, Generalized Canonical Analysis, Procuste Analysis, Generalized Procuste Analysis, Multidimensional Scaling, Factor Analysis, Co-inertia Analysis, Orthogonal Rotation.

ACT-STATIS, AFM, ACOM, CANDECOMP/PARAFAC, INDSCAL, TUCKALS-3, MAXBET, MAXDIFF, PCA-SUP, SUMPCA, ORTCP-A, ORTCP-B, ORTCP, Three modes-scaling, INDORT.

Singular value Decomposition theorem, Principal Components, Altearnating least squares, Iterative Majorization, Cauchy-Schwartz inequality, Trace Optimization.

Structure de l'ensemble des analyses multivariées des tableaux de données à trois entrées : Eléments théoriques et appliqués

Mohamed HANAFI

Mots clés :

Three mode data, Three way data, Linear Multivariate Analysis, Canonical Analysis, Generalized Canonical Analysis, Procuste Analysis, Generalized Procuste Analysis, Multidimensional Scaling, Factor Analysis, Co-inertia Analysis, Orthogonal Rotation.

ACT-STATIS, AFM, ACOM, CANDECOMP/PARAFAC, INDSCAL, TUCKALS-3, MAXBET, MAXDIFF, PCA-SUP, SUMPCA, ORTCP-A, ORTCP-B, ORTCP, Three modes-scaling, INDORT.

Singular value Decomposition theorem, Principal Components, Altearnating least squares, Iterative Majorization, Cauchy-Schwartz inequality, Trace Optimization.

1997

Sommaire

Chapitre I.3 : Approche méthodologique

Chapitre I.4 : Approche algorithmique

Chapitre I.5 Composante de liens

I.6. Conclusions et perspectives

II. DEUXIÈME PARTIE : Eléments appliqués

II.0 Introduction

Chapitre II.1 : Etude comparative entre les deux sous-univers

Chapitre II.2 : Analyse simultanée de K-couples de tableaux par l'analyse STATICO

II.3. Conclusions et perspectives

Références

Résumé

Ce travail s'inscrit dans le cadre des analyses multivariées des tableaux de données à trois entrées (K-tableaux et multitableaux). Il s'articule sur les récents travaux qui concernent les deux questions suivantes :

(**a**) la première est celle de la recherche pour ces analyses d'un cadre mathématique qui offre un théorème jouissant des mêmes propriétés intrinsèques que le théorème de la décomposition en valeurs singulières d'une matrice dans le cadre des analyses des tableaux à deux entrées,

(**b**) la deuxième est celle de l'analyse des liens entre ces analyses à différents niveaux (algorithmique, méthodologique, hiérarchique).

La question posée ici consiste à s'interroger sur la possibilité d'une formulation mathématique générale et unifiée des analyses des tableaux à trois entrées. La solution proposée nécessite l'étude de la structure de l'ensemble des analyses existant. Ce travail se consacre à cette étude en intégrant deux approches complémentaires, l'une théorique et l'autre appliquée. Plus exactement ce travail :

— introduit des notions pour exprimer cette problématique. "univers" : ensemble des analyses à trois entrées, "sous-univers" : un sous-ensemble de l'univers, "composante de liens" : sous-ensemble d'analyses liées simultanément selon trois approches différentes, géométrique, méthodologique et algorithmique;

— passe en revue des analyses de tableaux à trois entrées, sur la base des travaux de synthèse existant (chapitre I.1);

— introduit la notion de comportement géométrique d'une analyse. Cette nouvelle notion permet de lier les analyses indépendamment des critères utilisés par celles-ci. En particulier, deux comportements géométriques sont dégagés (chapitre I.2);

— complète le système d'analyses existant par la proposition de nouvelles analyses (chapitre I.3);

— reprend deux techniques pour construire des algorithmes convergents associés à l'optimisation de fonctions et propose un algorithme de résolution général pour la famille des nouvelles analyses (chapitre I.4);

— construit une composante de liens et montre que celle-ci contient comme cas particuliers différentes propositions d'unification (chapitre I.5);

— présente l'intérêt de ces analyses dans un champ expérimental, ici l'écologie (Introduction II.0);

— effectue, sur la base d'exemples, une étude comparative entre deux analyses,, chacune étant représentative d'un fonctionnement théorique (géométrique et algorithmique) propre (chapitre II.1);

— propose une analyse spécifique pour le couplage de deux tableaux à trois entrées, appelée analyse STATICO (chapitre II.2);

— permet d'avancer une hypothèse sur l'état réel de la structure de l'univers et propose un certain nombre de perspectives.

Introduction générale

Les données sont le support de base de la recherche en analyse multivariée. Elles se présentent sous forme de tableaux et elles ont au moins deux dimensions, l'une théorique, l'autre appliquée.

-- théorique : un tableau de données peut être considéré comme un objet mathématique abstrait (matrice, opérateur, tenseur) qui existe par lui-même, il n'est pertinent que par sa capacité à caractériser quelques propriétés mathématiques remarquables.

— appliquée : ce même tableau de données résulte d'un protocole expérimental et reflète une image réduite de la réalité avec sa nature très souvent complexe, qu'elle soit biologique, sociologique ou économique.

Analyser ces données, objectif principal de l'analyse multivariée, consiste à dégager simultanément des règles générales quant aux structures inscrites dans les données et aux fonctionnements qui gèrent celles-ci, qu'ils soient théoriques, ils consistent ainsi à caractériser l'objet mathématique associé à ces données, ou qu'ils soient expérimentaux, ils consistent à étudier et éventuellement à contrôler une certaine réalité expérimentale inscrite dans ces données.

Il parait naturel que les données qui représentent l'objet et l'élément irréductible de la recherche en analyse multivariée, soient le moteur et la base commune de l'interrelation entre d'une part, l'analyse multivariée et différents champs expérimentaux (approche appliquée), et d'autre part, l'analyse multivariée et les mathématiques (approche théorique) avec ses différentes sphères (algébrique, géométrique, numérique, topologique, combinatoire).
La réalité de la recherche en analyse multivariée est alors de nature interactive : elle interfère entre au moins deux approches complémentaires, l'une théorique et l'autre appliquée, même si chacune des deux approches est caractérisée par sa logique propre, ses objectifs propres, et surtout une réalité très différente.

Un tableau de données est classiquement défini selon deux entrées qui sont respectivement les lignes et les colonnes. Diverses analyses multivariées ont été proposées spécifiquement pour les données qui se présentent sous forme de tableaux à deux entrées. Il s'agit entre autres de l'analyse en composantes principales, des analyses factorielles de correspondances,... Ces analyses offrent des solutions adaptées pour l'exploration de ce type de données.
Dans une vision théorique, le cadre mathématique pour ces analyses est l'algèbre linéaire. Dans ce cadre, le théorème fondamental, celui de la Décomposition en Valeurs Singulières d'une matrice rectangulaire, permet l'unification de ces diverses analyses en

termes de fonctionnement théorique. Ce résultat permet d'établir un réel échange entre les approches théorique et appliquée. Le logiciel constitue l'expression concrète de cet échange; il passe ainsi du simple statut d'outil technique à celui de point de rencontre entre les deux approches.

Cependant, d'une part, il arrive qu'un tableau de données à deux entrées soit associé à des questions expérimentales qui impliquent une structuration précise de ce tableau (voir introduction de la deuxième partie, II.0). Cette structuration consiste à intégrer une entrée supplémentaire pour générer, à partir de ce tableau, plusieurs sous tableaux. D'autre part, si mathématiquement un tableau de données à deux entrées est synonyme d'une matrice, il est possible de s'intéresser à caractériser l'objet mathématique issu d'une partition selon ses lignes ou selon ses colonnes.
Dans les deux cas, on introduit une entrée supplémentaire dans un tableau à deux entrées et le tableau résultant est alors dit "tableau à trois entrées".

Les analyses des tableaux à deux entrées ne sont pas adaptées à la description de ce type de données. Pour leur traitement, il s'avère indispensable de recourir au développement d'analyses spécifiques.

Cette question a eu une part importante dans la littérature multivariée où de nombreuses propositions existent offrant un paysage caractérisé par sa diversité. En effet, l'absence de lien direct entre ces analyses, les problèmes algorithmiques qu'elles génèrent, leurs comportements et leurs fonctions exactes dans une vision appliquée, et surtout leurs lien intrinsèque aux analyses des tableaux à deux entrées, ont suscité de nombreux travaux. Ceci reflète la richesse des approches multivariées consacrées à l'étude des analyses des tableaux à trois entrées.
Un certain nombre de ces travaux peuvent être considérés comme des contributions indirectes à la question générale suivante :

Peut-on réaliser une formulation mathématique unifiée pour les analyses des tableaux à trois entrées ?

Ce travail se définit dans le cadre de cette question générale. L'idée que l'on propose pour l'étudier s'articule autour de deux étapes :

— la première étape consiste à étudier la structure de l'ensemble des analyses à trois entrées existantes.

— la deuxième étape consiste à trouver un cadre mathématique approprié pour exprimer cette structure.

Cette idée résulte de nombreux travaux récents d'unification des analyses des tableaux à trois entrées, ainsi que de multiples tentatives de généralisation de la décomposition en valeurs singulières.

Ce travail se consacre uniquement à la première étape qu'on estime nécessaire pour apporter des éléments de réponses à la question générale. On appelle "univers" l'ensemble des analyses des tableaux à trois entrées. L'étude de la structure de cet univers consiste à dégager ou à détecter les fonctionnements théoriques irréductibles qui gèrent ces analyses, ainsi que la fonction de chaque fonctionnement théorique vis-à-vis d'une question appliquée. Le point de vue adopté est celui d'une approche exploratoire; il a l'avantage de ne pas préjuger du cadre mathématique de ces analyses

Ce travail est constituée donc de deux parties :

— la première partie est théorique : elle a pour objectif de formuler cette question d'une manière plus explicite et d'apporter des éléments de réponse théoriques sur l'état réel de cette structure.

— la seconde partie est appliquée : elle est aussi nécessaire pour comprendre cet état. On considère l'écologie comme un champ expérimental, avec la question centrale de l'étude de la relation entre les êtres vivants et leur environnement qui est en lien avec l'analyse multivariée via les analyses de couplage.

La formulation de la question générale et le schéma poursuivi pour y répondre constituent les contributions majeures de ce travail. Citons également les contributions importantes suivantes :

— lier différentes démarches d'unification existantes;

— proposer des analyses nouvelles sur la base des travaux existants;

— mettre l'accent sur les points qu'on estime importants pour avoir une vision théorique claire de ce type d'analyses;

— explorer, dans une vision appliquée, le comportement de ces analyses vis à vis d'une question expérimentale et permettre de proposer des analyses adaptées dans un champ expérimental.

Chacune des deux parties comporte une introduction dans laquelle seront présentés ses objectifs et les chapitres qui la constituent. On trouve aussi à la fin de chaque partie une conclusion et des perspectives. Ce schéma est le même dans chaque chapitre.

Définitions et notations

L'objectif ici est de donner quelques définitions et notations qui concernent les tableaux de données à trois entrées qui seront utilisées par la suite.

1. Tableaux à trois entrées

Un tableau de données $\mathbf{X}$ met en jeu deux indices ou deux entrées, l'un pour les lignes (individus) et l'autre pour les colonnes (variables). Le tableau $\mathbf{X}$ est dit tableau à deux indices ou tableau à deux entrées. Si par exemple, on partitionne les lignes du tableau $\mathbf{X}$ en K blocs, on obtient trois indices : la ligne, la colonne, et le bloc. On dit que le tableau $\mathbf{X}$ est un tableau à trois entrées ou aussi un tableau à trois modes. Les tableaux à trois entrées sont définis selon trois indices, classiquement ces trois indices sont appelés : individus, variables et occasions (Pour des exemples voir II.0).

Plusieurs tentatives de classification de ce type de tableaux de données peuvent être rencontrées dans la littérature (par exemple, Carroll & Arabie (1980)). Il n'existe pas encore une terminologie unique, ou du moins, elle n'est pas encore bien établie. Ce n'est pas vraiment un inconvénient, l'essentiel étant d'adopter une définition qui soit claire, d'autant plus que le caractère "trois entrées" des données n'est pas un terme absolu à celles-ci, mais il est relatif à une attitude propre de l'investigateur ou de l'utilisateur lorsqu'il vise ces données. Dans ce sens, on ne s'intéresse pas ici à la discussion de la définition la plus appropriée, mais on en adopte une, celle utilisée en outre par Kiers (1988) qui distingue les tableaux de données à trois entrées en deux familles :

(F1) "multiple data sets " : par souci de respecter la convention en analyse multivariée qui consiste à mettre les individus en lignes et les variables en colonnes, ce type de données, appelées ici les K-tableaux, sont de deux sortes :

— les premiers sont les K-tableaux horizontaux, c'est le cas lorsque la partition concerne les variables, les sous-tableaux formés ont les mêmes individus (Figure 1. a).

— les seconds sont les K-tableaux verticaux, la partition concerne les individus, les sous-tableaux formés ont les mêmes variables (Figure 1. b).

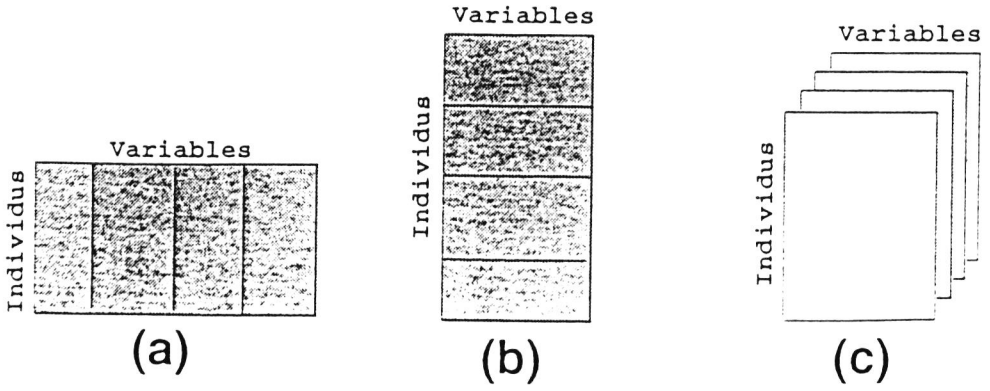


Figure 1. Les types de tableaux à trois entrées

(F2) "three data ways" appelées ici les multitableaux : les sous-tableaux ont les mêmes individus et les mêmes variables. On les appelle aussi cubes de données (Figure 1. c).

2. Etudes à trois entrées

A un tableau de données $\mathbf{X}$ à deux entrées on peut associer des questions pratiques différentes qui dépendent de l'objectif poursuivi, qui amènent à des analyses différentes. Citons par exemple, l'analyse en composantes principales (ACP) centrée, doublement centrée ou normée, l'analyse factorielle des correspondances (AFC), ou l'analyse des correspondances multiples (ACM) qui utilise les indicatrices des classes et les schémas de Tenenhaus & Young (1985). Autrement dit, à un tableau de données $\mathbf{X}$ correspondent plusieurs triplets $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ différents qui reflètent des objectifs différents mais qui concernent toujours le même tableau $\mathbf{X}$, de dimension $n \times p$, dont les lignes sont des observations sur n individus, et dont les colonnes sont les mesures sur p variables. $\mathbf{Q}$ est une matrice $p \times p$ définie positive utilisée pour mesurer les distances entre individus dans l'espace vectoriel $\mathbb{R}^p$. $\mathbf{D}$ est une matrice diagonale $n \times n$ à éléments positifs de trace unité, dite matrice des poids associés aux individus, et utilisée pour mesurer les distances entre variables dans l'espace vectoriel $\mathbb{R}^n$.

Pour tenir compte de cette diversité, on utilisera aussi la notion d'étude à trois entrées. On définit une étude K-tableaux horizontale comme un quadruplet statistique $(\mathbf{X}, \mathbf{Q}, \mathbf{D}, \Pi)$ constitué à partir de K études statistiques $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ $(1 \leq k \leq K)$ portant sur les mêmes n individus et K groupes de variables, comptant respectivement $p_1, p_2, \ldots, p_K$ variables. On note par p la somme des p_k. Ce quadruplet statistique $(\mathbf{X}, \mathbf{Q}, \mathbf{D}, \Pi)$ est construit de la manière suivante (Figure 2, pp. 14) :

-- $\mathbf{X}$ le K-tableaux horizontal de dimension $n \times p$ obtenu par juxtaposition horizontale des K tableaux $\mathbf{X}_k$,

-- Π la matrice diagonale constituée des poids π_k, π_k est un poids positif attribué *a priori* au tableau $\mathbf{X}_k$,

-- $\mathbf{Q}$ est la matrice bloc diagonale constituée des matrices $\mathbf{Q}_k$.

D'une manière analogue, on définit une étude K-tableaux verticale comme un quadruplet statistique $(\mathbf{X}, \mathbf{Q}, \mathbf{D}, \Pi)$ constitué à partir de K études statistiques $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$ $(1 \leq k \leq K)$ portant sur les mêmes p variables et K groupes d'individus, comptant respectivement $n_1, n_2, \ldots, n_K$ individus. On note par n la somme des n_k. Ce quadruplet statistique $(\mathbf{X}, \mathbf{Q}, \mathbf{D}, \Pi)$ est défini de la manière suivante :

-- $\mathbf{X}$ le K-tableaux vertical de dimension $n \times p$ est obtenu par juxtaposition verticale des K tableaux $\mathbf{X}_k$,

-- Π la matrice diagonale est constituée des poids π_k, π_k est un poids positif attribué *a priori* au tableau $\mathbf{X}_k$,

-- $\mathbf{D}$ est la matrice bloc diagonale constituée des matrices $\mathbf{D}_k$.

Enfin on définit une étude multitableaux comme un quadruplet statistique $(\mathbf{X}, \mathbf{Q}, \mathbf{D}, \Pi)$ constitué à partir de K études statistiques $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D})$ $(1 \leq k \leq K)$ portant sur les mêmes n individus et sur les mêmes p variables.

— $\mathbf{X}$ le multitableaux est constitué à partir des tableaux $\mathbf{X}_k$,

-- Π est la matrice diagonale constituée des poids π_k, π_k un poids positif attribué *a priori* à au tableau $\mathbf{X}_k$.

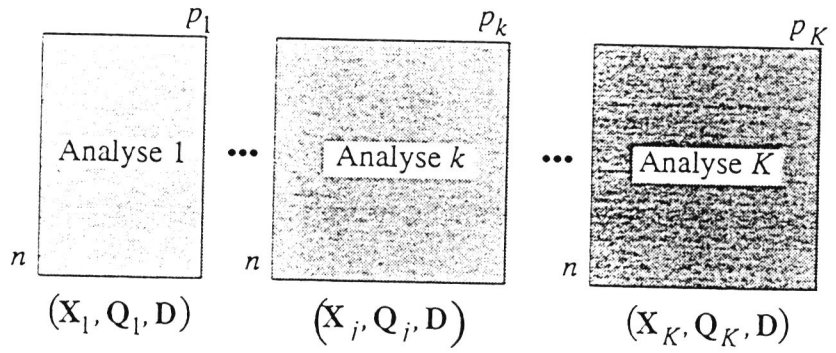


Figure 2. Construction d'une étude K-tableaux horizontale.

Première Partie

Structure de l'univers : Eléments théoriques

I.0. Définition du problème et objectifs

1. Fonction mathématique et fonction multivariée

L'algèbre linéaire est fondée sur les notions d'espace vectoriel euclidien, de la dimension d'un sous espace vectoriel, d'orthogonalité, d'application linéaire, et de rang d'application linéaire. Ces notions offrent le théorème de Décomposition en Valeurs Singulières (DVS) qui est à la base des analyses multivariées des tableaux à deux entrées.

L'apport de ce théorème réside dans deux fonctions au moins, l'une purement mathématique, et l'autre multivariée.

La fonction mathématique de la Décomposition en Valeurs Singulières est son aspect très interactif, dans le sens qu'elle exprime une résolution simultanée de plusieurs problèmes de nature théorique différente (Young & Housseholder, 1938; Darroch, 1965, Rutishauser, 1969; Chen, 1974; Rao, 1979). Citons notamment deux problèmes :

— *algorithmique :* les solutions de la décomposition en valeurs singulières sont les solutions de plusieurs problèmes d'optimisation différents (figure 1. a).

— *géométrique* : les solutions de la décomposition en valeurs singulières sont l'expression analytique de la question de l'étude du lien entre deux géométries, chacune est définie par un nuage de vecteurs, l'un pour les lignes (figure 1. b1, b2), l'autre pour les colonnes (figure 1. b1). Par la géométrie d'un nuage de vecteurs, on sous-entend son état typologique.

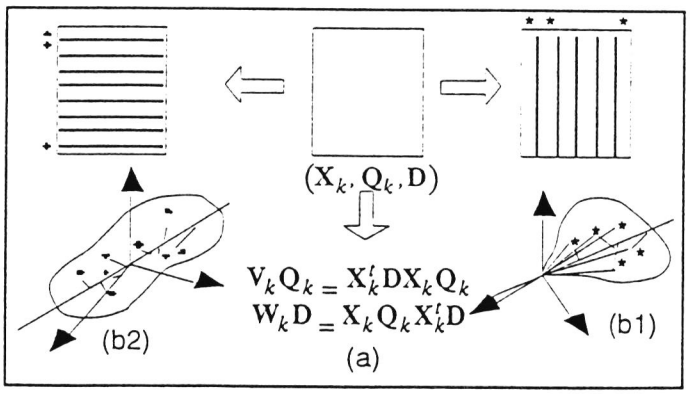


Figure 1. Un principe de fonctionnement unique pour les analyses des tableaux à deux entrées.

Le champ multivarié des tableaux à deux entrées est caractérisé par la diversité de ses analyses. Cette diversité a amené à s'interroger sur son lien à des fonctionnements intrinsèquement différents. La DVS donne la réponse à la question et révèle que le cœur de toutes ces analyses se rapporte, à des métriques près, à un seul principe de fonctionnement. Celui-ci consiste en un seul aboutissement géométrique (deux nuages de vecteurs) et en un seul algorithme de résolution (figure 1).

Dans ce sens, la DVS a une fonction multivariée d'une autre nature que sa fonction mathématique. Cette fonction multivariée réside dans son aspect unificateur de la diversité

des analyses des tableaux à deux entrées. Cette diversité est alors uniquement liée à des objectifs pratiques différents; sa prise en compte dans une unification équivaut à une traduction analytique des différents objectifs. Celle-ci s'effectue par un prétraitement des données et par l'introduction des métriques $\mathbf{Q_k}$ et $\mathbf{D}$ (Escoufier, 1977).

2. Eléments de motivation

Deux questions paraissent naturelles lorsqu'on passe des analyses des tableaux à deux entrées à celles des tableaux à trois entrées; ces deux questions visent :

— Premièrement, la recherche d'un cadre théorique pour les analyses des tableaux à trois entrées, qui aurait la même résonance que l'algèbre linéaire étant elle-même le cadre théorique des analyses des tableaux à deux entrées.
Le constat que les tableaux à deux entrées définissent des formes bilinéaires et que les tableaux à trois entrées (les multitableaux) définissent des formes trilinéaires, a fait de l'algèbre multilinéaire, plus exactement de l'algèbre tensorielle, le cadre théorique potentiel, d'autant plus que l'algèbre multilinéaire est une extension naturelle de l'algèbre linéaire (Franc, 1992).

— Deuxièmement, la recherche d'un théorème qui permettrait d'exprimer une fonction multivariée pour ces analyses; d'une manière analogue en algèbre linéaire, le théorème de Décomposition en Valeurs Singulières a une fonction multivariée dans le cadre des analyses des tableaux à deux entrées.
Dans le cadre de l'algèbre multilinéaire, cette question a amené à s'interroger sur la possibilité d'avoir un équivalent théorique qui soit une extension naturelle du théorème de décomposition en valeurs singulières. En d'autres termes :
peut-on exprimer la même fonction mathématique (résolution simultanée de plusieurs problèmes provenant d'horizons théoriques différents) pour les tenseurs ou bien cette fonction mathématique est-elle caractéristique des applications linéaires ?

La notion de rang d'une matrice est un concept théorique fondamental. Kruskal (1989) propose une revue des définitions de cette notion pour les tenseurs. Il note que, si du point de vue purement théorique, l'extension de cette notion aux tenseurs est concevable, des différences majeures et des complications surgissent par rapport au cas d'une matrice. Il continue en donnant une liste de six différences relatives à la notion de rang, entre les matrices - considérées comme des tenseurs d'ordre deux - et les tenseurs d'ordre trois (multitableaux) et plus.
Les démarches de Franc (1989) et de Denis & Dhorne (1989) sont plus directes. Elles ont le point commun de montrer que l'association "rang <--> existence d'une décomposition orthogonale" qui est possible dans le cas des matrices (fonction mathématique de la DVS)

et fondamentale dans les analyses des tableaux à deux entrées (fonction multivariée de la DVS), n'est pas possible dans le cas des tenseurs (multitableaux). Dans des cas simples, elles présentent des démonstrations basées sur des contre-exemples.

Dans cette sphère, les travaux cités ci-dessus révèlent d'une part, la complexité de l'objet multitableaux et d'autre part, sa différence fondamentale par rapport à un tableau à deux entrées. L'invariance de toutes les propriétés théoriques rencontrées dans le théorème de la Décomposition en Valeurs Singulières, invariance qui caractérise sa fonction mathématique, n'est pas vérifiée pour les tenseurs, *a priori* pour les tableaux à trois entrées. Il en découle que les réalisations simultanées de toutes ces propriétés sont une caractéristique des matrices, *a fortiori* des tableaux à deux entrées.

Dans une autre sphère, présentée souvent sans lien avec les résultats exposés précédemment, la richesse et la diversité des propositions existantes (analyses des tableaux à trois entrées) ont suscité la question de leurs liens éventuels. Cette question a été étudiée à plusieurs niveaux :

— au niveau méthodologique : elle consiste en une traduction analytique des objectifs pratiques liés aux analyses des tableaux à trois entrées (Van de Geer, 1984), un peu à l'image des métriques dans le cas des analyses des tableaux à deux entrées.

— au niveau algorithmique : elle consiste, d'une part, à établir que les analyses reviennent à optimiser diverses fonctions sous diverses contraintes (Ten Berge, 1977; Meyer, 1991; Sabatier, 1992), d'autre part, à proposer des principes généraux de résolution pour ces problèmes d'optimisation (Groenen, 1993; Kiers, 1995; Heiser 1995).

— au niveau hiérarchique : elle consiste à décrire la complexité des analyses les unes par rapport aux autres (Carroll & Wish, 1974; Kiers, 1991).

Ces travaux ont permis différentes contributions : compléter des analyses existantes, offrir des algorithmes convergents, comprendre l'articulation des analyses et proposer de nouvelles perspectives.

3. Définition du problème

On voit dans les résultats négatifs établis dans la première sphère (absence d'un théorème présentant une fonction mathématique interactive) une justification et une motivation de la deuxième sphère (liens entre les analyses existantes). En effet, les travaux de la première sphère montrent que la structure de l'ensemble des analyses existantes est complexe. Ceci provient d'une part, de la complexité des objets que les

analyses manipulent (tableaux à trois entrées) et d'autre part, de l'absence, pour ces objets (tenseurs), d'un théorème qui présenterait la même fonction mathématique que le théorème de Décomposition en Valeurs Singulières.
Par conséquent, on ne peut considérer l'ensemble de ces analyses sous forme d'une seule unité théorique, selon la logique : "fonction mathématique absente --> fonction multivariée absente". On entend ici par unité théorique, un ensemble d'analyses qui présentent le même fonctionnement théorique.

L'importance de la deuxième sphère peut résider dans l'exploration de la structure de l'ensemble des analyses existantes, en réalisant à des niveaux différents mais restreints des unités théoriques, à l'image de la fonction multivariée de la DVS qui permet d'exprimer l'ensemble des analyses des tableaux à deux entrées sous forme d'une seule unité théorique (figure 2).

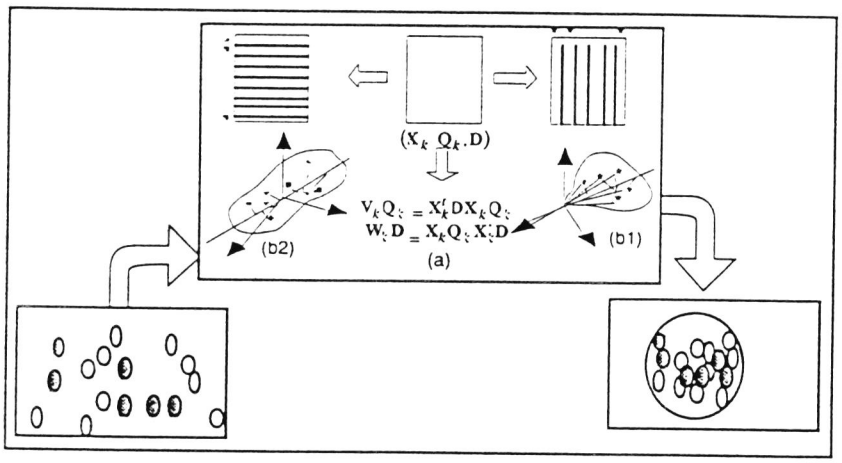


Figure 2. Unité théorique des analyses des tableaux à deux entrées : diversité de l'ensemble des analyses des tableaux à deux entrées (à gauche), chaque point est une analyse. Fonction mathématique de la DVS, elle est le cœur de ces analyses (au centre). Fonction multivariée de la DVS, les analyses sont liées par un seul fonctionnement théorique, pour former une seule unité théorique (à droite).

Si la première sphère nous permet d'avancer l'hypothèse de l'existence de divergences implicites et propres à ces analyses (absence d'une fonction mathématique interactive), la deuxième sphère nous permet de supposer la possibilité de construire des unités théoriques mais à des niveaux restreints. Les deux sphères prises simultanément supposent alors l'existence de plusieurs unités théoriques qui reconstituent l'ensemble de toutes les propositions existantes.

La question posée est de trouver le nombre minimal de ces unités théoriques au sein de l'ensemble de toutes les propositions existantes. On convient d'appeler cet ensemble "Univers", les unités théoriques "Composantes de liens" et le nombre minimal d'unités théoriques reconstituant l'univers " Dimension de l'univers" (figure 3).

Déterminer la dimension de cet univers suppose en particulier la définition d'une composante de liens, autrement dit, la définition d'un fonctionnement théorique. Il est naturel d'établir cette définition à l'image du principe de fonctionnement des analyses des tableaux à deux entrées (méthodologique, géométrique, algorithmique). Plus exactement, on propose de définir une composante de liens comme un sous-ensemble de l'univers. Les éléments de ce sous-ensemble, qui sont des analyses de tableaux à trois entrées, doivent simultanément :

— avoir le même comportement (fonctionnement) géométrique,

— permettre de tenir compte d'un certain nombre de principes méthodologiques,

— admettre des résolutions qui peuvent être calculées par un seul algorithme numérique.

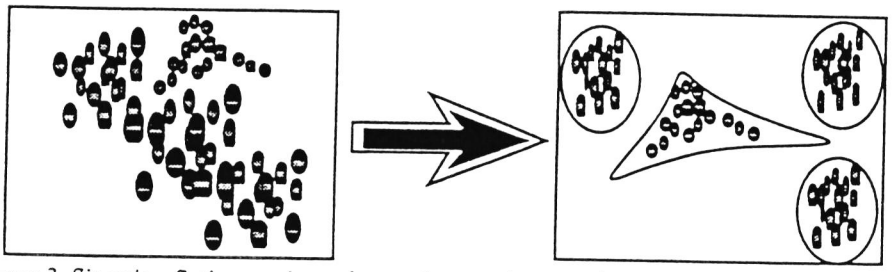


Figure 3. Situation fictive : univers des analyses existantes de tableaux à trois entrées (à gauche), 4 composantes de liens (à droite), la dimension de l'univers est égale à 4

4. Objectifs de cette partie

Dans ce travail, le problème qui consiste à déterminer la dimension de cet univers restera ouvert; par contre, on propose de construire une composante de liens selon la définition présentée ci-dessus.

Comme l'étude qui va être développée dans ce chapitre porte sur cet univers, il est naturel de le décrire. C'est l'objectif du premier chapitre de cette partie.

Pour construire une unité théorique, on considère alors trois approches complémentaires (figure 4) :

— la première approche est géométrique, son objectif est d'introduire une notion de comportement géométrique des éléments de l'univers. On propose deux comportements géométriques propres. Ceci permet d'extraire de l'univers deux sous-ensembles susceptibles de présenter des unités théoriques. C'est le contenu du deuxième chapitre.

— la seconde approche est méthodologique, son objectif est d'analyser la structure de chacun des deux sous-ensembles extraits dans l'approche géométrique. Ceci amène à

compléter les deux sous-ensembles choisis. En particulier, cette complétion permet la proposition de nouvelles analyses. Voici l'objet du troisième chapitre.

— la troisième et dernière approche est algorithmique, pour l'un des deux sous-ensembles résultant de l'opération de complétion dans la section méthodologique, l'aboutissement en termes de problèmes à optimiser, revient à maximiser une famille de fonctions sous diverses contraintes.
Sur la base d'une technique générale pour la construction d'algorithmes convergents, dite "Iterative Majorization" (Heiser, 1995), on construit un algorithme de résolution unique pour maximiser cette famille de fonctions. C'est l'objectif du quatrième chapitre.

Finalement, l'un des deux sous-ensembles est proposé comme composante de liens; c'est l'objectif du cinquième et dernier chapitre. On montre que la composante de liens construite contient comme cas particuliers d'autres démarches proposées dans la littérature et les complète. On considère l'absence d'un algorithme de résolution unique pour l'autre sous-ensemble, comme la perspective d'une deuxième composante de liens.

Au cours du développement de chacun des quatre premiers chapitres, des contributions seront présentées, ainsi que leurs liens avec un certain nombre de résultats existants. Un résumé de l'apport de chaque approche figure à la fin du chapitre correspondant.

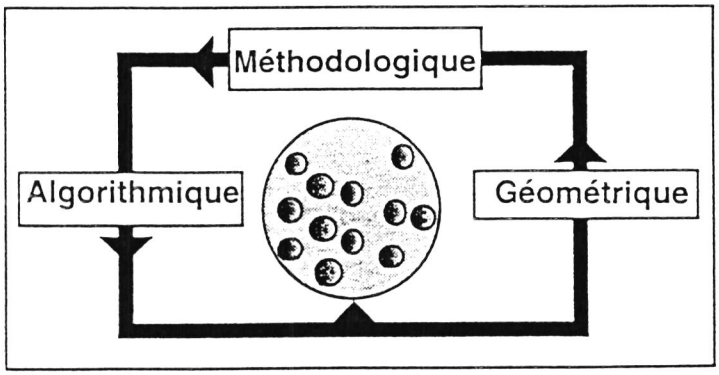


Figure 4. Résumé de la stratégie adoptée dans cette partie pour construire une composante de liens.

Chapitre I.1

Description de l'univers

0. Introdution

On rappelle que l'univers est l'ensemble des analyses des tableaux à trois entrées. Dans ce chapitre, il s'agit de décrire cet univers. On ne prétend pas faire l'examen exhaustif des analyses publiées à ce jour, elles sont très nombreuses, mais on propose de passer en revue les analyses les plus connues, à savoir : ACT-STATIS, CANDECOMP/PARAFAC, INDSCAL, TUCKALS-3, les analyses canoniques généralisées, les analyses de concordance, et les analyses par rotation.
On se restreint principalement à une revue des critères et des contraintes étant à la base des analyses qui sont les éléments de cet univers.

Dans une première section, on passe en revue les critères et contraintes des analyses canoniques généralisées. Dans une seconde section on passe en revue les critères et les contraintes des analyses par rotations qui contiennent les analyses procustes, et des analyses de concordance.

Pour ces deux premières sections, on se place dans le cadre d'une présentation évolutive des critères et des contraintes (Kettenring, 1971; Ten Berge, 1977; Van de Geer, 1984; Ten Berge & Knol 1984; Ten Berge 1986, 1988).

Dans la troisième section, on passe en revue des analyses telles que : ACT-STATIS, CONDECOMP, INDSCAL, et TUCKALS-3. Pour cela, on se place dans le cadre unifié établi par Kiers (1991).

Le choix de cette présentation sur la base des travaux qui ont un caractère synthétique et unifié, va dans le sens de notre motivation principale qui est de construire des unités théoriques. Pour l'examen d'autres aspects des analyses considérées ici ou encore d'autres analyses, on pourra consulter la thèse de Ten Berge (1977), le livre de synthèse de Van de Geer (1986), la thèse de Glaçon (1981), la thèse de Franc (1992) et la synthèse bibliographique de Kronenberg (1995).

Sauf mention contraire, les notations utilisées sont restreintes aux études munies des métriques identités $\left(\mathbf{X}, \mathbf{I}_p, \mathbf{I}_n, \mathbf{I}_K\right)$ (voir pour la définition pp.12-14). On considère le K-tableaux horizontal $\mathbf{X} = \left[\mathbf{X}_1 | \mathbf{X}_2 | \ldots | \mathbf{X}_K\right]$ à n lignes et $p = \sum_{k=1}^{K} p_k$ colonnes, formé à partir des K tableaux $\mathbf{X}_k$ de dimension $n \times p_k$.

1. Analyses canoniques généralisées

1.1 Analyse canonique d'un 2-tableaux horizontal

On note $\mathbf{c}_k$ une combinaison linéaire des colonnes du tableau $\mathbf{X}_k$. L'analyse canonique de Hotteling (1935, 1936) a été proposée pour exploiter les liaisons linéaires entre deux groupes de variables (K = 2). Elle consiste à extraire une combinaison linéaire

des variables c_1 de l'un des deux groupes, dite "composante synthétique", réalisant une corrélation maximale avec une composante synthétique de l'autre groupe, notée c_2. Elle se traduit analytiquement par le problème d'optimisation suivant :

$$\text{Maximiser } \text{Corr}(c_1, c_2)$$

$\text{Corr}(c_1, c_2)$ désigne le coefficient de corrélation linéaire entre les composantes c_1 et c_2.

Une fois trouvé un premier couple de composantes (c_1^1, c_2^1), on continue la recherche en résolvant le même problème de maximisation avec des contraintes supplémentaires. Celles-ci sont telles que les composantes du deuxième couple recherché (c_1^2, c_2^2) sont orthogonales aux composantes déjà trouvées (c_1^1 orthogonale à c_1^2, et c_2^1 orthogonale à c_2^2). Une telle recherche est dite recherche pas à pas ou "successive".

1.2 analyses canoniques d'un K-tableaux horizontal

D'après Kettenring (1971), Vinograde (1950) fut le premier à s'intéresser à la question de la généralisation de l'analyse de Hotteling à plus de deux groupes de variables. Kettenring relate que la proposition de Vinograde n'a pas été suivie car dans le cas de deux tableaux, elle ne redonne pas l'analyse de Hotteling. Plusieurs généralisations qui ne présentent pas cet inconvénient ont été effectuées par la suite. Ainsi, l'analyse dite variance généralisée de Steel (1951) (voir aussi Anderson, 1958 : chapitre 12) consiste à minimiser le déterminant de la matrice de corrélation des composantes synthétiques, ce qui revient analytiquement à minimiser la fonction suivante :

$$\omega_1(c_1, c_2, .., c_K) = \text{Déterminant}\left[\text{corr}(c_k, c_j)\right]$$

Horst (1961, 1965) propose deux analyses. La première est basée sur la maximisation de la somme des éléments de la matrice de corrélation des composantes synthétiques, elle est dite "méthode de la corrélation maximale" :

$$\omega_2(c_1, c_2, .., c_K) = \sum_{k \neq j \; k,j=1}^{K} \text{corr}(c_k, c_j)$$

La deuxième consiste à maximiser la plus grande valeur propre de la matrice de corrélation des composantes synthétiques, c'est ce qu'il a nommé "méthode d'approximation de rang un":

$$\omega_3(c_1, c_2, .., c_K, z) = \sum_{k=1}^{K} \left(\text{corr}(c_k, z)\right)^2$$

$\mathbf{z}$, est un vecteur de $\mathbb{R}^n$ dit vecteur auxiliaire.

McKoen (1966), McDonald (1968) et Carroll (1968) sont arrivés à la "méthode d'approximation de rang un" de Horst, en utilisant différentes approches.

Dans son article de synthèse, Kettenring (1971) ajoute deux propositions nouvelles basées sur deux critères; l'une consiste à maximiser la somme des carrés des éléments de la matrice de corrélation, l'autre à minimiser la plus petite valeur propre de la matrice de corrélation :

$$\omega_4(\mathbf{c}_1,\mathbf{c}_2,..,\mathbf{c}_K) = \sum_{k\neq j\ k,j=1}^{K} \left(\mathrm{corr}(\mathbf{c}_k,\mathbf{c}_j)\right)^2$$

$$\omega_3(\mathbf{c}_1,\mathbf{c}_2,..,\mathbf{c}_K,\mathbf{z}) = \sum_{k=1}^{K} \left(\mathrm{corr}(\mathbf{c}_k,\mathbf{z})\right)^2$$

Finalement, Lafosse (1989) propose le critère suivant, comme intermédiaire entre ceux de Horst et Carroll; il consiste à maximiser :

$$\omega_5(\mathbf{c}_1,\mathbf{c}_2,..,\mathbf{c}_K) = \sum_{k=1}^{K} \left(\mathrm{corr}\left(\mathbf{c}_k, \sum_{j\neq k\ j=1}^{K} \mathbf{c}_j \right) \right)^2$$

1.3. Critères des analyses canoniques généralisées et applications

Dans des contextes différents, plusieurs auteurs ont fait appel à l'ensemble de ces critères ou seulement à une partie d'entre eux.

— Dans le contexte des données qualitatives, Masson (1974) s'est intéressé à un critère de Kettenring, alors que Saporta (1975) a développé le critère proposé par Carroll avec des applications aux données qualitatives. Cette recherche a dégagé la notion de codage des variables qualitatives; elle a permis aussi de dégager l'idée de l'analyse canonique de Hotteling comme cadre théorique général des analyses des tableaux à deux entrées.

Meyer (1989, 1991, 1992) utilise les critères des analyses canoniques généralisées pour proposer des extensions de l'analyse des correspondances de deux variables qualitatives à celles de plusieurs variables qualitatives. Cette extension est basée sur la remarque que l'analyse des correspondances est un cas particulier de l'analyse de Hotteling sur les indicatrices des classes relatives aux variables qualitatives. Sans faire le lien avec les travaux de Saporta, la démarche de Meyer complète implicitement celle de Saporta à l'ensemble des critères des analyses canoniques. En effet, il utilise tous les critères en

question sauf celui de Lafosse et celui de Carroll. Il propose de maximiser le critère de Steel, alors qu'initialement, Steel propose de minimiser son critère.

— Dans le contexte des généralisations de l'analyse en composantes principales à plusieurs groupes de variables, Krzanowski (1979, 1988) propose de minimiser un critère qui revient au second critère de Kettenring (voir aussi Flury, 1995 : pp.16). Casin (1995) utilise un critère de Horst pour présenter une analyse dite Analyse Discriminante des Tableaux (ADT).

— Dans le contexte de la dimension infinie, Dauxois & Pousse (1976) font une étude synthétique et proposent des extensions des analyses canoniques généralisées en dimension infinie. On peut consulter un travail récent dans Dauxois & Romain & Viguier (1993). Dans le domaine de l'analyse multivariée non linéaire, on peut se référer à Gifi (1990), ou Coppi & Balasco (1989, chapitre 3).

2. Analyses par rotation et analyses de concordance

Pas toujours en lien direct avec les préoccupations de l'analyse canonique de Hotteling et ses généralisations, l'analyse par rotation orthogonale ou oblique, qui contient l'analyse Procuste (voir ci-dessous) fonctionne sur un concept différent. En effet, les analyses par rotation ont pour objectif de transformer un nuage vers un autre en déformant le moins possible leur géométrie.

2.1 Analyses par rotation d'un 2-tableaux

Dans le cas de deux tableaux, la thèse de Ten Berge (1977.a) effectue une approche synthétique et montre que les analyses proposées sont basées essentiellement sur l'optimisation de l'un des deux critères suivants :

$$f(\mathbf{T}_1,\mathbf{T}_2) = \mathrm{tr}\left(\mathbf{T}_1^t\mathbf{X}^t\mathbf{Y}\mathbf{T}_2\right)$$

$$g(\mathbf{T}_1,\mathbf{T}_2) = \mathrm{tr}\left(\left(\mathbf{T}_1^t\mathbf{X}^t - \mathbf{T}_2^t\mathbf{Y}^t\right)\left(\mathbf{X}\mathbf{T}_1 - \mathbf{Y}\mathbf{T}_2\right)\right) = \left\|\mathbf{X}\mathbf{T}_1 - \mathbf{Y}\mathbf{T}_2\right\|_{HS}^2$$

$\mathrm{tr}(\mathbf{A})$ désigne la trace de la matrice carrée $\mathbf{A}$. Le critère f est dit critère de produit scalaire, et le critère g est dit critère procuste (Mulaik, 1972). Plus exactement, l'optimisation de ces deux critères consiste à maximiser le critère f et à minimiser le critère g sous diverses contraintes sur $\mathbf{T}_1$ et $\mathbf{T}_2$.

Par la suite, on cite quelques unes de ces contraintes. On considère uniquement le cas de deux tableaux de même dimension $n \times p$ et de même rang égal à p (cas symétrique au sens de Ten Berge & Knol, 1984). Ten Berge (1977.b) a également traité le cas où le

nombre des colonnes est différent dans les deux tableaux (cas asymétrique au sens de Ten Berge). Ces contraintes sont les suivantes :

(•) $\mathbf{T}_k^t\mathbf{T}_k = \mathbf{I}_p \quad k = 1,2,$

(••) $\mathbf{T}_k, \quad k = 1,2$ ont des colonnes normées.

Dans le cas de la première contrainte (•), la rotation est dite orthogonale. Dans le cas de la seconde (••), la rotation est dite oblique. Cette dernière rotation donne généralement des résultats triviaux. C'est la raison pour laquelle, dans ce type de rotation, il est préférable d'utiliser d'autres contraintes comme la suivante :

(•••) les colonnes de la matrice $\mathbf{X}_k\mathbf{T}_k$ sont orthogonales $k = 1,2.$

Van de Geer (1971) fut parmi les premiers à montrer que l'optimisation des deux critères f et g sous la contrainte (•••), revient à optimiser ces mêmes critères sous les contraintes $\mathbf{T}_k^t\mathbf{T}_k = \mathbf{I}_p \quad k = 1,2$, sous réserve de remplacer les deux tableaux $\mathbf{X}_k$ par les projecteurs associés $\mathbf{P}_k \quad k = 1,2.$

On peut aussi fixer le rôle de l'un des deux tableaux et effectuer une rotation orthogonale ou oblique vers l'autre. Cela est possible en rendant l'une des deux contraintes inactive par la relation : $\mathbf{T}_1 = \mathbf{I}_p$, ou $\mathbf{T}_2 = \mathbf{I}_p$. Le critère g s'écrit :

$$h(\mathbf{T}) = \mathrm{tr}\Big(\big(\mathbf{T}_1^t\mathbf{X}^t - \mathbf{Y}^t\big)\big(\mathbf{X}\mathbf{T}_1 - \mathbf{Y}\big)\Big).$$

L'un des cas particuliers du critère h remonte jusqu'à Mosier (1939). On trouve dans Gower (1984) une synthèse des analyses basées sur la minimisation du critère h sous différentes contraintes.

Finalement, on peut selon les cas, sous les contraintes citées, construire les solutions d'une manière successive (pas à pas) ou simultanée.

