REVUE DE STATISTIQUE APPLIQUÉE

F. TORRE
D. CHESSEL

Co-structure de deux tableaux totalement appariés

Revue de statistique appliquée, tome 43, n° 1 (1995), p. 109-121
<http://www.numdam.org/item?id=RSA_1995__43_1_109_0>

Rev. Statistique Appliquée, 1995, **XLIII** (1), 109-121

CO-STRUCTURE DE DEUX TABLEAUX TOTALEMENT APPARIÉS

F. Torre (1) et D. Chessel (2)

(1) Ecologie des Systèmes Fluviaux, URA CNRS 1451, 1, rue Parmentier, 13200 Arles.
(2) URA CNRS 1451, Bât 403, Université Lyon I, 69622 Villeurbanne Cedex.

RÉSUMÉ

Deux triplets statistiques **(X, Q, D)** et **(Y, Q, D)** sont totalement appariés s'ils portent sur les mêmes unités statistiques (n lignes) et les mêmes variables (p colonnes). La note définit l'axe principal de co-inertie comme le vecteur unique de $\mathbb{R}^p$ maximisant la covariance des coordonnées des projections des deux nuages dans le même espace euclidien.

La somme des inerties des deux analyses est décomposée canoniquement entre un terme de co-inertie et un terme de différence entre les deux tableaux. On explicite les liens entre analyse de co-inertie, analyse de différence et analyses inter-classes et intra-classes. Une illustration et un logiciel sont proposés.

Mots-clés : *tableaux totalement appariés, analyse inter-classes, intra-classes, analyse inter-batterie, analyse de co-inertie, régression aux moindres carrés partiels.*

SUMMARY

Two statistical triplets **(X, Q, D)** and **(Y, Q, D)** are totally matched when the involved arrays (**X** and **Y**) are composed of the same individuals (n rows) and the same variables (p columns). This paper defines the main co-inertia axis as the unique vector of $\mathbb{R}^p$ which maximizes the covariance of the projected coordinates of the two multidimensional arrays in the same Euclidean space.

The sum of inertia resulting from the two analyses may be broken up into its term which characterizes the co-inertia and its term which integrates the difference between the two arrays. We give a short review about co-inertia analysis, analysis of difference and between- and within classes analyses. Furthermore, we give an example and we propose a software to make the co-inertia analysis of two totally matched tables.

Keywords : *totally matched tables, between classes analysis, within classes analysis, inter-battery analysis, co-inertia analysis, partial least square regression.*

1. Introduction

On dit que deux tableaux sont appariés quand ils décrivent les mêmes unités statistiques. Deux tableaux sont dits *totalement appariés* quand ils décrivent les

mêmes individus à l'aide des mêmes variables. On reprend là le vocabulaire utilisé par R. LAFOSSE, notamment dans (LAFOSSE 1985a; 1985b; 1989).

Lorsque l'appariement ne porte que sur les lignes des deux tableaux, les méthodes d'analyses utiles sont nombreuses (synthèse bibliographique pour l'écologie dans MERCIER 1991). L'analyse de co-inertie (CHESSEL & MERCIER, 1993) est une approche géométrique qui synthétise l'analyse inter-batterie de TUCKER (1958), l'analyse des correspondances de tableaux de profils écologiques de ROMANE (1972, *cf.* MERCIER *et al.* 1992) et l'analyse canonique sur variables qualitatives de CAZES (1980). Ces méthodes d'analyse symétrique utilisent des schémas du type (**X, Q, D**), (**Y, R, D**) et (**X**^{t}**DY, R, Q**) et repose sur la recherche d'axes dits de co-inertie maximisant la covariance entre les coordonnées des projections des lignes de chacun des deux tableaux. **Q** et **R** sont des produits scalaires diagonaux associés à des pondérations des variables. L'analyse inter-batterie est à la base de la régression aux moindres carrés partiels (PLS regression), méthode fondamentale en chimiométrie (HÖSKULDSSON 1988, STONE & BROOKS, 1990) dans l'analyse du couple structure-activité symétrique du couple espèces-milieu en écologie (LEBRETON *et al.*, 1991).

La principale propriété de cette famille d'ordination à deux tableaux est qu'elle effectue une analyse d'inertie de chacun des deux tableaux simultanément. Ces deux analyses sont compatibles en ce sens qu'on maximise le produit des trois quantités : inertie projetée sur un axe dans un espace, inertie projetée sur un axe dans l'autre et carré de la corrélation des coordonnées factorielles.

Les individus de chaque tableau n'étant pas décrits par les mêmes variables, ils ne sont pas représentables dans le même espace : on détermine donc des couples d'axes de co-inertie, et les axes composant chaque couple sont éléments de deux espaces différents. On peut amener les deux nuages dans un même espace par l'analyse canonique vue par CASIN & TURLOT (1986), mais l'introduction des métriques de Mahalanobis $\mathbf{Q} = (\mathbf{X}^t\mathbf{D}\mathbf{X})^-$ et $\mathbf{R} = (\mathbf{Y}^t\mathbf{D}\mathbf{Y})^-$ pose souvent plus de problèmes qu'elle n'en résout.

L'analyse d'un couple de tableaux totalement appariés a été jusqu'à présent abordé par l'analyse procustéenne. L'analyse de communauté (LAFOSSE, 1985a; 1985b; 1989) considère les deux nuages de n points de $\mathbb{R}^p$ et opère une rotation procuste orthogonale d'un nuage vers l'autre, puis une dilatation définies de telle façon qu'elle rapproche les points appariés des deux nuages. La rotation procuste orthogonale est définie par le passage d'un système d'axes à un autre. LAFOSSE (1989) montre que ces axes sont les axes définis dans l'analyse inter-batterie de TUCKER (1958). Ce sont également les axes de co-inertie.

Le problème d'optimisation en jeu dans l'analyse de communauté rejoint celui de l'analyse de co-structure et ne prend donc pas en compte l'appariement des colonnes des deux tableaux. Seule la dilatation opérée ensuite tire parti de cette propriété. On est alors amené à la question posée dans la figure 1.

Dans ce qui suit, on considère deux triplets statistiques (**X, Q, D**) et (**Y, Q, D**) où **X** et **Y** sont deux tableaux à n lignes (ou individus) et p colonnes (ou variables), **Q** est la matrice d'un produit scalaire de $\mathbb{R}^p$ dans la base canonique et **D** une matrice diagonale contenant les poids des n lignes. **D** est inversible et de trace unité. **X** et **Y** sont **D**-centrés par colonne. Les deux triplets représentent indifféremment des

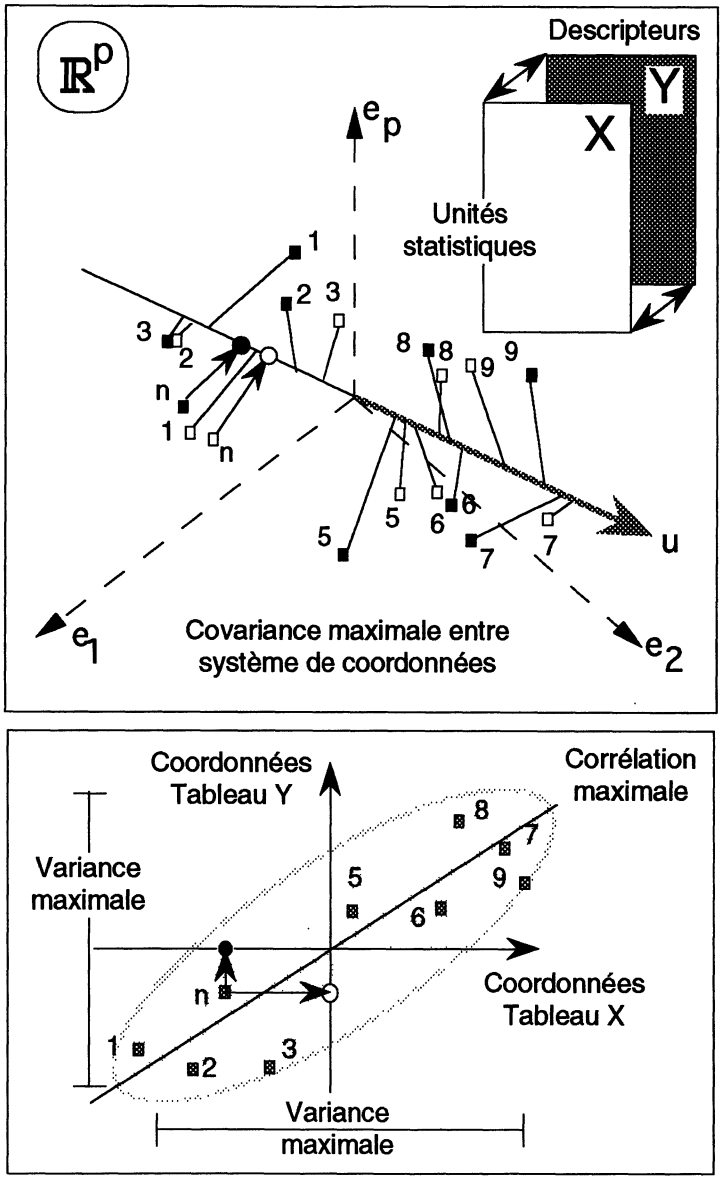

FIGURE 1
Question posée dans l'analyse d'un couple de deux tableaux complètement appariés : existe t'il un axe de co-inertie commun aux deux nuages ?

analyses en composantes principales ou des analyses des correspondances simples ou multiples dans l'esprit d'ESCOUFIER (1982, 1985, 1987) et TENENHAUS & YOUNG (1985).

L'appariement total des deux tableaux nous autorise à considérer les individus des deux tableaux comme les éléments d'un seul et même espace.

2. Analyse de co-structure entre (X,Q,D) et (Y,Q,D)

2.1. Proposition

Soit $\mathbf{u}$ un vecteur $\mathbf{Q}$-normé de $\mathbb{R}^p$. La covariance entre coordonnées des projections des deux nuages sur $\mathbf{u}$:

$$c(\mathbf{u}) = \text{Cov}\,(\mathbf{XQu}, \mathbf{YQu}) = (\mathbf{XQu}|\mathbf{YQu})_{\mathbf{D}}$$

est maximum pour le premier vecteur propre $\mathbf{Q}$-normé de la matrice $\mathbf{Q}$-symétrique :

$$\widetilde{\mathbf{C}} = \frac{1}{2}(\mathbf{Y}^t\mathbf{DXQ} + \mathbf{X}^t\mathbf{DYQ})$$

Démonstration

On note d'abord que $\langle \widetilde{\mathbf{C}}\mathbf{u}, \mathbf{Qu}\rangle = \text{Cov}(\mathbf{XQu}, \mathbf{YQu})$