Les choix sur les tableaux (tableaux initiaux, projecteurs) et les choix possibles des contraintes (orthogonale, oblique, active pour les deux tableaux, inactive pour l'un des deux tableaux) peuvent générer jusqu'à 24 analyses possibles. On trouve dans Ten Berge (1977.b) les solutions algorithmiques aux problèmes d'optimisation associés à ces analyses.

2.2. Généralisation à un K-tableaux ou un multitableaux

La généralisation des analyses ci-dessus à plus de deux tableaux était une question naturelle, parallèle à celle qui s'effectue dans le cadre des analyses canoniques généralisées. Deux critères ont été proposés :

— le premier généralise le critère f, donné par :

$$\tilde{f}(\mathbf{T}_1,\mathbf{T}_2,\ldots,\mathbf{T}_K)=\sum_{k<l}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)$$

— le second généralise le critère g, défini comme :

$$\tilde{g}(\mathbf{T}_1,\mathbf{T}_2,\ldots,\mathbf{T}_K)=\sum_{k<l}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'-\mathbf{T}_l'\mathbf{X}_l'\right)\left(\mathbf{X}_k\mathbf{T}_k-\mathbf{X}_l\mathbf{T}_l\right).$$

Dans un premier temps et dans le cas symétrique, Ten Berge & Knol (1977) montrent que les deux critères $\tilde{f}$ et $\tilde{g}$ sont équivalents. Dans un deuxième temps, ils traitent le cas asymétrique et ils retrouvent le critère proposé par Gower (1975) dans le cadre de l'analyse procuste généralisée.

Dans les deux cas (symétrique et asymétrique), la motivation principale des auteurs a été de proposer des extensions des analyses par rotation au cas de plus de deux tableaux, ainsi que d'apporter des résolutions algorithmiques. Ten Berge & Knol (1984) mettent d'avantage l'accent sur les analyses par rotations orthogonales. Dans le cadre de l'analyse procuste à plusieurs tableaux, on peut trouver un exposé plus détaillé dans Gower (1995).

Pratiquement au même moment, Van de Geer (1984, 1986) considère une famille de critères pour l'étude de la relation entre K groupes de variables. Il discute à quoi ces critères ramènent dans le cas $K=2$ et comment ils peuvent être généralisés dans le cas $K\geq 3$. Il traite uniquement la situation où les tableaux de données ont même nombre de lignes et même nombre de colonnes $m=p_1=p_2=\cdots=p_K$ (cas symétrique au sens de Ten berge & Knol). Dans sa démarche, Van de Geer va directement au but en proposant ce qu'il appelle "basic choices" qu'on traduit ici par "choix de base". Le point de vue de Van de Geer est intéressant car en dehors de tout argument algorithmique, il justifie explicitement la diversité des critères et des contraintes. Il explique comment ces choix sont associés à des préoccupations souvent divergentes (Van de Geer, 1984, pp.80).
Van de Geer propose quatre critères. Les plus connus d'entre eux sont MaxBet et MaxDiff. Les problèmes MaxBet et MaxDiff se définissent analytiquement comme deux problèmes d'optimisation :

Le problème MaxBet consiste à maximiser le critère suivant :

$$\psi_1(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k,l=1}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right),$$

alors que le problème MaxDiff consiste à maximiser le critère :

$$\psi_2(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k\neq l\ k,l=1}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right),$$

avec $\mathbf{T}_k$ des matrices d'ordre $m \times m$. On note par $\mathbf{T}$ la matrice définie par :

$$\mathbf{T}^t = \left[\mathbf{T}_1^t \quad \mathbf{T}_2^t \quad \cdots \quad \mathbf{T}_K^t\right].$$

Ces maximisations sont soumises aux trois contraintes suivantes :

(c1) $\mathbf{T}$ a une seule colonne de norme égale à K,

(c2) $\mathbf{T}_k$ a une seule colonne de norme égale à 1,

(c3) $\mathbf{T}_k$ doit être orthogonale.

Les matrices $\mathbf{X}_k$ désignent, soit les tableaux d'origine, soit les projecteurs associés.

Ten Berge (1986, 1988) a unifié le plus possible les deux problèmes MaxBet et MaxDiff. Il a traité le cas général où les matrices ont des nombres de colonnes différents (les nombres p_i ne sont pas nécessairement égaux). Il a montré aussi que (Ten Berge, 1986) les trois contraintes proposées par Van de Geer peuvent être unifiées sous forme d'une seule contrainte générale :

$$\mathbf{T}_k^t \mathbf{T}_k = \mathbf{I}_r, \; k = 1, 2, \ldots, K, \; r \leq \min\{p_k \quad k = 1, 2, \ldots, K\}.$$

Explicitement, MaxBet avec la contrainte (c1) revient au cas K=1 appliqué à la matrice globale $\mathbf{X}$; (c2) et (c3) sont trivialement réunies sous la contrainte $\mathbf{T}_k^t \mathbf{T}_k = \mathbf{I}_r$. De plus, Ten Berge a offert des algorithmes pour résoudre le problème général MaxBet. Le problème parallèle MaxDiff a été, lui aussi, résolu par Ten Berge et Knol (1984), mais dans un cadre un peu moins général que MaxBet car il ne traite pas le cas $\mathbf{t}^t\mathbf{t} = K$.

On reviendra plus en détails sur les choix de Van de Geer et les contributions de Ten Berge dans le troisième chapitre I.3 (approche méthodologique) de cette partie.

3. Analyses de type PCA-SUP

On se restreint ici aux études K-tableaux et multitableaux munies des métriques identités, c'est-à-dire aux quadruplets $\left(\mathbf{X}, \mathbf{I}_p, \mathbf{I}_n, \mathbf{I}_K\right)$. Pour une matrice $\mathbf{A} = \left[a_{ij}\right]$ d'ordre $n \times p$, on note par $\mathrm{vec}(\mathbf{A})$ le vecteur constitué à partir des colonnes de $\mathbf{A}$ par juxtaposition verticale de celles-ci. Pour une matrice $\mathbf{B}$ d'ordre $m \times q$, on note par $\mathbf{A} \otimes \mathbf{B}$ la matrice d'ordre $nm \times pq$, constituée à partir des sous matrices $a_{ij}\mathbf{B}$. La matrice $\mathbf{A} \otimes \mathbf{B}$ est dite produit de Kronecker des deux matrices $\mathbf{A}$ et $\mathbf{B}$. Pour K matrices $\mathbf{Z}_k$ de dimension $n \times p$, on note par $\hat{\mathbf{Z}}$ une matrice de dimension $np \times K$ définie de la manière suivante :

$$\hat{\mathbf{Z}} = \left[\text{vec}(\mathbf{Z}_1) \quad \text{vec}(\mathbf{Z}_2) \quad \cdots \quad \text{vec}(\mathbf{Z}_K)\right].$$

Dans une approche unificatrice, Kiers (1991) montre qu'une grande partie des analyses existantes optimisent un seul critère sous diverses contraintes. Il distingue deux cas, les multitableaux et les K-tableaux. On respecte ici cette distinction. Par contre, le critère et les contraintes seront présentés d'une manière sensiblement différente de celle de Kiers. On commence par considérer le cas d'une étude multitableaux; on envisagera ensuite celui d'une étude K-tableaux.

3.1. Analyses des multitableaux

Dans le cas d'une étude multitableaux, Kiers montre que la base d'une partie des analyses existantes est le problème d'optimisation (**P**) suivant :

$$\text{Minimiser } PCASUP\left(\hat{\mathbf{X}}, \hat{\mathbf{F}}, \mathbf{C}\right) = \left\|\hat{\mathbf{X}} - \hat{\mathbf{F}}\mathbf{C}'\right\|_{HS}^2 \text{ sous les contraintes } \mathbf{C} \in \Xi,\ \hat{\mathbf{F}} \in \Sigma$$

L'idée de Kiers consiste à lier des analyses par un jeu de contraintes. On cite par la suite les contraintes ainsi que les noms des analyses associées au problème (**P**). On préfère une présentation des contraintes sous forme d'ensembles, initialement l'auteur définit plus explicitement ces contraintes. Les contraintes portent sur deux ensembles, le premier est l'ensemble Ξ, le deuxième est l'ensemble Σ. Plus exactement, l'ensemble Ξ consiste en deux ensembles qu'on note Ξ_1 et Ξ_2. L'ensemble Σ consiste en plusieurs ensembles qu'on note Σ_i (voir ci-dessous). On note par l'entier r le rang de la matrice $\hat{\mathbf{X}}$.

L'ensemble Ξ consiste en deux contraintes :

— la première est l'ensemble des matrices d'ordre $K \times r$, qu'on note par $\mathbb{R}^{K \times r}$, autrement dit $\Xi_1 = \mathbb{R}^{K \times r}$.

— la deuxième est un sous ensemble Ξ_2 de $\mathbb{R}^{K \times r}$ défini comme suit :

$\mathbf{C} \in \Xi_2$ si et seulement si la matrice $\mathbf{C}$ a ses K lignes identiques.

Une telle matrice $\mathbf{C}$ s'écrit alors sous la forme : $\mathbf{C}' = [\mathbf{x} \quad \mathbf{x} \quad \cdots \quad \mathbf{x}]$, $\mathbf{x}$ étant un vecteur ligne de la matrice $\mathbf{C}$.

Pour retrouver des analyses existantes générées par le problème d'optimisation (**P**), Kiers fixe la première contrainte $\Xi_1 = \mathbb{R}^{K \times r}$. L'ensemble Σ peut être alors l'une des contraintes suivantes :

-- $\Sigma_1 = \mathbb{R}^{np \times r}$: ensemble des matrices de dimension $np \times r$.

Le problème (**P**) associé est l'analyse en composantes principales de la matrice $\hat{\mathbf{X}}$. C'est le cas dans le calcul du compromis de l'analyse de Jaffrenou (1978), initialement proposée par Tucker (1966).

-- $\Sigma_2 = \left\{(\mathbf{A}\otimes\mathbf{B})\hat{\mathbf{G}} \quad \mathbf{A}\in\mathbb{R}^{n\times r_1}, \mathbf{B}\in\mathbb{R}^{p\times r_2}, \hat{\mathbf{G}}\in\mathbb{R}^{r_1 r_2\times r}\right\}$, avec $r_1 < n,\ r_2 < p$.

Le problème (**P**) associé est à la base de l'analyse dite de Tuckals-3 qui est présentée dans Tucker (1966); pour une résolution algorithmique, on peut voir (Kronenberg & De Leeuw, 1980), pour une synthèse sur les travaux liés à cette analyse, on peut consulter (Kronenberg, 1983; Franc, 1992).

Le problème (**P**) relatif aux deux contraintes Ξ_1 et Σ_2 est équivalent au problème (**P**) relatif aux contraintes d'orthogonalité Σ_2', Σ_2'', $\Sigma_2' \cap \Sigma_2''$, et Ξ_1' (voir ci-dessus). Plus exactement, si on remplace dans le problème (**P**) la contrainte Ξ_1 par la contrainte Ξ_1', et la contrainte Σ_2 par l'une des trois contraintes Σ_2', Σ_2'', ou $\Sigma_2' \cap \Sigma_2''$, la valeur minimale du critère ne change pas (Kiers, 1991, pp.453). Les contraintes Σ_2', Σ_2'', et $\Sigma_2' \cap \Sigma_2''$ sont les suivantes :

$$\Sigma_2' = \left\{(\mathbf{A}\otimes\mathbf{B})\hat{\mathbf{G}} \quad \mathbf{A}\in\Omega_1, \mathbf{B}\in\mathbb{R}^{p\times r_2}, \hat{\mathbf{G}}\in\mathbb{R}^{r_1 r_2\times r}\right\}, \text{ avec } \Omega_1 = \left\{\mathbf{A}\in\mathbb{R}^{n\times r_1}, \mathbf{A}^t\mathbf{A} = \mathbf{I}_{r_1}\right\}$$

$$\Sigma_2'' \left\{(\mathbf{A}\otimes\mathbf{B})\hat{\mathbf{G}} \quad \mathbf{A}\in\mathbb{R}^{n\times r_1}, \mathbf{B}\in\Omega_2, \hat{\mathbf{G}}\in\mathbb{R}^{r_1 r_2\times r}\right\}, \text{ avec } \Omega_2 = \left\{\mathbf{B}\in\mathbb{R}^{p\times r_2}, \mathbf{B}^t\mathbf{B} = \mathbf{I}_{r_2}\right\}.$$

$$\Sigma_2' \cap \Sigma_2'' = \left\{(\mathbf{A}\otimes\mathbf{B})\hat{\mathbf{G}} \quad \mathbf{A}\in\Omega_1, \mathbf{B}\in\Omega_2, \hat{\mathbf{G}}\in\mathbb{R}^{r_1 r_2\times r}\right\}$$

$$\Xi'' = \left\{\mathbf{C}\in\mathbb{R}^{K\times r} \quad \mathbf{C}^t\mathbf{C} = \mathbf{I}_r\right\}$$

$$-- \ \Sigma_3 = \left\{\left[\text{vec}\left(\mathbf{a}_1\mathbf{b}_1^t\right) \quad \text{vec}\left(\mathbf{a}_2\mathbf{b}_2^t\right) \quad \cdots \quad \text{vec}\left(\mathbf{a}_r\mathbf{b}_r^t\right)\right] \quad \mathbf{A}\in\mathbb{R}^{n\times r}, \mathbf{B}\in\mathbb{R}^{p\times r}\right\}$$

Les vecteurs $\mathbf{a}_j \in \mathbb{R}^n$ sont les vecteurs colonnes de la matrice $\mathbf{A}$, les vecteurs $\mathbf{b}_j \in \mathbb{R}^p$ sont les vecteurs colonnes de la matrice $\mathbf{B}$.

Le problème (**P**) associé est à la base de l'analyse proposée simultanément par Carroll & Chang (1970) qui nomment cette analyse "CANDECOMP" (CANonical DECOMPosition), alors que Harshmann (1970) la nomme "PARAFAC" (PARAllel FACtor), elle est dite CANDECOMP/PARAFAC. On trouve dans Franc (1992) une synthèse des travaux liés à cette analyse.

Dans l'analyse de TUCKALS-3, les contraintes d'orthogonalité Σ_2', Σ_2'' et $\Sigma_2' \cap \Sigma_2''$ laissent invariant le problème (**P**), ce n'est pas le cas de l'analyse CANDECOMP/PARAFAC, ce qui a permis de proposer des variantes de l'analyse CANDECOMP/PARAFAC sous des contraintes d'orthogonalité. Plus explicitement, soient les deux ensembles :

$$O_1 = \left\{ \mathbf{A} \in \mathbb{R}^{n \times r} \quad \mathbf{A}^t\mathbf{A} = \mathbf{I}_r \right\} \text{ et } O_2 = \Sigma_3 \cap \left\{ \mathbf{B} \in \mathbb{R}^{p \times r} \quad \mathbf{B}^t\mathbf{B} = \mathbf{I}_r \right\},$$

on définit les contraintes suivantes :

-- $\Sigma_4 = \left\{ \left[\text{vec}\left(\mathbf{a}_1\mathbf{b}_1^t\right) \quad \text{vec}\left(\mathbf{a}_2\mathbf{b}_2^t\right) \quad \cdots \quad \text{vec}\left(\mathbf{a}_r\mathbf{b}_r^t\right) \right] \quad \mathbf{A} \in O_1, \mathbf{B} \in \mathbb{R}^{p \times r} \right\}$

-- $\Sigma_4' = \left\{ \left[\text{vec}\left(\mathbf{a}_1\mathbf{b}_1^t\right) \quad \text{vec}\left(\mathbf{a}_2\mathbf{b}_2^t\right) \quad \cdots \quad \text{vec}\left(\mathbf{a}_r\mathbf{b}_r^t\right) \right] \quad \mathbf{A} \in \mathbb{R}^{n \times r}, \mathbf{B} \in O_2 \right\}$

-- $\Sigma_5 = \left\{ \left[\text{vec}\left(\mathbf{a}_1\mathbf{b}_1^t\right) \quad \text{vec}\left(\mathbf{a}_2\mathbf{b}_2^t\right) \quad \cdots \quad \text{vec}\left(\mathbf{a}_r\mathbf{b}_r^t\right) \right] \quad \mathbf{A} \in O_1, \mathbf{B} \in O_2 \right\}$

Les variantes de CANDECAMP/PARAFAC définies par le problème (**P**) relatif respectivement aux contraintes Σ_4, Σ_4' et Σ_5 sont appelées par Kiers respectivement : ORTCP-A, ORTCP-B, ORTCP.

Kiers définit PCA-SUM comme le critère dont les analyses associées correspondent à l'analyse en composantes principales du tableau-somme des K tableaux, ceux-ci constituant le multitableaux. Un exemple de ces analyses est l'analyse du compromis de ACT-STATIS (cas multitableaux) après pondération. Il montre que ces analyses sont aussi des cas particuliers du problème (**P**). En effet, elles sont associées au problème (**P**) relatif aux deux contraintes suivantes :

$$\Xi_2 \text{ et } \Sigma_5 = \left\{ \left[\text{vec}\left(\mathbf{a}_1\mathbf{b}_1^t\right) \quad \text{vec}\left(\mathbf{a}_2\mathbf{b}_2^t\right) \quad \cdots \quad \text{vec}\left(\mathbf{a}_r\mathbf{b}_r^t\right) \right] \quad \mathbf{A} \in O_1, \mathbf{B} \in O_2 \right\}$$

3.2. Analyses des K-tableaux

On passe maintenant au cas des K-tableaux; les analyses suivent la même logique sous réserve de remplacer la matrice $\hat{\mathbf{X}}$ par la matrice $\hat{\mathbf{W}}$,

$$\hat{\mathbf{W}} = \left[\text{vec}(\mathbf{W}_1) \quad \text{vec}(\mathbf{W}_2) \quad \cdots \quad \text{vec}(\mathbf{W}_K) \right]$$

avec $\mathbf{W}_k = \mathbf{X}_k\mathbf{X}_k^t$ dans le cas d'un K-tableaux horizontal, $\mathbf{W}_k = \mathbf{X}_k^t\mathbf{X}_k$ dans le cas d'un K-tableaux vertical. L'entier r désigne maintenant le rang de la matrice $\hat{\mathbf{W}}$, ce qui implique que l'on prend, par rapport au cas des multitableaux, les égalités suivantes :

$$p = n \text{ et } r' = r_1 = r_2.$$

Le problème (**P**) devient dans le cas des K-tableaux le problème (**P'**)

Minimiser $PCASUP\left(\hat{\mathbf{W}}, \hat{\mathbf{F}}, \mathbf{C}\right) = \left\| \hat{\mathbf{W}} - \hat{\mathbf{F}}\mathbf{C}^t \right\|^2$ sous les contraintes $\mathbf{C} \in \Xi$, $\hat{\mathbf{F}} \in \Sigma$

D'une manière plus explicite, l'ensemble $\overline{\Xi}$ consiste en deux contraintes $\overline{\Xi}_1 = \Xi_1$, $\overline{\Xi}_2 = \Xi_2$. L'ensemble $\overline{\Sigma}$ consiste en plusieurs ensembles que l'on note $\overline{\Sigma}_i$; ils sont définis ci-dessous :

— $\overline{\Sigma}_1 = \mathbb{R}^{n^2 \times r}$ l'ensemble des matrices de dimension $n^2 \times r$.

Le problème (**P'**) associé est l'analyse en composantes principales de la matrice $\hat{\mathbf{W}}$. C'est le cas de la première étape dite étape interstructure dans la méthode ACT-STATIS.

— $\overline{\Sigma}_2 = \left\{ (\mathbf{A} \otimes \mathbf{A})\hat{\mathbf{G}} \quad \mathbf{A} \in \mathbb{R}^{n \times r}, \hat{\mathbf{G}} \in \mathbb{R}^{r^2 \times r} \right\}$

Le problème (**P'**) associé est à la base de l'analyse dite 3-modes-scaling, qui est la version Tuckals-3 appliquée aux matrices $\mathbf{W}_k$.

— $\overline{\Sigma}_3 = \left\{ \left[\mathrm{vec}\left(\mathbf{a}_1\mathbf{a}_1^t\right) \quad \mathrm{vec}\left(\mathbf{a}_2\mathbf{a}_2^t\right) \quad \cdots \quad \mathrm{vec}\left(\mathbf{a}_r\mathbf{a}_r^t\right) \right] \quad \mathbf{A} \in \mathbb{R}^{n \times r} \right\}$

les vecteurs $\mathbf{a}_j \in \mathbb{R}^n$ sont les vecteurs colonnes de la matrice $\mathbf{A}$. Le problème (**P'**) associé est à la base de l'analyse INDSCAL qui est une version de CANDECAMP/PARAFAC dans le cas des K-tableaux.

De la même manière que ORTCP-A, ORTOCP-B et ORTOCP sont des variantes sous contraintes d'orthogonalité de CONDECOM/PARAFAC, le problème (**P'**) peut être soumis au contraintes d'orthogonalité pour donner des variantes INDSCAL. Plus explicitement, soit l'ensemble $\overline{\mathrm{O}} = \left\{ \mathbf{A} \in \mathbb{R}^{n \times r} \quad \mathbf{A}^t\mathbf{A} = \mathbf{I}_r \right\}$, on définit la contrainte :

— $\overline{\Sigma}_4 = \left\{ \left[\mathrm{vec}\left(\mathbf{a}_1\mathbf{a}_1^t\right) \quad \mathrm{vec}\left(\mathbf{a}_2\mathbf{a}_2^t\right) \quad \cdots \quad \mathrm{vec}\left(\mathbf{a}_r\mathbf{a}_r^t\right) \right] \quad \mathbf{A} \in \overline{\mathrm{O}}_1 \right\}$

Le problème (**P'**) associé à cette contrainte, est dit INDORT.

3.3 Analyses et représentation pyramidale

Une ordination des contraintes permet de comprendre l'articulation des analyses les unes par rapport aux autres; ainsi, les contraintes peuvent être ordonnées comme suit :

$$\Xi \times \Sigma_1 \supseteq \Xi \times \Sigma_2 \supseteq \Xi \times \Sigma_3 \supseteq \begin{Bmatrix} \Xi \times \Sigma_4 \\ \Xi \times \Sigma_4' \end{Bmatrix} \supseteq \Xi \times \Sigma_5 \supseteq \Xi' \times \Sigma_5$$

$$\overline{\Xi} \times \overline{\Sigma}_1 \supseteq \overline{\Xi} \times \overline{\Sigma}_2 \supseteq \overline{\Xi} \times \overline{\Sigma}_3 \supseteq \overline{\Xi} \times \overline{\Sigma}_4 \supseteq \overline{\Xi}' \times \overline{\Sigma}_4$$

En particulier, cette ordination des contraintes permet une représentation pyramidale des analyses (figure 1), chacune est une version sous contraintes selon cet ordre

d'inclusion jusqu'à SUMPCA. Les contraintes sont présentées de la plus faible contrainte (base de la pyramide) à la plus forte contrainte (apex de la pyramide).

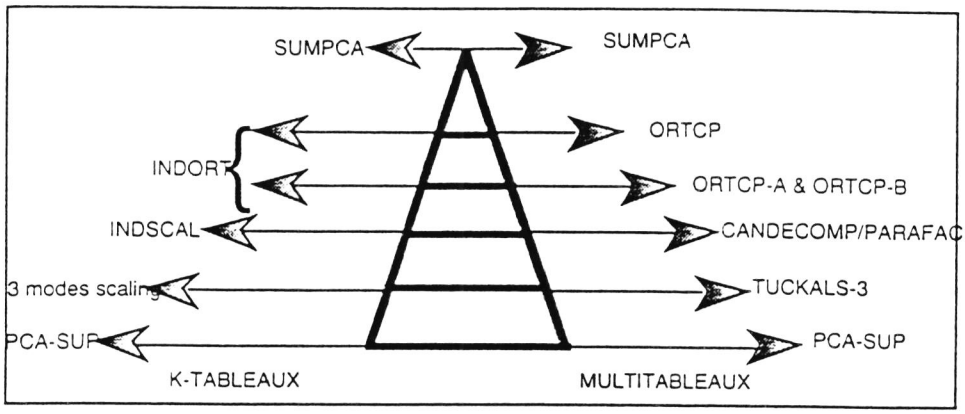


Figure 1. Ordination des analyses selon les contraintes utilisées dans l'optimisation du critère PCASUP. Horizontalement : type des études (K-tableaux ou multitableaux). Verticalement : ordination des contraintes.

4. Conclusion

En conclusion de ce chapitre, un certain nombre d'analyses ont été rappelées sur la base des critères et des contraintes qu'elles utilisent. On estime qu'elles peuvent donner une idée assez représentative tant de l'univers que de la diversité des éléments qui le constituent.

Dans cette présentation, on a insisté surtout sur les travaux de synthèse tels que ceux de Kettenring (1971), Van de Geer (1984), Ten berge (1986, 1988) et Kiers (1991). La raison principale en est que, lors des prochains chapitres, ces travaux constitueront un support des discussions autour de l'objectif principal de cette partie : celui de la construction des composantes de liens.

Chapitre I.2

Approche géométrique

0. Introduction

Dans ce chapitre, on considère le cas d'une étude K-tableaux horizontale, la démarche utilisée reste valable dans le cas d'une étude K-tableaux verticale ou d'une étude multitableaux exprimée ici par deux études K-tableaux horizontale et verticale.
On propose un équivalent géométrique de la problématique liée aux analyses de ces études, sous forme d'un problème de comparaison de plusieurs géométries définies par plusieurs nuages non nécessairement situés dans le même espace euclidien.

Le choix d'une démarche euclidienne (nuage de vecteurs, métrique, espace euclidien) renvoit à une autre question, celle de la recherche d'objets comparables. Trois solutions spécifiques et différentes pour cette question seront considérées. Deux d'entre elles sont des pratiques connues telles que :

— le passage aux opérateurs d'inertie (matrices des produits scalaires) comme par exemple dans la méthode ACT-STATIS,

— la recherche des sous espaces appropriés en analyse de concordance comme par exemple le problème MaxBet.

La troisième solution est considérée comme une alternative concurrente à celle qui s'appuie sur le passage aux opérateurs d'inertie.

On discute dans quelle mesure le problème initial : la comparaison de plusieurs géométries, est équivalent à celui de l'étude de la géométrie définie par les nuages des objets comparables précédemment mentionnés. Selon l'aboutissement aux nuages d'objets, on propose de distinguer deux comportements (fonctionnements) géométriques. Ceci permet d'extraire de l'univers deux sous-ensembles, les éléments de chacun d'eux suivent un comportement propre.

1. Problème géométrique associé à l'analyse d'une étude à trois entrées

Géométriquement, un triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ génère deux nuages de vecteurs (figure 1). L'un a n vecteurs, il est situé dans $\mathbb{R}^{p_k}$ muni de la métrique $\mathbf{Q}_k$, on l'appelle nuage des individus; l'autre a p_k vecteurs, il est situé dans $\mathbb{R}^n$ muni de la métrique D, on l'appelle nuage des variables.

Ce qu'on appelle ici géométrie d'un nuage de vecteurs, c'est l'état typologique de ce nuage. En ce sens, l'analyse d'un triplet consiste en un lien entre la géométrie définie par le nuage des individus et la géométrie définie par le nuage des variables. Il s'agit d'une association types-variables par types-individus.

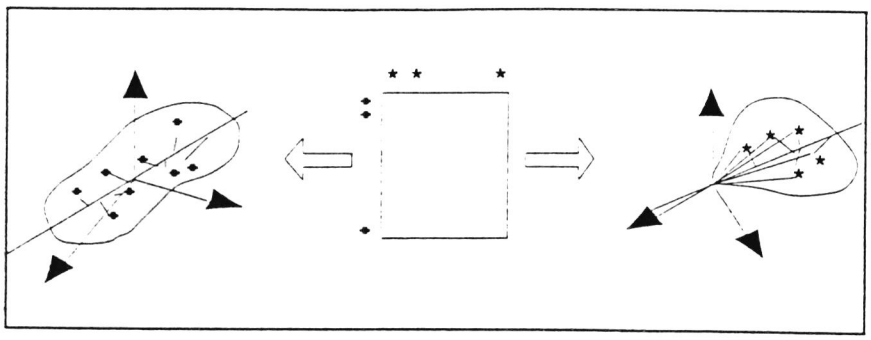


Figure 1. Situation géométrique dans l'analyse d'un triplet statistique $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$.

Une étude K-tableaux horizontale $(\mathbf{X}, \mathbf{Q}, \mathbf{D}, \Pi)$ qui est formée à partir de K triplets $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ génère K nuages d'individus situés dans K espaces euclidiens différents $\mathbb{R}^{p_k}$ (figure 2) et K nuages de variables situés dans le même espace euclidien $\mathbb{R}^n$ (figure 3).

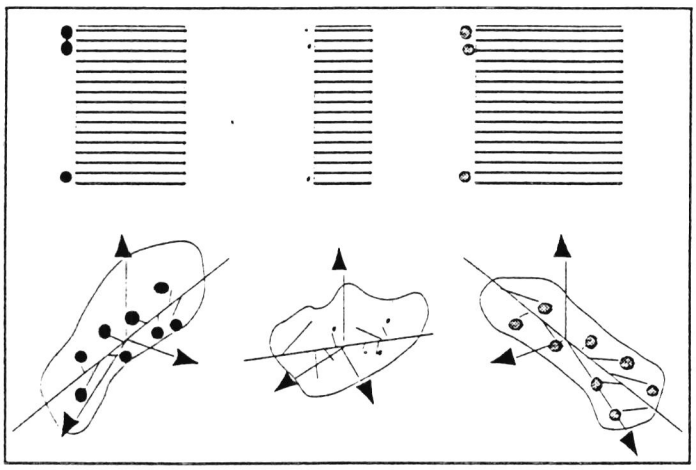


Figure 2. Nuages des individus définis par chaque K triplet dans $\mathbb{R}^{p_k}$.

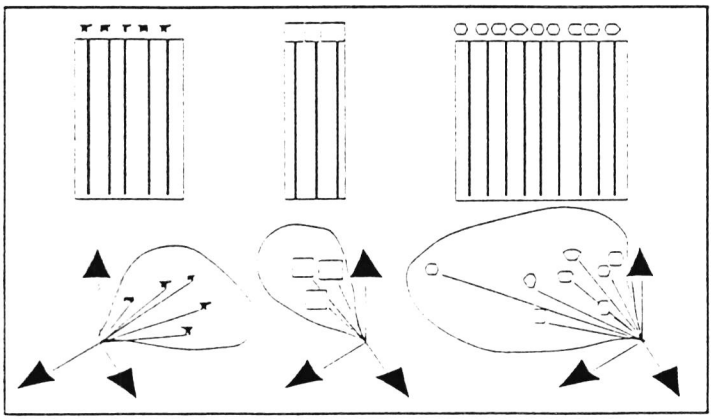


Figure 3. Nuages des variables définis par chaque triplet dans $\mathbb{R}^n$.

Les K analyses définies par les K triplets $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ sont exprimées par K géométries définies par les K nuages d'individus, K géométries définies par les K nuages de variables et K liens "géométrie des individus - géométrie des variables" des K triplets.

1.1. Problème A & problème A'

L'existence de K géométries permet de se poser naturellement la question de leurs liens éventuels, au sens de la définition suivante : "deux géométries sont liées si elles définissent relativement les mêmes types". Cette question des liens entre K géométries peut se formuler *a priori* :

— soit au niveau des K géométries des nuages d'individus,

— soit au niveau des K géométries des nuages de variables,

cette alternative revient à l'un des deux problèmes suivants :

Problème **A**

Étant donné K nuages appariés (les nuages ont le même nombre de vecteurs) mais situés dans K espaces vectoriels, chacun étant muni d'une métrique propre :

peut-on comparer ces K nuages sur la base de leurs géométries respectives ?

Problème **A'**

Étant donné K nuages situés dans le même espace vectoriel euclidien et non appariés deux à deux (les nuages n'ont pas le même nombre de vecteurs) :

peut-on comparer ces K nuages sur la base de leurs géométries respectives ?

1.2. Question de la constructions d'objects euclidiens comparables.

L'approche privilégiée ici est une approche euclidienne. Dans une telle approche, on entend par "comparaison des nuages", la construction d'objets mathématiques de nature vectorielle (application linéaire, rang, ou autres) tels que leur comparaison permet de comparer les nuages (soit au niveau des individus, soit au niveau des variables), autrement dit, les géométries. En effet, les nuages sont les expressions vectorielles de ces géométries.

A première vue, vouloir comparer directement les K nuages (d'individus ou de variables) se heurte à une incompatibilité (ambiguité), due au fait que :

— les K nuages d'individus sont situés dans des espaces vectoriels euclidiens différents (problème **A**),

— les K nuages de variables sont situés dans le même espace vectoriel mais présentent chacun un nombre différent de vecteurs (problème A').

Cette incompatibité est fictive dans le problème A, mais bien réelle dans le problème A'. En effet, le nuage est situé dans un espace vectoriel, mais il est porté par un objet euclidien qui n'est pas dans le même espace vectoriel. Le nuage des individus est situé dans l'espace euclidien $\mathbb{R}^{p_k}$ mais l'objet euclidien porteur de sa géométrie se trouve lui dans l'espace euclidien $\mathbb{R}^n$. De même, le nuage des variables est situé dans l'espace euclidien $\mathbb{R}^n$ mais l'objet euclidien porteur de sa géométrie se trouve dans l'espace euclidien $\mathbb{R}^{p_k}$.
Une comparaison des nuages est associée à des objets de même nature donc comparables (géométries <--> nuages <--> objets comparables); c'est le cas des K nuages d'individus. Par contre, une comparaison des nuages des variables n'est pas possible, au moins d'une manière naturelle car elle implique des objets vectoriellement différents ($\mathbb{R}^{p_k}$).
De ce fait, toute comparaison des K nuages des variables se heurte naturellement à un problème de dimension, d'autant plus que ce problème de dimension n'est pas extérieur à la nature de l'étude K-tableaux mais qu'il est généré par celle-ci et par celle-ci seulement.

1.3 Solutions à la question de la construction d'objets comparables.

Pour construire des objets comparables : condition nécessaire pour résoudre le problème A dans une approche euclidienne, trois solutions spécifiquement différentes seront présentées. Après cela, on expliquera dans quel sens la comparaison d'objets générés par chaque solution implique une comparaison des nuages d'individus.

1.3.1 Première solution

Elle est largement utilisée dans la littérature des étude K-tableaux. On cite à titre d'exemple : ACT-STATIS, INDSCAL. Elle correspond au passage d'un triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ à son opérateur d'inertie $\mathbf{W}_k\mathbf{D}$ qui est défini comme suit :

$$\mathbf{W}_k\mathbf{D} = \mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k'\mathbf{D}$$

Ce passage est appelé par Kiers (1991) "Fitting derived data", dénomination plus précise, selon l'auteur, que celle qu'emploient Harshman & Lundy (1984) qui nomment ce passage "Indirect Fitting Data".

1.3.2 Deuxième solution

Elle correspond à un passage sensiblement différent de celui de la première solution. Ce passage est basé sur la décomposition en valeurs singulières. Plus exactement, cette

solution consiste à passer d'un triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ à un opérateur qu'on note $\mathbf{H}_k\mathbf{D}$, $\mathbf{D}$ symétrique et semi-défini positif, de la manière suivante :

On considère la décomposition en valeurs singulière de $\mathbf{X}_k$, ($p_k < n$) comme suit :

$$\mathbf{X}_k = \mathbf{V}_k\mathbf{D}_k\mathbf{U}_k^t,$$

avec $\mathbf{V}_k^t\mathbf{D}\mathbf{V}_k = \mathbf{U}_k^t\mathbf{Q}_k\mathbf{U}_k = \mathbf{I}_{p_k}$, r_k le rang de $\mathbf{X}_k$, et $\mathbf{D}_k$ est une matrice diagonale contenant les valeurs singulières σ_k^j rangées dans un ordre décroissant. On définit $\mathbf{H}_k\mathbf{D}$ par l'expression suivante :

$$\mathbf{H}_k\mathbf{D} = \mathbf{X}_k\mathbf{Q}_k\tilde{\mathbf{U}}_k\tilde{\mathbf{V}}_k^t\mathbf{D}$$

Le passage aux opérateurs $\mathbf{H}_k\mathbf{D}$ est proposé ici comme une alternative au passage aux opérateurs $\mathbf{W}_k\mathbf{D}$.

1.3.3 La troisième solution

Elle consiste aussi en un passage. Pour un triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ et pour un sous espace vectoriel de $\mathbb{R}^{p_k}$ de dimension $r \leq \min\{r_k, 1 \leq k \leq K\}$, ses vecteurs de base sont stockés par colonne dans une matrice $\mathbf{T}_k$, r_k désigne le rang de la matrice $\mathbf{X}_k$. On passe du triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ à la matrice suivante :

$$\mathbf{X}_k\mathbf{Q}_k\mathbf{T}_k.$$

L'opération géométrique associée à ce passage correspond à la projection $\mathbf{Q}_k$ orthogonale dans l'espace $\mathbb{R}^{p_k}$ du nuage d'individus du triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ sur le sous espace engendré par les colonnes de la matrice $\mathbf{T}_k$. Ce passage est implicitement utilisé en analyse de concordance ou en analyse procuste.

Ces solutions ont un point commun : elles génèrent des objets comparables. On note déjà un autre point commun entre la première et la deuxième solution : c'est un passage à des opérateurs $\mathbf{D}$ symétriques.

2. Formulation géométrique du problème A

2.1 Nuages des objets comparables et problème A

Pour comparer des géométries, on se ramène à la comparaison des nuages et pour comparer ces nuages, on se ramène à la comparaison d'objets (vectoriellement comparables). Par cette démarche, on espère qu'une comparaison des objets permettra effectivement une comparaison des géométries qui est le problème A à résoudre.

Comparer des objets vectoriels consiste à étudier la géométrie qu'ils définissent via le nuage de ces objets. Dans cette logique, il parait naturel que ces objets jouissent d'un contexte euclidien, c'est à dire qu'ils soient considérés comme des vecteurs d'un espace euclidien; c'est le cas pour les trois solutions, comme on va le développer ci-dessous.
En effet pour les deux premières solutions, on considère l'espace vectoriel des opérateurs **D** symétriques, muni du produit scalaire de Hilbert-Schmidt classiquement connu. Ainsi, on peut mesurer l'adéquation de deux objets (de deux géométries d'individus) sur la base du produit scalaire défini; ceci a permis à Escoufier & Robert (1973) d'introduire une mesure de la liaison entre deux triplets statistiques. Plus explicitement, au k^{ème} triplet statistique $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$, on définit une mesure de la géométrie d'individus associée, par :

$$Vav(\mathbf{X}_k) = tr\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k'\mathbf{D}\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k'\mathbf{D}\right) = \sum_i^n \lambda_i^2$$

L'ajustement de deux géométries est alors défini par :

$$Covv(\mathbf{X}_k, \mathbf{X}_l) = tr(\mathbf{W}_k\mathbf{D}\mathbf{W}_l\mathbf{D}) = \sum_i^n \mu_i^2 .$$

Il s'en suit qu'on peut mesurer la corrélation entre deux triplets k et j par le coefficient de corrélation vectoriel qui s'écrit :

$$Rv(\mathbf{X}_k, \mathbf{X}_j) = \frac{Covv(\mathbf{X}_k, \mathbf{X}_j)}{\sqrt{Vav(\mathbf{X}_k)}\sqrt{Vav(\mathbf{X}_j)}}$$

De la même manière, on introduit les mesures de liaisons entre géométries basées sur le deuxième passage :

$$Vav_{\mathbf{H}}(\mathbf{X}_k) = tr\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{U}_k\mathbf{V}'\mathbf{D}\mathbf{X}_k\mathbf{Q}_k\mathbf{U}_k\mathbf{V}'\mathbf{D}\right) = \sum_i^n \lambda_i$$

$$Covv_{\mathbf{H}}(\mathbf{X}_k, \mathbf{X}_j) = tr\left(\mathbf{H}_k\mathbf{D}\mathbf{H}_j\mathbf{D}\right) = \sum_i^n \mu_i$$

$$Rv_{\mathbf{H}}(\mathbf{X}_k, \mathbf{X}_j) = \frac{Covv_{\mathbf{H}}(\mathbf{X}_k, \mathbf{X}_j)}{\sqrt{Vav_{\mathbf{H}}(\mathbf{X}_k)}\sqrt{Vav_{\mathbf{H}}(\mathbf{X}_j)}}$$

La troisième solution utilise le même principe, mais dans ce cas, la structure vectorielle est différente, c'est celle des applications linéaires de $\mathbb{R}^r$ dans $\mathbb{R}^n$. $\mathbb{R}^r$ est muni de la métrique canonique et $\mathbb{R}^n$ est muni de la métrique **D**. On mesure la liaison entre deux objets par le produit scalaire de Hilbert-Schmidt associé à cette nouvelle structure vectorielle. Plus explicitement :

$$Vav(\mathbf{X}_k, \mathbf{T}_k) = tr\left(\mathbf{T}_k^t \mathbf{Q}_k \mathbf{X}_k^t \mathbf{D} \mathbf{X}_k \mathbf{Q}_k \mathbf{T}_k\right)$$

$$Covv\left((\mathbf{X}_k, \mathbf{T}_k), (\mathbf{X}_j, \mathbf{T}_j)\right) = tr\left(\mathbf{T}_j^t \mathbf{Q}_j \mathbf{X}_j^t \mathbf{D} \mathbf{X}_k \mathbf{Q}_k \mathbf{T}_k\right)$$

$$Rv\left((\mathbf{X}_k, \mathbf{T}_k), (\mathbf{X}_j, \mathbf{T}_j)\right) = \frac{Covv\left((\mathbf{X}_k, \mathbf{T}_k), (\mathbf{X}_j, \mathbf{T}_j)\right)}{\sqrt{Vav(\mathbf{X}_k, \mathbf{T}_k)}\sqrt{Vav(\mathbf{X}_j, \mathbf{T}_j)}}$$

Ces démarches communes sont liées à un objectif commun, celui de rester dans une logique euclidienne : construction d'objets comparables --> plongement de ces objets dans des structures euclidiennes --> nuages de ces objets.

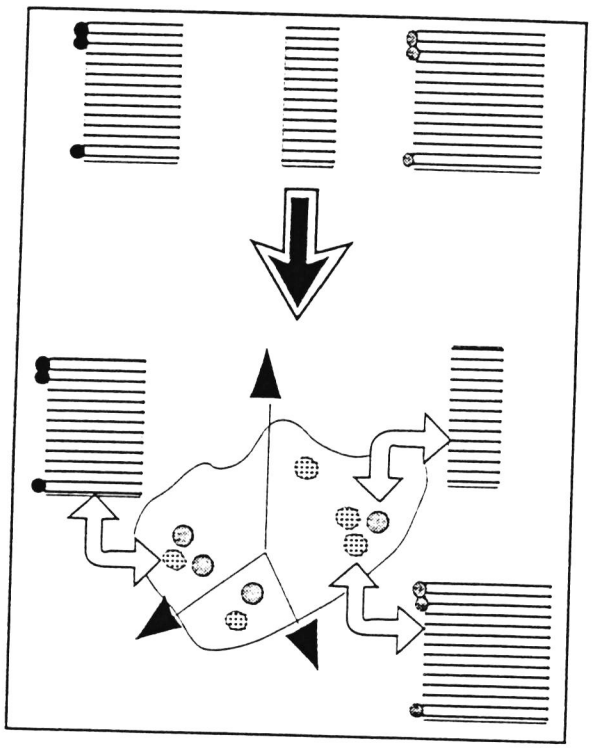


Figure 4. De la comparaison de K géométries d'individus à la géométrie des objets comparables.

Pour les deux premières solutions l'aboutissement en espace euclidien est le même, celui des opérateurs $\mathbf{D}$ symétriques. Leur seule différence est de générer deux nuages d'objets $\{\mathbf{H}_k\mathbf{D} \quad k = 1, 2, \ldots, K\}$ et $\{\mathbf{W}_k\mathbf{D} \quad k = 1, 2, \ldots, K\}$ qui sont distincts mais situés dans le même espace euclidien. Chaque nuage génère un état typologique de ces objets, c'est à dire une géométrie d'objets. Les géométries des deux nuages sont distinctes car les mesures de liaison associées aux objets sont spécifiquement distinctes.

A l'opposé des deux premières, la troisième solution renvoie à une structure euclidienne différente, l'espace vectoriel des applications linéaires de $\mathbb{R}^r$ dans $\mathbb{R}^n$ et elle génère dans le même espace, une famille infinie de nuages d'objets.

En effet, soit $F_k(r)$ l'ensemble des matrices de dimension $p_k \times r$ et de rang r, on considère la famille des nuages associés qui est l'ensemble :

$$\left\{\left\{\mathbf{X}_k\mathbf{Q}_k\mathbf{T}_k \quad k=1,2,....,K\right\} \quad \mathbf{T}_k \in F_k(r) \quad k=1,2,....,K\right\}.$$

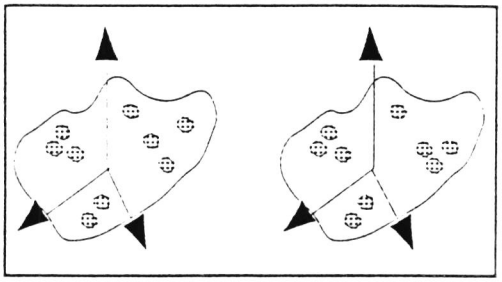


Figure 5. Les deux premières solutions génèrent deux nuages d'objets distincts dans le même espace.

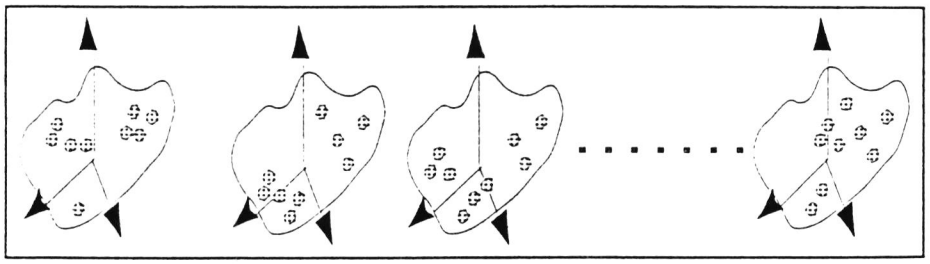


Figure 6. La troisième solution génère une infinité de nuages dans le même espace. Elle en génère autant dans des espaces différents, si on fait varier r entre 1 et r.

Si les trois solutions correspondent à des possibilités théoriques pour construire des objets comparables, leur aboutissement géométrique (nuage d'objets) est conçu de manière différente. Les unes (première et deuxième solutions) conçoivent la question comme une extension naturelle de la logique de l'analyse en composantes principales, avec, pour chacune d'entre elles, un aboutissement à un seul nuage d'objets, l'autre (troisième solution) voit la question par la considération d'une famille infinie de nuages.

Même si les trois solutions utilisent des mesures associées qui sont intrinsèquement identiques (issues du produit scalaire de Hilbert-Schmidt), leur aboutissement typologique (nuages d'objets) qui constitue le point important et l'objectif en analyse multivariée, présente des différences.

2.2. Etats typologiques des nuages d'objets comparables & problème A

Il reste à vérifier que la typologie du nuage d'objets implique la typologie des géométries d'individus; c'est-à-dire que l'on veut établir une équivalence entre le problème A et l'étude de la géométrie définie par les nuages d'objets. Autrement dit, on espère au moins que deux objets de même type dans le nuage des objets définissent la même géométrie au niveau des nuages d'individus. On explique par la suite que les deux

premières solutions et la troisième conçoivent le type des géométries d'individus de deux façons différentes.