La matrice $\widetilde{\mathbf{C}}$ est $\mathbf{Q}$-symétrique. En effet :

$$\begin{aligned}
\langle \widetilde{\mathbf{C}}\mathbf{u}, \mathbf{Qv}\rangle &= \frac{1}{2}\langle \mathbf{Y}^t\mathbf{DXQu}, \mathbf{Qv}\rangle + \frac{1}{2}\langle \mathbf{X}^t\mathbf{DYQu}, \mathbf{Qv}\rangle \\
&= \frac{1}{2}\langle \mathbf{YQv}, \mathbf{DXQu}\rangle + \frac{1}{2}\langle \mathbf{XQv}, \mathbf{DYQu}\rangle \\
&= \frac{1}{2}\langle \mathbf{XQu}, \mathbf{DYQv}\rangle + \frac{1}{2}\langle \mathbf{YQu}, \mathbf{DXQv}\rangle \\
&= \frac{1}{2}\langle \mathbf{X}^t\mathbf{DYQv}, \mathbf{Qu}\rangle + \frac{1}{2}\langle \mathbf{Y}^t\mathbf{DXQv}, \mathbf{Qu}\rangle \\
&= \langle \widetilde{\mathbf{C}}\mathbf{v}, \mathbf{Qu}\rangle
\end{aligned}$$

La $\mathbf{Q}$-symétrie de $\widetilde{\mathbf{C}}$ garantit l'existence d'une base de vecteurs propres $\mathbf{Q}$-orthonormée de l'image de $\widetilde{\mathbf{C}}$ dans $\mathbb{R}^p$. Le premier vecteur $\mathbf{u}_1$, associée à la plus grande valeur propre λ_1, donne :

$$\text{Max}_{||\mathbf{u}||_Q=1}\langle \widetilde{\mathbf{C}}\mathbf{u}, \mathbf{Qu}\rangle = \lambda_1 = \text{Cov}\,(\mathbf{XQu}_1, \mathbf{YQu}_1)$$

2.2. Remarques

1) Sous contrainte d'orthogonalité à $\mathbf{u}_1$, le second vecteur $\mathbf{u}_2$, associée à la seconde valeur propre λ_2, maximise la même quantité et ainsi de suite. On appellera analyse de co-inertie totalement appariée entre les triplets (**X,Q,D**) et (**Y,Q,D**) la

recherche des vecteurs $\mathbf{u}_1, \ldots, \mathbf{u}_q$ de $\mathbb{R}^p$ et la projection des nuages sur les plans qu'ils définissent.

2) La quantité maximisée n'est pas une inertie et la matrice $\widetilde{\mathbf{C}}$ n'est pas positive. Il ne semble pas que le plan défini par $\mathbf{u}_1$ et $\mathbf{u}_2$ ait une propriété d'optimalité équivalente à celle d'un plan d'inertie optimale.

3) Si $\mathbf{Q}$ est diagonale, comme dans les analyses élémentaires classiques, on diagonalisera pratiquement $1/2\mathbf{Q}^{1/2}(\mathbf{Y}^t\mathbf{D}\mathbf{X} + \mathbf{X}^t\mathbf{D}\mathbf{Y})\mathbf{Q}^{1/2}$. Si $\mathbf{Q}$ n'est pas diagonale, comme en analyse discriminante, on utilisera une décomposition de Choleski ou une diagonalisation préliminaire de $\mathbf{Q}$.

4) Sur les axes principaux de cette analyse de co-inertie, il est possible de projeter, outre les deux nuages de points, leurs axes principaux respectifs et leurs axes de co-inertie respectifs (ici au sens de deux tableaux simplement appariés sur les lignes). On observe qu'ils sont distincts mais peuvent être très proches.

5) La relation simple :

$$\text{Cov}\,(\mathbf{XQu}, \mathbf{YQu}) = \text{Corr}\,(\mathbf{XQu}, \mathbf{YQu})\sqrt{\text{Var}\,(\mathbf{XQu})}\sqrt{\text{Var}\,(\mathbf{YQu})}$$

montre qu'on obtient un compromis entre les deux analyses simples et l'analyse canonique des deux tableaux, qui d'un certain point de vue sont exécutées simultanément avec un axe unique. C'est pourquoi on peut appeler poids canoniques les composantes des vecteurs trouvés.

6) Les coordonnées des projections sur deux axes de cette analyse vérifient :

$$\text{Cov}\,(\mathbf{XQu}_j, \mathbf{YQu}_k) + \text{Cov}\,(\mathbf{XQu}_k, \mathbf{YQu}_j) = 0 \qquad 1 \leqslant j \neq k \leqslant p$$

7) On peut appeler régression PLS totalement appariée l'utilisation des coordonnées de $\mathbf{XQu}_1$ pour modéliser par régression $\mathbf{D}$-pondérée le tableau $\mathbf{Y}$ et, réciproquement, l'utilisation des coordonnées de $\mathbf{YQu}_2$ pour modéliser par régression $\mathbf{D}$-pondérée le tableau $\mathbf{X}$.

Dans la logique de la régression PLS on recommence la recherche du premier axe de co-inertie avec les tableaux résidus des régressions.

3. Analyse de co-inertie et approches voisines

3.1. Analyse inter-classe

Dans $\mathbb{R}^p$, une ligne du tableau $\mathbf{X}$ est appariée à une ligne du tableau $\mathbf{Y}$ pour former un vecteur lié. Les milieux de ces vecteurs, dont les coordonnées sont dans $1/2(\mathbf{X}+\mathbf{Y})$ définissent une analyse d'inertie dite analyse inter-classes (Figure 2) qui utilise le triplet $1/2(\mathbf{X} + \mathbf{Y}), \mathbf{Q}, \mathbf{D})$. On peut rapprocher cette analyse de l'analyse de co-inertie en remarquant que son inertie totale (I_B, B pour between) vaut :

$$4I_B = \text{Tr}\,((\mathbf{X}^t + \mathbf{Y}^t)\mathbf{D}(\mathbf{X} + \mathbf{Y})\mathbf{Q}) = I_X + 2\text{Tr}\,(\widetilde{\mathbf{C}}) + I_Y$$

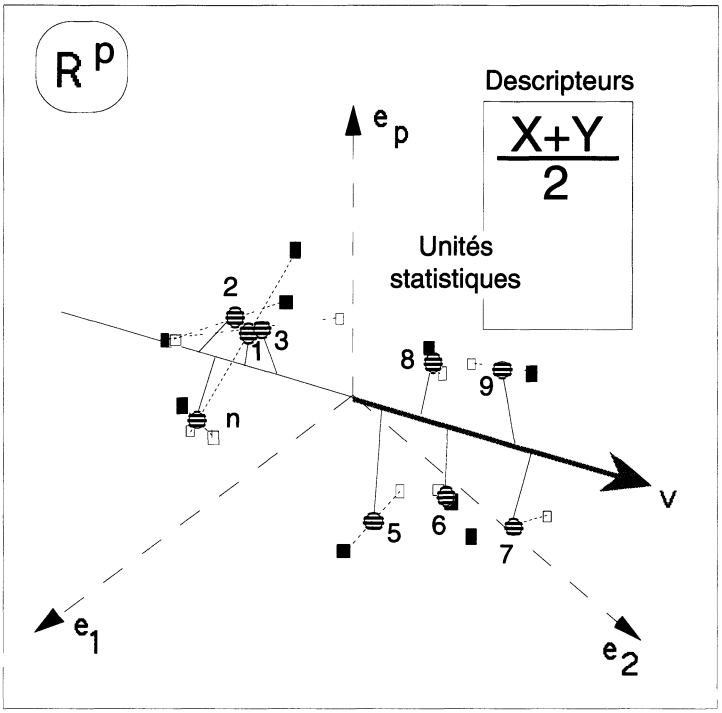

FIGURE 2
Axe principal de l'analyse inter-classes associé à deux tableaux appariés

où I_X et I_Y désignent les inerties totales des analyses de $\mathbf{X}$ et $\mathbf{Y}$.

On notera que l'axe principal $\mathbf{v}$ de l'inter-classe maximise (au facteur 4 près) :

$$\text{Var}\,(\mathbf{XQv}) + \text{Var}\,(\mathbf{YQv}) + 2\,\text{Cov}\,(\mathbf{XQv}, \mathbf{YQv})$$

3.2. Analyse des différences et analyse intra-classes

L'analyse du triplet statistique $(\mathbf{Y} - \mathbf{X}, \mathbf{Q}, \mathbf{D})$ permet d'expliciter ce en quoi les deux tableaux sont différents. Dans cette analyse, on prend comme référence la description des individus fournie par le tableau $\mathbf{X}$ et on cherche à comprendre dans quelle mesure la description des individus fournie par le tableau $\mathbf{Y}$ s'en écarte. Les axes de cette analyse sont les mêmes que ceux du triplet $(\mathbf{X} - \mathbf{Y}, \mathbf{Q}, \mathbf{D})$ et la représentation garde encore une symétrie parfaite entre les deux tableaux. Cette remarque renvoie à l'analyse intra-classes associée à l'inter-classe du paragraphe précédent.

On considère alors, par appariement sur les variables, le tableau :

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}$$

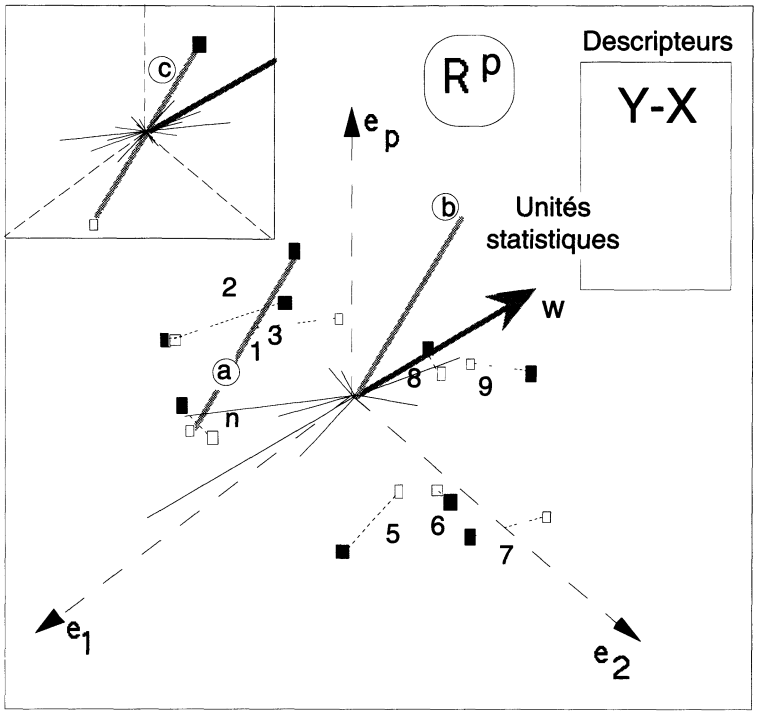

FIGURE 3
Analyse des différences entre deux tableaux complètement appariés. L'axe principal du tableau $\mathbf{Y} - \mathbf{X}$ *est aussi celui de l'analyse intra-classe (en haut, à gauche) associée à l'analyse inter-classes de la figure 2*