L'exemple le plus simple, de deux objets définissant le même type dans un nuage d'objets donné, est le cas extrême où deux objets sont confondus. Nous allons voir que ce cas extrême s'exprime au niveau des géométries d'individus de deux manières différentes pour les trois solutions présentées.

Pour le nuage d'objets généré par la première solution, ce cas extrême signifie qu'il y a égalité entre deux opérateurs d'inertie; on suppose qu'ils sont associés à deux triplets qu'on note $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$, $(\mathbf{X}_l, \mathbf{Q}_l, \mathbf{D})$. En effet, si les opérateurs associés aux triplets $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$, $(\mathbf{X}_l, \mathbf{Q}_l, \mathbf{D})$ sont confondus, alors :

$$\mathbf{W}_k\mathbf{D} = \mathbf{W}_l\mathbf{D}.$$

Cette égalité implique que les deux triplets ont les mêmes composantes principales, qu'ils ont les mêmes inerties associées pour chaque composante principale et que les nuages d'individus sont situés dans des espaces différents mais de même dimension. Autrement dit, on a la même typologie d'individus (même état typologique <--> même géométrie) mais les nuages de ceux-ci sont situés dans des espaces vectoriels différents.

Pour la deuxième solution, ce cas extrême s'exprime par une égalité de deux opérateurs :

$$\mathbf{H}_k\mathbf{D} = \mathbf{H}_l\mathbf{D}$$

cette égalité génère les mêmes conclusions que le cas précédent.

Pour la troisième solution, l'existence de deux objets confondus ne se traduit pas comme dans les deux cas précédents. En effet, si on considère l'ensemble $G_k(r) = \{\mathbf{X}_k\mathbf{Q}_k\mathbf{T}_k \quad \mathbf{T}_k \in F_k(r)\}$, ensemble de tous les sous nuages du nuage des individus du triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ obtenus par projection sur les sous espaces engendrés par les colonnes des $\mathbf{T}_k$, une égalité entre deux objets signifie qu'il existe un autre triplet qu'on note $(\mathbf{X}_l, \mathbf{Q}_l, \mathbf{D})$ tel que :

$$G_l(r) \cap G_k(r) \neq \varnothing$$

Cela signifie qu'il existe deux matrices qu'on note $\tilde{\mathbf{T}}_k$ et $\tilde{\mathbf{T}}_l$ telles que :

$$\tilde{\mathbf{T}}_k \in F_k(r), \text{ et } \tilde{\mathbf{T}}_l \in F_l(r) \;\; \mathbf{X}_k\mathbf{Q}_k\tilde{\mathbf{T}}_k = \mathbf{X}_l\mathbf{Q}_l\tilde{\mathbf{T}}_l$$

Il existe alors un sous nuage qui s'obtient par projection des deux nuages initiaux sur deux sous espaces vectoriels. Ce sous nuage définit un état typologique, c'est à dire une géométrie. Cette géométrie associée à ce sous nuage commun s'exprime dans les deux géométries de départ.

Contrairement aux deux premières solutions, pour les nuages d'objets générés par la troisième, le fait que deux objets définissent le même type ne signifie pas nécessairement qu'ils définissent la même géométrie d'individus des deux triplets $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$, et $(\mathbf{X}_l, \mathbf{Q}_l, \mathbf{D})$.

On parle alors de deux conceptions de la géométrie d'individus, associées aux solutions de la question de la construction d'objets comparables. L'une est associée aux deux premières solutions, l'autre est associée à la troisième.

L'analyse typologique relative à l'une des deux conceptions de la géométrie d'individus ne correspond pas toujours à l'analyse typologique relative à l'autre; une comparaison de géométries peut donc viser les typologies initiales (première conception), comme elle peut viser une co-typologie commune (deuxième conception).

On peut alors définir deux géométries liées ou relativement de même type selon deux conceptions. Selon la première, deux géométries sont relativement les mêmes ou de même type si elles définissent relativement les mêmes états typologiques des deux nuages associés aux individus des deux triplets de départ, considérés séparément.

La deuxième définit deux géométries comme étant relativement les mêmes ou de même type s'il existe un état typologique qui s'exprime relativement dans les deux.

2.3 Notion de comportement géométrique d'une solution du problème A

La formulation des objectifs des analyses des études à trois entrées comme l'étude des liens entre K géométries, a amené à considérer le problème **A**. La résolution de ce problème met l'accent sur la question de la construction d'objets comparables. Plus exactement, la capacité à résoudre le problème **A** dépend *a priori* du potentiel des solutions qu'on peut générer pour résoudre la question de la construction d'objets comparables.

Une solution possible pour cette question (dans ce chapitre, trois solutions possibles) génère alors une solution pour le problème **A**. Le principe est de suivre, pour chacune des solutions possibles, une logique commune. Cette logique consiste d'une part, à générer à partir de cette solution, K objets qui sont comparables, chacun des objets étant

associé à un triplet, et d'autre part, à plonger ces objets dans un espace vectoriel muni d'une structure euclidienne.

Ainsi, pour chacune des solutions possibles, on aboutit à des nuages d'objets situés dans des espaces vectoriels de structures euclidiennes intrinsèquement identiques (produit scalaire d'Hilbert-Schmidt). En conséquence, le problème **A** est équivalent au problème de l'analyse des géométries définies par ces nuages d'objets. Cela montre que la résolution du problème **A** n'admet pas une solution unique. En effet, d'une solution (question de la construction d'objets comparables) à une autre, les nuages d'objets sont différents et leur différence est d'ordre quantitatif (mesure de l'adéquation entre les objets). Par conséquent, les états typologiques de ces nuages d'objets (leurs géométries) sont différents.

Deux raisons au moins, nous permettent de considérer que les résolutions du problème **A** peuvent être partitionnées en deux catégories :

— la première raison est l'aboutissement aux nuages d'objets, ces derniers étant soit en nombre fini, soit en nombre infini. Cette raison est naturelle car théoriquement, le traitement d'un problème infini suit une démarche divergente et souvent une logique plus complexe que celui d'un problème fini. Comme on l'a expliqué plus haut, la première et la deuxième solutions ont le point commun de générer un nombre fini de nuages d'objets (un seul), alors que la troisième solution génère un nombre infini de nuages d'objets.

— la deuxième raison est que les objets générés par les solutions pour la question de la construction d'objets comparables, ne prennent pas en compte de la même façon l'état typologique d'un nuage d'individus.

En effet, comme on l'a déjà expliqué, les objets comparables générés à partir de la première et de la deuxième solutions prennent en compte l'état typologique d'un nuage d'individus de la même façon, celle qui consiste à le considérer comme une entité unique. Ce n'est pas le cas des objets générés par la troisième solution. Plus exactement, l'état typologique, selon cette solution, est considéré comme un ensemble, celui des états typologiques des sous nuages obtenus par projection du nuage d'individus sur des sous espaces. Ici, c'est l'état typologique de chaque sous nuage qui est considéré comme une entité unique.

La résolution du problème **A** est alors relative à la définition qu'on se donne de l'état typologique d'un nuage d'individus. Dans ce sens :

— le problème **A** admet au moins deux solutions, si on considère la définition selon laquelle l'état typologique dans un nuage de vecteurs est une entité unique. C'est cette définition qu'utilisent les objets générés par la première et la deuxième solution.

— le problème **A** admet une solution, si on considère que l'état typologique n'est pas unique dans un nuage de vecteurs, plus exactement si on considère que dans un nuage de vecteurs il existe un ensemble d'états typologiques possibles; c'est ce qui est envisagé par les objets générés par la troisième solution.

Les arguments ci-dessus distinguent deux définitions de la géométrie d'un nuage d'individus et deux aboutissements en termes de nuages d'objets en nombre fini ou infini. On parle alors de deux comportements géométriques. Ces deux comportement sont valables quel que soit le type de données (K-tableaux ou multitableaux).

3. Univers et comportements géométriques

Il s'agit de soumettre l'univers au problème **A**. Cela consiste à lier les analyses qui présentent le même comportement géométrique. Selon les deux comportements définis, il existe alors deux pôles, chaque pôle présente un comportement géométrique propre. En effet, la présence de ces deux comportements géométriques signifie qu'il existe deux sous-ensembles de l'univers, susceptibles de présenter chacun une unité théorique.Chacun de ces deux sous-ensemble de l'univers est dit ici sous-univers.

Plus exactement, la soumission des éléments de l'univers à ces deux comportements permet d'extraire de cet univers deux sous-ensembles :

(a) un premier sous-ensemble peut contenir au moins les analyses suivantes : STATIS, CANDECOMP, INDSCAL et TUCKALS-3.

(b) un deuxième sous-ensemble peut contenir au moins, les analyses canoniques généralisées et les analyses par rotation.

Toutes ces analyses sont présentées dans le chapitre précédent (I.2).

3.1. Premier sous-univers extrait de l'univers

Dans le cas des multitableaux, Kiers montre qu'un certain nombre d'analyses existantes optimisent le critère suivant :

$$PCASUP\left(\hat{\mathbf{X}},\hat{\mathbf{F}},\mathbf{C}\right)=\left\|\hat{\mathbf{X}}-\hat{\mathbf{F}}\mathbf{C}'\right\|_{HS}^{2} \text{ selon les contraintes } \mathbf{C}\in\Xi,\ \hat{\mathbf{F}}\in\Sigma$$

alors que dans le cas des K-tableaux, d'autres analyses optimisent le critère suivant :

$$PCASUP\left(\hat{\mathbf{W}},\hat{\mathbf{F}},\mathbf{C}\right)=\left\|\hat{\mathbf{W}}-\hat{\mathbf{F}}\mathbf{C}'\right\|_{HS}^{2} \text{ selon les contraintes } \mathbf{C}\in\overline{\Xi},\ \hat{\mathbf{F}}\in\overline{\Sigma}.$$

Les contraintes Ξ, Σ, $\overline{\Xi}$, $\overline{\Sigma}$ sont définies dans la premier chapitre de cette partie (I.1).

Dans les deux cas, multitableaux ou K-tableaux, on peut reformuler le critère de ces analyses comme :

$$PCASUP(\hat{\mathbf{Z}}, \hat{\mathbf{F}}, \mathbf{C}) = \left\| \hat{\mathbf{Z}} - \hat{\mathbf{F}}\mathbf{C}' \right\|_{HS}^2 \text{ selon les contraintes } \mathbf{C} \in \Xi_{\hat{\mathbf{Z}}},\ \hat{\mathbf{F}} \in \Sigma_{\hat{\mathbf{Z}}}.$$

Selon la matrice $\hat{\mathbf{Z}}$:

— dans le cas des multitableaux :

$$\hat{\mathbf{Z}} = \hat{\mathbf{X}},\ \Xi_{\hat{\mathbf{Z}}} = \Xi \text{ et } \Sigma_{\hat{\mathbf{Z}}} = \Sigma,$$

— dans le cas des K tableaux :

$$\hat{\mathbf{Z}} = \hat{\mathbf{W}},\ \Xi_{\hat{\mathbf{W}}} = \overline{\Xi} \text{ et } \Sigma_{\hat{\mathbf{W}}} = \overline{\Sigma}.$$

Dire que les analyses définies à partir de l'optimisation du critère PCASUP ont le même comportement géométrique signifie :

— Premièrement, que leur aboutissement est fini ou infini en termes de nuage d'objets.

Pour la question de la construction d'objets comparables, on considère les deux solutions suivantes; elles consistent à générer deux nuages d'objets à savoir :

pour les multitableaux, $\{\mathbf{X}_k \quad k = 1, 2, \ldots, K\}$, pour les K-tableaux $\{\mathbf{W}_k \quad k = 1, 2, \ldots, K\}$

Les analyses basées sur l'optimisation du critère PCASUP considèrent un seul nuage d'objets, chacun correspond à un cas, multitableaux ou K-tableaux.

.— Deuxièmement, que les objets du nuage utilisent une définition propre de l'état typologique d'un nuage d'individus.

Dans le cas des multitableaux, la définition d'un type d'objets s'exprime par le cas extrême que deux tableaux du nuage $\{\mathbf{X}_k \quad k = 1, 2, \ldots, K\}$ sont égaux. En conséquence, cela implique au niveau de la géométrie d'individus que la définition prise de cet état typologique est celle qui consiste à définir un état typologique comme une entité unique. Dans le cas des K-tableaux ce cas a été discuté dans la section précédente.

Il en découle que les analyses STATIS, CANDECOMP, INDSCAL et TUCKALS-3, qui sont liées à l'optimisation de ce critère présentent un même comportement géométrique.

3.2. Deuxième sous-univers extrait de l'univers

Un deuxième sous-ensemble peut contenir au moins, les analyses canoniques généralisées et les analyses par rotation qui sont présentées dans le chapitre précédent (I.1). En termes de nuage d'objets, leur aboutissement est infini. Ces analyses ont le point commun de suivre, dans leur construction des nuages d'objets, la troisième solution présentée ci-dessus. Du point de vue de leur comportement géométrique, leurs différences sont liées uniquement aux critères utilisés.

En conséquence, dans cette troisième section, la soumission de l'univers aux deux comportements géométriques, permet d'extraire de l'univers, deux sous-ensembles susceptibles de présenter des unités théoriques.

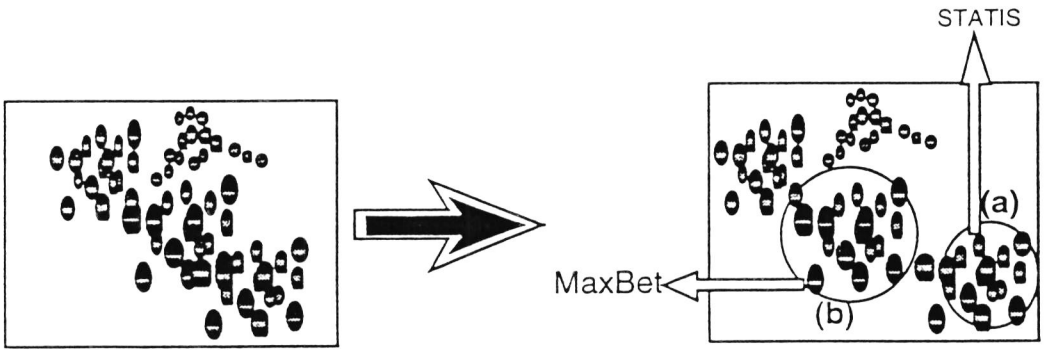


Figure 7. Définition de deux sous-univers de l'univers, chacun présente un comportement géométrique propre.

4. Problème A, comportements géométriques et critères

On vient de soumettre l'univers aux deux comportements géométriques définis précédemment. Ceci a amené à extraire deux sous-ensembles de l'univers qu'on a appelés sous-univers, chacun correspondant à un comportement géométrique propre. Selon la définition du comportement géométrique proposé, on a noté aussi que cette extraction est indépendante des critères utilisés.

Maintenant, on propose de passer à un autre niveau dans la résolution du problème A avec une autre question, celle de l'exploration des géométries des nuages d'objets dans chacun de ces deux sous-univers. On se restreint au point de vue qui consiste à considérer chaque élément des deux sous-univers comme une solution à cette question.

Plus explicitement, on considère que la question d'exploration des géométries des nuages d'objets pose le problème de la recherche d'un critère approprié pour explorer ces géométries. Les critères qui sont à la base des éléments de ces deux sous-univers sont alors considérés comme des critères possibles pour explorer la géométrie des nuages d'objets.

Comme on va l'expliquer ci-dessous, ce point de vue permet en particulier de comprendre la fonction de chacun des deux sous-univers par rapport à la question de l'exploration des géométries des nuages d'objets. Autrement dit, cette question est dépendante des deux comportements géométriques, plus exactement des deux aboutissements géométriques en termes de nuages d'objets.

4.1. Exploration du nuage d'objets par le premier sous-univers

Pour le premier sous-univers, la reformulation de l'approche de Kiers sous forme du problème (**P''**) montre que les éléments de ce sous-univers optimisent un seul critère sous diverses contraintes. En particulier, cet aboutissement sous forme d'un seul problème d'optimisation de ces éléments, a permis de choisir les éléments de ce sous-univers comme ayant le même aboutissement géométrique en terme de nuage d'objets. Par rapport à la question de l'exploration de ce nuage d'objets, il permet aussi d'affirmer que cette question est résolue par un seul critère PCASUP.

4.2. Exploration du nuage d'objets par le deuxième sous-univers

Pour le deuxième sous-univers, la question d'explorer les géométries des nuages d'objets est fondamentalement différente; elle pose deux problèmes qui proviennent essentiellement de l'aboutissement géométrique des élements de ce sous-univers, en nombre infini de nuages d'objets

— Le premier problème est la présence d'un nombre infini de nuages, il vise :

comment peut-on explorer une infinité de nuages ?

— Le deuxième problème est l'exploration de la géométrie de chaque nuage, les nuages considérés étant en nombre infini.

Ces deux problèmes sont résolus d'une manière simultanée et unique par tous les éléments de ce sous-univers. La réponse est inscrite dans le fonctionnement de ces éléments, elle est simple et elle consiste en ceci :

l'exploration de ce nombre infini de nuages d'objets s'effectue par le choix d'un seul d'entre eux.

On peut donner une réponse plus générale : l'exploration de ce nombre infini de nuages s'effectue par le choix d'un nombre fini d'entre eux.

Dans ce sous-univers, le critère a un double rôle, d'une part, effectuer ce choix fini de nuages d'objets, qui résout le problème de la présence de ce nombre infini de nuages et d'autre part, explorer chacun des nuages choisis qui sont en nombre fini.

Le critère pour explorer le nuage d'objets pour les éléments du deuxième sous-univers n'est pas unique, contrairement au cas des élements du premier sous-univers. En effet, les analyses canoniques généralisées par exemple, s'appuient sur plusieurs critères (voir chapitre I.1).

On fait uniquement remarquer ici que dans le cas du premier sous-univers, l'aboutissement géométrique est associé à un seul critère. Dans le cas du deuxième sous-univers, l'aboutissement géométrique est associé à plusieurs critères. Ceci laisse croire que dans le cas d'un seul nuage, on utilise un seul critère, tandis que dans le cas d'un nombre infini de nuages on utilise plusieurs critères. Ainsi, le critère est une conséquence du comportement géométrique défini dans chaque sous-univers. Cette remarque permet d'une part, de renforcer l'idée que les deux sous-univers sont distincts au niveau de leur fonction par rapport à la question de l'exploration de la géométrie des nuages, d'autre part, de donner peut-être le comportement géométrique comme une justification de la diversité des critères qui sont à la base de ces éléments.

5. Conclusion

En conclusion, dans ce chapitre, une approche géométrique a été considérée, son objectif étant de pouvoir établir une discussion autour des liens entre les analyses existantes indépendamment d'une part des critères et des contraintes qui sont à la base de ces analyses, d'autre part des types de tableaux (multitableaux, K-tableaux).

En particulier, on a proposé une reformulation de ces analyses en problème géométrique, ce qui a amené à introduire la notion de comportement géométrique pour lier des analyses existantes. Cette notion a permis de distinguer deux sous-ensembles de cet univers, chacun ayant un comportement géométrique propre. La distinction des deux sous-univers sur la base de ces deux comportements géométriques est considérée ici comme une proposition nouvelle.

La question de la recheche des solutions pour la question de la contruction d'objets comparables révèle son importance comme un passage théorique indispensable pour d'une part, générer des solutions au problème issu de cette formulation géométrique et d'autre part, pouvoir saisir les définitions possibles de ce qu'est l'état typologique d'un nuage de vecteurs. Cet état typologique est à la base de la notion de comportement (fonctionnement) géométrique.

Autrement dit, la recherche d'autres comportements géométriques, au sens de la définition présentée ci-dessus, peut passer par la recherche d'autres solutions possibles pour la question de la construction d'objets comparables.

À un autre niveau et selon le point de vue considéré, la question de l'exploration de la géométrie des nuages permet de comprendre la fonction du critère dans l'analyse et revèle encore plus la distinction basée sur les notions de comportement géométrique.

Chapitre I.3

Approche méthodologique

0. Introduction

Dans l'approche géométrique, on a soumis l'univers aux deux comportements géométriques que l'on a définis. Ceci a amené à extraire deux sous-univers, chacun correspondant à un comportement géométrique propre. On a noté aussi que cette extraction est indépendante des critères utilisés, conformément à la définition de la notion de comportement géométrique qui a été proposée.
À la fin du précédent chapitre, on a donné une fonction géométrique à ces critères, ceux-ci étant des solutions possibles à la question de l'exploration des géométries des nuages d'objets. En outre, les critères qui sont à la base des éléments du deuxième sous-univers ne sont pas uniques, contrairement au cas des éléments du premier sous-univers.

Dans ce chapitre, on passe à une discussion qui concerne l'analyse de chacun des deux sous-univers. Cette analyse consiste à étudier la structure de chacun d'eux sur la base des critères et des contraintes utilisés par leurs éléments (analyses). Ces éléments sont liés aux questions pratiques auxquelles ils sont censés apporter des solutions.

Implicitement, les expressions "objectif pratique" ou "question pratique" supposent qu'en général, on ne peut associer à des questions différentes, des analyses basées sur un seul critère et une seule contrainte. Ces questions peuvent être d'essence expérimentale.

0.1. Etat de la structure de chacun des deux sous-univers

Les deux sous-univers sont caractérisés par la diversité de leurs critères ou de leurs contraintes ou bien des deux à la fois. En effet, les éléments du premier sous-univers sont basés sur un seul critère et diverses contraintes, alors que les éléments du second sont basés sur divers critères et diverses contraintes.

La structure du premier sous-univers est relativement simple par rapport à celle du second. En effet, ses éléments sont basés sur un seul critère mais sous diverses contraintes. Dans le premier sous-univers, c'est la diversité des contraintes qui exprime une prise en compte par ses éléments de la diversité des objectifs pratiques.

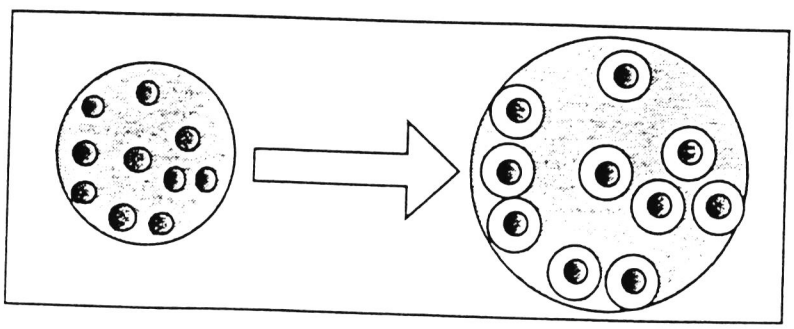


Figure 1. À droite : état du premier sous-univers, chaque point est une analyse. À gauche : état de sa structure, chaque élément correspond à un objectif pratique. Le contour de chaque élément illustre une prise en compte de l'objectif pratique par une analyse

La structure du deuxième sous-univers est plus complexe. En effet, ici, la diversité des critères s'exprime de deux façons différentes. La première reflète diverses solutions pour une même question pratique, la seconde est liée à une diversité des objectifs pratiques.

Ce n'est pas la même chose, en effet, la première suppose l'existence de plusieurs solutions pour la même question, la seconde suppose l'existence d'une solution spécifique pour une question spécifique.

C'est pourquoi, pour analyser la structure du deuxième sous-univers, on propose de distinguer deux sortes de critères : des critères qui se situent au même niveau, et d'autres qui se situent à des niveaux différents.

— Les critères qui se situent au même niveau, sont des critères qui ont la même fonction.

Les critères d'analyses canoniques généralisées sont proposés pour résoudre une question commune, celle du lien entre plusieurs groupes de variables. C'est pourquoi ils ont été situés ici au même niveau. En effet, la question qui consiste à savoir quel est le critère le plus approprié pour résoudre cette question commune n'a pas encore une réponse claire dans la littérature si d'ailleurs elle n'en est pas absente.

— Les critères qui se situent à des niveaux différents, ont des fonctions spécifiques.

Ces fonctions reflètent des objectifs pratiques différents et divergents, les critères reflétant des solutions plus appropriées à ces fonctions. Citons par exemple, les critères MaxBet et Maxdiff proposés par Van de Geer (1984) dans les analyses de concordance.

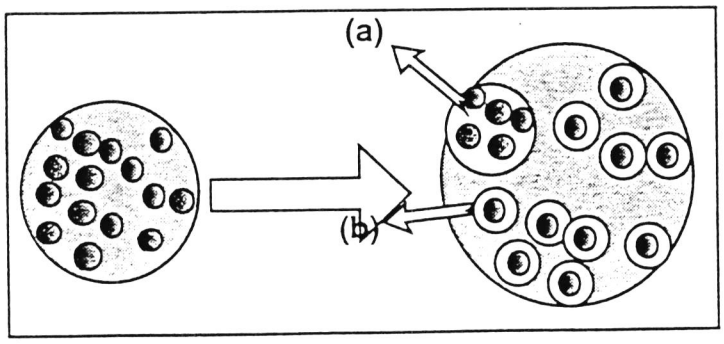


Figure 2. Structure du deuxième sous-univers. (a) analyses associées aux critères qui se situent au même niveau. (b) analyses associées aux critères qui se situent à des niveaux différents.

0.2.Objectifs

Dans le deuxième sous-univers, horizontalement, on a des critères qui ont la même fonction, verticalement, on a des critères qui définissent des fonctions spécifiques. Ceci permet d'une manière naturelle de se poser la question de générer pour chacun des

premiers critères un ensemble de critères qui ont des fonctions spécifiques. Autrement dit, ceci permet de proposer pour une question pratique, un ensemble de critères qui se situent au même niveau à l'image des critères d'analyses canoniques qui sont proposés pour résoudre une question pratique donnée. On parle d'une possibilité de réaliser une complétion de ce deuxième sous-univers.

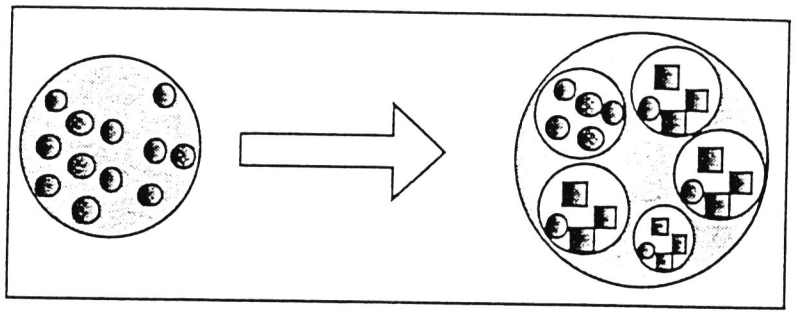


Figure 2. Illustration de l'objectif de la complétion

Parallèlement, on se pose la question de la complétion du premier sous-univers. Contrairement à celle du deuxième sous-univers, la complétion du premier ne peut être basée sur la diversité des critères existants, du fait que les éléments du premier sous-univers sont basés sur un seul critère. Cela n'empêche pas que l'on peut le compléter sur une base géométrique, à savoir, l'utilisation des opérateurs $\mathbf{H}_k$ introduits dans le chapitre précédent.

On note ici que le but de la complétion correspond à réaliser simultanément deux objectifs, le premier étant de proposer des nouvelles analyses, le deuxième étant de représenter d'une manière homogène et structurée, chacun des deux sous-univers.

Ce chapitre est constitué de deux sections. Dans une première section, on s'occupe de compléter le deuxième sous-univers, dans une seconde section, on complètera le premier.

1. Complétion du deuxième sous-univers

Pour compléter le deuxième sous-univers, on emprunte la voie définie dans le cadre du travail de Van de Geer. Plus exactement, cette complétion s'articule autour de trois points :

— Premièrement, on propose une revue des choix de base proposés par Van de Geer (1984) qui reflètent une prise en compte analytiques de la diversité des objectifs pratiques. On rappelle ensuite, l'étude menée par Ten Berge (1986, 1988) qui concerne l'un des critères proposés par Van de Geer. Sur la base de cette étude, on montre que les

choix de base peuvent se reformuler d'une manière plus générale. On convient d'appeler les choix de base résultant de cette reformulation "principes méthodologiques". Ils sont extérieurs à toute approche théorique proprement dite et ils reflètent une prise en compte d'un certain nombre d'objectifs pratiques.

— Deuxièmement, sur la base de ces principes méthodologiques, on se pose la même question que Van de Geer, celle qui l'a amené à introduire ses critères. Cela nous amène naturellement à introduire d'autres critères par modifications et généralisations de critères existants.

— Troisièmement, pour expliquer comment ces critères sont modifiés et généralisés, on considère d'abord un cas particulier que l'on généralise ensuite.

1.1. Principes méthodologiques

On considère le K-tableaux horizontal $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2|...|\mathbf{X}_K]$ à n lignes et $p = \sum_{k=1}^{K} p_k$ colonnes, formé à partir de K matrices $\mathbf{X}_k$ de dimension $n \times p_k$. On note $\mathbf{T}^t = [\mathbf{T}_1^t|\mathbf{T}_2^t|...|\mathbf{T}_K^t]$ avec $\mathbf{T}_k$ de dimension $p_k \times r$ et $k = 1,2,..,K$. $r \leq \min_{1 \leq k \leq K}(r_k)$. L'entier r_k est le rang du tableau $\mathbf{X}_k$, $k = 1,2,..,K$.

Van de Geer (1984, pp.80) propose trois choix de base et explique comment ces choix sont associés à des préoccupations souvent divergentes. Ces trois choix sont les suivants :

(i) What to analyze?

C'est un choix qui concerne la prise en compte ou non des structures internes des tableaux, analytiquement cela revient à choisir :

— soit l'analyse des matrices d'origines $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_K$, qui correspond à une prise en compte des structures internes des tableaux;

— soit l'analyse des matrices définies par les projecteurs associés à chaque tableau.

Cette analyse est appelée par l'auteur "Analysis of $\mathbf{P}$" car il note ces projecteurs : $\mathbf{P}_k$. Van de Geer explique comment l'analyse des projecteurs consiste en une extraction des inerties des tableaux. Les inerties associées à chacun des tableaux sont l'expression de la variabilité de chacun d'eux.
Dans le langage des triplets statistiques, cela revient à choisir ou non pour chaque tableau, la métrique de Mahalanobis $\left(\mathbf{X}_k^t \mathbf{D} \mathbf{X}_k\right)^-$. Pour l'utilisateur de l'analyse multivariée, les analyses fondées sur les projecteurs ou implicitement, les métriques de Mahalanobis

portent toutes, les limitations : stabilité numérique, difficulté d'interprétation (Kettenring (1985)). C'est la raison pour laquelle Van den Wollenberg et Tucker proposent chacun une analyse comme alternative à l'analyse de Hotteling. Celle de Van den Wollenberg (1967) est dite "Analyse de Redondance" et celle de Tucker (1958) est dite "Analyse Interbatterie".

Le choix " What to analyse" est intrinsèquement pris en compte pour proposer des alternatives telles que : la régression multiple, alternative dans la régression PLS (Tenenhaus, 1995) ou dans la régression sur composantes (Obadia, 1978), l'analyse discriminante, alternative dans l'analyse inter-classe, l'analyse canonique, alternative dans les analyses sur variables instrumentales (Sabatier, 1987)) ou dans l'analyse de co-inertie (Chessel & Mercier, 1993).

(ii) Fairness and orthogonality constraints

Ce choix est lié à la question de la pondération des groupes; il consiste à choisir entre deux objectifs :

— L' un correspond à la recherche de solutions qui tolèrent la prise en compte d' une partie des groupes en ignorant les autres. Selon cet objectif, le vecteur $\mathbf{t}$ est contraint d'être normé; le vecteur $\mathbf{t}$ est constitué à partir des vecteurs $\mathbf{t}_k$

— L' autre consiste plutôt à chercher des solutions telles que les groupes jouent *a priori* le même rôle. Selon cet objectif, on contraint les vecteurs des poids $\mathbf{t}_k$ à être normés.

Ces deux contraintes sont liées à la question de la pondération des groupes qui se veut d'équilibrer le rôle joué par chacun d'eux. La question est résolue ici par le choix de la contrainte "vecteurs $\mathbf{t}_k$ normés". On rejoint ainsi la logique de l'analyse factorielle multiple (Escofier & Pagès; 1984, 1985). Celle-ci résout la question de la pondération des groupes par une approche différente, en imposant des pondérations extérieures. En effet, chaque groupe est pondéré par la plus grande valeur propre de l'analyse de chaque tableau.

Dans une approche successive et simultanée, Van de Geer considère uniquement le cas des matrices $\mathbf{T}_k$ orthonormales de dimension $p_k \times p_k$.

(iii) Variance bias

Ce choix ne concerne que les objectifs liés à la recherche des solutions qui tiennent compte de la structure internes des tableaux; c'est à dire, au niveau du choix (i) on choisit l'analyse des tableaux d'origine et non pas *des projecteurs associés.*

Le choix (iii) peut se traduire comme la recherche des solutions avec un degré variable d'insistance dans leur prise en compte de la structure interne. On peut illustrer plus clairement ce choix en prenant un intervalle (-1, 1). En effet, on suppose que :

— le point -1 correspond aux solutions associées aux analyses des projecteurs. Au voisinage de ce point, les solutions reflètent l'optimalité des relations linéaires sans prendre en compte la structure interne des tableaux.

— Le point 1 représente les solutions associées aux analyses séparées de chaque tableau. Ces solutions signifient l'optimalité de la prise en compte de la structure interne des tableaux.

— les solutions qui prennent en compte la structure interne des tableaux dans leur analyse simultanée, sont situées au voisinage du point 0.

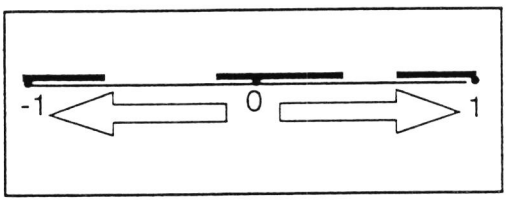


Figure 3. Illustration du troisième choix. (-1) Analyses des projecteurs (choix (i)). (0) Analyses des tableaux (choix (i)). (1) Analyses séparées.

Le choix (iii) correspond alors à des solutions qui sont au voisinage du point 1. C'est à dire que le choix (iii) correspond à des solutions qui tiennent davantage compte de la structure interne des tableaux que des relations linéaires entre ceux-ci.

Du point de vue des choix établis par Van de Geer, les généralisations de Ten Berge (1986, 1988) se sont focalisées sur des extensions utilisant les choix (i) et (ii), ce qui permet de les reformuler d'une manière plus générale en principes méthodologiques.

L'approche unificatrice de Ten Berge (1986, 1988) pour résoudre les problèmes MaxBet et MaxDiff a des contributions multiples. On reprend ici les deux contributions relatives aux choix (i) et (ii) (voir aussi Ten Berge, 1986, pp.83) : la première consiste à généraliser le choix (i) (pour tout type de tableaux). La deuxième consiste à résumer le choix (ii) (formulation des contraintes sous forme d'une seule contrainte générale).

Ces deux contributions nous permettent de reformuler les choix de bases de Van de Geer sous une forme plus générale que l'on convient d'appeller "principes méthodologiques". Ces principes méthodologiques sont :

Premier Principe

Il consiste en l'une des deux considérations suivantes : la première est le choix des tableaux $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K$, la deuxième est le choix des projecteurs "analysis of $\mathbf{P}$"

Dans les deux cas, les tableaux de données $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K$ ont des nombres de colonnes $p_1, p_2, \ldots, p_K$ différents ou non. Autrement dit, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K$ sont l'un des deux types, K-tableaux ou multitableaux .

Deuxième principe

Il consiste en une formulation des contraintes (c1), (c2), (c3) d'une manière plus unifiée sous forme de l'une des deux contraintes générales suivantes :

— la première est :

$$\mathbf{T}_k'\mathbf{T}_k = \mathbf{I}_r, \ k = 1, 2, \ldots, K, \ r \leq \min(r_k).$$

Comme Ten Berge (1986) fait pour introduire le problème MaxBet généralisé.

— La deuxième est :

$$\mathbf{T}'\mathbf{T} = \mathbf{I}_r, \ r \leq \min(r_k)$$

La contrainte $\mathbf{T}'\mathbf{T} = \mathbf{I}_r, \ r \leq \min(r_k)$ est proposée ici comme une généralisation de la contrainte $\mathbf{t}'\mathbf{t} = K$ considérée par Van de Geer, et qui n'a pas été généralisée par Ten Berge.

Troisième principe

Il correspond au troisième choix de Van de Geer. Plus Analytiquement, il revient à chercher un critère pour évaluer la concordance entre les lignes des solutions exprimées par les matrices $\mathbf{X}_1\mathbf{T}_1, \mathbf{X}_2\mathbf{T}_2, \ldots, \mathbf{X}_K\mathbf{T}_K$ (voir aussi Van de Geer, 1984, pp. 82).

1.2 Complétion

Pour compléter le deuxième sous-univers, différentes approches avec prise en compte variable de la structure interne des tableaux (troisième principe), seront considérées. Plus exactement, on propose de modifier et de généraliser cinq critères existants qui ne tiennent compte que des liens entre les groupes sans s'occuper de la structure interne des tableaux. Cette modification/généralisation consiste en trois points :

— premièrement, les critères sont modifiés pour que les solutions tiennent compte de la structure interne des tableaux (premier principe);

— deuxièmement, ces critères sont généralisés pour qu'ils tiennent compte de données générales (premier principe);

— troisièmement, ces critères sont généralisés pour qu'ils tiennent compte des contraintes générales : $\mathbf{T}_k'\mathbf{T}_k = \mathbf{I}_r$ et $\mathbf{T}'\mathbf{T} = \mathbf{I}_r$ (deuxième principe).

Voici les critères que l'on propose, ils incluent MaxBet et MaxDiff comme suit :

$$\psi_1(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k,l=1}^{K}\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{X}_k^t\mathbf{X}_l\mathbf{T}_l\right)$$

$$\psi_2(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k\neq l}^{K}\sum_{k,l=1}\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{X}_k^t\mathbf{X}_l\mathbf{T}_l\right)$$

$$\psi_3(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k,l=1}^{K}\left(\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{X}_k^t\mathbf{X}_l\mathbf{T}_l\right)\right)^2$$

$$\psi_4(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k\neq l}^{K}\sum_{k,l=1}\left(\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{X}_k^t\mathbf{X}_l\mathbf{T}_l\right)\right)^2$$

$$\psi_6(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k=1}^{K}\left(\left(\sum_{l=1}^{K}\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{X}_k^t\mathbf{X}_l\mathbf{T}_l\right)\right)\right)^2$$

$$\psi_5(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K,\mathbf{T}_{K+1})=\sum_{k=1}^{K}\left(\mathrm{tr}\left(\mathbf{T}_{K+1}^t\mathbf{X}_k\mathbf{T}_k\right)\right)^2$$

$$\psi_6(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k=1}^{K}\left(\left(\sum_{l=1}^{K}\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{X}_k^t\mathbf{X}_l\mathbf{T}_l\right)\right)\right)^2$$

$$\psi_7(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k=1}^{K}\left(\left(\sum_{k\neq l}^{K}\sum_{l=1}\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{X}_k^t\mathbf{X}_l\mathbf{T}_l\right)\right)\right)^2$$

Par la suite, on repose la même question que celle qui a conduit Van de Geer à introduire ses critères. Cette question concerne les liens entre les K groupes de variables; elle consiste à chercher des solutions exprimées par $\mathbf{X}_1\mathbf{T}_1,\mathbf{X}_2\mathbf{T}_2,\ldots,\mathbf{X}_K\mathbf{T}_K$ qui réalisent simultanément les deux conditions suivantes :

(a1) Les solutions $\mathbf{X}_1\mathbf{T}_1,\mathbf{X}_2\mathbf{T}_2,\ldots,\mathbf{X}_K\mathbf{T}_K$ reflètent des liens entre les groupes $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_K$;

(a2) Chaque solution $\mathbf{X}_k\mathbf{T}_k$ doit tenir compte de la structure interne des tableaux $\mathbf{X}_k$.

Pour expliquer en quoi ces critères sont liés à des critères existants, et aussi en quoi ils sont différents. On considère en premier lieu le cas particulier qui suit.

1.2.1 Cas particulier

On se restreint au cas particulier : tous les tableaux $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_K$ sont centrés et $r=1..$

Dans ce cas, le problème à résoudre consiste à construire des solutions qui tiennent compte simultanément des deux conditions suivantes :

(a1) Les solutions doivent exprimer des liens entre les K groupes de variables $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K$.

(a2) Les solutions doivent tenir compte de la structure interne des tableaux $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K$. Autrement dit, les solutions doivent être corrélées avec les variables des tableaux de données $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K$.

Réaliser la condition (a1) revient à construire des solutions liant les K groupes de variables en considérant des combinaisons linéaires de chaque groupe de variables $\mathbf{X}_1\mathbf{t}_1, \mathbf{X}_2\mathbf{t}_2, \ldots, \mathbf{X}_K\mathbf{t}_K$. On espère que les liens entre les K vecteurs $\mathbf{X}_1\mathbf{t}_1, \mathbf{X}_2\mathbf{t}_2, \ldots, \mathbf{X}_K\mathbf{t}_K$ reflètent les liens entre les K groupes de variables $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K$.

Ce problème est classiquement étudié par les analyses canoniques généralisées (Kettenring, 1971; Lafosse, 1989). Les analyses canoniques généralisées consistent à optimiser l'ensemble de critères que l'on a présenté dans la première section du premier chapitre (I.1). On se restreint ici aux critères suivants :

$$\omega_1(\mathbf{t}_1, \mathbf{t}_2, .., \mathbf{t}_K) = \sum_{k \neq l \; k,l=1}^{K} \mathrm{corr}(\mathbf{X}_k\mathbf{t}_k, \mathbf{X}_l\mathbf{t}_l)$$

$$\omega_2(\mathbf{t}_1, \mathbf{t}_2, .., \mathbf{t}_K) = \sum_{k \neq l \; k,l=1}^{K} \left(\mathrm{corr}(\mathbf{X}_k\mathbf{t}_k, \mathbf{X}_l\mathbf{t}_l)\right)^2$$

$$\omega_3(\mathbf{t}_1, \mathbf{t}_2, .., \mathbf{t}_{K+1}) = \sum_{k=1}^{K} \left(\mathrm{corr}(\mathbf{X}_k\mathbf{t}_k, \mathbf{t}_{K+1})\right)^2$$

$$\omega_4(\mathbf{t}_1, \mathbf{t}_2, .., \mathbf{t}_K) = \sum_{k=1}^{K} \left(\mathrm{corr}\left(\mathbf{X}_k\mathbf{t}_k, \sum_{l \neq k \; l=1}^{K} \mathbf{X}_l\mathbf{t}_l \right) \right)^2$$

$\mathrm{Corr}(\mathbf{X}_k\mathbf{t}_k, \mathbf{X}_l\mathbf{t}_l)$ désigne le coefficient de corrélation linéaire entre $\mathbf{X}_k\mathbf{t}_k$ et $\mathbf{X}_l\mathbf{t}_l$.

Les critères ω_1, ω_2, et ω_3 correspondent respectivement aux critères SUMCOR, SSQCOR, et MAXVAR (voir Kettenring, 1971, pp.434). Le critère ω_4 est un critère proposé par Lafosse (Lafosse, 1989, pp.273).

Les critères ω_i $(i = 1, 2, 3, 4)$ sont tous maximisés pour offrir les solutions des analyses canoniques généralisées. Ces solutions ne satisfont pas nécessairement la condition (a2) (pour plus de détails, voir Van de Geer, 1984, pp.80).

Pour réaliser simultanément les deux conditions (a1) et (a2), on propose d'adapter les critères ω_i $(i = 1, 2, 3, 4)$ spécifiques aux analyses canoniques généralisées pour qu'ils tiennent compte de (a2). L'idée consiste à remplacer dans les critères ω_i $(i = 1, 2, 3, 4)$, la corrélation $\mathrm{Corr}(\mathbf{X}_k\mathbf{t}_k, \mathbf{X}_l\mathbf{t}_l)$ entre $\mathbf{X}_k\mathbf{t}_k$ et $\mathbf{X}_l\mathbf{t}_l$ par la covariance entre $\mathbf{X}_k\mathbf{t}_k$ et $\mathbf{X}_l\mathbf{t}_l$. On

rejoint indirectement la démarche de Tucker (1958) qui maximise la covariance au lieu de la corrélation, pour proposer son analyse.

On note par $\mathrm{cov}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}_l\mathbf{t}_l)$ la covariance entre $\mathbf{X}_k\mathbf{t}_k$ et $\mathbf{X}_l\mathbf{t}_l$, par $\mathrm{var}(\mathbf{a})$ la variance d'une variable $\mathbf{a}$.

L'écriture :

$$\mathrm{cov}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}_l\mathbf{t}_l)=\sqrt{\mathrm{var}(\mathbf{X}_k\mathbf{t}_k)}\sqrt{\mathrm{var}(\mathbf{X}_l\mathbf{t}_l)}\mathrm{Corr}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}_l\mathbf{t}_l),$$

permet de donner deux interprétation :

— Premièrement, la présence de la corrélation $\mathrm{Corr}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}_l\mathbf{t}_l)$ signifie qu'on prend en compte la condition (a1), c'est à dire le lien entre $\mathbf{X}_k\mathbf{t}_k$ et $\mathbf{X}_l\mathbf{t}_l$.

— Deuxièmement, la présence des expressions des variances $\mathrm{var}(\mathbf{X}_k\mathbf{t}_k)$ et $\mathrm{var}(\mathbf{X}_l\mathbf{t}_l)$ des solutions $\mathbf{X}_k\mathbf{t}_k$ et $\mathbf{X}_l\mathbf{t}_l$, signifie que, dans la réalisation de ce lien, l'on prend en considération des solutions $\mathbf{X}_k\mathbf{t}_k$ et $\mathbf{X}_l\mathbf{t}_l$ qui tiennent compte de la structure interne des tableaux $\mathbf{X}_k$, $\mathbf{X}_l$.

La covariance est alors un critère compromis qui tient compte simultanément des deux conditions (a1) et (a2).

En remplaçant $\mathrm{Corr}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}_l\mathbf{t}_l)$ par $\mathrm{cov}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}_l\mathbf{t}_l)$ dans les critères ω_i $(i=1,2,3,4)$, on obtient dans cet ordre les critères $\tilde{\omega}_i$ $(i=1,2,3,4)$ qui s'écrivent :

$$\tilde{\omega}_1(\mathbf{t}_1,\mathbf{t}_2,..,\mathbf{t}_K)=\sum_{k\neq l\,k,l=1}^{K}\mathrm{cov}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}_l\mathbf{t}_l)$$

$$\tilde{\omega}_2(\mathbf{t}_1,\mathbf{t}_2,..,\mathbf{t}_K)=\sum_{k\neq l\,k,l=1}^{K}\left(\mathrm{cov}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}_l\mathbf{t}_l)\right)^2$$

$$\tilde{\omega}_3(\mathbf{t}_1,\mathbf{t}_2,..,\mathbf{t}_{K+1})=\sum_{k=1}^{K}\left(\mathrm{cov}(\mathbf{X}_k\mathbf{t}_k,\mathbf{t}_{K+1})\right)^2$$

$$\tilde{\omega}_4(\mathbf{t}_1,\mathbf{t}_2,..,\mathbf{t}_K)=\sum_{k=1}^{K}\left(\mathrm{cov}\left(\mathbf{X}_k\mathbf{t}_k,\sum_{k\neq l\,l=1}^{K}\mathbf{X}_l\mathbf{t}_l\right)\right)^2$$

Les critères $\tilde{\omega}_i$ $(i=1,2,3,4)$ sont proposés pour tenir compte simultanément des deux conditions (a1) et (a2). Les solutions sont obtenues par la maximisation de ces critères sous les contraintes $\mathbf{t}_k^t\mathbf{t}_k=1$

1.2.2 Généralisation

Pour que ces critères génèrent des solutions basées sur les les trois principes méthodologiques, on propose les généralisations suivantes :

On s'intéresse tout d'abord aux généralisation basées sur le troisième principe. Les solutions obtenues sur la base des critères $\tilde{\omega}_i \ (i=1,2,3,4)$ garantissent dans un certain mesure la condition (a2), mais ne correspondent pas nécessairement aux solutions basées sur le troisième principe. En effet, celles-ci doivent insister davantage sur la prise en compte des structures internes que sur les relations uniquement linéaires.