et la partition des lignes en n groupes de 2 définis par l'appariement sur les individus. L'analyse intra-classes est celle du triplet :

$$\left(\begin{bmatrix} \mathbf{X} - 1/2(\mathbf{X}+\mathbf{Y}) \\ \mathbf{Y} - 1/2(\mathbf{X}+\mathbf{Y}) \end{bmatrix}, \mathbf{Q} \begin{bmatrix} 1/2\mathbf{D} & 0 \\ 0 & 1/2\mathbf{D} \end{bmatrix}\right)$$

On diagonalise ici l'opérateur :

$$\frac{1}{4}[(\mathbf{X}-\mathbf{Y})^t(\mathbf{Y}-\mathbf{X})^t] \begin{bmatrix} 1/2\mathbf{D} & 0 \\ 0 & 1/2\mathbf{D} \end{bmatrix} \begin{bmatrix} (\mathbf{X}-\mathbf{Y}) \\ (\mathbf{Y}-\mathbf{X}) \end{bmatrix} \mathbf{Q} = \frac{1}{4}(\mathbf{X}-\mathbf{Y})^t\mathbf{D}(\mathbf{X}-\mathbf{Y})\mathbf{Q}$$

L'opérateur en jeu dans l'analyse intra-classe est identique à celui en jeu dans l'analyse des différences au facteur 4 près. On note donc que les inerties associées I_D (D pour différences et I_W (W pour within) vérifient :

$$I_D = 4I_W$$

3.3. Liens entre analyse de la co-inertie et analyse de la différence

Comme $(\mathbf{X}-\mathbf{Y})^t\mathbf{D}(\mathbf{X}-\mathbf{Y})\mathbf{Q}+2\widetilde{\mathbf{C}}=\mathbf{Y}^t\mathbf{D}\mathbf{Y}\mathbf{Q}+\mathbf{X}^t\mathbf{D}\mathbf{X}\mathbf{Q}$, on a :

$$I_X+I_Y=2\text{Tr}\,(\widetilde{\mathbf{C}})+I_D$$

On vérifie de plus que :

$$I_X+I_Y=2I_B+2I_W=2I_B+\frac{1}{2}I_D$$

On obtient deux décomposition de la somme des inerties des triplets d'origine. Les termes intervenant dans chacune de ces deux décompositions sont à leur tour décomposés suivant les axes principaux de chaque analyse.

Les relations liant les traces proviennent de relations analogues entre variables du type $\mathbf{x}, \mathbf{y}, (\mathbf{x}-\mathbf{y})$ et $(\mathbf{x}+\mathbf{y})$:

$$\forall(j,k)\in\{1,\ldots,p\}^2 \quad \text{Var}\,(\mathbf{x}^j)+\text{Var}\,(\mathbf{y}^k)=\text{Var}\,(\mathbf{x}^j-\mathbf{y}^k)+2\text{Cov}\,(\mathbf{x}^j,\mathbf{y}^k)$$

En particulier, lorsqu'on considère deux variables se correspondant d'un tableau à l'autre (cas où $j=k$),

$$\forall j\in\{1,\ldots,p\} \quad \text{Var}\,(\mathbf{x}^j)+\text{Var}\,(\mathbf{y}^j)=\text{Var}\,(\mathbf{x}^j-\mathbf{y}^j)+2\text{Cov}\,(\mathbf{x}^j,\mathbf{y}^j)$$

On peut donc envisager de décomposer les traces variable par variable et évaluer ainsi l'importance de chacune des p variables dans la co-inertie et la différence entre triplets.

Le comportement de ces pratiques sur des données complexes sont à explorer. Dans certains cas, les axes d'inertie sont conservés et les nuages déformés; dans d'autre cas le nuage peut être conservé avec une rotation des axes. On peut toujours comparer les valeurs des critères optimisés, en particulier inertie projetée sur un axe (analyses simples), covariance des coordonnées avec deux axes de projection (analyse de co-inertie simple), co-inertie avec un seul axe de projection (analyse des co-inertie totalement appariée). Ces analyses fonctionnant sur des schémas de dualité quelconque, après vérification des cohérences des pondérations utilisées, sont disponibles dans le logiciel ADE (CHESSEL & DOLÉDEC, 1993).

4. Illustration

On utilise un jeu de données (Tableau 1) préparé et étudié par LAFOSSE (1985a). Il concerne la mortalité par accident sur la voix publique selon la classe d'âge, le sexe et la catégorie d'usagers (données cumulées des années 1958, 1959 et 1960). $n=14$ classes d'âge, d'amplitude égale à 5 ans, de 0 à 70 ans sont considérées comme individus statistiques, $p=5$ catégories d'usagers étant interprétées comme

les variables. Les deux tableaux donnent la répartition en pourcentage par catégorie d'usagers (somme par ligne unité). Le premier tableau concerne les femmes ($\mathbf{X}$) et le second concerne les hommes ($\mathbf{Y}$). On utilise la pondération uniforme ($\mathbf{D} = (1/14)\mathbf{I}_{14}$) et la métrique canonique ($\mathbf{Q} = \mathbf{I}_5$). Les deux tableaux sont centrés par colonnes.

TABLEAU 1

Données traitées par l'analyse de co-inertie totalement appariée (LAFOSSE 1985a)
P = piéton, C = cycliste, M = motard, A = automobiliste,
Autres = autres catégories d'usagers

	Femmes (X)				
Age	P	C	M	A	Autres
[0,5[	,473	,008	,034	,454	,031
[5,10[	,576	,045	,030	,322	,027
[10,15[	,307	,176	,116	,365	,036
[15,20[	,126	,127	,468	,252	,027
[20,25[	,086	,062	,450	,377	,025
[25,30[	,092	,054	,323	,502	,029
[30,35[	,097	,059	,267	,547	,030
[35,40[	,106	,063	,245	,549	,037
[40,45[	,134	,076	,243	,507	,040
[45,50[	,169	,089	,224	,481	,037
[50,55[	,200	,098	,200	,461	,041
[55,60[	,270	,106	,150	,430	,044
[60,65[	,384	,090	,097	,387	,042
[65,70[	,636	,039	,021	,271	,033
Moyennes	*,261*	*,078*	*,205*	*,422*	*,034*

	Hommes (Y)				
Age	P	C	M	A	Autres
[0,5[	,552	,011	,035	,375	,027
[5,10[	,630	,096	,028	,224	,022
[10,15[	,247	,378	,141	,201	,033
[15,20[	,045	,124	,681	,126	,024
[20,25[	,046	,045	,611	,250	,048
[25,30[	,050	,051	,537	,305	,057
[30,35[	,063	,057	,466	,347	,067
[35,40[	,069	,064	,440	,358	,069
[40,45[	,086	,079	,440	,335	,060
[45,50[	,105	,096	,434	,304	,061
[50,55[	,128	,119	,415	,283	,055
[55,60[	,154	,141	,419	,270	,016
[60,65[	,201	,160	,339	,256	,044
[65,70[	,346	,182	,221	,219	,032
Moyennes	*,194*	*,115*	*,372*	*,275*	*,044*

Parmi les multiples figures qu'il est possible de faire avec un tel jeu de données, nous n'en garderons qu'une seule (Figure 4). En effet, on peut penser que l'analyse de co-inertie totalement appariée, ici décrite, est «généraliste», en ce sens qu'elle permet d'orienter le dépouillement vers telle ou telle approche plus spécifique en autorisant un point de vue central. La proposition II.1 donne deux vecteurs normés de $\mathbb{R}^5$.

$$\mathbf{u}_1 = (-.7427, -.0274, +.6633, +.0851, +.0217)$$
$$\mathbf{u}_2 = (+.1836, -.3977, +.3314, -.8317, -.0810)$$

Ils sont normés et orthogonaux. La première question est «quelle valeur ont ces vecteurs en terme d'inertie projetée?». Pour $\mathbf{X}$, l'inertie projetée sur $\mathbf{u}_1$ vaut .04876 pour un maximum possible de .05013 (première valeur propre de l'ACP de $\mathbf{X}$) et l'inertie projetée sur $\mathbf{u}_2$ vaut .0094 pour un valeur de .0099 pour la deuxième valeur propre de l'ACP de $\mathbf{X}$. Le plan P engendré par $(\mathbf{u}_1, \mathbf{u}_2)$ donne donc une inertie projetée de 0.0582 pour un maximum (plan 1-2 de l'ACP de $\mathbf{X}$) de 0.0600. On peut dire que le plan P est très voisin de l'optimum, ce qui se voit en projetant les axes d'inertie de $\mathbf{X}$ sur P (Figure 4, vecteur 1F et 2F dans le cercle unité).

Pour $\mathbf{Y}$, l'inertie projetée sur $\mathbf{u}_1$ vaut .06652 pour un maximum possible de .06762 (première valeur propre de l'ACP de $\mathbf{Y}$) et l'inertie projetée sur $\mathbf{u}_2$ vaut .0076 pour une valeur de .0097 pour la deuxième valeur propre de l'ACP de $\mathbf{Y}$. Le plan P donne donc une inertie projetée de 0.0741 pour un maximum (plan 1-2 de l'ACP de

$\mathbf{Y}$) de 0.0773. Le plan P est assez voisin de l'optimum, ce qui se voit en projetant les axes d'inertie de $\mathbf{Y}$ sur P (Figure 4, vecteur 1H et 2H dans le cercle unité). On notera que l'axe 1 de l'ACP de $\mathbf{Y}$ a été automatiquement changé de signe pour optimiser la covariance, ce qui est souvent le cas dans les analyses de co-inertie. Les projections des deux nuages sur P sont donc sensiblement les plans 1-2 des ACP de base.