Pour proposer des critères permettant de prendre en considération le troisième principe, on place la discussion autour des deux critères MaxBet et MaxDiff.

Dans le cas particulier considéré, le critère MaxDiff donne le critère $\tilde{\omega}_1$, ce qui permet au critère MaxBet de donner le critère $\tilde{\omega}_5$:

$$\tilde{\omega}_5(\mathbf{t}_1,\mathbf{t}_2,..,\mathbf{t}_K)=\sum_{k,l=1}^{K}\operatorname{cov}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}_l\mathbf{t}_l)=\sum_{k=1}^{K}\operatorname{var}(\mathbf{X}_k\mathbf{t}_k)+\tilde{\omega}_1(\mathbf{t}_1,\mathbf{t}_2,..,\mathbf{t}_K)$$

La présence de l'expression $\sum_{k=1}^{K}\operatorname{var}(\mathbf{X}_k\mathbf{t}_k)$ montre que les solutions générées par le critère MaxBet donnent plus d'importance à la structure interne des tableaux, que les solutions générées par le critère MaxDiff (voir Ten Berge, 1988, pp.487). Le critère MaxBet correspond au troisième principe.

On peut interpréter MaxBet comme une variante de MaxDiff dont les solutions tiennent compte du troisième principe. Par la suite, on utilise ce raisonnement pour proposer à partir des deux critères $\tilde{\omega}_2$, $\tilde{\omega}_4$ respectivement les deux critères $\tilde{\omega}_6$, $\tilde{\omega}_7$.

Les deux critères $\tilde{\omega}_6$ et $\tilde{\omega}_7$ sont proposés pour tenir compte du troisième principe, ils se définissent comme :

$$\tilde{\omega}_6(\mathbf{t}_1,\mathbf{t}_2,..,\mathbf{t}_K)=\sum_{k,l=1}^{K}\left(\operatorname{cov}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}_l\mathbf{t}_l)\right)^2;$$

$$\tilde{\omega}_7(\mathbf{t}_1,\mathbf{t}_2,..,\mathbf{t}_K)=\sum_{k=1}^{K}\left(\operatorname{cov}(\mathbf{X}_k\mathbf{t}_k,\mathbf{X}\mathbf{t})\right)^2$$

Ainsi les critères $\tilde{\omega}_i \ (i=1,2,\ldots,7)$ offrent des solutions qui vérifient les conditions (a1) et (a2), ainsi que des solutions basées sur le troisième principe.

On considère maintenant le deuxième principe. Afin que l'on puisse générer des solutions basées sur ce principe, on généralise les critères pour qu'ils tiennent compte des contraintes qui expriment celui-ci. Les critères $\tilde{\omega}_1$, $\tilde{\omega}_2$, $\tilde{\omega}_3$, et $\tilde{\omega}_4$ donnent dans cet ordre les critères suivants :

$$\psi_2(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k\neq l}^{K}\sum_{k,l=1}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)$$

$$\psi_4(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k\neq l}^{K}\sum_{k,l=1}\left(\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)\right)^2$$

$$\psi_5(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K,\mathbf{T}_{K+1})=\sum_{k=1}^{K}\left(\mathrm{tr}\left(\mathbf{T}_{K+1}'\mathbf{X}_k\mathbf{T}_k\right)\right)^2$$

$$\psi_7(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k=1}^{K}\left(\left(\sum_{k\neq l}^{K}\sum_{l=1}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)\right)\right)^2$$

et les critères $\tilde{\omega}_5$, $\tilde{\omega}_6$ et $\tilde{\omega}_7$ donnent les critères suivants :

$$\psi_1(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k,l=1}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)$$

$$\psi_3(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k,l=1}^{K}\left(\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)\right)^2$$

$$\psi_6(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k=1}^{K}\left(\left(\sum_{l=1}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)\right)\right)^2$$

En effet, il suffit de remarquer que dans le cas $r=1$, la quantité $\mathrm{tr}\left(\mathbf{T}_i'\mathbf{X}_i'\mathbf{X}_j\mathbf{T}_j\right)$ devient $\mathrm{cov}\left(\mathbf{X}_i\mathbf{t}_i,\mathbf{X}_j\mathbf{t}_j\right)$.

Les tableaux sont supposés centrés, on peut prendre des types de tableaux plus généraux pour proposer les critères ψ_i $(i=3,4,5,6,7)$ dans des contextes différents tels que ceux de "matching factor" et "configurations matrices", ainsi que celui des analyses canoniques généralisées (pour plus de détails voir Ten Berge, 1988, pp488).

La maximisation de chacun des critères ψ_i $(i=3,4,5,6,7)$ sous chacune des contraintes $\mathbf{T}_k'\mathbf{T}_k=\mathbf{I}_r$, $(i=1,2,...,K)$ ou $\mathbf{T}'\mathbf{T}=\mathbf{I}_r$, avec $r\leq\min(r_k)$ est alors considérée comme la proposition d'une analyse nouvelle.

L'ensemble des nouvelles analyses définies sur la base des critères et contraintes proposés apporte des extensions dans plusieurs directions et permet une présentation plus synthétique de différentes approches existantes.
En effet, la démarche présentée ici reformule d'une manière plus générale les choix de Van de Geer (1984) en principes méthodologiques; d'autre part, elle offre des extensions à d'autres critères et d'autres contraintes que ceux que propose Van de Geer. De même, elle généralise l'approche de Ten Berge (1986, 1988) de deux critères MaxBet et MaxDiff, à l' ensemble des critères ψ_i $(i=3,4,5,6,7)$. Dans le cadre des analyses

canoniques généralisées (Kettenring, 1971, Lafosse,1989), elle étend ces analyses pour qu'elles tiennent compte des trois principes méthodologiques.

Du point de vue de la structure du deuxième sous-univers, la disparité des critères rend la structure de celui-ci hétérogène, dans le sens que pour certaines questions pratiques, il existe plusieurs solutions alors que pour d'autres, il n'en existe qu'une seule. En laissant s'exprimer le plus possible la diversité des critères, on produit une structuration de ce sous-univers. Le sous-univers issu de la complétion est désormais homogène. En effet, chaque question pratique est maintenant associée à un ensemble de critères situés au même niveau.

2. Complétion du premier sous-univers

La diversité des critères qui sont à la base des éléments du deuxième sous-univers a permis de compléter celui-ci. La complétion du premier sous-univers suit une démarche différente, du fait que ses éléments sont basés sur un seul critère.
C'est la diversité des solutions relatives à la question de la contruction d'objets comparables introduite dans le chapitre précédent (I.2), qui permet de compléter le premier sous-univers. Néanmoins, la complétion de celui-ci dépend de la nature des tableaux, elle concerne ici uniquement les K-tableaux. En effet, les éléments du deuxième sous-univers sont basées sur le problème (**P''**) suivant :

Minimiser $PCASUP\left(\hat{\mathbf{Z}},\hat{\mathbf{F}},\mathbf{C}\right)=\left\|\hat{\mathbf{Z}}-\hat{\mathbf{F}}\mathbf{C}^{t}\right\|_{HS}^{2}$ sous les contraintes $\mathbf{C}\in\Xi_{\hat{\mathbf{Z}}}$, $\hat{\mathbf{F}}\in\Sigma_{\hat{\mathbf{Z}}}$

Dans le cas des multitableaux $\hat{\mathbf{Z}}=\hat{\mathbf{X}}$, $\Xi_{\hat{\mathbf{Z}}}=\Xi$, $\Sigma_{\hat{\mathbf{Z}}}=\Sigma$, dans le cas des K-tableaux $\hat{\mathbf{Z}}=\hat{\mathbf{W}}$ et $\Xi_{\hat{\mathbf{W}}}=\overline{\Xi}$, $\Sigma_{\hat{\mathbf{W}}}=\overline{\Sigma}$.(voir le chapitre I.1)

Pour compléter le premier sous-univers, on utilise la solution géométrique pour construire des objets comparables, qui le passage aux opérateurs **HD** que l'on substitue au passage des opérateurs **WD**

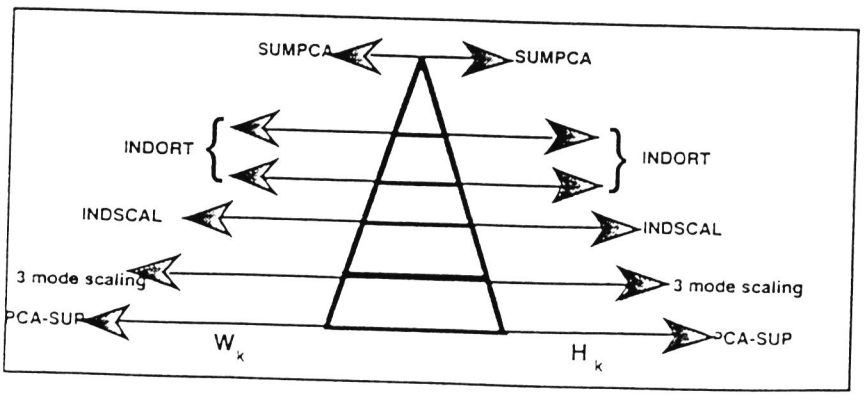


Figure 3. Complétion du premier sous-univers

On définit ainsi, des nouvelles analyses associées au problème (**P''**) telles que $\hat{Z} = \hat{H}$ et $\Xi_{\hat{H}} = \overline{\Xi}$, $\Sigma_{\hat{H}} = \overline{\Sigma}$. Par conséquent, on propose une alternative à l'approche unificatrice de Kiers sur la base des opérateurs **HD.**

3. Conclusion

En conclusion, dans ce chapitre, la discussion autour des critères d'une partie des analyses existantes amène naturellement à se poser la question de leur complétion. Deux complétions ont été proposées pour les deux sous-univers, l'une est basée sur des choix méthodologiques, l'autre est basée sur une solution géométrique.

Ces opérations de complétion ont permis de proposer un certain nombre d'analyses nouvelles. Ces propositions nouvelles sont essentiellement basées sur des travaux existants. Le choix particulier qui consiste à se baser sur des travaux existants est simultanément lié à deux objectifs, d'une part celui d'analyser la structure de chacun des deux sous-univers, d'autre part, à celui de proposer des nouvelles analyses.

Cette démarche révèle deux points que l'on estime importants; le premier est son aspect unificateur de travaux existants, le deuxième est que chacune des deux structures associées aux deux sous-univers est marquée par la diversité de ces propositions.

Cette diversité s'exprime à deux niveaux, une diversité des solutions pour une même question et une diversité des propositions pour des questions différentes.

Selon les moyens théoriques dont chaque sous-univers resultants dispose, la structure de chacun d'eux exprime alors, chacune à sa façon cette diversité. Les deux sous-univers résultants se mettent d'accord sur le fait que les solutions proposées (analyses) ne sont uniques dans aucun des deux sous-univers.

Chapitre I.4

Approche algorithmique

0. Introduction

On se consacre uniquement aux analyses nouvelles introduites dans la section précédente, celles qui sont issues des opérations de complétion. Il s'agit maintenant de passer à une discussion tout aussi nécessaire que les précédentes, celle du calcul effectif dans chacun des deux sous-univers, des solutions des problèmes d'optimisation associés à ces analyses. Quant aux analyses existant déjà, on peut se référer aux travaux cités ou à leurs bibliographies pour obtenir des algorithmes de résolution.

Pour les analyses qui complètent le premier sous-univers, il suffit de mettre à jour les opérateurs **WD** par les opérateurs **HD** dans des algorithmes existants pour calculer des solutions du problème (**P**).

La question de construire des algorithmes pour les analyses qui complètent le deuxième sous-univers, nécessite une démarche plus élaborée.

Afin de proposer des algorithmes de résolution pour les problèmes d'optimisation associés à ces analyses, on peut construire des algorithmes spécifiques à chacun d'entre eux. Mais ici, on choisit d'aborder cette question dans le cadre unifié établi jusqu'à présent, en proposant un seul algorithme général de résolution pour ces problèmes d'optimisation. C'est l'objectif de cette section. Cela va aussi dans le sens du principal objectif de cette partie qui est de contruire une composante de liens.

Tout d'abord, on rappelle quelques notions de base de la théorie d'optimisation des fonctions. C'est le contenu de la première section.

Dans le cadre de la construction d'algorithmes convergents pour l'optimisation des fonctions en analyse multivariée, on évoque deux techniques (Heiser, 1995) à savoir : "Alternating Least Squares (ALS)" et "Iterative Majorization IM". C'est le contenu de la seconde section.

Dans une troisième section, sur la base la technique "IM", on propose de résoudre un problème général de maximisation.

Dans une quatrième section on construit un algorithme de résolution unique pour ce problème général de maximisation.

Dans une cinquième section on montre que cet algorithme général permet le calcul des solutions des problèmes de maximisation associés aux analyses qui complètent le deuxième sous-univers.

1. Quelques repères concernant l'optimisation des fonctions

Un problème d'optimisation consiste à maximiser ou minimiser des fonctions sous diverses contraintes. Formellement, il peut être énoncé de la manière suivante :

"maximiser ou minimiser" $f(x_1, x_2, \ldots, x_n)$ sous la contrainte $x = (x_1, x_2, \ldots, x_n) \in \tilde{\Omega}$ (*)

La fonction f est une fonction d'un domaine Ω et à valeurs réelle, $\tilde{\Omega}$ un sous ensemble du domaine Ω. Généralement, l'ensemble $\tilde{\Omega}$ est dit "ensemble contrainte". Dans la plupart des cas, il est exprimé analytiquement par une fonction dite "fonction contrainte" par la relation :

$$\tilde{\Omega} = \{x \in \Omega \quad C_i(x) = 0 \quad i = 1, 2, .. p\},$$

les fonctions $C_i \quad i = 1, 2, \ldots, p$ peuvent être des fonctions à valeurs réelles.

Deux questions principales sont classiquement liées à la résolution du problème (*) :

— La première est celle de l'existence, elle consiste à établir que le problème (*) admet au moins une solution. Sous réserve d'une réponse positive à la question de l'existence, celle ci est généralement suivie d'une autre question qui consiste à établir l'unicité ou la non-unicité des solutions.

— La deuxième question concerne le calcul effectif d'une solution du problème (*), c'est ce qu'on entend ici par algorithmie. Cette question dépend trivialement de la première; en effet, si la solution n'existe pas, on ne peut la calculer. De ce fait, la question d'algorithmie s'impose en général, si on a déjà établi une réponse positive à la question de l'existence.

L'algorithmie est nécessaire lorsqu'on veut aller au-delà de l'argument formel et souvent implicite qui caractérise les démarches pour établir l'existence des solutions au problème (*). L'algorithmie a pour objectif un calcul effectif des solutions et, dans les meilleurs des cas, elle vise l'établissement d'une formule explicite de ces solutions.

La question de l'algorithmie n'est pas seulement un problème numérique, mais parfois, elle peut être considérée aussi comme une question purement théorique située à un autre niveau. En effet, un calcul effectif de la solution revient souvent à la recherche d'une propriété intrinsèque de la fonction ou de la contrainte qui définissent le problème d'optimisation.

Il n'existe pas de principe général pour résoudre le problème (*) au niveau de l'existence, et encore moins au niveau de l'algorithmie. Le problème (*) ne présente généralement pas un caractère intrinsèquement invariant, du fait qu'il est très sensible à la perturbation (changement) de la fonction f, de la contrainte C, ou du domaine Ω. Plus exactement, les difficultés à résoudre le problème (*) viennent du fait que les solutions dépendent des propriétés très spécifiques, soit de la fonction, soit de la contrainte, soit du

domaine, celui-ci étant considéré en général comme un espace avec une structure topologique.

Un exemple simple est celui du calcul effectif de la moyenne d'un vecteur dans deux espaces topologiques de nature différente; par exemple, soient les deux problèmes suivants :

$$\text{Minimiser } \sum_{i=1}^{n}(t-x_i)^2$$

$$\text{Minimiser } \sum_{i=1}^{n}|t-x_i|$$

où $|t|$ désigne la valeur absolue du nombre réel t.

La première fonction à minimiser est régulière au sens qu'elle est indéfiniment différentiable, la seconde est simplement continue. Les deux fonctions sont intrinsèquement les mêmes, c'est une moyenne relative à deux espaces de nature topologique différente, l'une est euclidienne, l'autre ne l'est pas. L'un des deux problèmes admet une solution unique, l'autre admet selon la parité des composantes, une ou deux solutions. Cette attitude ne dépend *a priori* d'aucune approche interactive entre des notions intrinsèques.

L'un répond à la question de l'existence et de l'unité, l'autre n'a généralement pas la même réponse. Calculer une moyenne dans un sens comme dans un autre revient intrinsèquement au même, mais l'optimisation des deux fonctions ci-dessus met en jeu des propriétés très spécifiques, ici au niveau de la fonction et de l'espace topologique associés.

Dans le cas des espaces de dimension finie, la question de l'existence est généralement facile à établir grâce à leurs propriétés topologiques très favorables comme "la compacité", sous réserve que la régularité de la fonction soit acquise (la fonction à optimiser et la fonction contrainte doivent être au moins continues).

2. Techniques d'optimisation en analyse multivariée

Le lien de l'analyse multivariée avec les problèmes d'optimisation est largement évoqué dans la littérature (voir ci-dessous); en effet, les analyses proposées se ramènent à des problèmes d'optimisation sous diverses contraintes.

Si la question de l'existence ne pose généralement pas de problèmes, l'unicité n'est pas souvent garantie et la motivation principale en analyse multivariée reste la question d'algorithmie. Cette dernière est évidement nécessaire pour effectuer à terme les analyses.

Etablir des principes permettant de générer des algorithmes aux moins convergents (voir ci-dessous), est une démarche naturelle; celle-ci est renforcée par la nature très proche des fonctions à optimiser en analyse multivariée.

Plusieurs travaux ont été menés pour établir des principes généraux de résolution d'une famille de problèmes d'optimisation. D'après Heiser (1995), on peut distinguer au moins deux techniques générales, très pratiques pour l'optimisation des fonctions en analyse multivariée. La première est dite "Alternating Least Squares" (ALS) et la deuxième est dite " Iterative majorization " (IM).

Ces deux techniques ont un point commun, celui de construire des algorithmes convergents; en effet, on peut les considérer comme deux possibilités différentes de résoudre le problème suivant :

1.1 Problème de mise à jour

Soit une intialisation $\left(x_1^0, x_2^0, \ldots, x_n^0\right) \in \tilde{\Omega}$, trouver explicitement $\left(x_1^1, x_2^1, \ldots, x_n^1\right) \in \tilde{\Omega}$ dite mise à jour de $\left(x_1^0, x_2^0, \ldots, x_n^0\right)$ telle que :

$$f\left(x_1^0, x_2^0, \ldots, x_n^0\right) < f\left(x_1^1, x_2^1, \ldots, x_n^1\right) \quad (**)$$

La condition (**) est une condition suffisante pour construire des algorithmes convergents associés à la maximisation de fonctions.

1.2. Première technique "ALS"

Pour résoudre le problème de mise à jour, l'ALS procède par la résolution d'une série de problèmes simples, comme suit :

On cherche un élément x_1^1 solution du problème de maximisation suivant :

$$\text{Maximiser } f\left(z_1, x_2^0, \ldots, x_n^0\right) \text{ sous la contrainte } \left(z_1, x_2^0, \ldots, x_n^0\right) \in \tilde{\Omega}.$$

On cherche un élément x_2^1 solution du problème de maximisation suivant :

$$\text{Maximiser } f\left(x_1^1, z_2, \ldots, x_n^0\right) \text{ sous la contrainte } \left(x_1^1, z_2, \ldots, x_n^0\right) \in \tilde{\Omega}.$$

Ainsi de suite, jusqu'à n, on cherche x_n^1 solution de :

$$\text{Maximiser } f\left(x_1^1, x_2^1, \ldots, z_n\right) \text{ sous la contrainte } \left(x_1^1, x_2^1, \ldots, z_n\right) \in \tilde{\Omega}$$

D'après ce procédé, il en découle que :

$$f\left(x_1^0, x_2^0, \ldots, x_n^0\right) < f\left(x_1^1, x_2^0, \ldots, x_n^0\right) < f\left(x_1^1, x_2^1, \ldots, x_n^0\right) < \ldots < f\left(x_1^1, x_2^1, \ldots, x_n^1\right)$$

Ce procédé suppose que les problèmes de maximisation ci-dessus ont des résolutions algorithmiques simples. Pour utiliser cette technique, il est préférable de découper le problème de maximisation en plusieurs problèmes simples.

L'ALS a été largement utilisée ces dernières années. Par exemple, dans le cadre des généralisations de l'Analyse en Composantes Principales pour les K-tableaux verticaux, Millsap & Meredith (1988) proposent une analyse dite "analyse en composante principale simultanée". Kiers & Ten Berge (1989) proposent deux algorithmes convergents pour la résolution de ce problème de minimisation en utilisant la technique ALS.

1.3. Deuxième technique "IM"

La deuxième technique est dite "Iterative Majorization IM", elle est basée sur une idée plus intuitive et très pertinente à la fois. Pour construire une mise à jour, elle considère un recouvrement de la fonction à optimiser f, par une famille de fonctions, dites fonctions auxilliaires, qu'on note $\{\mu_y(x) \quad y \in \Omega\}$ (voir Heiser, 1995). L'écriture $\mu_y(x)$ signifie aussi que chaque fonction $\mu_y(x)$ dépend des deux variables x et y. Ce recouvrement de fonctions doit vérifier :

(a) pour y fixé on a l'inégalité :

$$\mu_y(x) \leq f(x) \quad \forall x \in \Omega$$

(b) pour x fixé, soit par exemple $\tilde{x}$ la valeur fixée de x dans le domaine Ω alors :

$$\{f(\tilde{x})\} \cap \{\mu_y(\tilde{x}) \quad y \in \Omega\} = \{\mu_y(\tilde{x}) \quad y = \tilde{x}\}.$$

C'est à dire que seule la fonction auxilliaire indexée par $\tilde{x}$ doit interpeler $f(\tilde{x})$, et de plus, pas en n'importe quelle valeur de x mais seulement en $\tilde{x}$.

(c) on suppose aussi que pour y fixé la fonction auxilliaire $\mu_y(x)$ admet un maximum unique.

Si la construction de ce recouvrement est possible, alors il permet de résoudre le problème de mise à jour, de la manière suivante :

on a :

$$f(x^0) = \mu_{x^0}(x^0),$$

μ_{x^0} admet un maximum unique alors il existe x^1 tel que :

$$\mu_{x^0}(x^0) \leq \mu_{x^0}(x^1).$$

Alors on peut considérer l'ensemble $\{\mu_y(x^1) \quad y \in \Omega\}$. On sait qu'il intersecte f uniquement dans l'ensemble $\{f(\tilde{x})\} \cap \{\mu_y(\tilde{x}) \quad y \in \Omega\} = \{\mu_y(\tilde{x}) \quad y = \tilde{x}\}$.

En particulier, pour $\tilde{x} = x^1$, il en découle que :

$$f(x^0) = \mu_{x^0}(x^0) < \mu_{x^1}(x^1) \leq f(x^1)$$

À titre d'exemples de l'utilisation de cette technique, on cite :

— Dans le contexte des généralisations de l'analyse procuste de Mosier (1939), Ten Ten Berge & Nevels(1977), Ten Berge (1991) qui utilisent cette technique pour minimiser la fonction générale suivante :

$$f(\mathbf{t}) = \|\mathbf{At} - \mathbf{v}\|^2 + \alpha\left(\|\mathbf{t}\|^2 - \delta\right)^2.$$

— Dans différents contextes, Kiers (1990, 1991, 1995) utilise aussi à plusieurs reprises cette technique pour maximiser sous diverses contraintes d'orthogonalité les fonctions suivantes :

$$f(\mathbf{X}) = c + tr(\mathbf{AX}) + \sum_{k=1}^{K} tr\left(\mathbf{X}^t\mathbf{B}_k\mathbf{X}\mathbf{C}_k\mathbf{X}^t\right)$$

$$f(\mathbf{X}) = \sum_{k=1}^{K} tr\left(\mathbf{X}^t\mathbf{A}_k\mathbf{X}\right)\left(\mathbf{X}^t\mathbf{C}_k\mathbf{X}\right)^{-1}$$

$$f(\mathbf{X}) = \sum_{k=1}^{K} tr\left(\mathbf{X}^t\mathbf{A}_k\mathbf{X}\right)\left(\mathbf{X}^t\mathbf{A}_k^t\mathbf{X}\right)\left(\mathbf{X}^t\mathbf{C}_k\mathbf{X}\right)^{-1}$$

— Dans le contexte "Multidimensional Scaling", on peut également se référer à De Leeuw & Heiser (1980), Meulman (1986), De Leeuw (1988)), ainsi qu'à Groenen (1993) qui a consacré sa thèse à cette technique.

3. Formulation générale

L'approche présentée ici peut s'appuyer sur la technique "Iterative Majorization", la propriété (a) est prise en compte par l'inégalité (4) ci-dessous (voir aussi inégalité (9)).

3.1. fonction générale

Comme on va le montrer et l'expliquer par la suite, un certain nombre de critères, en particulier ceux qui sont introduits dans le chapitre précédent, peuvent être considérés comme des cas particuliers d'une fonction générale qui est la suivante :

$$\Psi(\mathbf{T}) = \Phi(\mathbf{T}, \mathbf{T}), \tag{1}$$

avec $\mathbf{T}$ une matrice inconnue de dimension $p \times r$, supposée partitionnée ou non en K sous matrices $\mathbf{T}_k$ de dimension $p_k \times r$. La fonction Φ est définie de la manière suivante :

$$\Phi(\mathbf{T}, \mathbf{W}) = \mathrm{tr}\left(\mathbf{T}'\mathbf{A}_{\mathbf{W}}\mathbf{T}\right), \tag{2}$$

avec $\mathbf{W}$ une matrice de dimension $p \times r$, et $\mathbf{A}_{\mathbf{W}}$ une matrice semi-définie positive de dimension $p \times p$, qui dépend continûment de $\mathbf{W}$. De plus, cette fonction Φ est supposée avoir les propriétés suivantes :

$$\Phi(\mathbf{T}, \mathbf{T}) \geq 0 \tag{3}$$

$$\Phi(\mathbf{T}, \mathbf{W}) \leq \left(\Phi(\mathbf{T}, \mathbf{T})\right)^{\frac{1}{2}} \left(\Phi(\mathbf{W}, \mathbf{W})\right)^{\frac{1}{2}}, \tag{4}$$

(4) est une inégalité de type Cauchy-Schwartz.

3.2. Problème général de maximisation

Il s'agit ici de construire des algorithmes convergents pour résoudre le problème général suivant :

Maximiser $\Psi(\mathbf{T})$ sous la contraintes $\mathbf{T}_k'\mathbf{T}_k = \mathbf{I}_r,\ k = 1, 2, \ldots, K$

4. Résolution algorithmique

On note d'entrée que la maximisation de la fonction générale ne change pas si on ajoute une matrice $\alpha_{\mathbf{W}}\mathbf{I}$ à la matrice $\mathbf{A}_{\mathbf{W}}$, avec $\alpha_{\mathbf{W}}$ une fonction qui dépend de $\mathbf{W}$ et à valeurs réelle. De même, on peut supposer la matrice $\mathbf{A}_{\mathbf{W}}$ symétrique définie positive. En effet, si la matrice $\mathbf{A}_{\mathbf{W}}$ n'est pas symétrique, elle est remplacée par sa partie symétrique sans que le problème de maximisation de la fonction $\Psi(\mathbf{T})$ soit affecté.

L'objectif ici est construire un algorithme général pour la maximisation de la fonction générale $\Psi(\mathbf{T})$ sous les contraintes $\mathbf{T}_k'\mathbf{T}_k = \mathbf{I}_r$ $k = 1, 2, \ldots, K$. On démontrera ensuite, que la maximisation de ce problème exprime comme cas particuliers les problèmes d'optimisation qui sont à la base des analyses introduites dans le chapitre précédent. C'est pourquoi, dans un premier temps, on considère un car particulier de ce problème général qui sera suivi par un traitement dans son cadre général.

3.1.1. Cas particulier

Le cas particulier ici consiste à construite un algorithme pour le problème suivant :

Problème (PP) : Maximiser $\Psi(\mathbf{T})$ sous la contrainte $\mathbf{T}'\mathbf{T} = \mathbf{I}_r$,

Ce problème est un cas particulier du problème général. En effet, il suffit de remaquer que la maximisation de la fonction $\Psi(\mathbf{T}) = tr\left(\mathbf{T}^t\mathbf{A}_\mathbf{T}\mathbf{T}\right)$ sous la contrainte $\mathbf{T}^t\mathbf{T} = \mathbf{I}_r$, revient à la maximisation de la fonction $\hat{\Psi}(\mathbf{W}) = tr\left(\mathbf{W}^t\mathbf{M}_\mathbf{W}\mathbf{W}\right)$ sous les contraintes $\mathbf{W}_k^t\mathbf{W}_k = \mathbf{I}_r \quad k = 1,2,\ldots,K$, avec $\mathbf{W}_k$ est une matrice d'ordre $p \times r, \quad k = 1,2,\ldots,K$. Ainsi, si on note par $\mathbf{W}$ la matrice $\mathbf{W}^t = \left[\mathbf{W}_1^t \quad \mathbf{W}_2^t \quad \cdots \quad \mathbf{W}_2^t\right]$ d'ordre $Kp \times r$, par $\mathbf{M}_\mathbf{W}$ une matrice d'ordre d'ordre $Kp \times Kp$ définie comme :

$$\mathbf{M}_\mathbf{W} = \begin{bmatrix} \mathbf{A}_\mathbf{W} & 0 \\ 0 & 0 \end{bmatrix}$$

Il en découle que le problème (PP) est un cas particulier du problème général.

Pour maximiser $\Psi(\mathbf{T})$ sous la contrainte $\mathbf{T}^t\mathbf{T} = \mathbf{I}_r$ on construit un algorithme dans lequel $\mathbf{T}$ est mise à jour itérativement. En effet, si on considère une matrice courante $\tilde{\mathbf{T}}$, et $\mathbf{Z}$ la mise à jour de cette matrice courante, alors construire un algorithme monotonement convergent qui maximise $\Psi(\mathbf{T})$ revient à construire une mise à jour $\mathbf{Z}$ en fonction de la matrice courante $\tilde{\mathbf{T}}$ telle que $\Psi\left(\tilde{\mathbf{T}}\right) \leq \Psi(\mathbf{Z})$. La résolution du problème qui suit permet de construire une telle matrice $\mathbf{Z}$:

Problème A1: Maximiser $\mathrm{tr}\left(\mathbf{T}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right)$ sous contraintes $\mathbf{T}^t\mathbf{T} = \mathbf{I}_r$.

Soit $\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}} = \mathbf{PDQ}^t$ la décomposition en valeurs singulières de la matrice $\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}$ de dimension $(p \times r)$, avec $\mathbf{P}^t\mathbf{P} = \mathbf{Q}^t\mathbf{Q} = \mathbf{QQ}^t = \mathbf{I}_r$ et $\mathbf{D}$ une matrice diagonale à éléments positifs rangés dans un ordre décroissant, alors la matrice solution du problème A1 est donnée par :

$$\mathbf{Z} = \mathbf{PQ}^t. \tag{5}$$

Pour la démonstration, voir Cliff (1966).

Comme $\mathbf{Z}$ est une solution de A1, alors on a :

$$\mathrm{tr}\left(\mathbf{T}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right) \leq \mathrm{tr}\left(\mathbf{Z}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right), \tag{6}$$

Pour toute matrice $\mathbf{T}$, vérifiant la contrainte, en particulier pour $\mathbf{T} = \tilde{\mathbf{T}}$. Ceci implique :

$$\mathrm{tr}\left(\tilde{\mathbf{T}}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right) \leq \mathrm{tr}\left(\mathbf{Z}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right) \leq \left(\mathrm{tr}\left(\mathbf{Z}^t\mathbf{A}_{\tilde{\mathbf{T}}}\mathbf{Z}\right)\right)^{\frac{1}{2}}\left(\mathrm{tr}\left(\tilde{\mathbf{T}}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right)\right)^{\frac{1}{2}} \tag{7}$$

La deuxième inégalité est basée sur l'inégalité de Cauchy-Schwartz. Il en découle que :

$$\mathrm{tr}\left(\tilde{\mathbf{T}}^t \mathbf{A}_{\tilde{\mathbf{T}}} \tilde{\mathbf{T}}\right) \leq \mathrm{tr}\left(\mathbf{Z}^t \mathbf{A}_{\tilde{\mathbf{T}}} \mathbf{Z}\right), \tag{8}$$

Autrement dit :

$$\Psi\left(\tilde{\mathbf{T}}\right) \leq \Phi\left(\mathbf{Z}, \tilde{\mathbf{T}}\right). \tag{9}$$

L'inégalité (4) implique :

$$\Psi\left(\tilde{\mathbf{T}}\right) \leq \Phi\left(\mathbf{Z}, \tilde{\mathbf{T}}\right) \leq (\Phi(\mathbf{Z}, \mathbf{Z}))^{\frac{1}{2}} \left(\Phi\left(\tilde{\mathbf{T}}, \tilde{\mathbf{T}}\right)\right)^{\frac{1}{2}}; \tag{10}$$

Il est immédiat que $\Psi\left(\tilde{\mathbf{T}}\right) \leq \Psi(\mathbf{Z})$.

Comme on vient de le montrer, la solution du problème A1 permet de faire croître monotonement la fonction $\Psi(\mathbf{T})$. Une solution peut être obtenue en mettant à jour itérativement la matrice $\tilde{\mathbf{T}}$ selon l'égalité (5). Cette mise à jour est effectuée jusqu'à ce qu'aucun accroissement de la fonction $\Psi(\mathbf{T})$ ne soit possible.

La procédure converge en un point où la fonction $\Psi(\mathbf{T})$ ne peut croître par (4), parce que $\Psi(\mathbf{T})$ est bornée et continue et que l'accroissement est strictement monotone. Cette situation est obtenue si et seulement si les inégalités (7) à (10) deviennent des égalités. La deuxième inégalité dans (7) devient une égalité; cela signifie que $\mathbf{A}_{\tilde{\mathbf{T}}} \tilde{\mathbf{T}} = \mathbf{A}_{\tilde{\mathbf{T}}} \mathbf{Z}$ alors que $\tilde{\mathbf{T}} = \mathbf{Z} = \mathbf{P}\mathbf{Q}^t$, ce qui implique :

$$\mathbf{A}_{\tilde{\mathbf{T}}} \tilde{\mathbf{T}} = \mathbf{P}\mathbf{D}\mathbf{Q}^t = \tilde{\mathbf{T}}\mathbf{Q}\mathbf{D}\mathbf{Q}^t = \tilde{\mathbf{T}}\Theta \tag{11}$$

Θ étant une matrice symétrique semi-définie positive.

Dans le cas d'une solution globale, aucun accroissement de la fonction $\Psi(\mathbf{T})$ n'est possible, ce qui permet à l'équation (11) d'être une condition nécessaire pour la matrice qui maximise la fonction générale $\Psi(\mathbf{T})$.

Sur la base de la procédure de mise à jour décrite précédemment, on construit pour la maximisation de la fonction $\Psi(\mathbf{T})$ sous la contrainte $\mathbf{T}^t\mathbf{T} = \mathbf{I}_r$, l'algorithme itératif suivant :

1. Choisir $\tilde{\mathbf{T}}$ (e.g. aléatoirement, telle que $\tilde{\mathbf{T}}^t\tilde{\mathbf{T}} = \mathbf{I}_r$), ε (e.g, 0.00001).

2. Calculer $\mathbf{A}_{\tilde{\mathbf{T}}}$; si nécessaire la rendre symétrique définie positive.

3. Effectuer la décomposition en valeurs singulières de $\mathbf{A}_{\tilde{\mathbf{T}}} \tilde{\mathbf{T}} = \mathbf{P}\mathbf{D}\mathbf{Q}^t$.

4. Poser $\mathbf{Z} = \mathbf{P}\mathbf{Q}^t$.

5. Si $(\Psi(\mathbf{Z}) - \Psi(\mathbf{T}) \geq \varepsilon)$, alors $\tilde{\mathbf{T}} = \mathbf{Z}$ aller à l'étape 2, sinon, considérer que l'algorithme a convergé.

4.2. Cas général

On considère le problème général de la maximisation de la fonction générale $\Psi(\mathbf{T})$ sous les contraintes $\mathbf{T}_k^t\mathbf{T}_k = \mathbf{I}_r \quad k = 1, 2, \ldots, K$. Un algorithme qui fait croître monotonement la fonction $\Psi(\mathbf{T})$ sous les contraintes $\mathbf{T}_k^t\mathbf{T}_k = \mathbf{I}_r$, $k = 1, 2, \ldots, K$, sera construit par la recherche d'une mise à jour $\mathbf{Z}^t = \left[\mathbf{Z}_1^t | \mathbf{Z}_2^t | \ldots | \mathbf{Z}_K^t\right]$ en fonction d'une matrice courante $\tilde{\mathbf{T}}^t = \left[\tilde{\mathbf{T}}_1^t | \tilde{\mathbf{T}}_2^t | \ldots | \tilde{\mathbf{T}}_K^t\right]$ qui satisfait $\mathbf{T}_k^t\mathbf{T}_k = \mathbf{I}_r$, $k = 1, 2, \ldots, K$, et telle que $\Psi\left(\tilde{\mathbf{T}}\right) \leq \Psi(\mathbf{Z})$. Cela est possible grâce au problème suivant :

Problème A2 : Maximiser $\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{A}_{k_{\tilde{\mathbf{T}}}}\tilde{\mathbf{T}}\right)$ sous les contraintes $\mathbf{T}_k^t\mathbf{T}_k = \mathbf{I}_r$,

avec k fixé et $\mathbf{A}_{k_{\tilde{\mathbf{T}}}} = \left[\mathbf{A}_{k1_{\tilde{\mathbf{T}}}} \quad \mathbf{A}_{k2_{\tilde{\mathbf{T}}}} \quad \cdots \quad \mathbf{A}_{kK_{\tilde{\mathbf{T}}}}\right]$ une matrice (p_k, p) .

$$\mathbf{A}_{\tilde{\mathbf{T}}} = \begin{bmatrix} \mathbf{A}_{11_{\tilde{\mathbf{T}}}} & \mathbf{A}_{12_{\tilde{\mathbf{T}}}} & \cdots & \mathbf{A}_{1K_{\tilde{\mathbf{T}}}} \\ \mathbf{A}_{21_{\tilde{\mathbf{T}}}} & \mathbf{A}_{22_{\tilde{\mathbf{T}}}} & \cdots & \mathbf{A}_{2K_{\tilde{\mathbf{T}}}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{K1_{\tilde{\mathbf{T}}}} & \mathbf{A}_{K2_{\tilde{\mathbf{T}}}} & \cdots & \mathbf{A}_{KK_{\tilde{\mathbf{T}}}} \end{bmatrix}.$$

Soit la décomposition en valeurs singulières de la matrice $\mathbf{A}_{k_{\tilde{\mathbf{T}}}}\tilde{\mathbf{T}} = \mathbf{P}_k\mathbf{D}_k\mathbf{Q}_k^t$, avec $\mathbf{P}_k^t\mathbf{P}_k = \mathbf{Q}_k^t\mathbf{Q}_k = \mathbf{Q}_k\mathbf{Q}_k^t = \mathbf{I}_r$ et $\mathbf{D}_k$ une matrice diagonale à éléments positifs rangés dans un ordre décroissant, alors la solution du problème A2 est la matrice suivante :

$$\mathbf{Z}_k = \mathbf{P}_k\mathbf{Q}_k^t \tag{12}$$

Une démonstration se trouve dans Ten Berge &Knol (1984).

Soit $\mathbf{Z}_k$ la solution du problème A2, alors :

$$\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{A}_{k_{\tilde{\mathbf{T}}}}\tilde{\mathbf{T}}\right) \leq \mathrm{tr}\left(\mathbf{Z}_k^t\mathbf{A}_{k_{\tilde{\mathbf{T}}}}\tilde{\mathbf{T}}\right) \quad k = 1, 2, \ldots, K.$$

Après sommation sur k, on a :

$$\mathrm{tr}\left(\mathbf{T}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right) \leq \mathrm{tr}\left(\mathbf{Z}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right).$$

Pour $\mathbf{T}$ satisfaisant les présentes contraintes, l'inégalité de Cauchy-Schwarz donne :

$$\mathrm{tr}\left(\tilde{\mathbf{T}}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right) \le \mathrm{tr}\left(\mathbf{Z}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right) \le \left(\mathrm{tr}\left(\mathbf{Z}^t\mathbf{A}_{\tilde{\mathbf{T}}}\mathbf{Z}\right)\right)^{\frac{1}{2}}\left(\mathrm{tr}\left(\tilde{\mathbf{T}}^t\mathbf{A}_{\tilde{\mathbf{T}}}\tilde{\mathbf{T}}\right)\right)^{\frac{1}{2}},$$

en d'autres termes :

$$\Psi\left(\tilde{\mathbf{T}}\right) \le \Phi\left(\mathbf{Z},\tilde{\mathbf{T}}\right),$$

de même, l'inégalité (4) implique :

$$\Psi\left(\tilde{\mathbf{T}}\right) \le \Phi\left(\mathbf{Z},\tilde{\mathbf{T}}\right) \le \left(\Phi(\mathbf{Z},\mathbf{Z})\right)^{\frac{1}{2}}\left(\Phi\left(\tilde{\mathbf{T}},\tilde{\mathbf{T}}\right)\right)^{\frac{1}{2}}.$$

Il en découle que $\Psi\left(\tilde{\mathbf{T}}\right) \le \Psi(\mathbf{Z})$.

Une mise à jour itérative de $\mathbf{T}$ comme on l'a décrite ci-dessus fait croître monotonement la fonction $\Psi(\mathbf{T})$. La convergence est atteinte si et seulement si les inégalités précédentes deviennent des égalités; autrement dit : $\tilde{\mathbf{T}} = \mathbf{Z}$ ce qui implique :

$$\mathbf{A}_{k_{\tilde{\mathbf{T}}}}\tilde{\mathbf{T}} = \mathbf{P}_k\mathbf{D}_k\mathbf{Q}_k^t = \tilde{\mathbf{T}}_k\mathbf{Q}_k\mathbf{D}_k\mathbf{Q}_k^t = \tilde{\mathbf{T}}_k\Theta_k,\ k = 1,2,\ldots,K \qquad (13)$$

Pour tout $k = 1,2,\ldots,K$, Θ_k étant symétrique semi-définie positive.

En effet, l'équation (13) est une condition nécessaire pour la solution maximisante de la fonction $\Psi(\mathbf{T})$. Il est noté que la dérivation de cette condition (13) est similaire à une dérivation établie par Ten Berge & Knol & Kiers (1988). On propose ainsi l'algorithme général suivant :

4.2.1. Algorithme général de résolution

1. Choisir $\tilde{\mathbf{T}} = \left[\tilde{\mathbf{T}}_1^t \quad \tilde{\mathbf{T}}_2^t \quad \ldots \quad \tilde{\mathbf{T}}_K^t\right]^t$ (e.g. aléatoirement, telle que pour tout $k = 1,2,..,K$ on a $\tilde{\mathbf{T}}_k^t\tilde{\mathbf{T}}_k = \mathbf{I}_r$), ε (e.g, 0.00001)

2. Calculer $\mathbf{A}_{\tilde{\mathbf{T}}}$, la rendre symétrique semi-définie positive.

3. Pour $k = 1,2,..,K$, considérer la décomposition en valeurs singulières de la matrice : $\mathbf{A}_{k_{\tilde{\mathbf{T}}}}\tilde{\mathbf{T}} = \mathbf{PDQ}^t$.

4. Pour $k = 1,2,..,K$, on pose $\mathbf{Z}_k = \mathbf{P}_k\mathbf{Q}_k^t$

5. Soit $\mathbf{Z} = \left[\mathbf{Z}_1^t \quad \mathbf{Z}_2^t \quad \ldots \quad \mathbf{Z}_K^t\right]^t$

6. Si ($\Psi(\mathbf{Z}) - \Psi(\mathbf{T}) \geq \varepsilon$), alors $\tilde{\mathbf{T}} = \mathbf{Z}$ aller à l'étape 2, sinon, l'algorithme peut être considéré comme ayant convergé.

5. Applications : fonction générale et deuxième sous-univers

A présent, l'objectif est de montrer que la maximisation des critères $\psi_q(\mathbf{T})$ $(q = 1,2,..,7)$, sous les contraintes : $\mathbf{T}'\mathbf{T} = \mathbf{I}_r$ ou $\mathbf{T}_k'\mathbf{T}_k = \mathbf{I}_r$ $k = 1,2,..,K$, sont des cas particuliers de la maximisation de la fonction générale $\Psi(\mathbf{T})$. En effet, la fonction $\Psi(\mathbf{T})$ est définie sur la base d'une autre fonction $\Phi(\mathbf{T},\mathbf{W})$ selon l'égalité (1). Il suffit alors d'associer à chaque fonction $\psi_q(\mathbf{T})$ une fonction $\phi_q(\mathbf{T},\mathbf{W})$ $(q = 1,2,..,7)$ et de montrer que ces fonctions sont des cas particuliers de la fonction $\Phi(\mathbf{T},\mathbf{W})$. Pour $(q = 1,2,..,7)$, on considère les fonctions $\phi_q(\mathbf{T},\mathbf{W})$ définies par :

$$\phi_1(\mathbf{T},\mathbf{W}) = \sum_{k,l=1}^{K} \mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)$$

$$\phi_2(\mathbf{T},\mathbf{W}) = \sum_{k\neq l \;\; k,l=1}^{K} \mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)$$

$$\phi_3(\mathbf{T},\mathbf{W}) = \sum_{k,l=1}^{K} \mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)\mathrm{tr}\left(\mathbf{W}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{W}_l\right)$$

$$\phi_4(\mathbf{T},\mathbf{W}) = \sum_{k\neq l \;\; k,l=1}^{K} \mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)\mathrm{tr}\left(\mathbf{W}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{W}_l\right)$$

$$\phi_5(\mathbf{T},\mathbf{W}) = \sum_{k=1}^{K} \mathrm{tr}\left(\mathbf{T}_{K+1}'\mathbf{X}_k\mathbf{T}_k\right)\mathrm{tr}\left(\mathbf{W}_{K+1}'\mathbf{X}_k\mathbf{W}_k\right)$$

$$\phi_6(\mathbf{T},\mathbf{W}) = \sum_{k=1}^{K}\left(\sum_{l=1}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)\right)\left(\sum_{l=1}^{K}\mathrm{tr}\left(\mathbf{W}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{W}_l\right)\right)$$

$$\phi_7(\mathbf{T},\mathbf{W}) = \sum_{k=1}^{K}\left(\sum_{l\neq k \;\; l=1}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)\right)\left(\sum_{l\neq k \;\; l=1}^{K}\mathrm{tr}\left(\mathbf{W}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{W}_l\right)\right)$$

Si, pour $(q = 3,4,..,7)$, les fonctions $\phi_q(\mathbf{T},\mathbf{W})$ $(q = 1,2,..,7)$ vérifient les propriétés (1) et (2), (voir 3.1 plus haut dans ce chapitre I.4), alors les fonctions $\phi_q(\mathbf{T},\mathbf{W})$ peuvent être considérées comme cas particuliers de la fonction $\Phi(\mathbf{T},\mathbf{W})$, ce qui implique que les fonctions $\psi_q(\mathbf{T})$ $(q = 1,2,..,7)$ sont des cas particuliers de la fonction $\Psi(\mathbf{T})$.