On a représenté sur ce plan P les projections des vecteurs de la base canonique de $\mathbb{R}^5$, ce qui a deux avantages. Le premier est de discuter de tous les éléments (lignes, colonnes et tableaux) dans un même espace. Le second est de faire du biplot (avec propriété d'averaging ligne-colonne) pour retrouver la généralisation de la représentation triangulaire en dimension quelconque (GOWER 1967), ce qui redonne un sens géométrique très simple au biplot de GABRIEL (1971). L'inertie totale de $\mathbf{X}$ étant de .0620, on représente sur P 94% de la variabilité du tableau $\mathbf{X}$, soit encore 97% de l'inertie représentée sur le plan 1-2 de l'ACP simple de $\mathbf{X}$. L'inertie totale de $\mathbf{Y}$ étant de .08234, on représente sur P 90% de la variabilité du tableau $\mathbf{Y}$, soit encore 96% de l'inertie représentée sur le plan 1-2 de l'ACP simple de $\mathbf{Y}$. On perd légèrement sur l'ACP simple *en utilisant un plan commun pour les deux tableaux*, ce qui est un avantage décisif.

Puisque le plan P restitue la quasi totalité de l'information il exprime la ressemblance comme la différence des deux tableaux, ce que justifie la notion de co-structure (structure et corrélation). Les corrélations entre les deux séries de coordonnées égalent respectivement .908 sur $\mathbf{u}_1$ et .879 sur $\mathbf{u}_2$. La seconde question est donc «quelle valeur ont ces vecteurs en terme de co-inertie?». Quand on choisit un axe pour chaque nuage la covariance optimale vaut .05263, constituée des variances .04992 et .0674 et d'une corrélation de .907. Avec le premier axe de co-inertie totalement appariée on obtient respectivement 0.05172 (.04876, .06652 pour les variances, .908 pour la corrélation). Pour les axes 2 de co-inertie la covariance des coordonnées vaut .07592 (corrélation de .904); pour l'axe 2 de co-inertie totalement appariée la covariance des coordonnées vaut .07432 (corrélation de .879). On est donc encore de ce point de vue très proche de l'optimum.

La figure 4 est explicite. Non seulement le plan P est celui de la variabilité des tableaux mais aussi celui de leur points communs et de leurs points spécifiques. Le vecteur $\mathbf{u}_1$ est encore presque confondu avec l'axe 1 (vecteur 1D de la figure 4) de l'analyse des différences et le plan P contient presque le vecteur qui relie les centres de gravité des deux nuages. Le second axe d'inertie du tableau des différences, presque perpendiculaire à P, n'a pas de signification très claire et représente moins de 5% de l'inertie initiale.

La figure 4 contient donc toute l'information de comparaison des deux tableaux. On y lit la translation induite par la catégorie motard après 15 ans entre hommes et femmes, l'évolution commune de la répartition des victimes des deux sexes (piétons et auto avant 10 ans, écart brusque entre 10 et 15 ans pour les garçons mais pas pour les filles, évolution continue et prédominance de la catégorie automobiliste à l'âge adulte, le retour aux catégories piétons et cyclistes après 55 ans). Il est plus que vraisemblable que les mêmes tableaux donneraient aujourd'hui des images bien différentes.

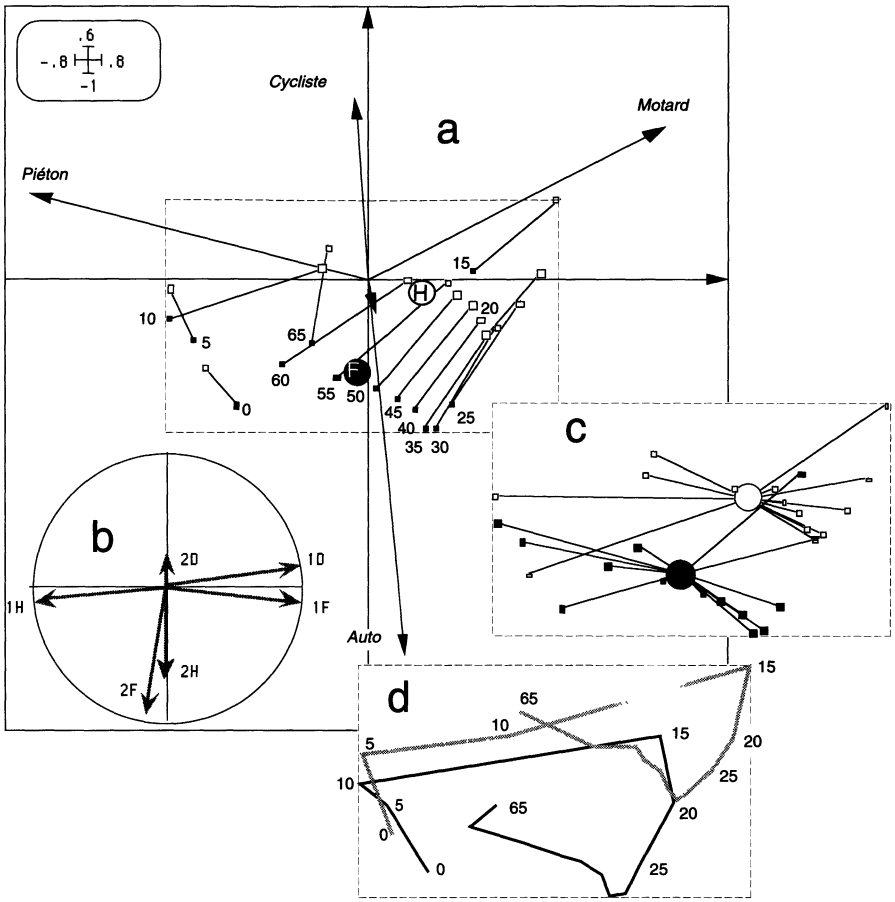

FIGURE 4

Analyse de co-inertie totalement appariée des tableaux hommes et femmes concernant la mortalité par accident de la route

a - Double biplot du plan de co-inertie. Projection de la base canonique de $\mathbb{R}^5$ *(vecteurs Piéton, ..., Auto). Projections des deux nuages non centrés (carrés noirs pour le tableau des femmes, carrés blancs pour le tableau des hommes) et de leur centre de gravité (F et H). Chaque point est au centre de gravité de la distribution à cinq composantes qui forme les données initiales.*

b - Projection sur le même plan dans l'intersection avec la sphère unité des axes principaux des nuages centrés (1H et 2H, 1F et 2F) et des axes principaux de l'analyse des différences des deux tableaux (1D et 2D).