On note que :

(i) Pour tout $(q = 1,2,..,7)$ $\psi_q(\mathbf{T}) = \phi_q(\mathbf{T},\mathbf{T})$, par conséquent, la propriété (1) est vérifiée.

(ii) Les fonctions $\phi_q(\mathbf{T},\mathbf{W})$ $(q=1,2,..,7)$ vérifient également la propriété (2). On démontrera cela seulement pour la fonction $\phi_6(\mathbf{T},\mathbf{W})$, Pour les autres fonctions on utilisera le même raisonnement que celui que l'on développe ci-dessous.

La fonction $\phi_6(\mathbf{T},\mathbf{W})$ est définie comme :

$$\phi_6(\mathbf{T},\mathbf{W}) = \sum_{k=1}^{K}\left(\sum_{l=1}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{T}_l\right)\right)\left(\sum_{l=1}^{K}\mathrm{tr}\left(\mathbf{W}_k'\mathbf{X}_k'\mathbf{X}_l\mathbf{W}_l\right)\right),$$

ce qui est équivalent à :

$$\phi_6(\mathbf{T},\mathbf{W}) = \sum_{k=1}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}\mathbf{T}\right)\mathrm{tr}\left(\mathbf{W}_k'\mathbf{X}_k'\mathbf{X}\mathbf{W}\right).$$

On note par $\mathbf{A}_k = \mathbf{X}_k'\mathbf{X}$, alors :

$$\phi_6(\mathbf{T},\mathbf{W}) = \sum_{k=1}^{K}\mathrm{tr}\left(\mathbf{T}_k'\mathbf{A}_k\mathbf{T}\right)\mathrm{tr}\left(\mathbf{W}_k'\mathbf{A}_k\mathbf{W}\right)$$

Soit $\mathbf{C}_{k_\mathbf{W}}$ une matrice de dimension (p_k, p) définie comme : $\mathbf{C}_{k_\mathbf{W}} = \mathrm{tr}\left(\mathbf{W}_k'\mathbf{A}_k\mathbf{W}\right)\mathbf{A}_k$, et $\mathbf{C}_\mathbf{W}$ une matrice de dimension (p,p) définie comme : $\mathbf{C}_\mathbf{W} = \left[\mathbf{C}_{1_\mathbf{W}}' \quad \mathbf{C}_{2_\mathbf{W}}' \quad \cdots \quad \mathbf{C}_{K_\mathbf{W}}'\right]'$, alors la fonction $\phi_6(\mathbf{T},\mathbf{W})$ s'écrit :

$$\phi_6(\mathbf{T},\mathbf{W}) = \mathrm{tr}\left(\mathbf{T}'\mathbf{C}_\mathbf{W}\mathbf{T}\right).$$

Un raisonnement analogue permet de montrer que les fonctions $\phi_q(\mathbf{T},\mathbf{W})$ peuvent s'écrire comme :

$$\phi_1(\mathbf{T},\mathbf{W}) = \mathrm{tr}\left(\mathbf{T}'\mathbf{A}\mathbf{T}\right)$$

$$\phi_2(\mathbf{T},\mathbf{W}) = \mathrm{tr}\left(\mathbf{T}'\hat{\mathbf{A}}\mathbf{T}\right)$$

$$\phi_3(\mathbf{T},\mathbf{W}) = \mathrm{tr}\left(\mathbf{T}'\mathbf{B}_\mathbf{W}\mathbf{T}\right)$$

$$\phi_4(\mathbf{T},\mathbf{W}) = \mathrm{tr}\left(\mathbf{T}'\hat{\mathbf{B}}_\mathbf{W}\mathbf{T}\right)$$

$$\phi_5(\mathbf{T},\mathbf{W}) = \mathrm{tr}\left(\mathbf{T}'\tilde{\mathbf{L}}_\mathbf{W}\mathbf{T}\right)$$

$$\phi_6(\mathbf{T},\mathbf{W}) = \mathrm{tr}\left(\mathbf{T}'\mathbf{C}_\mathbf{W}\mathbf{T}\right)$$

$$\phi_7(\mathbf{T},\mathbf{W}) = \mathrm{tr}\left(\mathbf{T}'\hat{\mathbf{C}}_\mathbf{W}\mathbf{T}\right)$$

Les expressions explicites des matrices $\mathbf{A}, \hat{\mathbf{A}}, \mathbf{B}_W, \hat{\mathbf{B}}_W, \mathbf{C}_W, \hat{\mathbf{C}}_W, \tilde{\mathbf{L}}_W$ sont données dans ce même chapitre (voir section 7. annexe).

On vient de montrer que les critères $\psi_q(\mathbf{T})$ $(q = 1,2,..,7)$ sont des cas particuliers de la fonction générale $\Psi(\mathbf{T})$.

6. Conclusion

En conclusion, dans ce chapitre, l'algorithme général que l'on propose peut être utilisé pour obtenir les solutions des analyses définies sur la base des critères $\psi_q(\mathbf{T})$ $(q = 1,2,..,7)$ et des contraintes générales $\mathbf{T}'\mathbf{T} = \mathbf{I}_r$ et $\mathbf{T}_k'\mathbf{T}_k = \mathbf{I}_r$. On établit ainsi un traitement algorithmique unifié de toutes ces analyses. Celui-ci généralise l'approche de Ten berge (1988), qui concerne seulement les critères $\psi_q(\mathbf{T})$ $(q = 1,2)$ et les contraintes $\mathbf{T}_k'\mathbf{T}_k = \mathbf{I}_r$. De plus, l'approche présentée ici généralise celle que propose Meyer (1991). En effet, les critères considérés par Meyer sont des cas particuliers de la fonction générale $\Psi(\mathbf{T})$, les matrices $\mathbf{X}_k$ étant remplacées par des bases orthonormales $\mathbf{P}_k$, et $r = 1$.

Les algorithmes proposés convergent monotonement en un point stationnaire mais la convergence vers une solution globale n'est pas garantie. Il est recommandé d'exécuter l'algorithme en utilisant plusieurs initialisations.

Les points stationnaires (solutions locales) sont caractérisés par les équations (11) ou (13), selon la contrainte choisie. Pour une grande partie des problèmes, ces conditions apparaissent comme nouvelles; pour les problèmes traités précédemment par Ten Berge (1988) et Meyer (1991), elles peuvent être considérées comme des cas particuliers de la maximisation de notre fonction générale. Spécifiquement, l'équation stationnaire établie par Ten Berge est un cas particulier de (13). Celui-ci est le cas où la matrice $\mathbf{A}_{k_{\tilde{\mathbf{T}}}}$ ne dépend pas de $\tilde{\mathbf{T}}$ (les matrices $\mathbf{A}_{k_{\tilde{\mathbf{T}}}}$ sont constantes). De même, les conditions stationnaires établies par Meyer sont des cas particuliers de (13) avec $r = 1$.

7. Annexe

$\mathbf{A}$ est une matrice d'ordre $\left(\sum_{k=1}^{K} p_k \times \sum_{k=1}^{K} p_k\right)$ définie par :

$$\mathbf{A} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \cdots & \mathbf{X}_1'\mathbf{X}_K \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & \cdots & \mathbf{X}_2'\mathbf{X}_K \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_K'\mathbf{X}_1 & \mathbf{X}_K'\mathbf{X}_2 & \cdots & \mathbf{X}_K'\mathbf{X}_K \end{bmatrix}$$

$\hat{\mathbf{A}}$ est une matrice d'ordre $\left(\sum_{k=1}^{K} p_k \times \sum_{k=1}^{K} p_k\right)$ définie par :

$$\hat{A} = \begin{bmatrix} 0 & X_1^t X_2 & \cdots & X_1^t X_K \\ X_2^t X_1 & 0 & \cdots & X_2^t X_K \\ \vdots & \vdots & \ddots & \vdots \\ X_K^t X_1 & X_K^t X_2 & \cdots & 0 \end{bmatrix}$$

$\hat{B}$ est une matrice d'ordre $\left(\sum_{k=1}^{K} p_k \times \sum_{k=1}^{K} p_k \right)$ définie par :

$$B_W = \begin{bmatrix} \mathrm{tr}\left(W_1^t X_1^t X_1 W_1\right) X_1^t X_1 & \mathrm{tr}\left(W_1^t X_1^t X_2 W_2\right) X_1^t X_2 & \cdots & \mathrm{tr}\left(W_1^t X_1^t X_K W_K\right) X_1^t X_K \\ \mathrm{tr}\left(W_2^t X_2^t X_1 W_1\right) X_2^t X_1 & \mathrm{tr}\left(W_2^t X_2^t X_2 W_2\right) X_2^t X_2 & \cdots & \mathrm{tr}\left(W_2^t X_2^t X_K W_K\right) X_2^t X_K \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{tr}\left(W_K^t X_K^t X_1 W_1\right) X_K^t X_1 & \mathrm{tr}\left(W_K^t X_K^t X_2 W_2\right) X_K^t X_2 & \cdots & \mathrm{tr}\left(W_K^t X_K^t X_K W_K\right) X_K^t X_K \end{bmatrix}$$

$\hat{B}$ est une matrice d'ordre $\left(\sum_{k=1}^{K} p_k \times \sum_{k=1}^{K} p_k \right)$ définie par :

$$\hat{B}_W = \begin{bmatrix} 0 & \mathrm{tr}\left(W_1^t X_1^t X_2 W_2\right) X_1^t X_2 & \cdots & \mathrm{tr}\left(W_1^t X_1^t X_K W_K\right) X_1^t X_K \\ \mathrm{tr}\left(W_2^t X_2^t X_1 W_1\right) X_2^t X_1 & 0 & \cdots & \mathrm{tr}\left(W_2^t X_2^t X_K W_K\right) X_2^t X_K \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{tr}\left(W_K^t X_K^t X_1 W_1\right) X_K^t X_1 & \mathrm{tr}\left(W_K^t X_K^t X_2 W_2\right) X_K^t X_2 & \cdots & 0 \end{bmatrix}$$

C_W est une matrice d'ordre $\left(\sum_{k=1}^{K} p_k \times \sum_{k=1}^{K} p_k \right)$ définie par :

$$C_W = \begin{bmatrix} \mathrm{tr}\left(W_1^t X_1^t X W\right) X_1^t X \\ \mathrm{tr}\left(W_2^t X_2^t X W\right) X_2^t X \\ \vdots \\ \mathrm{tr}\left(W_K^t X_K^t X W\right) X_K^t X \end{bmatrix}$$

Soit $\hat{X}_k = [X_1 \quad X_2 \quad \cdots \quad X_{k-1} \quad 0 \quad X_{k+1} \quad \cdots \quad X_K]$, alors $\hat{C}_W$ est une matrice d'ordre $\left(\sum_{k=1}^{K} p_k \times \sum_{k=1}^{K} p_k \right)$ définie par :

$$\hat{C}_W = \begin{bmatrix} \mathrm{tr}\left(W_1^t X_1^t \hat{X}_1 W\right) X_1^t \hat{X}_1 \\ \mathrm{tr}\left(W_2^t X_2^t \hat{X}_2 W\right) X_2^t \hat{X}_2 \\ \vdots \\ \mathrm{tr}\left(W_K^t X_K^t \hat{X}_K W\right) X_K^t \hat{X}_K \end{bmatrix}$$

L_W est une matrice d'ordre $\left(n \times \sum_{k=1}^{K} p_k \right)$ définie par :

$$\mathbf{L}_{\mathrm{W}} = \left[\left(\mathbf{W}_{K+1}^{t} \mathbf{X}_1 \mathbf{W}_1 \right) \mathbf{X}_1 \quad \left(\mathbf{W}_{K+1}^{t} \mathbf{X}_2 \mathbf{W}_2 \right) \mathbf{X}_2 \quad \cdots \quad \left(\mathbf{W}_{K+1}^{t} \mathbf{X}_K \mathbf{W}_K \right) \mathbf{X}_K \right]$$

$\tilde{\mathbf{L}}_{\mathrm{W}}$ est la matrice d'ordre $\left(\left(n + \sum_{k=1}^{K} p_k \right) \times \left(n + \sum_{k=1}^{K} p_k \right) \right)$ définie par :

$$\tilde{\mathbf{L}}_{\mathrm{W}} = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{L}_{\mathrm{W}} \end{bmatrix}$$

Chapitre I.5

Composante de liens

Il s'agit maintenant de définir le sous-ensemble de l'univers qui sera considéré comme la composante de liens que l'on propose. Avant de définir les éléments de cette composante de liens, on effectue une dernière généralisation.

1. Généralisation aux études à trois entrées

Les critères $\psi_q(\mathbf{T})$ $(q=1,2,..,7)$ et les deux contraintes $\mathbf{T}_k^t\mathbf{T}_k=\mathbf{I}_r$ $k=1,2,...,K$ et $\mathbf{T}^t\mathbf{T}=\mathbf{I}_r$ introduits, sont exprimés uniquement dans le cas particulier des études K-tableaux ou multitableaux, munies des métriques identité $(\mathbf{X},\mathbf{I}_p,\mathbf{I}_n,\mathbf{I}_K)$. Pour une étude K-tableaux ou multitableaux $(\mathbf{X},\mathbf{Q},\mathbf{D},\Pi)$, les critères et les contraintes seront alors exprimés sous les formes suivantes :

Les critères $\psi_q(\mathbf{T})$ $(q=1,2,..,7)$ deviennent :

$$\psi_1(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k,l=1}^{K}\pi_k\pi_l\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\mathbf{X}_l\mathbf{Q}_l\mathbf{T}_l\right)$$

$$\psi_2(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k\neq l\ k,l=1}^{K}\pi_k\pi_l\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\mathbf{X}_l\mathbf{Q}_l\mathbf{T}_l\right)$$

$$\psi_3(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k,l=1}^{K}(\pi_k\pi_l)^2\left(\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\mathbf{X}_l\mathbf{Q}_l\mathbf{T}_l\right)\right)^2$$

$$\psi_4(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k\neq l\ k,l=1}^{K}(\pi_k\pi_l)^2\left(\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\mathbf{X}_l\mathbf{Q}_l\mathbf{T}_l\right)\right)^2$$

$$\psi_5(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K,\mathbf{T}_{K+1})=\sum_{k=1}^{K}\pi_k^2\left(\mathrm{tr}\left(\mathbf{T}_{K+1}^t\mathbf{D}\mathbf{X}_k\mathbf{Q}_k\mathbf{T}_k\right)\right)^2$$

$$\psi_6(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k=1}^{K}\pi_k^2\left(\sum_{l=1}^{K}\pi_l\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\mathbf{X}_l\mathbf{Q}_l\mathbf{T}_l\right)\right)^2$$

$$\psi_7(\mathbf{T}_1,\mathbf{T}_2,..,\mathbf{T}_K)=\sum_{k=1}^{K}\pi_k^2\left(\sum_{k\neq l\ l=1}^{K}\pi_l\mathrm{tr}\left(\mathbf{T}_k^t\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\mathbf{X}_l\mathbf{Q}_l\mathbf{T}_l\right)\right)^2$$

Les deux contraintes deviennent :

$$\mathbf{T}_k^t\mathbf{Q}_k\mathbf{T}_k=\mathbf{I}_r\ \ i=1,2,...,K;$$

$$\mathbf{T}^t\mathbf{Q}\mathbf{T}=\mathbf{I}_r$$

2. Définition de la composante de liens

Les critères ci-dessus généralisent les critères introduits dans les sections prédédentes, aux différents types d'études (voir notations). Ceci permet en particulier l'utilisation dans différents contextes, des analyses basées sur la maximisation de chacun des critères $\psi_q(\mathbf{T})$ $(q = 1,2,..,7)$ sous l'une des deux contraintes mentionnées ci-dessus.

Comme on l'a expliqué au cours des trois précédents chapitres, ces analyses sont liées simultanément selon trois approches différentes à savoir :

— une approche géométrique : ces analyses ont le même comportement géométrique. Plus exactement, il s'agit du comportement géométrique associé à la solution (iii) du chapitre I.2

— une approche méthodologique : les analyses prennent en compte la diversité des objectifs pratiques, selon les principes méthodologiques définis dans le chapitre I.3.

— une approche algorithmique : un seul algorithme permet le calcul effectif des solutions des problèmes de maximisation qui sont à la base de ces analyses. On peut le constater en se référant au chapitre I.4.

On peut alors définir la composante de liens; ses éléments sont des analyses d'études K-tableaux ou multitableaux basées sur la maximisation des critères $\psi_q(\mathbf{T})$ $(q = 1,2,..,7)$ qui sont présentés ci-dessus sous l'une des deux contraintes $\mathbf{T}_k^t\mathbf{Q}_k\mathbf{T}_k = \mathbf{I}_r$ $i = 1,2,...,K$; ou $\mathbf{T}^t\mathbf{Q}\mathbf{T} = \mathbf{I}_r$.

3. Conclusion

Cette unité théorique donne lieu à plusieurs contributions relatives à chacune des trois approches; on les a résumées dans la conclusion de chacune des trois chapitres (I.2, I.3 et I.4). On rappelle ici une contribution que l'on estime particulièrement importante, à savoir que la composante de liens contient comme cas particuliers des approches existantes, soit au niveau méthodologique (voir la conclusion du chapitre I.3), soit au niveau algorithmique (voir la conclusion du chapitre I.4). L'approche géométrique parait alors comme une nouvelle alternative pour établir des liens entre des analyses existantes.

I.6. Conclusions et perspectives

1. Conclusions

Une approche théorique possible des analyses multivariées des tableaux à trois entrées a été proposée, elle consiste en l'étude de la structure de l'ensemble des analyses existantes sur la base de la complexité des liens entre celles-ci.

Des notions que l'on espère appropriées pour exprimer cette problématique ont été introduites : univers, sous-univers, dimension de l'univers, composante de liens, comportement géométrique, structure d'un sous-univers, structure de l'univers

En particulier, ces notions ont amené à formuler la question de l'étude de cette structure sous forme d'un problème, celui de la recherche de la dimension minimale de de l'univers. Explicitement, cette recherche vise à extraire le nombre de fonctionnements théoriques interactifs qui sont à la base des analyses existantes.
En effet, il s'agissait de construire des sous-ensembles de l'ensemble des analyses existantes, tels que les éléments de chacun d'eux, aient simultanément les mêmes propriétés, géométriques, méthodologiques et algorithmiques. Ces trois propriétés ont été introduites au cours de la présente partie. De tels sous-ensembles sont dits "composantes de liens" ou "unités théoriques".

Ainsi, une composante de liens contient des éléments qui obéissent à un même principe de fonctionnement à caractère interactif (géométrique, méthodologique, algorithmique). Celui-ci est une réalisation simultanée de plusieurs approches théoriques de nature différente, à l'image du principe de fonctionnement des analyses des tableaux à deux entrées qui ont le même principe de fonctionnement interactif.

Dans ce sens, trois approches ont été considérées et des contributions multiples ont été apportées.

Ainsi, l'approche géométrique a permis de dégager la notion de comportement géométrique. Sur la base de celle-ci, deux comportements géométriques ont été distingués. Ceci a permis d'extraire deux sous-ensembles susceptibles de présenter des composantes de liens. On les a appelés "sous-univers". Les analyses appartenant à chacun des deux sous-univers se trouvent géométriquement liées. Ce lien géométrique est établi indépendamment des critères qui sont à la base des analyses. (Pour plus de détails sur ces contributions, voir la fin du deuxième chapitre de cette partie - chapitre I.2).

L'approche méthodologique est située à un niveau moins implicite que l'approche géométrique. Elle est consacrée à l'analyse de la structure de chacun des deux sous-

univers, sur la base des critères que les éléments utilisent. On constate alors, que l'état de chacun de ces deux sous-univers ne reflète ni la diversité des solutions théoriques, ni la prise en compte des questions pratiques dont les analyses sont censées apporter des solutions. Ce constat a conduit naturellement à l'idée de compléter les deux sous-univers. Cette complétion a réalisé simultanément deux objectifs : le premier est de proposer des nouvelles analyses, le second est d'obtenir des sous-univers résultants plus structurés. (pour plus de détails, voir la fin du troisième chapitre- chapitre I.3)

L'approche algorithmique est certainement l'approche la plus explicite. Elle a montré que les éléments d'une partie du deuxième sous-univers sont liés aussi par un seul algorithme de résolution qui permet le calcul des solutions des problèmes d'optimisation étant à la base de ces éléments. En effet, un algorithme général de résolution a été proposé. Cette approche algorithmique a pour autre conséquence, l'obtention d'une expression analytique de ces liens sous forme d'équations algébriques, celles-ci étant exprimées par les conditions nécessaires qui vérifient les solutions de ces problèmes d'optimisation.

Ces trois approches permettent alors de proposer une composante de liens (chapitre I.5).

2 Perspectives

Etant donné le caractère interactif d'une approche qui vise l'étude de la structure de l'univers, on se restreint ici à deux perspectives dont on pense qu'elles méritent d'être approfondies.

— La première est l'absence d'un seul algorithme de résolution pour le premier sous-univers. Cette question reste ouverte. La réponse à celle-ci permettra de considérer le premier sous-univers comme une seconde composante de liens.

— La seconde concerne la structure de l'univers. En effet, l'état de la structure de chacun des deux sous-univers se révèle emboîté sur plusieurs niveaux. D'un niveau à l'autre, les solutions théoriques associées qui expriment des fonctionnements présentent un caractère demeurant marqué par une diversité permanente. Celle-ci est considérée ici comme un facteur structurant (voir chapitre I.3). On peut alors avancer que l'univers n'est qu'un niveau parmi d'autres dans ce système emboîté. Cela revient à considérer aussi une diversité des composantes de liens. On pense que cette hypothèse sur l'état de la structure de l'univers mérite un approfondissement.

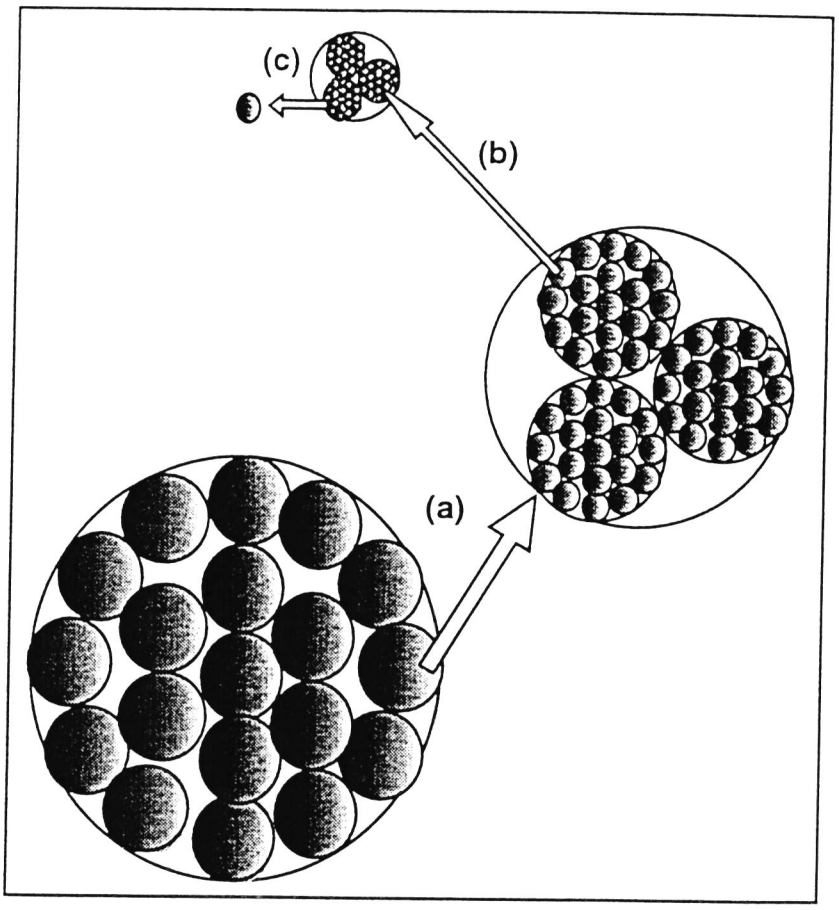


Figure 1. Illustration de l'hypothèse avancée sur l'état réel de la structure de l'univers. Cette figure montre trois niveaux emboités (a), (b) et (c). Chaque niveau exprime un fonctionnement théorique propre et à chacun d'eux, se manifeste une permanence de la diversité des solutions théoriques apportées.

Deuxième Partie

Structure de l'univers : Eléments appliqués

II.0. Introduction

1. Eléments de motivation appliquée : Exemples en écologie

En écologie, l'étude de la relation entre les êtres vivants (flore ou faune) et leur environnement est un problème central. Elle consiste à s'intéresser aux liens qui existent entre les deux compartiments fondamentaux de l'écosystème, la biocénose (êtres vivants) et le biotope (environnement). Cette étude considère alors l'écosystème comme une entité fonctionnelle où des processus physiques et biologiques sont en permanence en interaction; l'importance de cela est largement soulignée dans la littérature écologique (voir par exemple Crocker, 1952; Margalef, 1968).

Pour l'étude de cette relation, la campagne d'échantillonnage conduit généralement à l'acquisition d'un jeu de données récolté sur un ensemble de n stations (sites, relevés). Le jeu de données se présente sous forme de deux tableaux **X** et **Y**. Le tableau **X** dit "floro-faunistique" (figure 1) décrit la présence-absence, ou l'abondance (importance quantitative) de p taxons. Le tableau **Y** dit "mésologique" (figure 1) associé au tableau **X**, décrit l'environnement (milieu); il contient des mesures ou des évaluations de p paramètres descriptifs de l'environnement dans les mêmes stations. Un couple de tableaux ayant en commun les relevés (stations) est ainsi obtenu. Chacun d'entre eux a deux dimensions ou deux entrées (stations, taxons ou variables mésologiques).

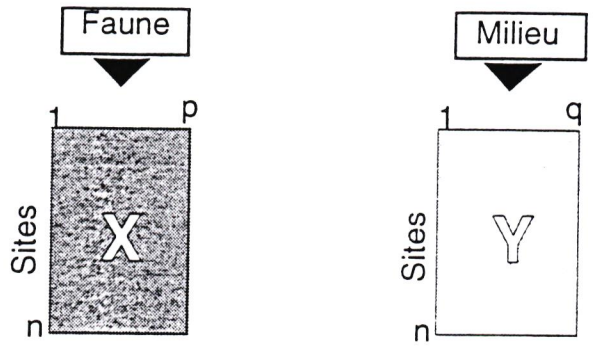


Figure 1. Couple de tableaux.

La question qui concerne la connaissance de l'importance quantitative ou qualitative des échelles telles que celle de l'espace, du temps, du groupe, ou de la région, est fondamentale pour la compréhension et le contrôle des phénomènes écologiques. Elle incite l'expérimentateur (écologue, biologiste) à intégrer, soit à l'un ou à l'autre des deux tableaux floro-faunistique ou mésologique, soit aux deux tableaux simultanément, une dimension supplémentaire (temps, espace, groupe, région). Ainsi, des situations très diverses peuvent être rencontrées, tant au niveau du tableau floro-faunistique **X**, qu'au niveau du tableau mésologique associé **Y**.

En effet, pour décrire le tableau floro-faunistique **X** qui contient les mesures d'abondance des taxons dans un ensemble de stations, plusieurs questions impliquent la considération

d'une troisième dimension qui génère une partition du tableau **X** en plusieurs tableaux donnant un K-tableaux ou un multitableaux. Ainsi à titre d'exemples, on cite les cas suivants :

— répartition des taxons du tableau **X** par groupes faunistiques (famille ou genre) (figure 2.A). Cette répartition est générée notamment lorsqu'on se pose la question de la valeur typologique des groupes faunistiques (Cazes & Chessel & Dolédec, 1988).

— répartition des stations (sites) en blocs selon leur appartenance à un ensemble régional ou continental (figure 2.B). Un exemple de cette répartition est considéré en biologie évolutive où est posée la question qui consiste à comparer la composition faunistique d'un peuplement de poissons dans trois continents différents (Winemiller, 1991).

— regroupement des stations (sites) par date. Ces stations peuvent faire l'objet d'une visite partielle ou totale à plusieurs reprises (figure 2.B). Ce regroupement est souvent effectué dans le cadre des enquêtes écologiques à composantes temporelles.

De même, pour décrire le tableau mésologique **Y** (Figure 2.C) qui contient des mesures qualitatives ou quantitatives de paramètres descriptifs de l'environnement, les sites peuvent être répartis selon les cas précédents.

Ainsi, chacun des deux tableaux floro-faunistique **X** et/ou mésologique **Y**, voit ses colonnes (variables, taxons) ou ses lignes (sites-stations-relevés) se structurer en fonction d'une troisième dimension. Cette structuration permet de générer à partir d'un tableau un K-tableaux ou un multitableaux.

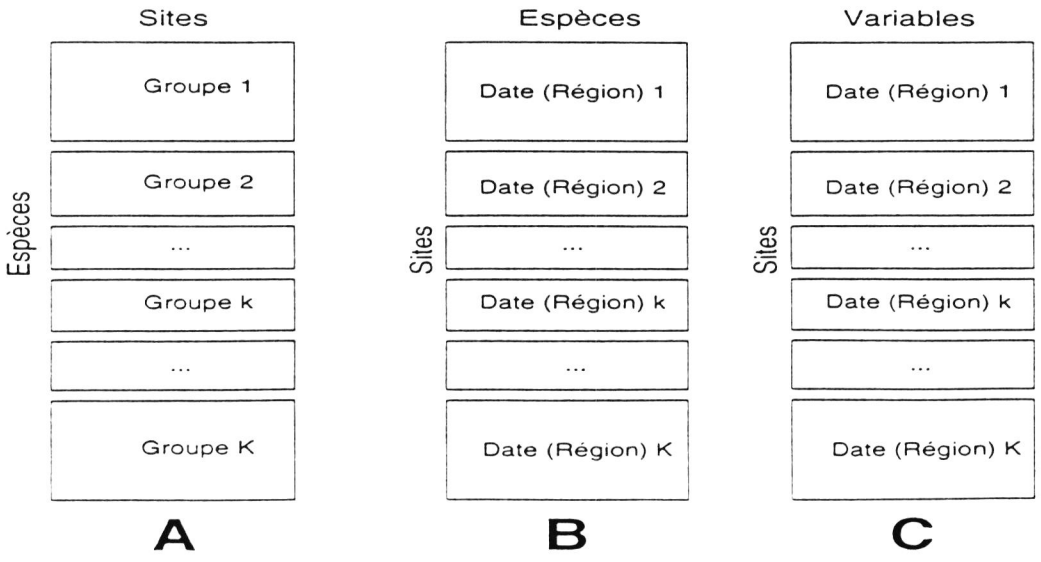


Figure 2. Situations expérimentales donnant un K-tableaux ou un multitableaux

La nouvelle structure en K-tableaux (ou multitableaux) définie par la troisième dimension (espace, date, groupe, région) sur l'un des deux tableaux (**X** ou **Y**) ou les deux simultanément, est alors prise en compte pour étudier naturellement la question centrale, celle de la relation : être vivant <-> environnement. Selon la structure imposée par cette troisième dimension sur les tableaux initiaux **X** et/ou **Y**, on distingue deux situations :

— dans la première situation, la structure en K-tableaux (ou multitableaux) se trouve simultanément dans les deux tableaux floro-faunistique et mésologique (figure 3.D). C'est souvent le cas dans le cadre des enquêtes écologiques à composantes spatiales ou temporelles (figure 2.B, 2.C).

— dans la deuxième situation, la structure en K-tableaux (ou multitableaux) se trouve dans l'un des deux tableaux et pas dans l'autre. C'est le cas pour la répartition du tableau faunistique **X** en groupes (figure, 2.A), cette partition du tableau faunistique **X** ne concerne pas le tableau mésologique **Y** (figure 3.E).

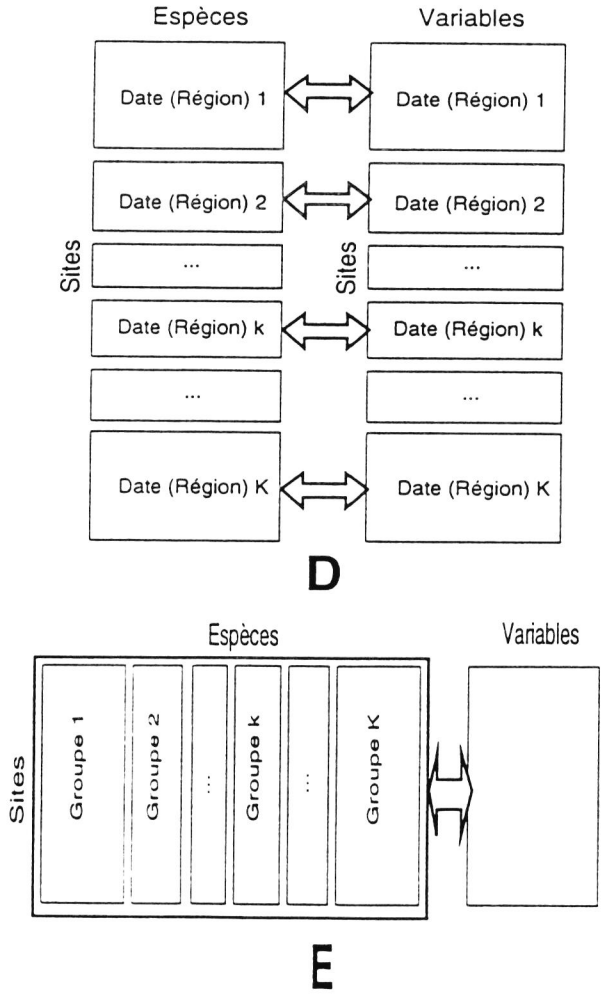


Figure 3. Deux situations d'extensions pour le couplage de deux tableaux

2. Formulations multivariées associées aux objectifs des exemples écologiques

2.1. Analyses des tableaux à deux entrées

Selon les objectifs poursuivis, pour chaque tableau du K-tableaux (ou multitableaux), on a un état typologique (relevés-taxons ou relevés-variables mésologiques). Celui-ci peut être exploré par des analyses telles que : l'analyse en composantes principales (ACP) centrée, doublement centrée ou normée, des analyses des correspondances multiples (ACM) qui utilisent les indicatrices des classes et les schémas de Tenenhaus et Young (1985), ou des analyses des correspondances floues (Chevenet et coll., 1994).

2.2. Analyses de couplage de deux tableaux

De même, l'étude des relations existant entre deux dates, deux régions, deux groupes, ou en général entre deux tableaux du K-tableaux (ou multitableaux), peut être effectuée par une analyse de couplage adaptée.

A l'image des propositions multivariées pour aborder la question du couplage entre deux tableaux (pour plus de détails théoriques, voir chapitre I.1), celle du couplage d'un tableau du K-tableaux faunistique et d'un tableau du K-tableaux mésologique est traitée par deux approches :

— une approche dissymétrique : il s'agit d'expliquer un état typologique faunistique des stations à l'aide de plusieurs variables de milieu (Ter Braak, 1987; Lebreton et al., 1991), ou inversement, de prédire un ensemble de variables de milieu à l'aide d'une structure faunistique (Ter Braak et Juggins, 1993). On peut consulter un récent travail de synthèse sur cette approche dans Lafosse (1997).

— une approche symétrique : il s'agit d'établir la concordance qui existe entre l'état typologique des stations du tableau floro-faunistique et celui des stations du tableau mésologique. Actuellement, les analyses de couplage de deux tableaux (analyse de co-inertie) permettent la mise en évidence d'un co-état typologique entre les relevés faunistiques et mésologiques pour différents types d'analyses de départ (Dolédec et Chessel, 1994).

2.3. Analyse des tableaux à trois entrées et extensions

La question qui concerne le K-tableaux (ou multitableaux) se veut plus générale que celles citées précédemment. Elle consiste à caractériser la stabilité ou la variabilité des relations qui existent entre :

(i)— les tableaux du K-tableaux (ou multitableaux) faunistique **X** ou mésologique **Y** (figure 2. A, B, C);

(ii)— les couples de tableaux générés simultanément par la troisième dimension au niveau des deux K-tableaux (ou multitableaux) faunistique **X** et mésologique **Y** (figure 3. D);

(iii)— le tableau mésologique **Y** et chacun des tableaux du K-tableaux (ou multitableaux) faunistique **X** (figure 3. E).

Dans le premier cas ((i)), on veut répondre à des questions telles que :
"Peut-on comparer les états typologiques des stations (individus) ou les états typologiques des taxons (variables) définies par chacun des tableaux issus du K-tableaux floro-faunistique ? "
Cette question reste valable pour le tableau mésologique.
"Y a t'il stabilité ou variabilité (temporelle ou régionale) de la séparation des individus. Certaines dates ou certaines régions sont-elles capables de reproduire le même état typologique au niveau des individus ou au niveau des variables ?"

Dans les deux autres cas ((i) et (iii)), on peut se poser la question suivante :
"Peut-on définir un co-état typologique reproductible et en mesurer la réalisation partielle ou totale dans certains groupes, dates, ou régions ?"

En écologie, sur le plan méthodologique, ces questions sont récentes et peu étudiées. On peut toutefois citer dans ce cadre, les analyses de co-inertie inter et intra-classes (Franquet et Chessel, 1994 ; Franquet et Coll., 1995), ou l'analyse triadique partielle (Chessel & Thioulouse (1987)).

3. Objectifs multivariés

L'objectif du point de vue de l'analyse multivariée consiste à proposer des analyses adaptées à la complexité des questions posées plus haut. Il vise d'une manière directe des propositions d'extension des analyses de couplage de deux tableaux à celles de l'analyse simultanée de plusieurs tableaux.
L'analyse de base considérée ici pour le couplage de deux tableaux est l'analyse symétrique de co-inertie. Une approche dissymétrique peut étre considérée en adoptant une analyse de base dissymétrique.

Les questions concernant les K-tableaux (ou multitableaux) évoquées plus haut ((i), (ii), (iii) et le choix de l'analyse de co-inertie comme analyse de couplage de base, permettent de préciser l'objectif multivarié appliqué associé qui consiste à :

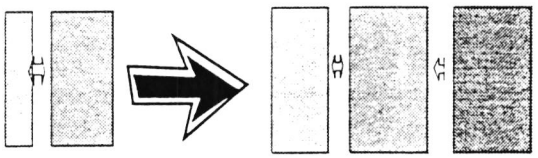
(1)— Généraliser l'analyse de co-inertie de deux tableaux à K tableaux avec K>2.

(2)— Généraliser l'analyse de co-inertie de deux tableaux à K couples de tableaux.

(3)— Généraliser l'analyse de co-inertie de deux tableaux au couplage entre un tableau et un K-tableaux.

4. Objectifs de cette partie

(1) *Généraliser l'analyse de co-inertie de deux tableaux à K tableaux avec K>2*



La question de la généralisation de l'analyse de co-inertie de deux tableaux à K tableaux avec K>2 n'admet pas une solution unique (première partie). En effet, il existe plusieurs solutions (généralisations) possibles obéissant soit au même fonctionnement théorique soit à des fonctionnements théoriques différents.

Cette situation pose un certain nombre de questions, parmi lesquelles on se restreindra à celle qui va dans le sens de l'objectif appliqué de cette partie. Il s'agit de la question de la comparaison entre deux généralisations possibles qui obéissent à des fonctionnements théoriques différents.

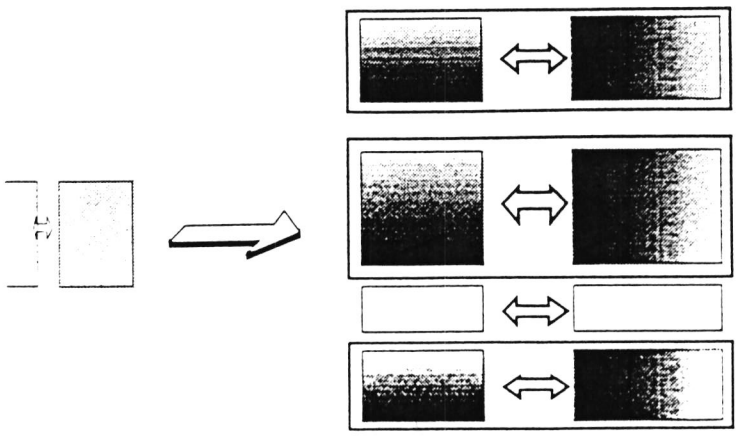
Dans un cadre général, cette question vise une comparaison entre les deux sous-univers discutés lors de la première partie. Elle ne sera réalisée ici qu'à travers deux analyses seulement, chacune appartenant à l'un des deux sous-univers. C'est le contenu du premier chapitre de cette partie.

La première analyse est connue, c'est ACT-STATIS (Analyse Conjointe de Tableaux-Structuration des Tableaux à trois indices de la Statistique), introduite par L'Hermier des Plantes (1976), elle est représentative du premier sous-univers.

La seconde est extraite des analyses nouvelles proposées lors de la complétion du deuxième sous-univers (Chpitre I.3), elle est appelée ici : Analyse de CO-inertie Multiple (ACOM); cette analyse est représentative du deuxième sous-univers.

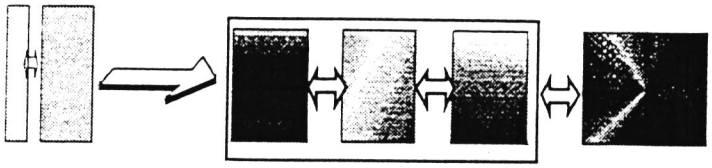
Ainsi le premier objectif de cette partie est une étude comparative entre ces deux analyses.

(2) Généraliser l'analyse de co-inertie de deux tableaux à K couples de tableaux



Cette question constitue le second objectif de cette partie, elle fait l'objet du second chapitre de cette partie. Plus exactement, une analyse spécifique sera proposée. On l'appelle STATICO, la solution intègre deux composantes : l'une issue de l'analyse de co-inertie de deux tableaux, l'autre issue de l'analyse ACT-STATIS. Un exemple d'illustration sera présenté.

(3). Généraliser l'analyse de co-inertie de deux tableaux au couplage entre un tableau et un K-tableaux



Cette question est proposée dans cette partie comme une perspective.

Dans cette partie, pour chacune des analyses discutées, un résumé des éléments théoriques de base, nécessaire pour la bonne compréhension de leur mise en œuvre, sera rappelé. Ces analyses ont été récemment intégrées dans le logiciel ADE-4 (Thioulouse & Coll., 1995), fusion des programmes ADE (Chessel & Doledèc, 1993), GraphMu et MacMul (Thioulouse 1989, 1990).

Chapitre II.1

Etude comparative entre les deux sous-univers

0. Introduction

Dans la première partie, il était question d'étudier la structure de l'ensemble des analyses existantes à trois entrées. En particulier, cette étude a dégagé deux fonctionnements théoriques qui ont été à la base de la distinction des deux sous-univers proposés. On a également expliqué comment les éléments de chacun des deux sous-univers, qui sont des analyses, obéissent à des logiques numériques (optimisation de fonctions différentes) et géométriques (comportement géométrique) propres.

Sur la base d'un exemple, ce chapitre vise une étude comparative entre les deux sous-univers. Ceux-ci comprenant un nombre très important d'analyses, on choisit de mener cette étude sur la base de deux analyses seulement, chaque sous-univers fournissant l'une d'entre elles. Plus exactement, on extrait du premier sous-univers une première analyse classiquement connue, il s'agit d'ACT-STATIS ("Analyse Conjointe de Tableaux - Structuration des Tableaux à Trois Indices de la Statistique"). On extrait la seconde parmi les nouvelles analyses que l'on a proposées lors de la complétion du deuxième sous-univers (I.3. approche méthodologique). Celle-ci a été introduite récemment dans la littérature par Chessel & Hanafi (1996) comme une généralisation à plus de deux tableaux de l'analyse de co-inertie (Doledèc & Chessel, 1994); on l'appelle "Analyse de Co-inertie Multiple" (ACOM).

Dans chacune des deux premières sections sera présentée une des deux analyses. Contrairement au premier chapitre qui considère une analyse uniquement par le critère et la contrainte qui sont à sa base, on privilégiera ici une présentation plus spécifique qui permettra de définir les outils que ces analyses génèrent lors de leur mise en œuvre. On expliquera également le choix de l'ACOM par rapport aux analyses introduites ou existantes qui constituent le deuxième sous-univers.

Dans une troisième section, on comparera à l'aide d'illustrations le comportement de ces deux analyses face à un même jeu de données associé à une question expérimentale dont on ne s'intéressera ici qu'à l'aspect multivarié.

On montrera alors, ce sera la conclusion principale de ce chapitre, qu'aucune des deux analyses n'a pu apporter d'indication qui aurait échappé à l'autre. Aucune non plus n'a mis en évidence de contradictions avec l'autre. Les différences théoriques entre les deux analyses reflètent alors uniquement des différences de moyens de mise en œuvre.

Le logiciel utilisé est ADE-4 (Thioulouse & Coll., 1996), fusion des programmes ADE (Chessel & Doledèc, 1993), GraphMu et MacMul (Thioulouse 1989, 1990). Il réunit les fonctions graphiques et les calculs associés aux deux analyses discutées ici. ADE-4 est en accès libre sur Internet (http://biomserv.univ-lyon1.fr/ADE-4.html).

Les deux analyses traitent les différents types d'études à trois entrées, mais pour les présenter, on se restreint à une étude horizontale $(\mathbf{X}, \mathbf{Q}, \mathbf{D}, \Pi)$. Celle-ci est formée à partir de K études $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ $(1 \leq k \leq K)$ portant sur les mêmes n individus et K groupes de variables, comptant respectivement $p_1, p_2, \ldots, p_K$ variables. L'entier p est la somme des p_k $(1 \leq k \leq K)$. Pour plus de détails sur les différents types d'études à trois entrées, on peut se référer aux définitions et notations (pp.17-19).

1. Présentation de l'Analyse Conjointe de Tableaux - Structuration des Tableaux à Trois Indices de la Statistique (ACT-STATIS)

L'analyse ACT-STATIS ("Analyse Conjointe de Tableaux - Structuration des Tableaux à Trois Indices de la Statistique") est présentée dans la thèse de L'Hermier des Plantes (1976), dans l'ouvrage de Lavit (1988) et plus récemment dans Lavit & Coll. (1994).

ACT-STATIS s'appuie sur le passage du triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ à son opérateur d'inertie $\mathbf{W}_k\mathbf{D} = \mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}$. De ce fait, comme on l'a expliqué dans le chapitre I.2, elle est considérée comme un élément du premier sous-univers.

Le nuage des opérateurs d'inertie $\mathbf{W}_k\mathbf{D} = \mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}$ est alors situé dans l'espace vectoriel des opérateurs $\mathbf{D}$-symétriques muni du produit scalaire d'Hilbert Schmidt (voir aussi le chapitre I.2). En conséquence, ACT-STATIS mesure l'état typologique du nuage des individus définis par le triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$, par la norme de Hilbert-Schmidt associée à son opérateur $\mathbf{D}$-symétrique $\mathbf{W}_k\mathbf{D} = \mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}$. Plus explicitement, au $k^{\text{ème}}$ triplet statistique $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$, l'analyse associe à la notion de mesure habituelle de l'état typologique du nuage d'individus qui est l'inertie totale I_t :

$$I_t = \text{Trace}\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\right) = \sum_i^n \lambda_i ,$$

la notion de variance vectorielle (Escoufier 1973) ou la norme d'opérateur :

$$\left\|\mathbf{W}_k\mathbf{D}\right\|_{\text{HS}}^2 = \text{Vav}(\mathbf{X}_k) = \text{Trace}\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\right) = \sum_i^n \lambda_i^2$$

où $\| \quad \|_{\text{HS}}$ désigne la norme d'Hilbert Schmidt.

A cette notion de variance vectorielle, l'analyse associe une seconde notion qui mesure l'adéquation de deux nuages d'individus définis par deux triplets $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ et $(\mathbf{X}_l, \mathbf{Q}_l, \mathbf{D})$, il s'agit de la covariance vectorielle qui est le produit scalaire issu de cette norme :

$$\text{Covv}(\mathbf{X}_k, \mathbf{X}_l) = \text{Trace}\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\mathbf{X}_l\mathbf{Q}_l\mathbf{X}_l^t\mathbf{D}\right) = \sum_{i=1}^n \mu_i$$

C'est exactement la co-inertie totale associée au couple $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ et $(\mathbf{X}_l, \mathbf{Q}_l, \mathbf{D})$ qui se décompose dans l'analyse de co-inertie des deux triplets (voir ci-dessous : section 2).
Il s'en suit qu'on peut associer une notion de corrélation entre deux triplets par le coefficient suivant dit "coefficient de corrélation vectorielle" :

$$\mathrm{Rv}(\mathbf{X}_k, \mathbf{X}_l) = \frac{\mathrm{Covv}(\mathbf{X}_k, \mathbf{X}_l)}{\sqrt{\mathrm{Vav}(\mathbf{X}_k)}\sqrt{\mathrm{Vav}(\mathbf{X}_l)}}$$

Relativement aux notions de mesure considérées ci-dessus (variance vectorielle, covariance vectorielle, corrélation vectorielle), ACT-STATIS utilise explicitement un schéma en trois étapes : Interstructure, Compromis et Intrastructure.

1.1. Etape Interstructure

Cette étape a pour objectif une représentation synthétique et globale des K études. Cela implique en particulier, la diagonalisation de la matrice de variance-covariance vectorielle et la conservation de plusieurs axes. En effet, du point de vue quantitatif, l'information concernant la comparaison entre les K études est prise en compte dans la matrice de variance-covariance vectorielle qu'on note $\Omega = \left[\pi_k \pi_l \mathrm{Covv}(\mathbf{X}_k, \mathbf{X}_l)\right]$, π_k étant un poids positif associé à $\mathbf{W}_k\mathbf{D}$. Classiquement, les poids utilisés sont les poids unitaires ou les inverses des normes des opérateurs d'inertie $\mathbf{W}_k\mathbf{D}$ $(1 \le k \le K)$. Dans le premier cas, on parle de la matrice de variance-covariance vectorielle, dans le second cas, de la matrice de corrélation vectorielle. Les racines carrées des poids π_k peuvent être rangées dans une matrice diagonale qui est la matrice Π évoquée dans la définition d'une étude à trois entrées.
Du point de vue géométrique, à la matrice Ω est associée une image euclidienne définie par un nuage pondéré par les poinds π_k dit "nuage des opérateurs $\mathbf{W}_k\mathbf{D}$" situé dans l'espace vectoriel des opérateurs $\mathbf{D}$-symétriques de $\mathbb{R}^n$ muni du produit scalaire de Hilbert-Schmidt. Pour plus de détails, on peut se referer à Lavit (1988, pp.99).
L'analyse du nuage des opérateurs est résolue classiquement par l'Analyse en Composantes Principales. Celle-ci définit un système d'axes d'inertie dont le premier a des propriétés très particulières comme tout premier axe d'une ACP non centrée.

1.2. Étape Compromis

Cette étape constitue l'idée fondamentale d'ACT-STATIS. Elle consiste à chercher un compromis entre les K études $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ $(1 \le k \le K)$. Plus exactement, l'objectif de cette étape vise à définir un état typologique moyen des individus entre les K états typologiques des individus définis respectivement par les K études. Dans l'esprit de l'analyse, cet état typologique moyen (compromis) est associé à une étude fictive qui fournit le meilleur état typologique possible. La prise en compte réelle de la construction de cet état typologique moyen, consiste à définir un opérateur compromis comme une

combinaison linéaire des K opérateurs d'inertie initiaux $\mathbf{W}_k\mathbf{D}$ $(1 \leq k \leq K)$. D'autre part, cette combinaison linéaire d'opérateurs a une variance vectorielle maximale.
L'opérateur compromis est alors la solution du problème d'optimisation suivant :

$$\text{Maximiser } \text{Trace}(\mathbf{WDWD}) \text{ sous la contrainte } \sum_{k=1}^{K} \alpha_k^2 = 1$$

avec $\mathbf{WD} = \sum_{k=1}^{K} \alpha_k \pi_k \mathbf{W}_k\mathbf{D}$.

Géométriquement, la résolution de ce problème découle de l'Analyse en Composantes Principales du nuage des opérateurs $\mathbf{W}_k\mathbf{D}$ et dépend du premier vecteur propre de la matrice Ω. En effet, les éléments de la matrice Ω sont tous positifs, les composantes notées β_k du premier vecteur propre de la matrice Ω sont alors de même signe, on peut les considérer toutes positives. Ainsi, le compromis s'exprime par $\mathbf{WD} = \sum_{k=1}^{K} \beta_k \pi_k \mathbf{W}_k\mathbf{D}$.
De ce fait, l'opérateur compromis est symétrique et semi-défini positif.

1.3. Étape Intrastructure

Cette étape consiste à analyser le nuage d'un tableau fictif dont l'opérateur d'inertie correspondant est l'opérateur compromis $\mathbf{WD} = \sum_{k=1}^{K} \beta_k \pi_k \mathbf{W}_k\mathbf{D}$. La prise en compte réelle de cette analyse consiste à définir des axes par la diagonalisation de l'opérateur compromis $\mathbf{WD}$.

2. Présentation de l'Analyse de CO-inertie Multiple (ACOM)

Dans un premier temps, on rappelle l'analyse de co-inertie de deux tableaux (Chessel et Mercier, 1993), qui sera suivie par la définition de l'analyse de co-inertie multiple. Dans un second temps, on explique le choix de cette analyse par rapport à celles introduites ou existantes qui constituent le deuxième sous-univers.

2.1. Analyse de co-inertie d'une études 2-tableaux horizontale

L'analyse de co-inertie de deux tableaux est le nom générique qui recouvre : l'analyse inter-batterie de Tucker (1958), l'analyse canonique sur variables qualitatives de Cazes (1980) et l'analyse des correspondances de tableaux de profils écologiques (Mercier et Coll. 1992). Sa stabilité numérique (Kazi-Aoual et Coll., 1995), sa facilité d'emploi en termes de double analyse d'inertie à coordonnées covariantes (Doledèc & Chessel, 1994; Prodon & Lebreton, 1994), son universalité en termes de conditions numériques, de types d'analyses de départ et ses propriétés de point de départ de la régression PLS (Tenenhaus et Coll., 1995), en font une bonne alternative à l'analyse canonique de

Hotelling (1936), qui est numériquement instable ou difficilement interprétable (Kettenring, 1985).

C'est Tucker (1958) qui a substitué à la maximisation d'un coefficient de corrélation entre variables canoniques (solution de l'analyse canonique de Hotteling), celle de la covariance entre combinaisons de variables. Optimiser une covariance (produit scalaire) impose de ne pas se désintéresser des variances des combinaisons, donc des inerties projetées. Pour plus de détails, on peut se référer à la section I.3 (principes méthodologiques) du premier chapitre.

L'analyse de co-inertie de deux triplets $(\mathbf{X}_1, \mathbf{Q}_1, \mathbf{D})$, $(\mathbf{X}_2, \mathbf{Q}_2, \mathbf{D})$ est l'analyse du triplet $(\mathbf{X}_2'\mathbf{D}\mathbf{X}_1, \mathbf{Q}_1, \mathbf{Q}_2)$. Du point de vue du problème d'optimisation associé (Chessel et Mercier, 1993), elle consiste en une maximisation successive (équivalente aussi à une maximisation simultanée) du critère suivant :

$$(\mathbf{X}_1\mathbf{Q}_1\mathbf{u}_1, \mathbf{X}_2\mathbf{Q}_2\mathbf{u}_2)_{\mathbf{D}} ;$$

sous les contraintes :

$$\|\mathbf{u}_1\|_{\mathbf{Q}_1} = \|\mathbf{u}_2\|_{\mathbf{Q}_2} = 1.$$

Géométriquement, elle revient à trouver un couple d'axes normés respectivement dans $\mathbb{R}^{p_1}$ et dans $\mathbb{R}^{p_2}$, dits "axes de co-inertie". Les solutions impliquent la diagonalisation des deux opérateurs dits "de co-inertie" $\mathbf{X}_2'\mathbf{D}\mathbf{X}_1\mathbf{Q}_1\mathbf{X}_1'\mathbf{D}\mathbf{X}_2\mathbf{Q}_2$ et $\mathbf{X}_1'\mathbf{D}\mathbf{X}_2\mathbf{Q}_2\mathbf{X}_2'\mathbf{D}\mathbf{X}_1\mathbf{Q}_1$.

2.2. Analyse de Co-inertie Multiple

Récemment introduite par Chessel & Hanafi (1996), l'Analyse de CO-inertie Multiple (ACOM) est proposée comme une généralisation de l'analyse de co-inertie à plus de deux tableaux. Elle se définit comme une recherche successive sous contraintes d'orthogonalité de (K+1)-uplet de vecteurs : K vecteurs $\mathbf{u}_k$ $\mathbf{Q}_k$-normés dans chaque espace $\mathbb{R}^{p_k}$, et un K+1-ième vecteur $\mathbf{v}$, $\mathbf{D}$-normé dans l'espace vectoriel $\mathbb{R}^n$, qui maximise la quantité :

$$g(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K, \mathbf{v}) = \sum_{k=1}^{K} \pi_k (\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k|\mathbf{v})_{\mathbf{D}}^2$$

Ces contraintes d'orthogonalité sont imposées sur les vecteurs $\mathbf{u}_k$ dans chacun des K espaces $\mathbb{R}^{p_k}$. Plus explicitement, l'ACOM au pas 1 se définit comme la recherche de k vecteurs $\mathbf{u}_k^1$ $\mathbf{Q}_k$-normés dans chaque espace $\mathbb{R}^{p_k}$, et d'une variable dite auxiliaire $\mathbf{v}^1$, $\mathbf{D}$-normée dans l'espace vectoriel $\mathbb{R}^n$, qui maximise :

$$g(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K, \mathbf{v}) = \sum_{k=1}^{K} \pi_k (\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k|\mathbf{v})_{\mathbf{D}}^2$$

Une fois qu'un premier (K+1)-uplet $\left(\mathbf{u}_1^1, \mathbf{u}_2^1, \ldots, \mathbf{u}_K^1, \mathbf{v}^1\right)$ de vecteurs est trouvé, on continue la recherche. Ainsi au pas 2, l'ACOM consiste à chercher (K+1) solutions : K vecteurs $\mathbf{u}_k^2$-$\mathbf{Q}_k$ normés dans chaque espace $\mathbb{R}^{p_k}$, et un vecteur $\mathbf{v}^2$ $\mathbf{D}$-normé dans $\mathbb{R}^n$, qui maximise la même quantité sous contrainte d'orthogonalité :

$$\left(\mathbf{u}_k^1 \middle| \mathbf{u}_k^2\right)_{\mathbf{Q}_k} = 0$$

Itérativement, au pas s, l'ACOM consiste à maximiser la même quantité sous les contraintes :

$$\left(\mathbf{u}_k^j \middle| \mathbf{u}_k^s\right)_{\mathbf{Q}_k} = 0 \quad (1 \le j < s, 1 \le k \le K).$$

L'Analyse de CO-inertie Multiple est à l'analyse de co-inertie ce qu'est l'Analyse Canonique Généralisée de Carroll (1968) à l'analyse canonique de Hotteling (1936). On trouve dans Saporta (1975) et Tenenhaus (1984) un exposé théorique de cette analyse de Carroll, ainsi que des applications aux variables qualitatives et aux mélanges de type de variables. En effet, l'analyse canonique généralisée au sens de Carroll (1968) est initialement proposée pour généraliser l'Analyse Canonique de Hotteling à plus de deux tableaux. Elle consiste à trouver d'abord au pas 1 des solutions $\mathbf{u}_k^1$ dans chaque espace $\mathbb{R}^{p_k}$, et une variable auxiliaire $\mathbf{v}^1$, $\mathbf{D}$-normée dans l'espace vectoriel $\mathbb{R}^n$, qui maximise la quantité :

$$f\left(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K, \mathbf{v}\right) = \sum_{k=1}^{K} \pi_k \mathrm{corr}^2\left(\mathbf{X}_k \mathbf{Q}_k \mathbf{u}_k, \mathbf{v}\right)$$

où corr(.,.) désigne la corrélation linéaire entre deux variables.

Au pas 2, on cherche des solutions $\mathbf{u}_k^2$ normées dans chaque espace $\mathbb{R}^{p_k}$, et une variable $\mathbf{v}^2$, dans $\mathbb{R}^n$ qui maximise la même quantité sous contrainte d'orthogonalité. Au pas s, la même quantité est maximisée sous les contraintes :

$$\left(\mathbf{v}^j \middle| \mathbf{v}^s\right)_{\mathbf{D}} = 0 \quad (1 \le j < s)$$

Pour l'utilisateur des analyses multivariées, l'analyse canonique et ses généralisations portent toutes les limitations associées aux analyses qui utilisent les projecteurs ou implicitement les métriques de Mahalanobis $\mathbf{Q}_k = \left(\mathbf{X}_k^t \mathbf{D} \mathbf{X}_k\right)^-$ (voir Van de Geer (1984) ou, ici-même, la section I.3, "principes méthodologiques"). En remplaçant les projecteurs par l'usage des opérateurs d'inertie $\mathbf{W}_k\mathbf{D} = \mathbf{X}_k \mathbf{Q}_k \mathbf{X}_k^t \mathbf{D}$, on définit l'analyse de co-inertie multiple.

2.3. Choix de l'ACOM dans le deuxième sous-univers

Dans le chapitre I.3, cet argument a permis d'une part, d'étendre les autres généralisations (Kettenring, 1971) de l'analyse canonique de Hotteling, et d'autre part, de proposer une formulation générale du critère qui est à la base de l'ACOM. En effet le critère de L'ACOM est le critère général Ψ_5 dans le cas particulier $r = 1$. S'impose alors la question du choix du critère étant à la base de la définition de l'ACOM, parmi les autres critères proposés ou existants qui constituent le deuxième sous-univers. Ce choix est motivé par les raisons pratiques suivantes qui sont vérifiées par le critère de l'ACOM :

— il utilise des variables auxiliaires,

— c'est un critère d'analyse en composantes principales,

— l'algorithme de calcul des solutions maximisantes de ce critère est de mise en œuvre aisée et permet le calcul d'une solution globale.

2.3.1. ACOM et variables auxiliaires

La notion de variables auxiliaires dans l'ACOM est issue de celle de l'Analyse Canonique Généralisée au sens de Carroll. Celle-ci est la seule généralisation de l'analyse canonique de Hotteling qui en fait usage. On voit une utilisation et une extension de cette notion, d'une part, dans l'analyse PRINQUAL de Tenenhaus (1977) qui consiste à maximiser, pour m fixé :

$$\sum_{k=1}^{K} \pi_k \sum_{j=1}^{m} \mathrm{corr}^2\left(\mathbf{X}_k \mathbf{Q}_k \mathbf{u}_k^1, \mathbf{v}_j\right),$$

et d'autre part, dans les analyses canoniques généralisées basées sur d'autres critères de corrélation (LAZRAQ et Coll.; 1992).

L'utilisation des variables auxiliaires dans l'ACOM a un apport important : celles-ci sont à la base d'une notion de moyenne sur laquelle on reviendra dans la prochaine section.

2.3.2. ACOM et analyse en composantes principales

Le critère de l'ACOM est celui de l'Analyse en Composantes Principales (ACP) du tableau suivant, qu'on appelle ici "tableau de synthèse" :

$$\mathbf{S}_{\mathbf{u}}^r = \left[\mathbf{X}_1\mathbf{Q}_1\mathbf{u}_1^r \quad \mathbf{X}_2\mathbf{Q}_2\mathbf{u}_2^r \quad \cdots \quad \mathbf{X}_K\mathbf{Q}_K\mathbf{u}_K^r\right]$$

d'une manière plus explicite :

$$S_{\mathbf{u}}^r = \begin{bmatrix} \left\langle \mathbf{x}_1^1, \mathbf{u}_1^r \right\rangle_{\mathbf{Q}_1} & \left\langle \mathbf{x}_1^2, \mathbf{u}_2^r \right\rangle_{\mathbf{Q}_2} & \cdots & \left\langle \mathbf{x}_1^K, \mathbf{u}_K^r \right\rangle_{\mathbf{Q}_K} \\ \left\langle \mathbf{x}_2^1, \mathbf{u}_1^r \right\rangle_{\mathbf{Q}_1} & \left\langle \mathbf{x}_2^2, \mathbf{u}_2^r \right\rangle_{\mathbf{Q}_2} & \cdots & \left\langle \mathbf{x}_2^K, \mathbf{u}_K^r \right\rangle_{\mathbf{Q}_K} \\ \vdots & \vdots & \ddots & \vdots \\ \left\langle \mathbf{x}_n^1, \mathbf{u}_1^r \right\rangle_{\mathbf{Q}_1} & \left\langle \mathbf{x}_n^2, \mathbf{u}_2^r \right\rangle_{\mathbf{Q}_2} & \cdots & \left\langle \mathbf{x}_n^K, \mathbf{u}_K^r \right\rangle_{\mathbf{Q}_K} \end{bmatrix}$$

$\mathbf{x}_i^k$ désigne l'individu i du tableaux kAinsi, au pas r, l'ACOM consiste en la recherche d'un triplet de synthèse $\left(\mathbf{S}_u^r, \mathbf{D}, \Pi\right)$ qui fournit au rang 1 la meilleure analyse possible.

Ceci est pratique. En effet, ce triplet de synthèse permet à chaque pas d'utiliser une logique classique d'interprétation d'un triplet statistique, ce qui a un apport dans le champ des utilisateurs.

2.3.3. ACOM et algorithme

L'algorithme général déjà proposé dans le chapitre I.4 n'assure pas le calcul d'une solution globale pour la maximisation du critère de l'ACOM, c'est pourquoi, on propose ici un second algorithme spécifique au critère de l'ACOM qui a l'avantage de calculer une solution globale. C'est l'objet du développement qui suit.

Solution d'ordre 1 de l'ACOM

Pour $\mathbf{v}$ un vecteur fixé, $\mathbf{D}$-normé dans $\mathbb{R}^n$, l'application de l'inégalité de Cauchy-Schwartz montre que la quantité $\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{u}_k|\mathbf{v}\right)_{\mathbf{D}}^2$ est majorée par $\left\|\mathbf{X}_k^t\mathbf{D}\mathbf{v}\right\|_{\mathbf{Q}_k}^2$. De plus, ce majorant est atteint pour $\mathbf{u}_k = \dfrac{\mathbf{X}_k^t\mathbf{D}\mathbf{v}}{\left\|\mathbf{X}_k^t\mathbf{D}\mathbf{v}\right\|_{\mathbf{Q}_k}}$.

Il s'ensuit que le vecteur $\mathbf{v}^1$, $\mathbf{D}$-normé dans l'espace vectoriel $\mathbb{R}^n$, qui maximise la fonction g, maximise également la quantité :

$$\sum_{k=1}^{K}\left\|\mathbf{X}_k^t\mathbf{D}\mathbf{v}\right\|_{\mathbf{Q}_k}^2 = \sum_{k=1}^{K}\left(\mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^t\mathbf{D}\mathbf{v}|\mathbf{v}\right)_{\mathbf{D}} = \sum_{k=1}^{K}\left(\mathbf{W}_k\mathbf{D}\mathbf{v}|\mathbf{v}\right)_{\mathbf{D}} = \mathbf{v}^t\mathbf{D}\left(\sum_{k=1}^{K}\mathbf{W}_k\mathbf{D}\right)\mathbf{v}$$

Donc $\mathbf{v}^1$ est la première composante $\mathbf{D}$-normée de l'ACP du tableau $\mathbf{X}$. Il en découle que les axes $\mathbf{u}_k^1$, $\mathbf{Q}_k$-normés dans $\mathbb{R}^{p_k}$, sont les vecteurs de $\mathbb{R}^{p_k}$ normalisés par bloc du premier axe principal du tableau $\mathbf{X}$.

Proposition

Les solutions d'ordre 2 existent et sont obtenues par les étapes suivantes :

— considérer les projecteurs $\mathbf{Q}_k$-orthogonaux notés $\mathbf{P}_k^1$ sur les sous-espaces vectoriels de $\mathbf{R}^{p_k}$ engendrés par le vecteur $\mathbf{u}_k^1$;

— définir le tableau $\mathbf{Z}$ de la manière suivante :

$$\mathbf{Z} = \left[\mathbf{Z}_1|\mathbf{Z}_2|\ldots|\mathbf{Z}_K\right] \text{ avec } \mathbf{Z}_k = \mathbf{X}_k - \mathbf{X}_k\mathbf{P}_k^1 ;$$

— calculer les solutions de rang 1 de la co-inertie multiple du tableau $\mathbf{Z}$.

Démonstration

on se restreint aux cas $\mathbf{Q}_k = \mathbf{I}_{p_k}$ et $\mathbf{D} = \mathbf{I}_n$ sans perte de généralité (il suffit de se placer par changement de base, dans des bases orthonormales), et on définit les deux problèmes suivants :

Problème 1 : Trouver $\mathbf{a}_j^1 \quad (1 \le j \le K)$ normés et $\mathbf{w}^1$ normé, solution du rang 1 de co-inertie multiple du tableau $\mathbf{Z} = [\mathbf{Z}_1 | \mathbf{Z}_2 | \ldots | \mathbf{Z}_K]$.

Problème 2 : Trouver $\mathbf{a}_j^2 \quad (1 \le j \le K)$ normés et $\mathbf{w}^2$ normé qui maximisent $\sum_{j=1}^{K} \left(\mathbf{X}_j \mathbf{u}_j | \mathbf{v}\right)^2$ sous les contraintes : $\left(\mathbf{v} | \mathbf{v}_1\right) = 0$ et $\left(\mathbf{u}_j^1 | \mathbf{u}_j\right) = 0 \quad (1 \le j \le K)$.

Les solutions du problème **1** et du problème **2** existent toujours car les fonctions à maximiser sont continues et les contraintes définissent des ensembles compacts. On va montrer que les solutions du problème **1** et du problème **2** sont identiques. Sachant que la solution du problème **1** est celle du paragraphe précédent, la solution du problème **2** sera alors acquise.

(a)— Commençons par constater que $\mathbf{w}^1$ est orthogonal à $\mathbf{v}^1$.

En effet, $\mathbf{v}^1$ est un vecteur propre de la somme des opérateurs d'inertie (voir ci-dessus) :

$$\left(\sum_{k=1}^{K} \mathbf{W}_k\right) \mathbf{v}^1 = \lambda \mathbf{v}^1.$$

Pour la même raison, en utilisant la symétrie des projecteurs, on a :

$$\left(\sum_{k=1}^{K} \mathbf{W}_k\right) \mathbf{w}^1 - \left(\sum_{k=1}^{K} \mathbf{X}_k \mathbf{P}_k \mathbf{X}_k^t\right) \mathbf{w}^1 = \mu \mathbf{w}^1$$

d'où, toujours grâce à la symétrie des opérateurs en jeu :

$$\left(\left(\sum_{k=1}^{K} \mathbf{W}_k\right) \mathbf{v}^1 \middle| \mathbf{w}^1\right) - \left(\left(\sum_{k=1}^{K} \mathbf{X}_k \mathbf{P}_k \mathbf{X}_k^t\right) \mathbf{v}^1 \middle| \mathbf{w}^1\right) = \mu \left(\mathbf{v}^1 \middle| \mathbf{w}^1\right).$$

D'autre part :

$$\left(\left(\sum_{k=1}^{K} \mathbf{W}_k\right) \mathbf{v}^1 \middle| \mathbf{w}^1\right) = \lambda \left(\mathbf{v}^1 \middle| \mathbf{w}^1\right),$$

de plus, d'après la solution au rang 1 et la définition de la projection sur un vecteur :

$$\mathbf{P}_k \mathbf{X}_k^t \mathbf{v}^1 = \left\| \mathbf{X}_k^t \mathbf{v}^1 \right\| \mathbf{u}_k^1 = \mathbf{X}_k^t \mathbf{v}^1,$$

d'où :

$$\mathbf{X}_k\mathbf{P}_k\mathbf{X}_k^t\mathbf{v}^1 = \mathbf{W}_k\mathbf{v}^1,$$

donc :

$$\left(\left(\sum_{k=1}^{K}\mathbf{X}_k\mathbf{P}_k\mathbf{X}_k^t\right)\mathbf{v}^1\middle|\mathbf{w}^1\right) = \left(\sum_{k=1}^{K}\mathbf{X}_k\mathbf{P}_k\mathbf{X}_k^t\mathbf{v}^1\middle|\mathbf{w}^1\right) = \sum_{j=1}^{K}\left(\mathbf{X}_k\mathbf{P}_k\mathbf{X}_k^t\mathbf{v}^1\middle|\mathbf{w}^1\right) = \sum_{j=1}^{K}\left(\mathbf{W}_k\mathbf{v}^1\middle|\mathbf{w}^1\right).$$

Finalement, en excluant le cas dégénéré $\mu = 0$:

$$\mu\left(\mathbf{v}^1\middle|\mathbf{w}^1\right) = 0 \Rightarrow \left(\mathbf{v}^1\middle|\mathbf{w}^1\right) = 0.$$

(b) — Montrons alors que, pour $1 \le k \le K$, $\mathbf{a}_k^1$ est orthogonal à $\mathbf{u}_k^1$.

D'après la solution de rang 1 du tableau $\mathbf{Z}$, on a :

$$\mathbf{a}_k^1 = \frac{\mathbf{Z}_k^t\mathbf{w}^1}{\left\|\mathbf{Z}_k^t\mathbf{w}^1\right\|}$$

avec $\mathbf{w}^1$ est la première composante principale du tableau $\mathbf{Z} = [\mathbf{Z}_1 | \mathbf{Z}_2 | \ldots | \mathbf{Z}_K]$.

Le caractère idempotent des projecteurs implique :

$$\left(\mathbf{I}_k - \mathbf{P}_k^1\right)\mathbf{a}_k^1 = \frac{\left(\mathbf{I}_k - \mathbf{P}_k^1\right)\mathbf{Z}_k^t\mathbf{w}^1}{\left\|\mathbf{Z}_k^t\mathbf{w}^1\right\|} = \frac{\left(\mathbf{I}_k - \mathbf{P}_k^1\right)\mathbf{X}_k^t\mathbf{w}^1}{\left\|\mathbf{Z}_k^t\mathbf{w}^1\right\|} = \frac{\mathbf{Z}_k^t\mathbf{w}^1}{\left\|\mathbf{Z}_k^t\mathbf{w}^1\right\|} = \mathbf{a}_k^1.$$

Le développement de $\mathbf{a}_k^1 = \left(\mathbf{I}_k - \mathbf{P}_k^1\right)\mathbf{a}_k^1$ donne alors :

$$\mathbf{a}_k^1 = \mathbf{a}_k^1 - \left(\mathbf{a}_k^1\middle|\mathbf{u}_k^1\right)\mathbf{u}_k^1 \Rightarrow \left(\mathbf{a}_k^1\middle|\mathbf{u}_k^1\right) = 0.$$

En conséquence, les solutions du problème 1 vérifient les contraintes du problème 2. Donc, par définition de la solution du problème 2, on a :

$$\sum_{k=1}^{K}\left(\mathbf{Z}_k\mathbf{a}_k^1\middle|\mathbf{w}^1\right)^2 = \sum_{k=1}^{K}\left(\mathbf{X}_k\mathbf{a}_k^1\middle|\mathbf{w}^1\right)^2 \le \sum_{k=1}^{K}\left(\mathbf{X}_k\mathbf{a}_k^2\middle|\mathbf{w}^2\right)^2.$$

De même, par définition, la solution du problème 2 vérifie :

$$\sum_{k=1}^{K}\left(\mathbf{Z}_k\mathbf{a}_k^2\middle|\mathbf{w}^2\right)^2 = \sum_{k=1}^{K}\left(\mathbf{X}_k(\mathbf{I} - \mathbf{P}_k)\mathbf{a}_k^2\middle|\mathbf{w}^2\right)^2 = \sum_{k=1}^{K}\left(\mathbf{X}_k\mathbf{a}_k^2\middle|\mathbf{w}^2\right)^2 \le \sum_{k=1}^{K}\left(\mathbf{X}_k\mathbf{a}_k^1\middle|\mathbf{w}^1\right)^2,$$

donc :

$$\sum_{k=1}^{K}\left(\mathbf{X}_k\mathbf{a}_k^2\middle|\mathbf{w}^2\right)^2=\sum_{k=1}^{K}\left(\mathbf{Z}_k\mathbf{a}_k^1\middle|\mathbf{w}^1\right)^2.$$

Solution d'ordre s

Au pas s, pour $s\geq 2$, la solution est alors obtenue en effectuant les trois étapes suivantes :

— considérer les projecteurs $\mathbf{Q}_k$-orthogonaux notés $\mathbf{P}_k^{s-1}$ sur les sous-espaces vectoriels de $\mathbb{R}^{p_k}$ engendrés par les systèmes orthonormés $\left\{\mathbf{u}_{k,}^1,\ldots,\mathbf{u}_k^{s-1}\right\}$;

— remplacer le tableau $\mathbf{X}$ par le tableau $\mathbf{X}^s$ de la manière suivante :

$$\mathbf{Z}=\left[\mathbf{X}_1-\mathbf{X}_1\mathbf{P}_1^{s-1}\middle|\mathbf{X}_2-\mathbf{X}_2\mathbf{P}_2^{s-1}\middle|\ldots\middle|\mathbf{X}_K-\mathbf{X}_K\mathbf{P}_K^{s-1}\right];$$

— calculer les solutions de rang 1 de co-inertie multiple du tableau $\mathbf{Z}$.

3. Illustrations comparées

3.1. Données et Objectifs

Les données utilisées sont publiées par Friday (1987). Dans $n = 16$ étangs, Friday a mesuré l'abondance de $p = 91$ espèces réparties en $K = 10$ groupes faunistiques comportant p_k taxons, avec respectivement $p_1 = 11$ (1-Hémiptera), $p_2 = 7$ (2-Odonata), $p_3 = 13$ (3-Trichoptera), $p_4 = 4$ (4-Ephemeroptera), $p_5 = 13$ (5-Coleoptera), $p_6 = 22$ (6-Diptera), $p_7 = 4$ (7-Hydracarina), $p_8 = 3$ (8-Malacostraca), $p_9 = 8$ (9-Mollusca) et $p_{10} = 6$ (10-Oligochaeta). Ces données faunistiques sont strictement celles de leur auteur (Friday, 1987; pp. 96 à 97) à l'exclusion près des groupes ne comportant qu'un ou deux taxons .

En écologie, pour décrire les données faunistiques, plusieurs questions sont soulevées parmi lesquelles celle de la valeur typologique des groupes faunistiques posée par les données présentées ci-dessus. Cette question consiste à quantifier chaque groupe faunistique pour refléter sa redondance à exprimer une information biologique produite par une partie ou l'ensemble des groupes. D'une autre manière, si on choisit de ne garder qu'une partie de l'ensemble des groupes, on cherche à estimer le degré de perte en terme d'information biologique généré par un tel choix.

La démarche pour résoudre la question de la valeur typologique peut consister tout d'abord à définir une information biologique commune à partir de l'ensemble des groupes, et ensuite à ordonner quantitativement ces groupes dans le but de refléter leur capacité à reproduire cette information biologique commune. Plus exactement, l'objectif

se traduit par chercher à caractériser la capacité de chaque groupe à reproduire l'état typologique des stations éventuellement induit partiellement ou totalement, par tout ou partie de l'ensemble des groupes.

ACT-STATIS par son utilisation du compromis et ACOM par son utilisation des variables auxilliaires sont deux analyses adaptées à l'exploration de la question posée. Mais le fait que ces deux analyses ont des fonctionnements théoriques différents permet de s'interroger naturellement sur leur comportement dans l'exploration de la question posée ci-dessus. C'est l'objectif ici, il privilégie un point de vue multivarié et ne fait donc pas de l'interprétation des données une priorité. Plus exactement :

Dans un premier temps, on met en œuvre chacune des deux analyses sur la base d'illustrations afin d'observer comment chacune d'elles génère ses moyens d'inteprétation.

Dans un second temps, on compare l'analyse du compromis d'ACT-STATIS et l'analyse de synthèse fournie par l'ACOM. On compare ainsi deux états typologiques communs aux études, chacun étant fourni par l'une des deux analyses. L'objectif est de répondre à l'alternative suivante :
face à la question de la valeur typologique des groupes faunistiques qui implique la définition d'un état typologique commun, les deux états typologiques définis par les deux analyses sont-ils relativement semblables ou au contraire, chacune des deux analyses définit-elle un état typologique commun qui lui est propre?

3.2. Illustrations

Les données se présentent sous forme de 10 tableaux regroupant un total de 91 taxons et 16 étangs. Il y a respectivement 11, 7, 13, 4, 13, 22, 4, 3, 8 et 6 taxons par groupe. Chaque tableau est identifié par le groupe de taxons utilisés ou un chiffre associé (voir ci-dessus). Les 16 étangs sont étiquetés par les lettres Q P R J E C D K B A G M L F H N. Les taxons de chaque groupe sont étiquetés par une lettre munie d'un indice. La lettre est déterminée par le numéro du groupe selon l'ordre lexicographique. Cette lettre étant déterminée, elle est indexée de 1 à p_k, p_k étant le nombre des taxons du groupe. Par exemple, le groupe 8 - Malacostraca correspond à la lettre H (8-ème lettre dans l'ordre lexicographique) ce groupe contient 3 taxons ($p_8 = 3$), ils seront étiquetés par H_1, H_2 et H_3.

L'Analyse en Composantes Principales centrée par variables (espèces) est utilisée sur les 10 tableaux séparément. Ainsi, à chaque groupe est associée une étude $\left(\mathbf{X}_k, \mathbf{I}_{p_k}, \frac{1}{16}\mathbf{I}_{16}\right)$. On note au passage que dans ADE-4, chacune des deux analyses (ACOM & ACT-STATIS) est programmée pour l'assemblage de triplets statistiques initiaux de nature arbitraire (variables quantitatives, qualitatives ou distributionnelles).
Une question peut se poser, celle qui touche la pondération des études. Elle consiste à éviter qu'une étude de forte inertie (tant par les normes des variables que par leur

nombre) ne prenne une importance disproportionnée dans la définition d'une part, du compromis s'il s'agit de l'analyse ACT-STATIS, d'autre part, des variables auxiliaires, s'il s'agit de l'ACOM. Les solutions envisagées dans ADE-4 consistent en :

— trois options pour l'ACOM qui sont respectivement, la pondération uniforme, la pondération par l'inverse de l'inertie totale ou la pondération par l'inverse de la première valeur propre de l'analyse séparée. La seconde a été choisie par défaut et a été utilisée ici dans l'illustration par l'ACOM.

— deux options pour ACT-STATIS sont classiquement proposées qui sont respectivement, l'analyse sur la base de la matrice des corrélations vectorielles ou l'analyse sur la base de la matrice des variances-covariances vectorielles. La première a été choisie par défaut et a été utilisée dans l'illustration par ACT-STATIS.

Ces considérations concernant la pondération des études étant prises, elles définissent d'une manière unique l'étude horizontale $\left(\mathbf{X}, \mathbf{I}_{\Sigma p_k}, \frac{1}{16}\mathbf{I}_{nK}, \Pi\right)$.

3.2.1. Utilisation d'ACT-STATIS

ACT-STATIS (Lavit et Coll., 1994) s'appuie sur le passage du triplet $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ à son opérateur d'inertie $\mathbf{W}_k\mathbf{D}$ élément de l'espace euclidien des opérateurs $\mathbf{D}$-symétriques de $\mathbb{R}^n$ muni du produit scalaire de Hilbert-Schmidt. L'analyse du nuage d'opérateurs, dans ce cas normé, définit un système d'axes d'inertie (interstructure, Figure 1).

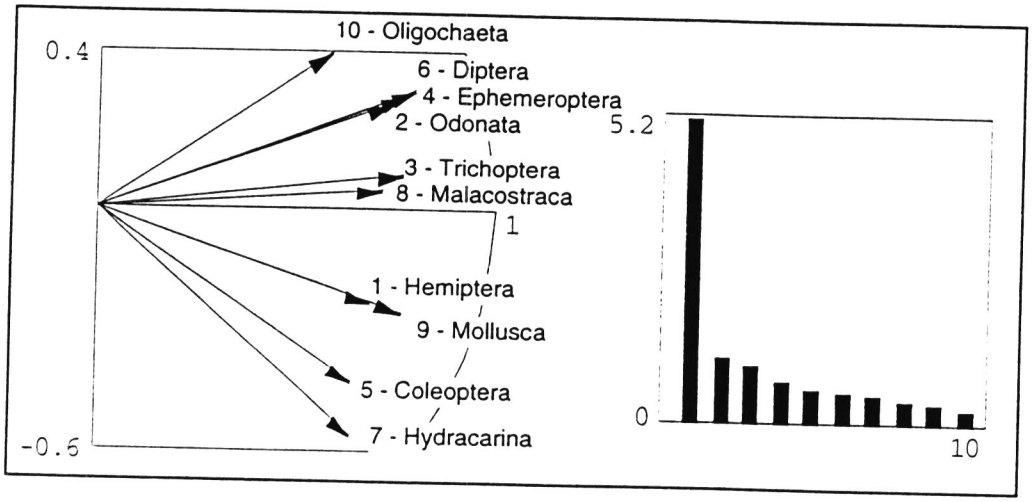


Figure1. Interstructure d'ACT- STATIS, à gauche, les valeurs propres de la matrice des coefficients Rv. A droite, l'image euclidienne des 10 opérateurs d'inertie normés.

Le tableau suivant donne pour chaque étude, le nombre de variables (Nvar), le carré de la norme de Hilbert-Schmidt des opérateurs associés (HS), la matrice de corrélation vectorielle (RV) entre études, (les éléments de la matrice sont multipliés par 1000), et le poids de l'étude dans la constitution du compromis (poids). Par les poids (dernière

colonne du tableau), on a une vision équilibrée du rôle des études dans la définition du compromis, le groupe 4 jouant un grand rôle par la norme.

k	N var	HS	RV										Poids
1	11	0,107	1000										0,305
2	7	0,232	442	1000									0,326
3	13	0,179	527	509	1000								0,341
4	4	1,507	439	571	543	1000							0,341
5	13	0,043	498	406	463	305	1000						0,285
6	22	0,296	428	640	614	624	451	1000					0,353
7	4	0,086	502	307	433	316	473	338	1000				0,284
8	3	0,643	347	432	510	596	310	494	418	1000			0,318
9	8	0,362	407	491	424	610	496	528	594	638	1000		0,339
10	6	0,536	389	410	442	407	254	514	284	335	242	1000	0,258

La diagonalisation de l'opérateur compromis (graphe des valeurs propres dans la figure B) donne un système de vecteurs propres de $\mathbb{R}^n$ (n = 16). On définit un plan par les deux premiers vecteurs propres. Si $\mathbf{z}_j$ est le *j-ème* vecteur propre **D**-normé de l'opérateur compromis **WD**, σ_j *la j-ème* valeur propre associée, les individus compromis sont positionnés par les coordonnées du vecteur $\sqrt{\sigma_j}\mathbf{z}_j$ (figure 2).

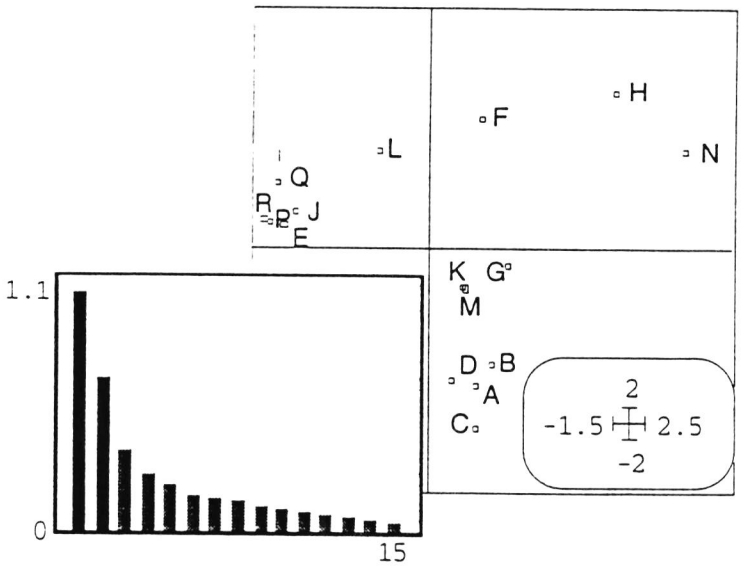


Figure 2. Graphe des valeurs propres et plan compromis des stations.

Ce plan compromis des stations (individus) s'accompagne de deux représentations :

— une première représentation vise l'expression de l'état typologique des individus compromis par les variables de chaque étude $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$. Cette représentation est liée à celle des individus compromis par les règles habituelles dans l'analyse d'un triplet. Plus explicitement, les vecteurs propres $\mathbf{z}_j$ sont situés dans l'espace euclidien $\mathbb{R}^n$ muni de la métrique **D**, de même que les K nuages des variables des K études $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$. On obtient une représentation simultanée des variables de chaque étude, par projection du

nuage des variables sur chacun des vecteurs $\mathbf{z}_j$, ce qui revient pour l'étude $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ à représenter les variables par les coordonnées du vecteur :

$$\mathbf{X}_k^t \mathbf{D} \mathbf{z}_j .$$

— Une seconde représentation vise les relations entre les analyses séparées et l'analyse du compromis (analyse commune). Elle a pour vocation de préciser ce qu'est l'analyse propre de chaque étude par rapport à l'analyse du compromis. On note que cette représentation a été proposée par Place (1980) (voir aussi Glaçon, pp. 37). Elle consiste à projeter les composantes principales (coordonnées de synthèse des individus) de chacune des K études sur le sous-espace défini par les vecteurs compromis $\mathbf{z}_j$.

Pour chacune des K études, se projettent sur le plan défini par l'analyse du compromis, son nuage des variables et ses deux premières composantes principales (figure 3).

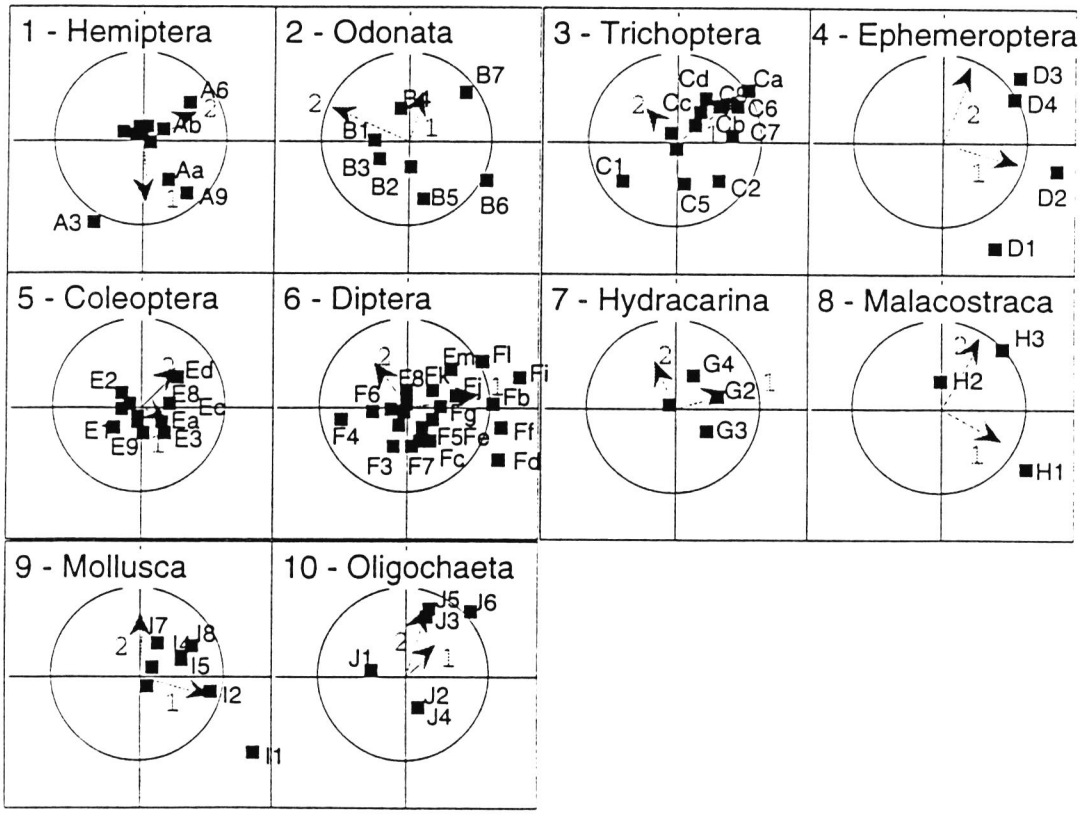


Figure 3. Projection des nuages des variables de chaque étude sur le plan compromis, ainsi que ses deux premières composantes principales.

La notion des trajectoires dans ACT-STATIS vise une représentation des stations par chacune des K études. Plus exactement, pour un individu donné, l'objectif consiste à établir une discussion sur la base d'une comparaison de ses multiples positions pour détecter des variations inter-études de cet individu. Les multiples positions de cet

individu sont définies par les différentes valeurs (positions) prises par les variables de chaque étude.
La solution proposée par les auteurs de l'analyse consiste à représenter ces trajectoires sur l'axe j relativement à une étude k par les coordonnées du vecteur :

$$\pi_k \mathbf{W}_k \mathbf{D}\mathbf{z}_j .$$

On note ici deux propriétés de cette représentation (Glaçon, pp.40) :

— La première est la reconstitution des produits scalaires entre opérateurs. C'est sur la base de celle-ci que Chessel & Dolédec (1996, fascicule 6) proposent un indice de la valeur typologique des groupes.

— La deuxième propriété consiste à situer un individu compromis i au barycentre de ses représentations pour chaque étude, pondérées par les poids β_k qui ont servi à la construction du compromis (Lavit pp. 99) (figure 2). La figure 4 illustre cette deuxième propriété, ainsi les barycentres des représentations du même individu redonnent sa position dans le compromis.