c - Reprise du cartouche central pour exprimer le point de vue inertie commune (nuages non centrés).

d - Reprise du cartouche central pour exprimer le point de vue évolution commune des deux statistiques

5. Conclusion

De manière générale, l'analyse de co-inertie l'emporte largement sur l'analyse canonique en terme de stabilité numérique et de facilité d'interprétation. Elle évite de fabriquer de la corrélation sans signification. L'analyse des correspondances des tableaux de Burt croisés, très pratiquée, a donc son équivalent pour tout type de variables. La régression PLS est d'abord un mode d'utilisation des axes de co-inertie. Quand deux tableaux sont totalement appariés, on a montré que la manipulation d'un seul système d'axes de co-inertie est possible. Comme dans le cas des analyses de co-inertie locales et spatiales utilisant des graphes de voisinage (CHESSEL & SABATIER 1994), on est amené à diagonaliser des opérateurs non positifs. Il est possible que l'intérêt de tels opérateurs soit largement sous-estimé en analyse des données.

Remerciements

Nous exprimons notre gratitude à P. CAZES pour la précision et la gentillesse de ses conseils, à R. SABATIER pour le dynamisme de ses conversations et à D. PONT pour avoir attiré notre attention sur le sujet.

Références

CASIN Ph. & TURLOT J.C. (1986). Une présentation de l'analyse canonique généralisée dans l'espace des individus. *Revue de Statistique Appliquée*, XXXV, 3, 65-75.

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*, V, 2, 145-161 & 387-406.

CHESSEL D. & DOLÉDEC D. (1993). ADE Version 3.6 : Hypercard© Stacks and Program library for the Analysis of Environmental Data. Ecologie des Eaux Douces et des Grands Fleuves, URA 1451, Université Lyon 1. Documentation, 8 fascicules, 750 p. Diffusion par FTP anonyme sur biom3.univ-lyon1.fr (134.214.100.42).

CHESSEL D. & MERCIER P. (1993). Couplage de triplets statistiques et liaisons espèces-environnement. In *Biométrie et environnement*, LEBRETON J.D. & ASSELAIN B. (Eds.), Masson, Paris, 15-44.

CHESSEL D. & SABATIER R. (1994). Couplage de triplets statistiques et graphes de voisinage. In *Biométrie et analyse des données spatio-temporelles*, ASSELAIN B. et Coll. (Eds.), Société Française de Biométrie, ENSA, Rennes, 28-37.

ESCOUFIER Y. (1982). L'analyse des tableaux de contingence simples et multiples. *Metron*, 40, 53-77.

ESCOUFIER Y. (1985). L'analyse des correspondances : ses propriétés et ses extensions. In *45th session of the International Statistical Institute*. Amsterdam. 28.2.1-28.2.16.

ESCOUFIER Y. (1987). The duality diagramm : a means of better practical applications. In *Development in numerical ecology*. LEGENDRE P. & LEGENDRE L. (Eds.) NATO advanced Institute, Serie G, Springer Verlag, Berlin, 139-156.

GABRIEL K.R. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrika*, 58, 453-467.

GOWER J.C. (1967). Multivariate analysis and multivariate geometry. *The statistician*, 17, 13- 28.

HOSKULDSSON A. (1988). PLS régression methods. *Journal of Chemometrics*, 2, 211-228.

LAFOSSE R. (1985a). *Analyses procustéennes de deux tableaux. Propositions d'une technique visant la détection de points originaux. Essai de présentation synthétique d'analyse de deux tableaux.* Thèse de 3° cycle. Université Paul Sabatier, Toulouse.

LAFOSSE R. (1985b). Une nouvelle analyse procustéenne de deux tableaux, appariement typique et atypique de deux nuages de points. In *Data Analysis and Informatics, IV*. Diday, E. & Coll. (Eds.), Elsevier Science Publishers, North-Holland, 407-414.

LAFOSSE R. (1989). Ressemblance et différences entre deux tableaux totalement appariés. *Statistique et Analyse des Données*. 14, 2, 1-24.

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.

MERCIER P. (1991). *Analyses des relations espèces-environnement et étude de la co-structure d'un couple de tableaux.* Thèse de doctorat, Université Lyon 1, 1-168.

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.

ROMANE F. (1972). Utilisation de l'analyse multivariable en Phytoécologie. *Investigación pesquera*, Barcelona, 36, 131-139.

STONE M. & BROOKS R.J. (1990). Continuum régression : cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components régression. *Journal of the Royal Statistical Society*, B, 52, 237- 269.

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.

TUCKER L.R. (1958). An inter-battery method of factor analysis. *Psychometrika*, 23, 2, 111-136.