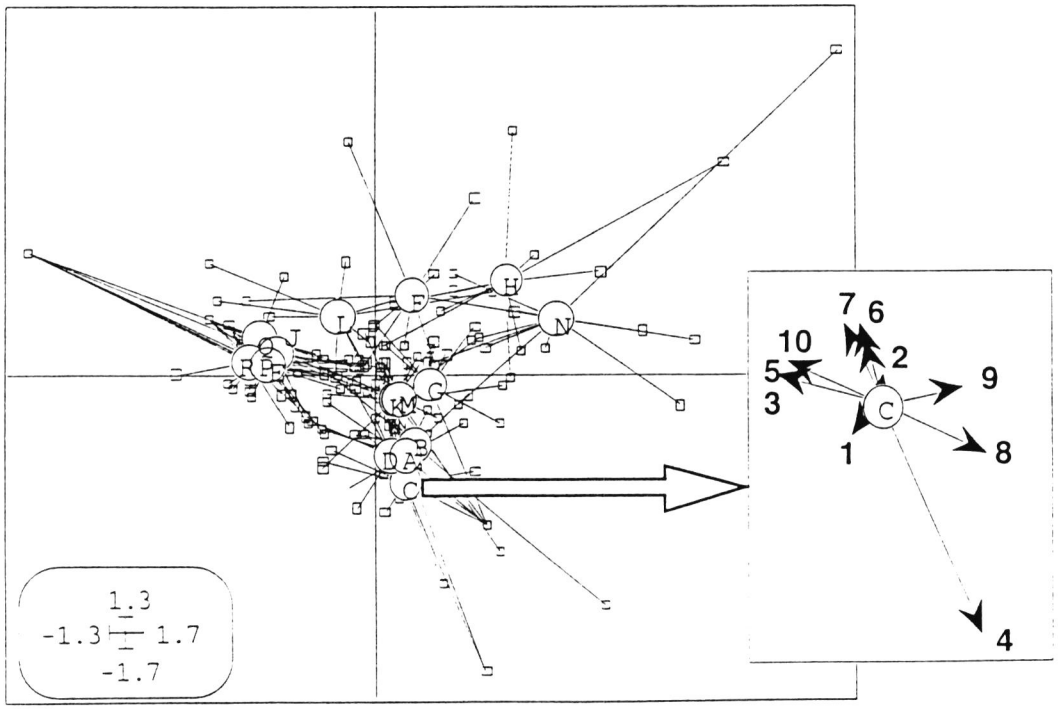


Figure 4. Représentation multiple des stations par la technique des trajectoires. En premier plan le detail pour la station C.

En conclusion de l'utilisation d'ACT-STATIS, on peut faire deux remarques, la première concerne les représentations simultanées des individus, la deuxième concerne la fonction de l'étape interstructure dans l'analyse.

Premièrement, l'objectif abordé par les trajectoires présente des difficultés, c'est ce que relate par exemple Glaçon : *"Il parait difficile de comparer pour un individu donné, les différentes valeurs que prennent les variables pour cet individu "*.

La proposition de Place (1980) de positionner les individus de chaque étude par la prévision des vecteurs propres du compromis par régression multiple sur chaque étude, ne peut pas résoudre la question car cette prévision est parfaite pour les études comptant plus de variables que d'individus.

L'absence de critère visant la représentation simultanée des nuages initiaux est alors un point faible dans ACT-STATIS. Géométriquement, le passage aux nuages des opérateurs a pour prix une impossibilité de revenir d'une manière naturelle et pertinente aux nuages initiaux. En effet, les composantes des vecteurs propres du compromis sont des codes numériques normés et non corrélés des stations, mais elles ne correspondent à aucune approche géométrique, même si leur pertinence globale ne fait pas de doute. Leur obtention n'implique pas d'une manière directe les nuages initiaux, ni par projection, ni par représentation euclidienne. Cela renforce l'idée de l'existence d'une différence entre une étude et son opérateur d'inertie.

Deuxièmement, en laissant faire l'interstructure (état typologique des opérateurs, Figure 1) avant la description du compromis (moyenne d'opérateurs) ACT-STATIS indique d'abord s'il est légitime de faire une moyenne d'études (Lavit, 1988, pp.96). Cette interrogation sur la validité du compromis n'est pas fondée, et surtout elle pose deux problèmes :

— le premier est d'ordre conceptuel : s'interroger sur la validité du compromis c'est s'interroger sur la notion de la moyenne. Il existe un opérateur moyen de K opérateurs d'inertie comme il existe une valeur moyenne de K mesures. Que l'on considère ensuite la dispersion autour de cette moyenne est un problème évidemment pertinent, mais qui ne remet pas en cause la notion même de la moyenne dont le rôle principal est de servir de référence pour établir une discussion typologique.

— Le deuxième est d'ordre pratique : présenter l'étape interstructure (typologie d'opérateurs) avant la description du compromis (intrastructure), laisse penser que l'analyse propose une approche typologique (interstructure) avant la notion plus simple de la moyenne (compromis), ce qui peut être une source de difficultés pour l'utilisateur.

L'étape Interstructure dans ACT-STATIS a alors une fonction ambiguë dans l'analyse, si elle ne contribue pas à cette ambiguité vis-à-vis de la fonction des deux autres étapes. En effet, par cette étape, même si l'analyse offre une lecture typologique entres études, d'une part, elle ne propose aucun moyen pour la décrire, et d'autre part, elle contredit son objectif en tant qu'analyse cherchant un état typologique reproductible d'une étude à une autre et défini par l'analyse du compromis.

3.2.2. Utilisation de l'ACOM

Dans une boucle, l'algorithme de l'ACOM est une décomposition en valeurs singulières utilisée au rang 1. Ainsi au pas s, pour chaque triplet de départ $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$, l'ACOM définit un système de s axes $\mathbf{Q}_k$-orthonormés dans $\mathbb{R}^{p_k}$, qu'on note $\mathbf{U}_k$; on les appelle "axes de co-inertie". On obtient alors K systèmes d'axes de co-inertie situés dans K espaces différents. Sur ces axes de co-inertie correspondent deux représentations :

— la première est obtenue par projection du nuage des individus de chaque étude sur le système d'axes correspondant. On obtient K analyses simples coordonnées. On remarquera que l'ACOM opère d'une manière directe dans les K nuages initiaux.

— la deuxième représentation vise la relation entre les analyses coordonnées obtenues et les analyses séparées. Elle consiste à projeter les axes d'inertie (principaux) situés dans $\mathbb{R}^{p_k}$ sur les axes de co-inertie $\mathbf{U}_k$ qui sont situés aussi dans le même espace. Dans ce sens, l'ACOM et ACT-STATIS opèrent différemment.
En effet, pour ACT-STATIS, il était question de projeter les composantes principales des études. D'autre part, cette projection s'effectue sur le même sous-espace, celui-ci étant défini par les vecteurs propres du compromis. Dans l'ACOM, ce sont les axes des analyses séparées qui sont projetés dans des sous-espaces séparés. Il s'agit là des axes de co-inertie.

Ainsi, sur chaque système d'axes de co-inertie, on projette le nuage des individus et les axes principaux de l'étude correspondante.

Comme indiqué dans le tableau suivant, on peut comparer la valeur des systèmes d'inertie et des systèmes de co-inertie par les inerties projetées. On n'a pas de répartition d'inertie globale, comme c'est le cas dans la diagonalisation de l'opérateur compromis, mais une discussion étude par étude de la valeur en terme d'inertie des axes de co-inertie.

k	1	2	3	4	5	6	7	8	9	10
Iner Max	0,241	0,324	0,372	1,005	0,144	0,426	0,195	0,702	0,561	0,536
Inertie	0,165	0,272	0,343	0,959	0,098	0,400	0,177	0,636	0,552	0,377
ProScal 2	0,107	0,231	0,247	0,879	0,068	0,344	0,125	0,499	0,449	0,165
Cos2	0,65	0,85	0,72	0,91	0,69	0,86	0,70	0,78	0,81	0,43

Quelques paramètres numériques associés au pas 1 de l'ACOM. k — *numéro du tableau. Iner Max — Inertie projetée sur le premier axe d'inertie. Inertie — Inertie projetée sur le premier axe de co-inertie. ProScal2 — Carré de la covariance entre la coordonnée sur le premier axe de co-inertie et la première variable auxiliaire. Cos2 — Carré de corrélation correspondant.*

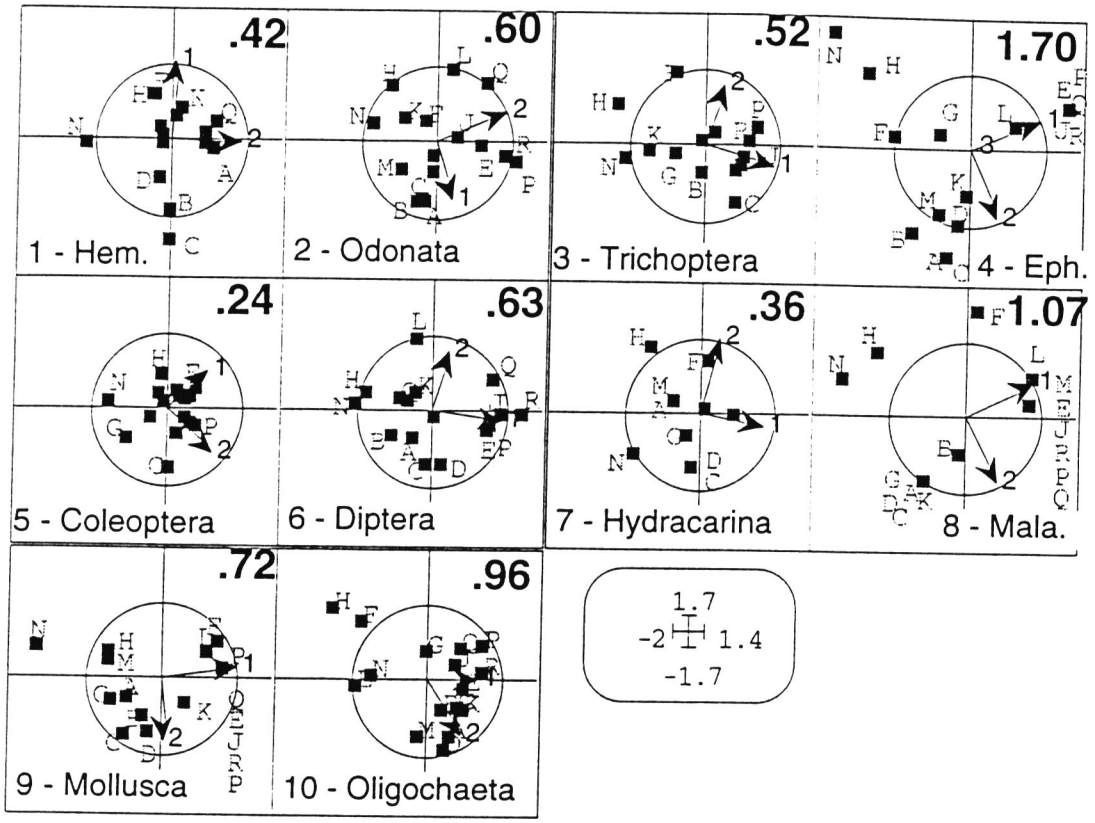


Figure 5. Plans de co-inertie, sur lesquels figurent les représentations des stations ainsi que les axes principaux par étude.

L'ACOM fournit les systèmes d'axes de co-inertie et elle fournit simultanément un système d'axes de variables auxiliaires $\mathbf{V}$ $\mathbf{D}$-orthonormé situés dans l'espace $\mathbb{R}^n$, où sont également situés les nuages des variables des K études. A ce système d'axes de variables auxiliaires correspondent deux représentations :

— la première s'appuie sur le triplet de synthèse considéré par l'ACOM. En effet, l'ACOM considère le tableau de synthèse $\mathbf{S}_{\mathbf{u}}^{r}$ à n lignes et K colonnes (coordonnées des nuages projetés sur un axe de co-inertie). L'entier r étant le pas de l'ACOM supposé fixé, on associe au tableau $\mathbf{S}_{\mathbf{u}}^{r}$ une étude $\left(\mathbf{S}_{\mathbf{u}}^{r}, \Pi, \mathbf{D}\right)$. Une colonne j contient les coordonnées du vecteur $\mathbf{X}_j\mathbf{Q}_j\mathbf{u}_j^r$ qui ont servi dans la représentation des projections du nuage des individus de l'étude $\left(\mathbf{X}_j, \mathbf{Q}_j, \mathbf{D}\right)$ sur l'axe de co-inertie $\mathbf{u}_j^r$. Cette colonne j est aussi une moyenne pondérée des variables du tableau $\mathbf{X}_j$. Comme l'ACOM consiste à trouver la première composante principale $\mathbf{v}^r$ de l'étude de synthèse $\left(\mathbf{S}_{\mathbf{u}}^{r}, \Pi, \mathbf{D}\right)$, une représentation d'un état typologique de synthèse des états typologiques obtenus par projection sur les axes de co-inertie correspondants, s'obtient selon les règles classiques de l'analyse d'un triplet. Ainsi une représentation des variables auxiliaires qui sont des coordonnées non corrélées des stations (voir la proposition dans la section précédente, **(a)**) exprime un état typologique de synthèse associé à une représentation de synthèse

des études. Cette dernière représentation est obtenue après identification de chaque étude à ses coordonnées factorielles de co-inertie.

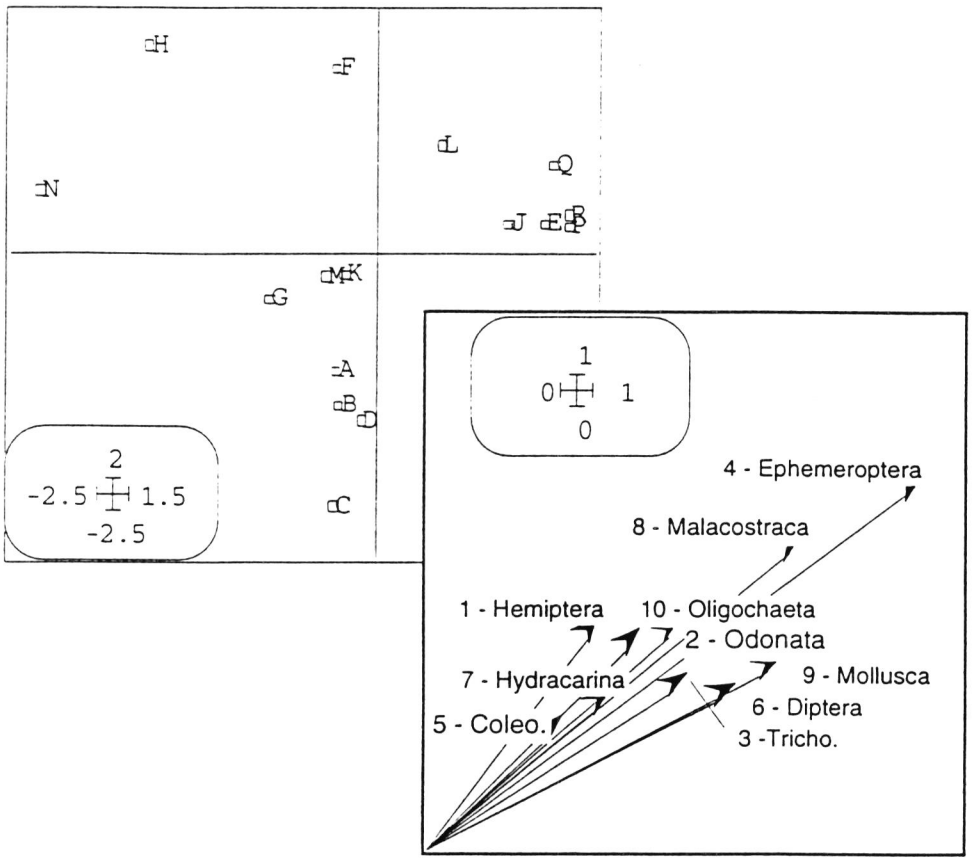


Figure 6. Plan de synthèse des plans de co-inertie et représentation des 10 études

— la deuxième représentation vise l'expression de l'état typologique des stations qui sont les coordonnées des variables auxiliaires par les variables de chaque étude $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$. Cette représentation est liée à celle des individus compromis par les règles habituelles dans l'analyse d'un triplet. Plus explicitement, les vecteurs auxiliaires $\mathbf{v}^r$ sont situés dans l'espace euclidien $\mathbb{R}^n$ muni de la métrique $\mathbf{D}$, de même que les K nuages des variables des K études $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$. On obtient une représentation simultanée des variables de chaque étude par projection du nuage des variables sur chacun des vecteurs $\mathbf{v}^r$, ce qui revient pour l'étude $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ à représenter les variables par les coordonnées du vecteur $\mathbf{X}_k^t \mathbf{D} \mathbf{v}^r$.

Les cartes factorielles de co-inertie dans chaque espace (figure 5) peuvent se superposer après normalisation des coordonnées avec les variables auxiliaires (figure 6) de même rang pour exprimer la part qui revient à la corrélation coordonnée/variable auxiliaire dans la covariance correspondante. En effet, optimiser un carré de covariance entre coordonnée et variable auxiliaire normée, revient à optimiser le produit de la variance de la coordonnée (inertie projetée) par le carré de la corrélation.

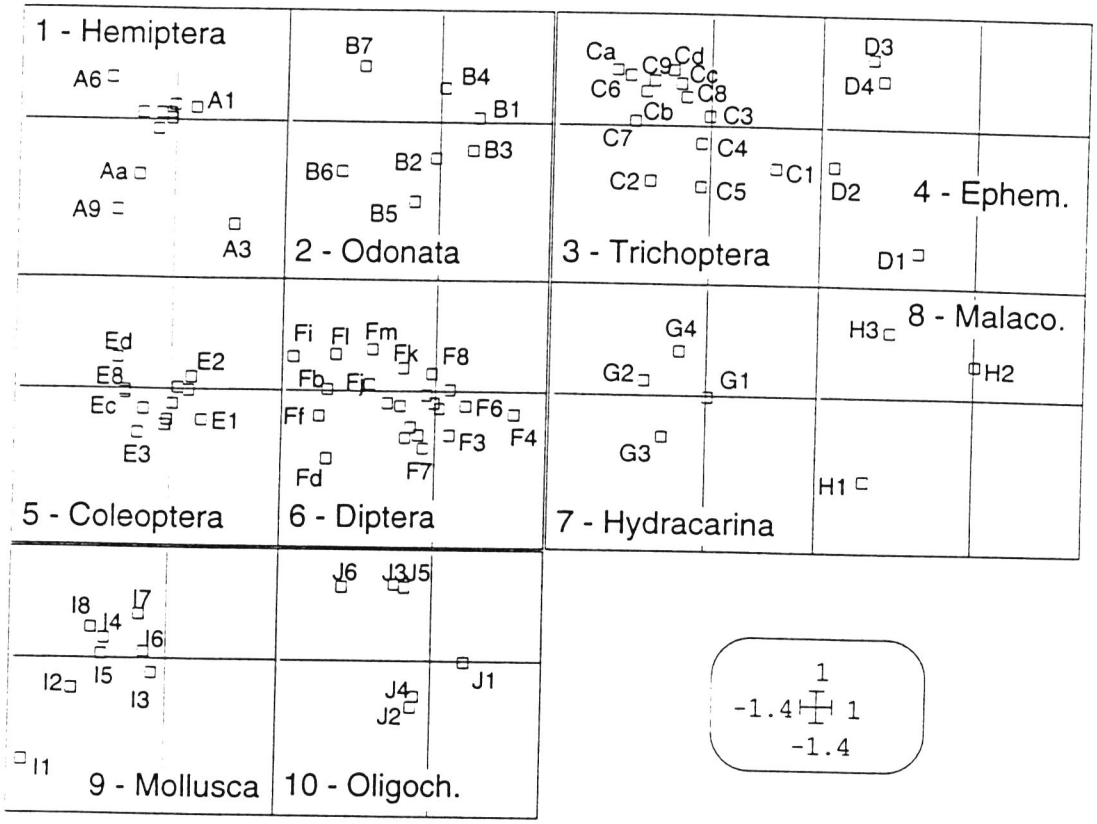


Figure 7. Projection du nuage des variables de chaque étude sur le plan défini par des deux premières variables auxilliaires.

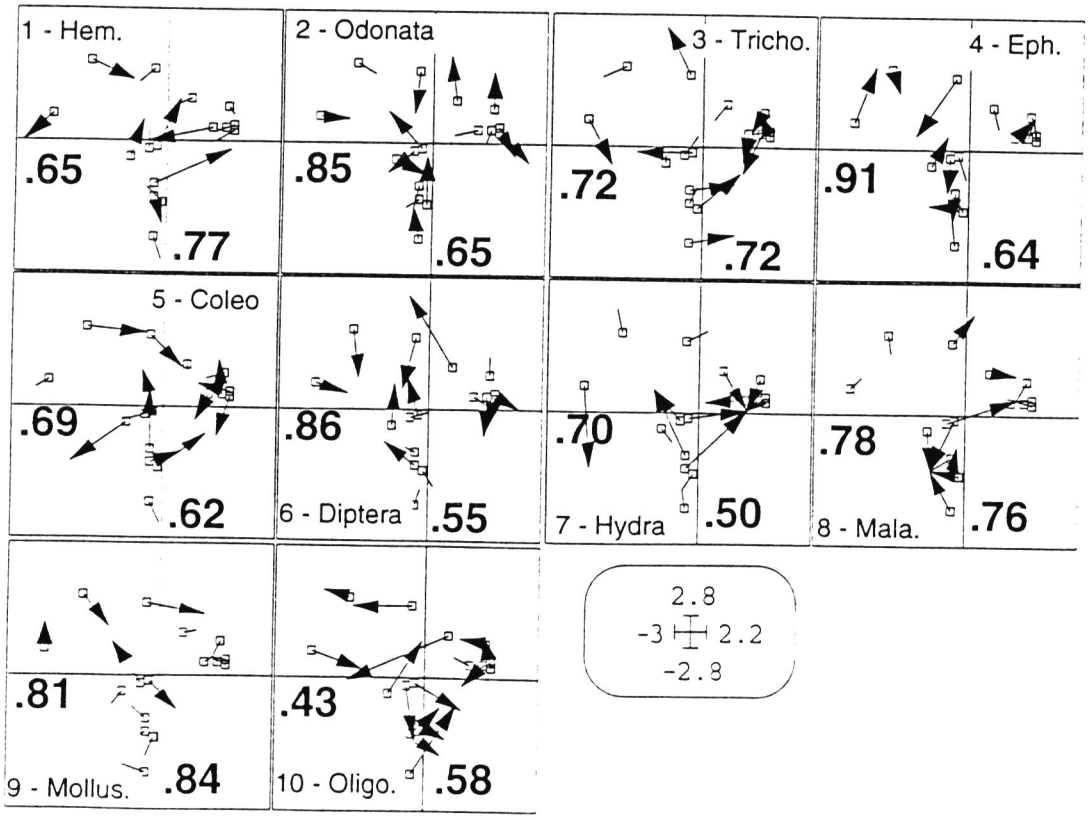


Figure 8. Superposition des plans de co-inertie normalisés sur le plan de synthèse défini par les variables auxiliaires. Les indications numériques correspondent aux carrés de corrélation axe par axe, pour chacune des 10 études.

En conclusion, l'ACOM fait des nuages des triplets initiaux son point de départ et utilise des variables auxilliaires. Ceci lui permet de reconstituer d'une manière fluide et cohérente ses moyens d'interprétation. Ainsi, elle ne pose aucune difficulté particulière d'interprétation pour un utilisateur des pratiques multivariées classiques.

2.2.3. ACT-STATIS, ACOM & état typologique commun

La fonction des deux analyses vis-à-vis de la question de la valeur typologique des groupes vise à définir un état typologique commun, puis à étudier sa reproductibilité par chacun des groupes. La valeur typologique associée à un groupe sera son degré à exprimer ou à reproduire totalement ou partiellement cet état typologique commun.

On ne s'intéresse pas ici à associer à chaque groupe faunistique une valeur numérique qui indiquera sa valeur typologique, à ce sujet on renvoit aux solutions proposées par (Chessel & Dolédec; 1996, fascicules 6 et 7). On s'intéresse plutôt à l'état typologique commun qui est à la base de la définition d'une valeur typologique. Plus exactement, a priori chacune des deux analyses utilise des moyens théoriques propres pour définir un état typologique commun aux études. Ainsi on obtient deux état typologiques communs, chacun étant issu d'une des deux analyses. Une alternative se présente alors :
ces deux états typologiques communs sont relativement les mêmes, ou au contraire, ces deux états typologiques présentent des différences importantes, du fait qu'ils sont issus de deux analyses qui mettent en avant des propriétés théoriques différentes.

La cohérence entre les deux états typologiques moyens des stations (figure 10) souligne l'accord des deux analyses dans leur définition d'un état typologique commun, bien que les deux analyses soient fondées sur des bases théoriques différentes.

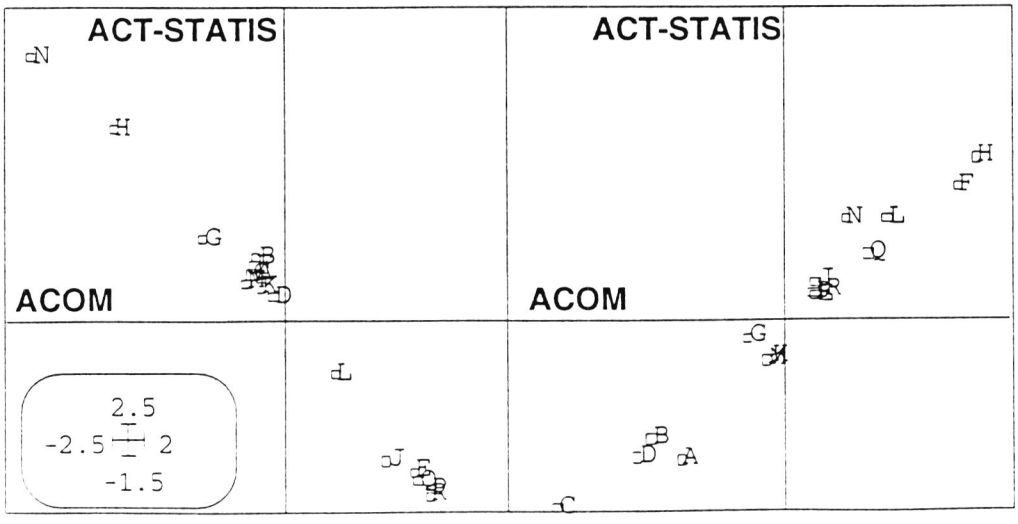


Figure 10. A gauche, représentation de rang 1; à droite, celle de rang 2. Les abscisses sont les coordonnées (stations) des variables auxilliaires de l'ACOM, les ordonnées sont les coordonnées des stations compromis dans ACT-STATIS.

3. Conclusion

En conclusion de ce chapitre, vis-à-vis de la question pratique posée, aucune des deux analyses n'a apporté d'indication qui aurait échappé à l'autre. Aucune non plus, n'a mis en évidence de contradiction avec l'autre. Chaque analyse est différente par sa base théorique, mais toutes deux partagent un même objectif qui est la définition d'un état typologique de référence, et la discussion de sa reproductibilité par chacune des études. Leur différences théoriques génèrent des différences de moyens d'interprétation et d'exploration.

Par les deux analyses considérées ici, à savoir ACT-STATIS et ACOM, on a montré qu'il est possible d'atteindre un objectif pratique inscrit dans les données par des analyses à fondements théoriques différents.

L'une des conséquences de cette conclusion est celle qui s'inscrit dans la logique de la discussion de la première partie. Plus exactement, il s'agit de la perspective de l'étude des liens entre les analyses à trois entrées non pas sur la base de leur fonctionnement théorique, mais sur la base de leur capacité à résoudre ou non une même question pratique.

Chapitre II.2

Analyse simultanée de K couples de tableaux par l'analyse STATICO

0. Introduction

Il s'agit maintenant d'étudier la deuxième question, celle qui consiste à généraliser l'analyse de co-inertie de deux tableaux, à l'analyse simultanée de K couples de tableaux. Ce chapitre qui propose une analyse de la description simultanée de K couples de tableaux. Dans son fondement, l'analyse que l'on propose combine deux composantes issues de deux analyses qui sont l'analyse de co-inertie et de l'analyse ACT-STATIS. On l'appelle STATICO (**STATIS** et **CO**-inertie). Un exemple d'illustration sera présenté sur des données écologiques.

On considère K couples de triplets $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$ $(1 \le k \le K)$, $\mathbf{X}_k$ est un tableau $n_k \times p$ dont les lignes sont des observations de n_k individus et dont les colonnes sont les mesures de p variables. $\mathbf{Y}_k$ est un tableau $n_k \times q$ portant sur les mêmes n_k individus que le tableau $\mathbf{X}_k$ et les colonnes de $\mathbf{Y}_k$ sont les mesures de q variables, pas nécessairement celles de $\mathbf{X}_k$. On obtient ainsi deux études K-tableaux verticales. La première étude K-tableaux $(\mathbf{X}, \mathbf{Q}, \mathbf{D}, \Pi)$ est formée à partir des K triplets $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$ $(1 \le k \le K)$. π_k est un poids positif attribué *a priori* au tableau $\mathbf{X}_k$. La deuxième étude K-tableaux $\left(\mathbf{Y}, \mathbf{R}, \mathbf{D}, \tilde{\Pi}\right)$ est formée à partir des K triplets statistiques $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$. $\tilde{\pi}_k$ est un poids positif attribué *a priori* au tableau $\mathbf{Y}_k$.

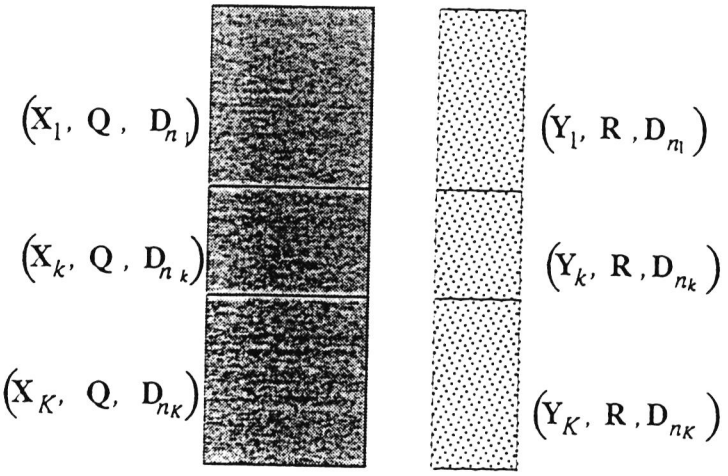


Figure 1. Deux études K-tableaux verticales

En présence de deux études K-tableaux verticales ou de K couples de triplets définis comme ci-dessus et selon les objectifs poursuivis, plusieurs possibilités d'exploration par des analyses spécifiques sont envisageables. On cite quelques unes de ces possibilités :

— pour chacun des K triplets $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$, respectivement $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$, un état typologique (individus, variables) est obtenu par l'analyse de ce triplet. Cette analyse peut correspondre, selon l'objectif poursuivi, à des analyses élémentaires telles que : l'analyse

en composantes principales (ACP) centrée, doublement centrée ou normée, des analyses des correspondances multiples (ACM) qui utilisent les indicatrices des classes et les schémas de Tenenhaus & Young (1985), ou des analyses des correspondances floues (ACF, Chevenet & Coll., 1994).

— Verticalement, l'analyse des liens entre les K relations (K états typologiques des variables) définies par l'analyse de l'étude K-tableaux verticale $(\mathbf{X}, \mathbf{Q}, \mathbf{D}, \Pi)$ (ou $\left(\mathbf{Y}, \mathbf{R}, \mathbf{D}, \tilde{\Pi}\right)$) ou l'analyse simultanée des K triplets $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$, respectivement $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$, peut être effectuée par une analyse adaptée. A titre d'exemple, on cite ACT-STATIS (Lavit & Coll., 1994), ACOM (Chessel & Hanafi, 1996) ou SCA (Kiers & Ten Berge 1994). Une présentation des deux premières analyses sur un exemple se trouve dans le chapitre II.1 de cette partie. Pour d'autres analyses possibles, on peut se référer au chapitre I.1 (Description de l'univers) de la première partie.

-- Horizontalement, pour chacun des K couples de triplets $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$, une analyse de la relation existant entre les deux groupes de variables définis par les deux triplets $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$, $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$ peut être effectuée par l'analyse de co-inertie (Chessel & Mercier, 1993) qui permet d'établir un co-état typologique d'individus entre les deux états typologiques de chacun de ces deux triplets de départ. Ainsi, en effectuant les K analyses de co-inertie sur les K couples de triplets $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$ $(1 \le k \le K)$, on peut établir K co-états typologiques d'individus, chacun étant défini par l'un des K couples de triplets $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$ $(1 \le k \le K)$

De même que la présence de K états typologiques définis par K triplets amène à s'interroger sur les liens qui existent entre ceux-ci, la présence de K co-états typologiques des K couples de triplets $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$, chacun étant exploré par l'analyse de co-inertie des deux triplets $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$, $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$, amène naturellement à se poser la question de leurs liens éventuels. Cette question est fréquemment évoquée dans le cadre des enquêtes écologiques à composantes temporelles avec la question centrale du couplage en écologie, comme on l'a expliqué dans l'introduction de cette partie.

L'objectif multivarié associé à cette question consiste à généraliser la question de l'étude des liens entre K états typologiques à celle du lien entre K co-états typologiques. Il constitue le contenu du présent chapitre.

La solution que l'on propose ici consiste en une analyse qu'on appelle STATICO (pour STATIs et CO-inertie). Elle se définit comme une combinaison de deux composantes, chacune d'entre elles étant issue d'une analyse propre. La première

composante est issue de l'analyse de co-inertie de deux triplets. Elle consiste en un passage, celui d'un couple $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$ au triplet croisé $(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{Q}, \mathbf{R})$ dont l'analyse définit l'analyse de co-inertie. Cette composante permet essentiellement de revenir au contexte d'une analyse simultanée de K triplets et de considérer le nuage des tableaux croisés.

Intervient alors la deuxième composante, celle issue d'ACT-STATIS. Elle est davantage liée à une stratégie d'analyse qu'à une analyse spécifique des K triplets croisés $(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{Q}, \mathbf{R})$.

STATICO utilise la stratégie en trois étapes d'ACT-STATIS. On propose cependant, de modifier légèrement la présentation classique des trois étapes, ceci ayant l'avantage de bien préciser la fonction de STATICO.

Dans la première section, sera présentée la composante issue de l'analyse de co-inertie, dans la seconde section sera présentée celle issue d'ACT-STATIS. Ces deux composantes constituent l'idée théorique centrale de STATICO et la définissent. Dans la troisième section, on illustrera sur la base d'un exemple une mise en œuvre de cette nouvelle analyse.

1. Composante de la co-inertie dans STATICO

La composante issue de l'analyse de co-inertie consiste en un passage du couple de triplets $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$ au triplet croisé $(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{Q}, \mathbf{R})$. On note que l'analyse du triplet croisé $(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{Q}, \mathbf{R})$ définit l'analyse de co-inertie. Ce passage a deux fonctions dans la définition de STATICO :

— La première est liée à l'objectif qui consiste à analyser simultanément les K couples $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$. Le passage au triplet croisé reflète une solution à la question de la prise en compte effective de la relation définie par le couple $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$.
En effet, toute l'information concernant les liens existant entre les deux triplets $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$ $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$ est prise en compte par la considération du triplet $(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{Q}, \mathbf{R})$. On privilégie alors le lien entre les deux triplets qui est l'objectif.

— La deuxième est liée aux analyses simultanées de K triplets. En effet, par ce passage, on ramène la question de l'analyse simultanée des K couples $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$ à la logique d'une analyse simultanée de K triplets, dans ce cas, les triplets croisés $(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{Q}, \mathbf{R})$ $(1 \leq k \leq K)$.

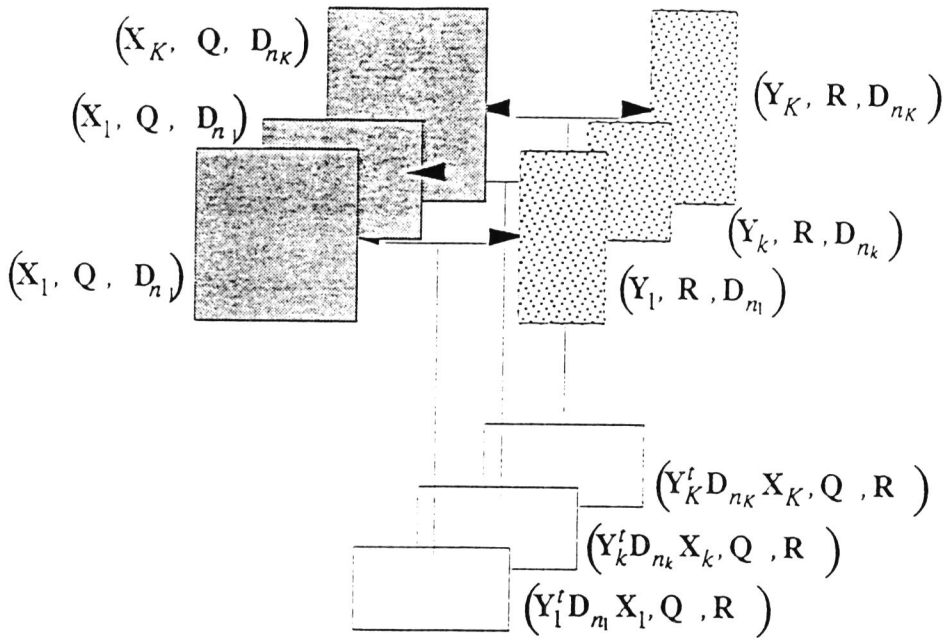


Figure 2. Composante de Co-inertie

Ce retour à la logique d'une analyse simultanée de K triplets peut s'effectuer selon deux points de vues.

Le premier est celui d'une analyse multitableaux. En effet, les K triplets croisés ont les mêmes lignes et les mêmes colonnes; ils impliquent la formation d'un multitableaux.

Le second point de vue est celui d'une analyse K-tableaux. Il s'agit là, de considérer le passage aux triplets croisés, démarche identique à celle qui s'effectue dans les analyses spécifiquement K-tableaux comme : ACT-STATIS (Lavit, 1988; Lavit & Coll., 1994), INDSCAL (Caroll & Chang, 1970) et d'autres (voir Chapitre I.1 : description de l'univers). Pour analyser K triplets $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$, ces analyses effectuent le passage de chacun des K triplets à son opérateur d'inertie.

Ce qui est au cœur de la distinction de ces deux points de vues, ce sont les deux possibilités géométriques qui sont à la base des analyses multitableaux. En effet, selon qu'on choisit une analyse gérée par l'une ou l'autre des deux possibilités géométriques, théoriquement, on n'effectue pas la même opération. Plus exactement, selon le premier point de vue, on peut choisir d'analyser le multitableaux formé à partir des K triplets croisés, par une analyse du deuxième sous-univers; ainsi, on combine deux fonctionnements géométriques. Le second point de vue amène implicitement à considérer le nuage des tableaux croisés. Le choix d'une analyse du premier sous-univers permet de confondre les deux points de vue, car en analysant un multitableaux ou en remplaçant le passage des opérateurs par les tableaux croisés, les deux logiques se confondent.

Théoriquement, d'une part la combinaison de deux fonctionnements géométriques ne permet pas d'attribuer à STATICO un statut dans les diverses approches unificatrices

existantes, d'autre part, le choix du fonctionnement géométrique qui définit le premier sous-univers implique que les deux points de vue évoqués plus haut se confondent.

C'est pourquoi, **STATICO**, selon le principe commun qui gère le premier sous univers, considère directement le nuage des tableaux croisés et le plonge dans un espace vectoriel euclidien approprié muni intrinsèquement du produit scalaire d'Hilbert-Schmidt.

Ainsi, pour k fixé, si on note $\mathbf{Z}_k = \mathbf{Y}_k^t \mathbf{D}_{n_k} \mathbf{X}_k$, le produit scalaire entre $\mathbf{Z}_k$ et $\mathbf{Z}_j$ est défini comme suit :

$$\left(\mathbf{Z}_k \middle| \mathbf{Z}_j\right)_{HS} = \text{Trace}\left(\mathbf{Z}_k^t \mathbf{R} \mathbf{Z}_j \mathbf{Q}\right)$$

Il en découle que la norme $\mathbf{Z}_k$ associée à ce produit scalaire de est définie comme suit :

$$\|\mathbf{Z}_k\|_{HS}^2 = Trace\left(\mathbf{X}_k \mathbf{Q} \mathbf{X}_k^t \mathbf{D}_{n_k} \mathbf{Y}_k \mathbf{R} \mathbf{Y}_k^t \mathbf{D}_{n_k}\right) = Trace\left(\mathbf{W}_\mathbf{X} \mathbf{D}_{n_k} \mathbf{W}_\mathbf{Y} \mathbf{D}_{n_k}\right)$$

$\mathbf{W}_\mathbf{X}$, $\mathbf{W}_\mathbf{Y}$ désignent respectivement l'opérateur d'inertie associé à $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$, $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$

La norme, ici la covariance vectorielle (Escoufier, 1973) entre les deux triplets $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$, $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$, mesure l'adéquation des deux états typologiques des individus des deux triplets. Le produit scalaire mesure l'adéquation des deux co-états typologiques définis par deux couples de triplets. Le produit scalaire permet alors d'étendre la notion de l'adéquation entre deux triplets à la notion de l'adéquation entre deux couples de triplets, ce qui va dans le sens de l'objectif qui se veut de privilégier la relation entre les couples de triplets.

Par la composante de co-inertie, STATICO effectue un passage et choisit un fonctionnement géométrique propre, celui qui consiste à considérer le nuage des tableaux croisés.

2. Composante d'ACT STATIS dans STATICO

La composante d'ACT-STATIS dans la définition de STATICO consiste en une stratégie d'analyse simultanée de K triplets, c'est celle d'ACT-STATIS qui est classiquement présentée en trois étapes (Interstructure -Compromis -Intrastructure). On adopte pour STATICO le principe de cette stratégie mais on propose de modifier légèrement sa présentation. Les raisons de cette modification ont été évoquées lors de la présentation d'ACT-STATIS, notamment la fonction ambiguë de l'étape interstructure. STATICO a une stratégie en trois étapes : définition d'une moyenne, analyse de la moyenne, stabilité de l'analyse moyenne.

La présentation de la composante de co-inertie dans STATICO amène à considérer le nuage des tableaux croisés, qui est considéré dans un espace vectoriel muni d'un produit scalaire comme ci-dessus. **STATICO** opère sur ce nuage selon un schéma en trois étapes qui sont les suivantes.

2.1. Définition d'une moyenne.

On appelle une combinaison linéaire des K triplets croisés $\left(\mathbf{Y}_k^t\mathbf{D}_k\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D}\right)$, un triplet $(\mathbf{Z}, \mathbf{R}, \mathbf{Q})$, avec $\mathbf{Z}$ une matrice obtenue comme combinaison linéaire des tableaux croisés $\mathbf{Y}_k^t\mathbf{D}_k\mathbf{X}_k$. Autrement dit, on a une matrice $\mathbf{Z}$ telle que :

$$\mathbf{Z} = \sum_{k=1}^{K} \alpha_k \mathbf{Y}_k^t\mathbf{D}_k\mathbf{X}_k$$

α_k $(1 \le k \le K)$ étant des nombres réels.

L'objectif de cette étape est la définition d'un triplet moyen. Celle-ci passe par la recherche parmi tous les triplets $(\mathbf{Z}, \mathbf{R}, \mathbf{Q})$, combinaisons linéaires des triplets croisés $\left(\mathbf{Y}_k^t\mathbf{D}_k\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D}\right)(1 \le k \le K)$, celui qui fournit la meilleure analyse relativement à la résolution du problème suivant :

$$\text{Maximiser } \|\mathbf{Z}\|_{HS}^2 = \text{Trace}\left(\mathbf{Z}^t\mathbf{R}\mathbf{Z}\mathbf{Q}\right) \text{ sous la contrainte } \sum_{k=1}^{K} \alpha_k^2 = 1$$

Cette résolution découle de l'analyse en composantes principales du nuage des tableaux croisés $\left(\mathbf{Y}_k^t\mathbf{D}_k\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D}\right)(1 \le k \le K)$. Elle implique la diagonalisation de la matrice des produits scalaires entre triplets croisés $\left(\mathbf{Y}_k^t\mathbf{D}_k\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D}\right)(1 \le k \le K)$. Les nombres α_k sont alors les composantes du premier vecteur propre normé de la matrice des produits scalaires. Les coordonnées α_k de ce vecteur propre sont tous de même signe, on les considère tous positifs.

Cette opération de diagonalisation de la matrice des produits scalaires est classiquement désignée comme l'**inter-structure** dans la stratégie d'ACT-STATIS. Elle a pour objectif une représentation globale des tableaux croisés. Dans STATICO, cette présentation est abandonnée dans le but de préciser la fonction de l'analyse qui ne correspond pas ici à étudier l'état typologique défini par le nuage des tableaux croisés. Pour plus de détails, on peut se référer au chapitre II.1 (pp. 138-139).

2.2. Analyse de la moyenne.

L'analyse de la moyenne est l'analyse du triplet moyen $\left(\sum_{k=1}^{K} \alpha_k \mathbf{Z}_k, \mathbf{D}_p, \mathbf{D}_q\right)$ défini dans l'étape précédente. Analyser ce triplet moyen est l'objectif de cette étape. Selon les règles classiques, l'analyse du triplet moyen fournit une suite de valeurs propres classiquement désignées comme valeurs de co-inertie (pseudo-valeurs propres) et deux

systèmes d'axes, l'un dans $\mathbb{R}^q$ $\mathbf{R}$-orthonormé qu'on note par $\mathbf{U}$, l'autre dans $\mathbb{R}^p$ $\mathbf{Q}$-orthonormé qu'on note par $\mathbf{V}$, classiquement désignés par axes de co-inertie. Dans STATICO, en reprenant l'idée conceptuelle d'ACT-STATIS, on suppose que cette analyse du triplet moyen est une analyse de co-inertie entre deux triplets fictifs.

Chaque ligne du tableau moyen reflète une prise en compte du lien moyen de chacune des variables q du K-tableaux $\mathbf{Y}$ (colonne) avec l'ensemble des p variables du K-tableaux $\mathbf{X}$. Il en va de même par transposition, pour les colonnes du tableau moyen.

Sur chacun des deux systèmes d'axes de co-inertie $\mathbf{U}$ et $\mathbf{V}$, on projette les deux nuages (lignes, colonnes).

Le nuage des lignes, respectivement le nuage des colonnes, est situé dans l'espace $\mathbb{R}^p$, respectivement dans l'espace $\mathbb{R}^q$, et se projette sur le système d'axes de co-inertie issu de l'analyse moyenne $\mathbf{U}$ situé dans $\mathbb{R}^q$, respectivement $\mathbf{V}$ situé dans $\mathbb{R}^p$. Selon les règles d'interprétation dans l'analyse d'un triplet, l'association des deux nuages projetés, définit ainsi une relation moyenne et de synthèse entre les deux groupes de variables du K-tableaux $\mathbf{X}$, et le K-tableaux $\mathbf{Y}$.

2.3. Stabilité de l'analyse moyenne.

La moyenne déjà construite (Définition de la moyenne) et analysée (Analyse de la moyenne) permet d'explorer une relation moyenne entre les deux groupes de variables.

L'objectif de cette dernière étape de STATICO est d'établir une discussion sur les réalisations partielles ou totales de cette analyse moyenne par chacun des K triplets croisés $\left(\mathbf{Y}_k^t\mathbf{D}_k\mathbf{X}_k, \mathbf{R}, \mathbf{Q}\right)$. Chacun de ceux-ci exprime une relation propre pertinente ou non, entre les deux groupes de variables, qui peut être explorée par son analyse.

Les éléments constitutifs de cette discussion sont conditionnés par le fonctionnement géométrique de l'analyse. En effet, ils découlent tous des deux systèmes d'axes de co-inertie $\mathbf{U}$ et $\mathbf{V}$ fournis lors de l'analyse de la moyenne. Ces éléments correspondent à deux objectifs :

— le premier objectif consiste à exprimer la relation moyenne définie par l'analyse de co-inertie de chacun des K couples de triplets $\left\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\right\}$ par l'analyse moyenne. L'expression de la relation entre les deux groupes de variables définis par l'analyse de co-inertie de chacun des K couples de triplets $\left\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\right\}$ par l'analyse moyenne, s'effectue selon la logique des trajectoires au sens de Place (1980).

En effet, l'analyse de chacun des K triplets croisés $\left(\mathbf{Y}_k^t\mathbf{D}_k\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D}\right)$ fournit également deux systèmes d'axes l'un noté $\mathbf{U}_k$ situé dans $\mathbb{R}^p$ $\mathbf{Q}$-orthonormé $\mathbb{R}^q$ $\mathbf{R}$-orthonormé, l'autre noté $\mathbf{V}_k$ situé dans $\mathbb{R}^q$ $\mathbf{R}$-orthonormé.

Les systèmes d'axes $\mathbf{U}$, $\mathbf{U}_k$ sont tous situés dans $\mathbb{R}^p$, les systèmes d'axes $\mathbf{V}$, $\mathbf{V}_k$.sont tous situés dans $\mathbb{R}^q$. On projette dans $\mathbb{R}^p$, respectivement dans $\mathbb{R}^q$, les axes $\mathbf{U}_k$ sur $\mathbf{U}$, respectivement $\mathbf{V}_k$ sur $\mathbf{V}$

On exprime ainsi ce que représente la relation moyenne à la relation propre de chacun des K couples de $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$ et ceci au niveau de chacun des deux triplets.

— Le deuxième objectif est à l'inverse du premier, il consiste à exprimer cette relation moyenne dans un vision propre de chacun des K couples de triplets $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$.

La relation moyenne est entièrement définie par l'association de deux états typologiques moyens, celui des lignes, et celui des colonnes.

L'expression de l'état typologique moyen des lignes par chacun des K triplets croisés s'effectue selon les règles classiques. En effet, les axes du système $\mathbf{V}$ dans $\mathbb{R}^q$ $\mathbf{R}$-orthonormé sont des codes numériques $\mathbf{R}$-normés des lignes. Ils sont situés dans le même espace que le nuage des colonnes de chacun des K triplets croisés $(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{R}, \mathbf{Q})$. La représentation associée à cet objectif correspond à la projection de chacun des nuages colonnes des K triplets sur ce système d'axes.

L'expression de l'état typologique moyen des colonnes par chacun des K triplets croisés suit la même démarche que celle effectuée pour les lignes. Le système d'axes $\mathbf{U}$ dans $\mathbb{R}^p$ $\mathbf{Q}$-orthonormé est situé dans le même espace que le nuage des lignes de chacun des K triplets croisés $(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{R}, \mathbf{Q})$. La représentation associée à cet objectif correspond à la projection de chacun des K nuages des lignes sur ce système d'axes.

Au niveau de chaque triplet croisé $(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{R}, \mathbf{Q})$ correspondent alors deux représentations, une pour les lignes, l'autre pour les colonnes. Rien ne permet d'associer les deux, mais les représentations des lignes de chacun des K triplets $(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{R}, \mathbf{Q})$ s'associent à la représentation des colonnes dans l'analyse de la moyenne. Inversement, les représentations des colonnes de chacun des triplets croisés s'associent à la représentation des lignes dans l'analyse de la moyenne.

L'expression de la relation moyenne au niveau des individus peut aussi s'effectuer selon une logique de co-inertie d'un couple de triplets.

Les nuages des individus des K triplets $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$ sont situés dans $\mathbb{R}^p$ où se trouve le système d'axes de co-inertie $\mathbf{U}$ issu de l'analyse moyenne. De même, les nuages des

individus des K triplets $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$ sont situés dans $\mathbb{R}^q$, où se trouve le système d'axes $\mathbf{V}$.

Par projection dans $\mathbb{R}^p$, respectivement $\mathbb{R}^q$, des nuages des individus des K triplets $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$, respectivement $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$, sur le système d'axes de co-inertie $\mathbf{U}$, respectivement $\mathbf{V}$, on exprime la relation moyenne au niveau des individus par chacun des K couples $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$. La concordance des deux nuages projetés associés à chacun des K couples $\{(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k), (\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)\}$ s'obtient par normalisation et superposition des deux nuages projetés, pour mesurer la part de réalisation de relation moyenne par chacun des K couples de triplets.

En résumé, **STATICO** se définit par la combinaison de deux composantes issues de deux analyses à fonctionnements géométriques différents, mais ne combine pas deux fonctionnements géométriques. Chacune des deux composantes a un rôle, celui de la co-inertie est lié à un objectif d'extension, celui d' ACT-STATIS est lié à une stratégie.

3. Illustration de l'analyse STATICO sur un exemple

On utilise ici les données de Pegaz-Maucet (1980). Ces données écologiques suscitent une question représentative d'un cadre d'application de **STATICO**. Cette question est celle de l'étude de la dynamique spatio-temporelle de la relation faune-milieu. Six (6) stations réparties dans la rivière Meaudret (voir figure 1) ont été visitées chacune une fois par saison (1-Printemps, 2-été, 3-automne, 4-Hiver). A chaque fois ont été mesurés dans chaque station 10 paramètres physiques du cours d'eau et identifiés les ephéméroptères (13 espèces) présents. On obtient donc un total de 24 relevés (6 stations * 4 saisons).

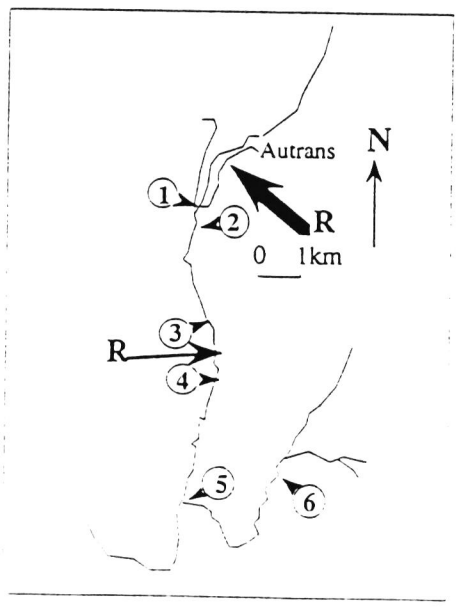


Figure 1.

Le tableau faunistique **X** (24, 13) est formé à partir de quatre tableaux $\mathbf{X}_k$ (6, 13). De même, le tableau mésologique **Y** (24, 10) est formé de quatre tableaux $\mathbf{Y}_k$ (6, 10). Les deux tableaux $\mathbf{X}_k$, $\mathbf{Y}_k$ correspondent à une saison.

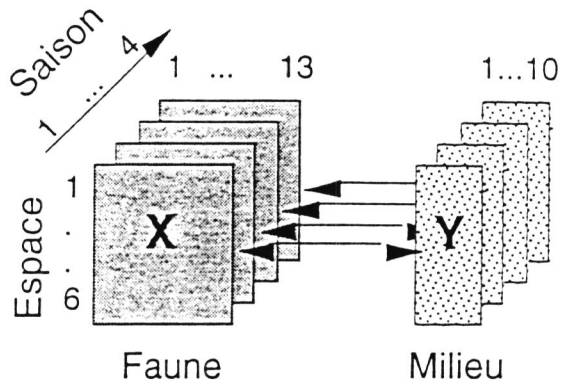


Figure 2. Situation expérimentale

Les espèces sont centrées par saison, les variables du milieu sont centrées par tableau, puis normalisées globalement (Bouroche, 1975). Cette normalisation globale permet de prendre en compte la variance intra-saison. Chacun de ces tableaux correspond à un triplet particulier $(\mathbf{X}_k, \mathbf{Q}, \mathbf{D}_k)$, $(\mathbf{Y}_k, \mathbf{R}, \mathbf{D}_k)$

Chacun des huit tableaux peut être exploré séparément pour décrire la diversité de la composition faunistique ou l'état du milieu. On peut aussi choisir d'analyser le multitableaux **X** pour mettre en évidence la part stable de la composition faunistique d'une saison à l'autre et étudier une dynamique temporelle de cette composition faunistique. Une illustration se trouve dans Thioulouse & Chessel (1987). A chaque saison k, le tableau faunistique $\mathbf{X}_k$ et le tableau du milieu $\mathbf{Y}_k$ portent sur les mêmes n_k stations. La définition de la relation faune-milieu s'effectue par l'analyse de co-inertie.

Il s'agit ici de décrire l'évolution de la relation faune-milieu d'une saison à une autre. On ne fait pas ici de l'interprétation des résultats une priorité, même si des éléments allant dans ce sens seront évoqués. L'objectif de cette illustration reste d'expliciter les moyens que l'analyse met en œuvre ainsi que la fonction de chacun d'eux dans la description de l'évolution de cette relation.

La composante de co-inertie dans STATICO consiste à passer aux triplets croisés $\left(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{R}, \mathbf{Q}\right)$; on construit ainsi 4 triplets croisés $\left(\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k, \mathbf{R}, \mathbf{Q}\right)$. Chacun des tableaux $\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k$ comporte 10 lignes (les variables environnementales) et 13 colonnes (les taxons), qui résument toute l'information concernant l'état de la relation faune-milieu à chacune des quatre saisons. Ces tableaux $\mathbf{Y}_k^t \mathbf{D}_k \mathbf{X}_k$ forment un nuage dans un espace vectoriel muni d'un produit scalaire.

La composante d'ACT-STATIS dans STATICO implique trois étapes. On définit une relation moyenne, et on discute la stabilité de cette relation moyenne pour chacune des 4 saisons. La définition de la moyenne entre les 4 tableaux $\mathbf{Y}_k^t\mathbf{D}_k\mathbf{X}_k$ implique la diagonalisation de la matrice des produits scalaires entre ces tableaux croisés (tableau 1) et la conservation du premier vecteur propre dont les composantes permettent de définir la combinaison linéaire. Un triplet moyen est alors obtenu comme combinaison linéaire des tableaux croisés pondérés par les composantes du premier vecteur propre $\mathbf{Z} = \sum_{k=1}^{K} \alpha_k \mathbf{Y}_k^t\mathbf{D}_k\mathbf{X}_k$. L'analyse du compromis est l'analyse du triplet $(\mathbf{Z},\mathbf{R},\mathbf{Q})$. Selon les règles classiques de l'analyse d'un triplet, celle de $(\mathbf{Z},\mathbf{R},\mathbf{Q})$ fournit simultanément deux systèmes d'axes de co-inertie et un système de valeurs propres. On choisit dans chacun des deux systèmes, deux axes sur lesquels on projette les lignes et les colonnes du tableau croisé moyen.

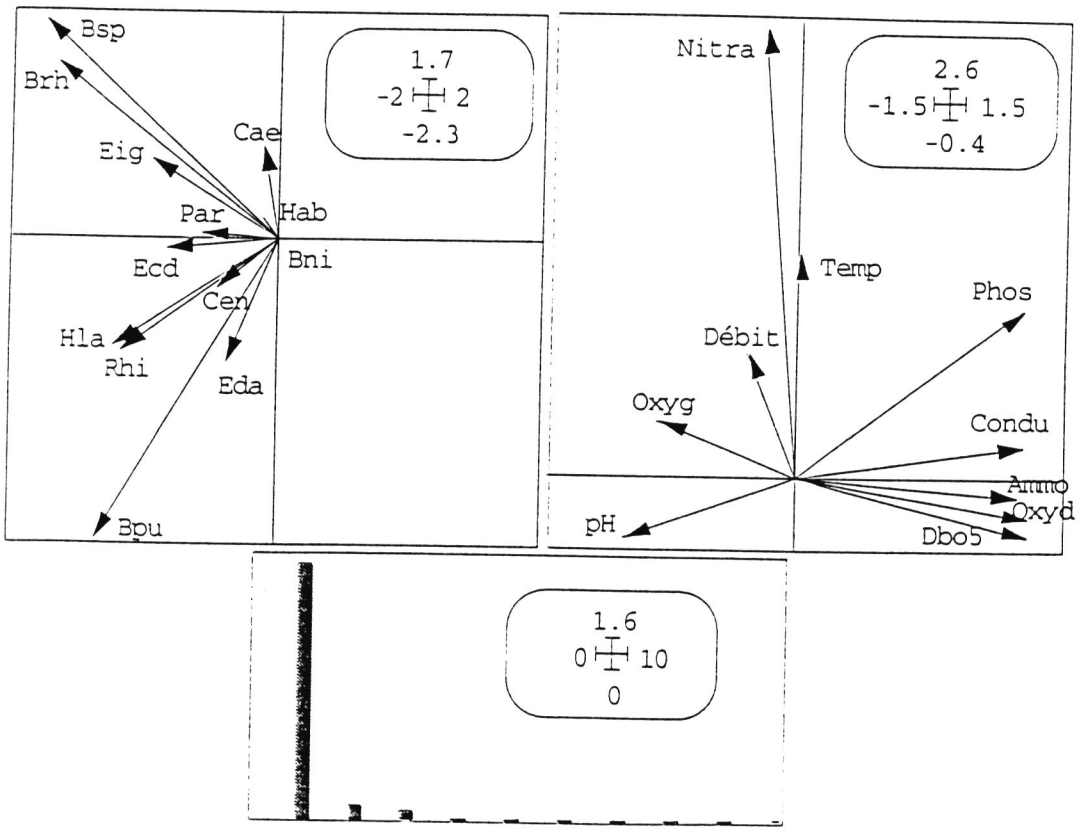


Figure 5. Analyse de la moyenne.

Deux états typologiques s'associent, définissant ainsi une relation moyenne faune-milieu.

Sur l'axe 1, les espèces caractérisent un effet taille en association avec les facteurs du milieu (Ammoniaque, Oxygène, DBO, Oxydabilité).

Sur l'axe 2, on a une répartition précise de la composition faunistique qui s'associe avec les facteurs du milieu déterminant l'axe 2 (Nitrates, débit, température). Selon la logique de l'analyse, cette relation moyenne est supposée majoritairement présente d'une

saison à une autre. Il s'agit de discuter de l'évolution de cette relation à travers une description de l'évolution des facteurs du milieu et de la composition faunistique. Autrement dit, on va discuter de l'évolution saisonnière de la relation faune-milieu. Les opérations géométriques correspondant aux représentations ci-dessus ont été expliquées lors de la présentation des trois étapes de STATICO.

— Au niveau faunistique, selon la logique de l'analyse, l'état typologique qui décrit la composition faunistique est supposé stable. L'expression de la relation moyenne est exprimée par chacune des saisons au niveau du tableau mésologique. Celle-ci permet de décrire l'évolution saisonnière des facteurs du milieu. Ici, d'une saison à l'autre, les variables du milieu (nitrates, température, débit) décrivent les facteurs qui s'impliquent dans l'évolution de la relation faune-milieu. De même, d'une saison à l'autre, on note une évolution de l'intensité des variables du milieu caractérisant l'axe 1.

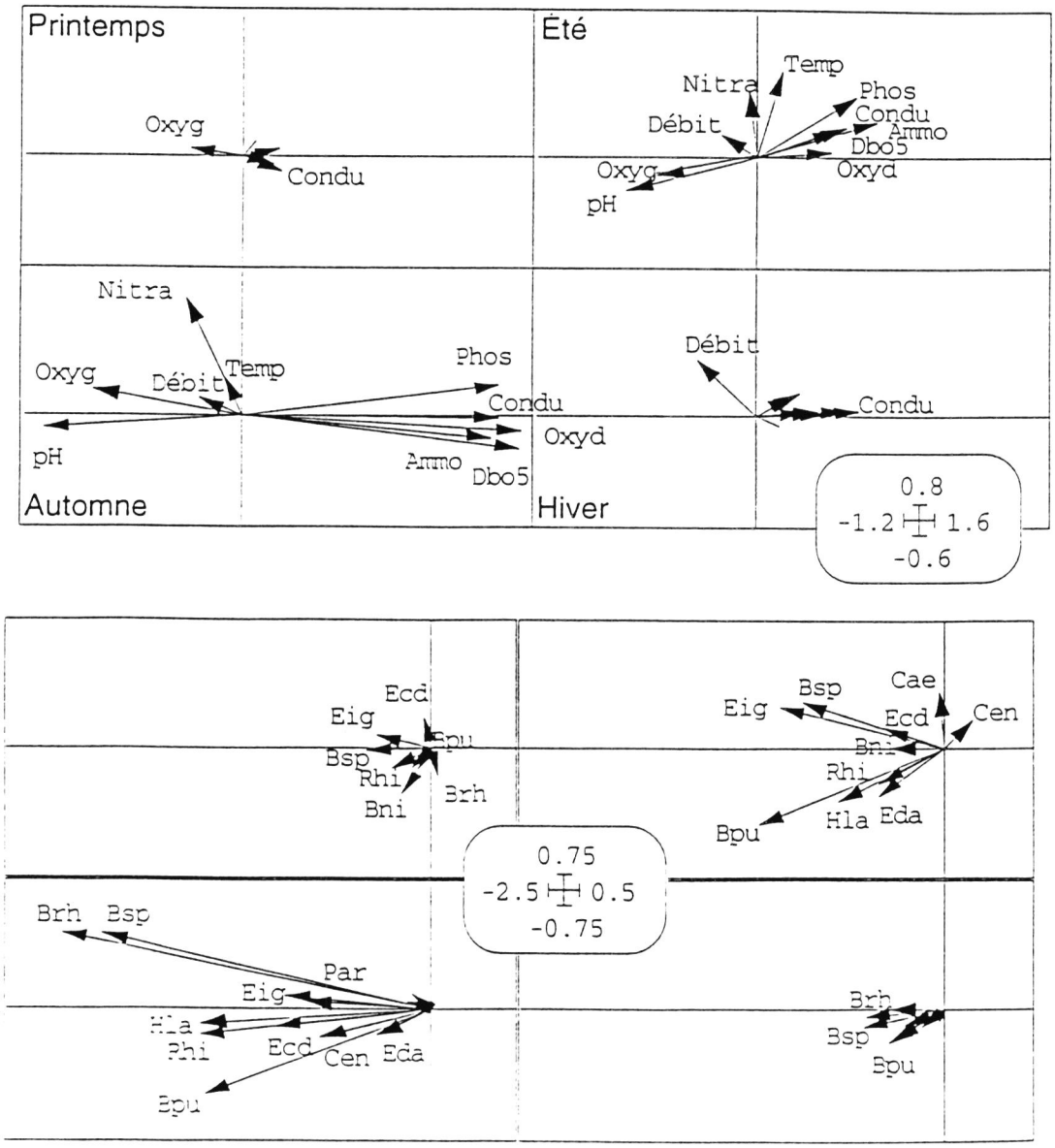


Figure 6. Expression de la relation moyenne au niveau faunistique et mésologique

— Au niveau mésologique, la relation moyenne est exprimée par chacune des saisons au niveau faunistique, permettant ainsi de décrire une évolution saisonnière de la composition faunistique.

— au niveau des stations, l'évolution de la relation moyenne d'une saison à l'autre passe par l'expression spatiale de celle-ci. Elle implique pour chacune des quatre saisons, l'expression du co-état typologique des stations entre le tableau faunistique et le tableau mésologique. On exprime ainsi la relation moyenne pour chacune des quatre saisons.

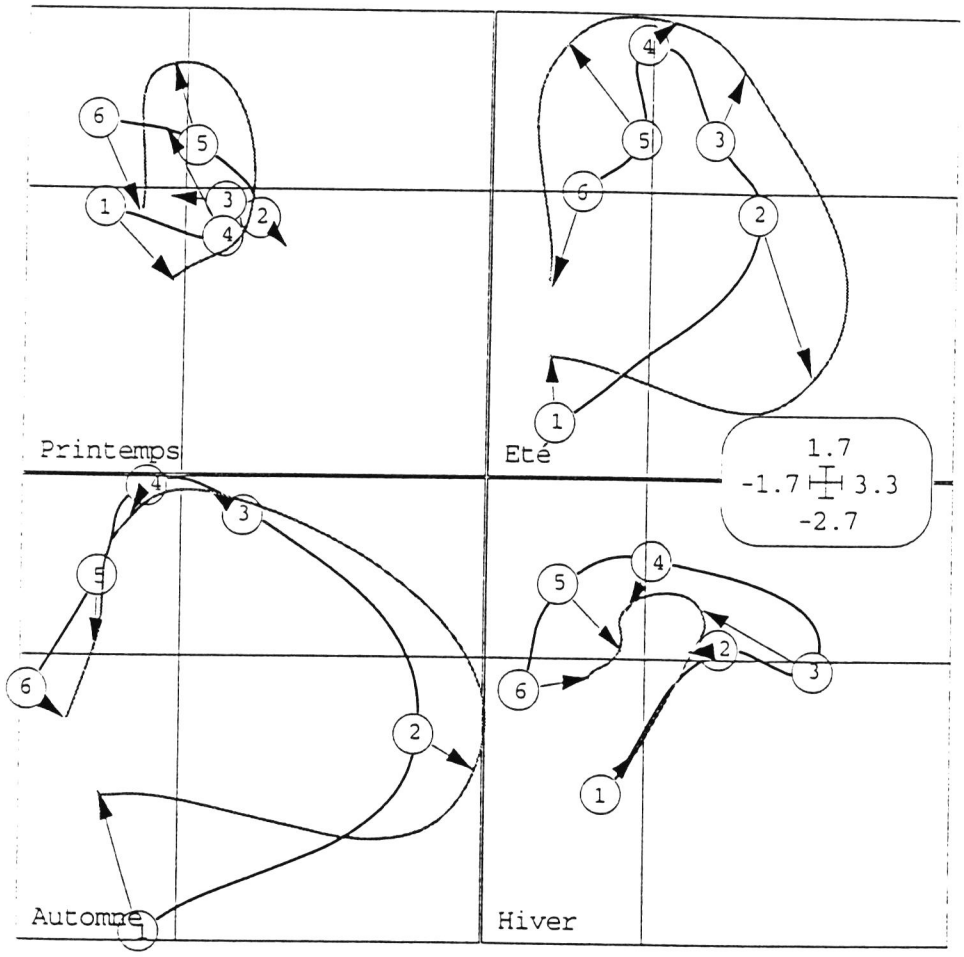


Figure 7. Par saison, au niveau des stations, expression de la relation moyenne

Ainsi, sur l'axe 1, à une augmentation des facteurs polluants correspond une diminution du nombre d'espèces. L'axe 1 décrit la pollution du cours d'eau, l'axe 2 correspond à une restauration de la rivière et la composition faunistique de l'amont est différente de celle l'aval. En effet, à une relation "pollution-restauration" correspond une dynamique spatiale de la composition faunistique selon deux pôles amont et aval.

En automne, l'importance des nitrates est plus grande et la station 5 est proche de la station 6 (référence non polluée sur la Bourne). La station 5 est pratiquement restaurée, elle est aussi proche de la station 1 (référence non polluée en amont du méaudret)

En été, la pollution en cours (tourisme) est plus forte et la restauration (en particulier dans la station 3) est faible mais la faune ne semble pas l'enregistrer totalement.

Le lien entre la relation moyenne définie par STATICO et la relation propre dans chaque saison passe par la projection des axes de co-inertie sur les axes de co-inertie compromis.

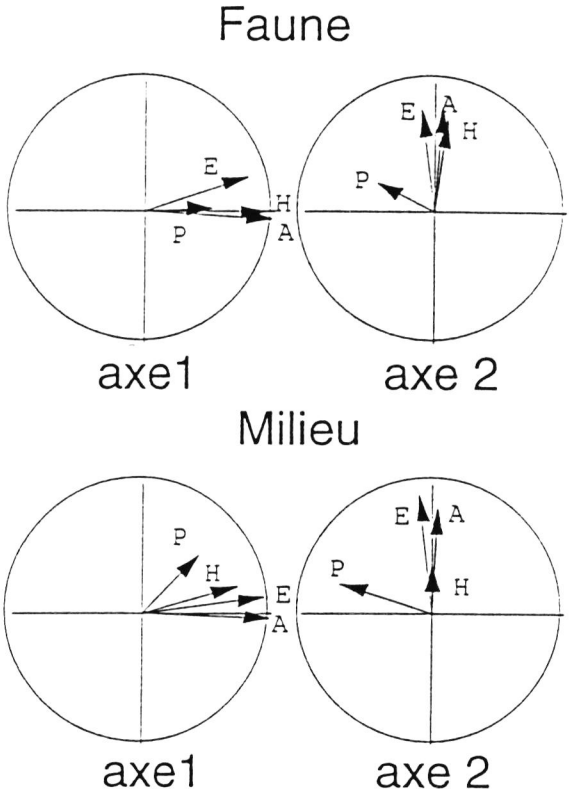


Figure 8. Projection des axes des analyses de co-inertie des 4 saisons

Ce dernier élément de comparaison est formé par la projection des axes de co-inertie du tableau faunistique, respectivement mésologique, de chaque saison sur le plan de co-inertie faunistique, respectivement mésologique, fourni dans l'analyse moyenne.

On voit que l'état typologique faunistique moyen est plus stable d'une saison à une autre que l'état typologique moyen mésologique.

4. Conclusion

L'analyse STATICO, combine deux composantes l'une issue de l'analyse de co-inertie de deux tableaux, l'autre issue d'ACT-STATIS. Elle appartient au premier sous univers. STATICO est présentée ici avec une illustration dans le cadre d'enquêtes écologiques à composantes temporelles, elle se généralise aisément à toute situation expérimentale pouvant se décrire sous la forme d'une série de K couples de tableaux. Elle permet de proposer une vision synthétique d'un co-état typologiques moyen sous d'une

analyse moyenne de co-inertie , et d'étudier stabilité de ce co-état typologique par chacun des K couples de tableaux. L'analyse STATICO est mise en oeuvre dans le cadre de la programmathèque ADE-4 (Daniel Chessel, Sylvain Dolédec et Jean Thioulouse), librement accessible sur internet (http://biomserv.univ-lyon1.fr/ADE-4.html)

II.3. Conclusions et perspectives

Dans cette deuxième partie, différentes contributions sont détaillées dans chacun les deux chapitres qui la constituent . On se restreint ici aux conclusions et perspectives qui concernent l'apport de cette partie appliquée à l'étude de la structure de l'univers.

1. Conclusions

Une approche appliquée a été considérée avec un champ expérimental représenté par l'écologie, des contributions ont été apportées à deux niveaux :

— au niveau de l'analyse multivariée appliquée : on a introduit en écologie des analyses adaptées, qu'elles soient existantes comme ACT-STATIS (chapitre II.1), ou nouvelles comme ACOM (chapitre II.1) et STATICO (chapitre II.1).

La mise en œuvre de ces analyses permet aux utilisateurs d'avoir accès à celles-ci. C'est le cas avec le logiciel ADE-4 qui est à la base des illustrations des analyses discutées dans cette partie.

Il parait important de mentionner à ce niveau que le champ expérimental ouvre de nouvelles perspective dans le cadre du développement méthodologique des analyses multivariées de tableaux de données à trois entrées.

— au niveau structurel : on a apporté deux éléments complémentaires aux apports théoriques pour l'étude de la structure de l'univers.

Le premier est l'étude comparative menée dans le chapitre II.1. En effet, dans la première partie, on a extrait de l'univers deux sous-univers. Ces deux sous-univers expriment deux solutions pour un même problème géométrique (chapitre I.2). Dans la deuxième partie, cette situation n'est plus valable. Ainsi, il est possible.que deux analyses à fonctionnement géométrique différent puissent résoudre une même question expèrimentale. Les différentes solutions géométriques expriment seulement différentes possibilités de mise en œuvre et d'interprétation.

Le second élément est la notion de statut d'une analyse dans une approche unificatrice. C'est le cas pour l'analyse STATICO qui a était choisi pour son statut d'élément du premier sous-univers.

2. Perspectives

La première perspective est une extension du schéma de l'étude comparative du chapitre II.1, aux différent niveaux de structure de l'univers. Cette extension vise des études comparatives entre :

— les différentes analyses qui constituent la composante de liens,

— les différentes analyses qui constituent le premier sous univers,

— les couples d'analyses, chaque couple étant défini par deux analyses, l'une issue du premier sous univers, l'autre issue du second.

La deuxième perspective consiste à généraliser la question du couplage de deux tableaux à celle du couplage entre K tableaux et un tableau de référence. Cette question est au centre des analyses basées sur une analyse de la moyenne et elle les généralise. Du point de vue structurel, l'intérêt de cette question peut résider dans le fait qu'elle combine deux logiques intrinsèquement différentes, la première consiste en un fonctionnement théorique unique, celui des analyses des tableaux à deux entrées, la deuxième consiste en la diversité de ses fonctionnements théoriques. Finalement, cette question est liée à celle, plus générale, consistant à décrire l'état typologique des nuages des tableaux.

Références

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, New York.

Bouroche, J.M. (1975) Analyse des données ternaires: La double analyse en composantes principales. Thèse de 3-ème cycle, Université de Paris VI.France.

Carroll, J. D, Chang, J. J (1970). Analysis of individual differences in multidimensional scaling via un n-way generalization of "Eckart-Young" decomposition. Psychometrika, 35, 283-319.

Carroll, J. D. & Arabie, P. (1980). Multidimensional scaling. Annual review of Psychology, 31, 607-649.

Carroll, J. D.& Wish, M. (1974). Models and methods for three-way multidimensional scaling. In D. H. Krantz, R. C. Atkinson, R. D. Luce, P. Suppes (Ads.) Contemporary developemnts in mathematical psychology, vol II: Measurement, psychophysics, and neural processing, 57-105. San Francisco : Freeman & Co.

Carroll, J.D. (1968). A generalization of canonical correlation analysis to three or more sets of variables. Proceeding of the 76th Convention of the American Psychological Association, 3, 227-228.

Casin, Ph. (1995). L'analyse discriminante de tableaux évolutifs. Revue de Statistiques Appliquées, XLIII, (3), 73-91.

Cazes, P. & Chessel, D. & Dolédec, S. (1988). L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie. Revue de Statistique Appliquée.36, 39-54.

Cazes, P. (1980). L'analyse de certains tableaux rectangulaires décomposé en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. I. Définitions et applications à l'analyse canonique des variables qualitatives. II Questionnaires : variantes des codages et nouveaux calculs de contributions. Les Cahiers de l'Analyse des Données, 5, 145-161 & 387-406.

Chen, W. Chao. (1974). An optimal property of principal components. Communications in statistics, 3, (10), 979-983.

Chessel, D. & Dolédec, S. (1996). ADE Version 4 : HyperCard © Stacks and Programme library for the Analysis of Environmental Data. Manuel d'utilisation. 10 fascicules. ESA-5023, Université Lyon 1, 69622 Villeurbanne cedex.

Chessel, D. & Hanafi, M. (1996). Analyses de la Co-inertie de K nuages de points. Revue de Statistiques Appliquées, XLVI, 2, 35-60.

Chessel, D. & Mercier, P. (1993). Couplage de triplets statistiques et liaisons espèces-environement. In : Biométrie et Environment. Lebreton, J. D. & ASSELAIN, B. (Eds.) Masson, Paris. 15-44.

Chevenet, F. & Dolédec, S. & Chessel, D. (1994). A fuzzy coding approach for the analysis of long-term ecological data. Freshwater Biology, 31, 295-309.

Cliff, N. (1966). Orthogonal rotation to congruence. Psychometrika, 31, 33-42.

Coppi, R. & Balasco, S. (1989). Multiway data Analysis. Amsterdam: Elsevier Science Publisher. North Holland.

Crocker, R. L. (1952). Soil genesis and the pedogenic factors. Quart. Rev. Biol., 27, 139-168.

Darroch, J. N. (1965). An optimal property of principal components. Anal. Math. Stat. 1579-1582.

Dauxois, J. & Pousse, A. (1976). Les analyses factorielles en calcul de probabilités et en statistiques : essai d'étude synthétique. Thèse d'état. Université de Toulouse. France.

Dauxois, J. & Romain, Y. & Viguier, Y.S. (1993). Comparison of two factor subspaces. Journal of Multivariate Analysis, 44, 160-178.

De Leeuw, J. & Heiser, W. J. (1980). Multidimentional Scaling with restrictions on the configuration. In P.R.Krishnaiah (Ed.). Multivariate Analysis V. Amsterdam: North Holland Publishing Company, 501-522

De Leeuw, J. (1988). Convergence of the majorization method for Multidimentional Scaling. Journal of Classification, 5, 163-180.

Denis, J. B. & Dhorne, Th. (1989). Orthogonal tensor decomposition of 3-way tables. In R.Coppi and S.Bolasco (Eds.), Multiway data analysis. Amsterdam: Elsevier Science Publisher, 269-276

Dolédec, S. & Chessel, D. (1994) Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology, 31, 277-294.

Escofier, B. & Pagès, J. (1984) L'analyse factorielle multiple : une méthode de comparaison de groupes de variables. In : *Data Analysis and Informatics III.* DIDAY, E. & Coll. (Eds.) Elsevier, North-Holland. 41-55.

Escofier, B. & Pagès, J. (1985). Mise en œuvre de l'analyse factorielle multiple pour les tableaux numériques qualitatifs ou mixtes. Rapport de recherche n°429. INRIA, Domaine de Voluceau-Rocquencourt, BP 105, 78153 Le chesnay cedex, France. 1-54 + annexes.

Escofier, B. & Pagès, J. (1994) Multiple factor analysis (AFMULT package). Computational Statistics and Data Analysis, 18, 121-140.

Escoufier, Y. (1973). Le traitement des variables vectorielles. Biometrics, 29, 750-760.

Escoufier, Y. (1977). Operators related to data matrix. Recents developments in statistics, BARRA (Ed.). North Holland, 125-131.

Flury, B. D. (1995). Developments in Principal Component Analysis. Recent Advances in Descriptive Multivariate Analysis. Oxford: Oxford University Press, (2), 14-33.

Franc, A. (1989). Multiway arrays : some algebras remarks. In R.Coppi and S.Bolasco (Eds.), Multiway data analysis. Amsterdam: Elsevier Science Publisher, 19-29.

Franc, A. (1992). Etude algébrique des multitableaux : Apports de l'algèbre tensorielle. Thèse de doctorat. Université de Montpellier II, France.

Franquet, E. & Chessel, D. (1994). Approche statistique des composantes spatiales et temporelles de la relation faune-milieu. C.R. Acad. Sci. Paris, Sciences de la vie, 317, 202-206.

Franquet, E., Dolédec, S. & Chessel, D. (1995). Using multivariate analyses for separating spatial and temporal effects within species-environment relationships. Hydrobiologia, 300-301, 425-431.

Friday, L. E. (1987). The diversity of macroinvertebrate and macrophyte communities in ponds. Freshwater Biology, 18, 87-104.

Gifi, A. (1990). Non linear multivariate Analysis. Wiley, New York.

Glaçon, F. (1981). Analyse conjointe de plusieurs matrices de données. comparaison de différentes méthodes. Thèse de 3-ème cycle. Université de Grenoble. France.

Gower, J. C. (1975). Generalized Procustes Analysis. Psychmetrika, 40, (1), 33-51

Gower, J. C. (1984). Multivariate Analysis: Ordination, multidimensional scaling and allied topics. In. EH.Lloyd (Ed.). Handbook of applicable mathematices: Wiley, Chichester, vol (VI), 727-781.

Gower, J. C. (1995). Orthogonal and Projection Procustes Analysis. Recent Advances in Descriptive Multivariate Analysis. Oxford: Oxford University Press, (6), 113-134.

Groenen, P. J. F. (1993). The majorization approach to Multidimensional Scaling. Leiden: DSWO press.

Harshman, R. A. & Lundy, M. E. (1984). The PARAFAC model for three-way factor analysis and multidimensional scaling. In H. G. Law Snyder Jr. J. A. Hattie, & R. P. McDonald(Eds.), Rsearch methods for multimode data analysis. New York, Praeger,122-215.

Harshmann, R. A. (1970). Foundations of the parafac procedure : models and conditions for an "exploratory"multi-modefactor analysis. UCLA Working Papers in Phonetics, 22, 31-44.

Heiser, W. J. (1995). Convergent Computation by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis. In Krzanowski (Ed.). Recent Advances in Descriptive Multivariate Analysis. Oxford University Press, 8, 157-189.

Horst, P. (1961). Relations among m sets of measures. Psychometrika, 26, 129-149.

Horst, P. (1965). Factor analysis of data matrices. New York: Holt, Rinehart & Winston.

Hotelling H. (1936). Relations between two sets of variates. Biometrika, 28, 321-377.

Hotelling, H. (1935). The most predictable criterion. Journal of Educational Psychology, 26, 139-142.

Jaffrenou, P. A. (1978). Sur l'analyse des familles finie de variables vectorielles. Thèse de 3ème cycle. Université Lyon I. France.

Kazi-Aoual, F.& Hitier, S. & Sabatier, R. & Lebreton, J. D. (1995). Refined approximations to permutation tests for multivariate inference. Computational Statistics and Data Analysis (in press).

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. Biometrika, 58, 433-451.

Kettenring, J. R. (1985). Canonical Analysis. In S. Kots & N. L. Johnson (Eds.) Encyclopedia of Statistical Sciences Wiley, New York, vol.1, 354-365.

Kiers, H. A. L. & Ten Berge, J. M. F. (1989). Alternating Least Squares algorithms for simultaneous Component Analysis with equal weight matrices for all populations. Psychmetrika, 54, 467-473.

Kiers, H. A. L. (1988). Comparison of "anglo-saxon" and " french" three-mode methods.Statistiques et analyse des données. 13, (3), 14-32.

Kiers, H. A. L. (1991). Hierarchical relations among three -way methods. Psychmetrika, 9, 449-470.

Kiers, H. A. L. (1995). Maximization of sums of quotients of quadratic forms and some generalizations. Psychmetrika, 60 (2), 221-245.

Kronenberg, P. M. & De Leeuw, M. E. (1980). Principal component analysis of three mode data by means of alternating least squares algorithms. Psychometrika, 45,96-97.

Kronenberg, P. M. (1983). Three mode principal component analysis : theory and applications. Leiden : DSWO press.

Kronenberg, P. M. (1995). Bibliography of three-way data analysis : part II & algorithm char for three-way methods. Departement of Education, Leiden University (Publication interne).

Kruskal, J. B. (1989). Rank, decomposition, and uniquenesse for 3-way and N-way. In R.Coppi and S.Bolasco (Eds.), Multiway data analysis (pp. 7-18). Amsterdam: Elsevier Science Publisher.

Krzanowski, W, J. (1979). Between-group compraison of principal component. Journal on Americain Statistical Association, 74, 703-707. (Correction note: 1981, 76, 1022 1988)

Krzanowski, W, J. (1988). Principales of multivariate analysis: A user's Perspective. Clarendon Press, Oxford.

L'Hermier des plantes, H.(1976). Structuration des tableaux à trois indices de la statistique. Thèse de 3-ème cycle. Université de Montpelier II. France.

Lafosse, R. (1989). Proposal for a generalised canonical analysis. In R.Coppi and S.Bolasco (Eds.), Multiway data analysis. Amsterdam: Elsevier Science Publisher, 269-276.

Lafosse, R. (1997). Analyse de concordance de deux tableaux : monogamies, simultanéités et découpages. Revue de Statistique Appliquée, n°45 (à paraitre).

Lavit, Ch. & Escoufier, Y. & Sabatier, R. & Traissac, P. (1994). The ACT (Statis method). Computational Statistics and Data Analysis, 18, 97-119.

Lavit, Ch. (1988) Analyse conjointe de tableaux quantitatifs. Masson, Paris.

Lazraq, A. & Cléroux, R. & Kiers, H. A. L. (1992). Mesures de liaison vectorielle et généralisation de l'analyse canonique. Revue de Statistique Appliquée, 39, 23-35.

Lebreton, J.D., Sabatier, R., Banco, G. & Bacou, A.M. (1991) Principal component and correspondence analyses with respect to instrumental variables : an overview of their role in studies of structure-activity and species- environment relationships. In : Applied Multivariate Analysis in SAR and Environmental Studies. Devillers, J. & Karcher, W. (Eds.) Kluwer Academic Publishers, 85-114.

Margalef, R. (1968). Perspectives in ecological theory. Univ. Chicago press, Chicago.

Masson, M. (1974) . Processus linéaires et analyse non linéaire des données. Thèse d'état, Université Paris VI. France.

Mc Koen, J. J. (1966). Canonical analysis : some relation between canonical correlation, factor analysis, discriminant analysis, and scaling theory. Psychometric monograph 13.

McDonald. R, P. (1968). A unified treatment of the weighting problem. Psychometrika, 26, 321-377.

Mercier, P. & Chessel, D. & Dolédec, S. (1992). Complete correspondence analysis of an ecological profile data table: a central ordination method. Acta Œcologica, 13, 25-44.

Meulman, J. J. (1986). A distance approach to nonlinear multivariate analysis. Leiden: DSWO.

Meyer, R. (1989). Extensions of correspondence analysis for the statistical exploration of multidimentional contingency tables. In: O.Opitz (ed). Conceptual and numerical Analysis of data . Spinger, Berlin, Heidelberg, 178-186.

Meyer, R. (1991). Canonical correlation analysis as starting point for extensions of correspondence analysis. Statistique et Analyse des données, 16, 55-77.

Meyer, R. (1992). Multidimensional scaling as a framework for correspondence analysis and its extensions.In: M. Schader (ed.). Analyzing and modeling Data and Knowledge. Spinger, Berlin, Heidelberg, 63-72.

Millsap, R. E. & Meredith, W. (1988). Component Analysis in cross-sectional and longitudinal data. Psychometrika, 53, 123-134.

Mosier, C. L. (1939). Determining a simple structure when loadings for certain items are knomn. Psychometrika, 4, 33-44.

Mulaik, S. A. (1972). The foundations of factor analysis. New York : MacGraw-Hill.

Pegaz-Maucet, D. (1980). Impact d'une perturbation d'origine organique sur la dérive des macro-invertébrés bentiques d'un cours d'eau. Comparaison avec le benthos. Thèse de 3ème cycle. Université Lyon I. France.

Rao, C.R. (1979). Separation for singular values of matrices and their applications in multivariate analysis. Journal of Multivariate Analysis, 9, 362-377.

Rutishauser, H. (1969). Computational aspects of F.L. Bauer's Simultanueous Iterartion Method. Numer.Math., 13,4-13.

Sabatier, R. (1993). Critères et contraintes pour l'ordination simultanée de K tableaux. In : Biométrie et Environnement. Lebreton, J.D. & Asselain, B. (Eds.) Masson, Paris, 101-121.

Saporta, G. (1975). Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse de 3-ème cycle. Université de Paris VI. France.

Steel, R. G. D. (1951). Maximum generalized variance for a set of linear functions . Ann. Math . Stat., 22, 456-460.

Ten Berge, J. M. F (1977). Optimizing factor invariance. Thèse de doctorat. Université de Groninguen. Pays Bas.

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, (1), 49-55.

Ten Berge, J. M. F. & Nevels, K. (1977). A general solution to Mosier's oblique procustes problem. Psychometrika, 42, (1), 593-600.

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.). Multidimentional Data Analysis. Leiden: DSWO Press, 81-87.

Ten Berge, J. M. F. (1988). Generalised approches to the MAXBET problem and the MAXDIFF problem with applications to canonical correlations. Psychometrika, 53, (4), 487-494.

Ten Berge, J. M. F. & Knol, D. L. & Kiers, H. A. L. (1988). A Treatment of the orthomax Rotation Family in Terms of Diagonalization, and a Re-Examination of a Singular Value Approach to Varimax Rotation. Computational Statistics Quartly, 3, 207-217.

Ten Berge, J. M. F. (1991). A general solution for a class of Weakly constrained linear regression problems. Psychometrika, 56, (4), 601-609.

Tenenhaus, M. & Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis ans other methods for quantifying categorical multivariate data. Psychometrika, 50, (1), 91-119.

Tenenhaus, M. (1977). Analyse en composantes principales d'un ensemble de variables nominales ou numériques. Revue de Statistique Appliquée, 25, 39-56.

Tenenhaus, M. (1984) L'analyse canonique généralisée de variables numériques, nominales ou ordinales par des méthodes de codage optimal. In : *Data Analysis and Informatics, III*. DIDAY, E. & Coll. (Eds.) Elsevier Science Publishers B.V., North-Holland. 71-84.

Tenenhaus, M., Gauchi, J.P. & Ménardo, C. (1995) Régression PLS et applications. Revue de Statistique Appliquée, 43, 7-63.

Ter Braak, C. J. F. & Juggins, S. (1993) Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia, 269/270, 485-502.

Ter Braak, C. J. F. (1987). Unimodal models to relate species to environment. Agricultural Mathematics Group, Box 100, NL-6700, AC Wageningen, The Netherlands, 1-152.

Thioulouse, J. & Chessel, D. (1987). Les analyses multi-tableaux en écologie factorielle. I De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. Acta Œcologica, Œcologia Generalis : 8, (4), 463-480.

Thioulouse, J. (1989). Statistical analysis and graphical display of multivariate data on the MacIntosh. Computer Applications in the BIOSciences : 5, 4, 287-292.

Thioulouse, J. (1990) MacMul and GraphMu : two Macintosh programms for the display and analysis of multivariate data. Computers and Geosciences : 8, 1235-1240.

Thioulouse, J., Devillers, J., Chessel,, D. & Auda, Y. (1991). Graphical techniques for multidimensional data analysis. In : Applied Multivariate Analysis in SAR and Environmental Studies. Devillers, J. & Karcher, W. (Eds.) Kluwer Academic Publishers, 153-205.

Thioulouse, J., Dolédec, S., Chessel, D. & Olivier, J.M. (1995). ADE Software: multivariate analysis and graphical display of environmental data. In : Software per l'ambiante. Gauriso, G. & Gozzili, A. (Eds.) Pàtron Editore, Bologna, 57-62.

Tucker, L. R. (1958). An interbattery method for factor invariance. Psychometrika, 23, 111-136.

Tucker, L. R. (1966) Some mathematical notes on three-mode factor analysis. Psychometrika, 31, 279-311.

Van de Geer, J. P. (1986). Relations among K sets of variables. In J. De Leeuw, W. J. Heiser, J. Meulman & F. Critchley (Eds.). Multidimentional Data Analysis. Leiden: DSWO Press, 67-79.

Van de Geer, J.P. (1971). Introduction to multivariate analysis for the social sciences. San Francisco : Freeman.

Van de Geer, J.P.(1984). Linear relations among K sets of variables. Psychometrika, 49, (1), 79-94.

Van de Geer, P. (1986). Introduction to linear multivariate analysis data analysis. Leiden: DSWO Press, (2 vols).

Van den Wollenberg, A. L. (1984). Redundancy analysis : an alternative for canonical correlation analysis. Psychometrika, 42, 297-220.

Vinograde, B. (1950). Canonical positive definite matrices under internal linear trasformations. Proc. Amer. Math. Soc., 1, 159-161.

Winemiller, K. O. (1991). Ecomorphological diversification in lowland freshwater fish assemblages from five biotic regions. Ecol. Monogr., 61, 343-365.

Young, G. & Housseholder, A.S. (1938). Discussion of a set of points in terms of their mutual distances. Psychometrika, 3, (1), 19-22.