Journal de la Société Française de Statistique
Vol. 156 No. 4 (2015)

L'Analyse Factorielle Discriminante de Tableaux Multiples

Title: Multiblock Linear Discriminant Analysis

Philippe CASIN[1]

Résumé : L'objet de cet article est de proposer une nouvelle technique, l'analyse factorielle discriminante de Tableaux Multiples, qui généralise l'analyse factorielle discriminante et l'analyse canonique généralisée, et s'applique à des tableaux dont les variables sont partitionnées en blocs et les individus sont partitionnés en groupes ; l'AFD-TM détermine une ou plusieurs variables synthétiques pour chacun des tableaux, de telle manière que les variables synthétiques des différents tableaux soient le plus liées entre elles tout en ayant un pouvoir discriminant le plus élevé possible pour la partition des individus donnée.

Abstract: The aim of this paper is to propose a new method, multiblock linear discriminant analysis which generalizes linear discriminant analysis and generalized canonical correlation analysis,and is a method for analyzing multiblock and multigroup data tables ; MLDA computes one or several new variables for each data table, such as these new variables take into account both relationships between sets of variables and canonical correlation between each data table and the partition of individuals.

Mots-clés : tableaux multiples, analyse factorielle discriminante, analyse canonique généralisée, tableau multi-blocs, tableau multi-groupes
Keywords: multivariate data table, linear discriminant analysis, generalized canonical analysis, multi-block data analysis, multi-group data analysis
Classification AMS 2000 : 62H30, 62h25, 62-07

1. Introduction

L'analyse factorielle discriminante (AFD) a été introduite par Fisher (1936) et appliquée à différents types d'iris pour les distinguer entre eux à partir de leurs caractéristiques physiques. Aujourd'hui, les applications de l'AFD sont nombreuses et concernent des domaines très variés : à partir d'un ensemble de données quantitatives ou qualitatives observées sur les mêmes n individus, il s'agit de caractériser la partition de ces n individus en g classes.

Une autre technique, l'analyse canonique (Hotelling, 1936), permet de mettre en évidence les relations linéaires existant entre deux ensembles de variables observées sur les mêmes individus. Les techniques qui généralisent l'analyse canonique (Kettenring, 1971 ; Carroll, 1968 ; Saporta, 1976 ; Casin, 1996 ; Cazes, 2004) décrivent les relations entre plusieurs tableaux de données ; l'ensemble des variables considéré est donc partitionné en K groupes, chaque groupe constituant un tableau de variables. Une partition des individus en groupes étant donnée, on considère dans cet article, non plus un seul, mais plusieurs tableaux de variables explicatives ; autrement dit, à la fois les individus sont divisés en g groupes distincts (comme en analyse discriminante) et les

[1] UFR Droit Economie Administration Ile du Saulcy 57000 Metz
E-mail : philippe.casin@univ-lorraine.fr

 ISSN: 2102-6238

variables sont divisées en K groupes, (comme dans le cas de l'analyse canonique, lorsque K=2, et de ses généralisations, lorsque K est supérieur à 2).

Les données sont donc bi-partitionnées, c'est-à-dire qu'il s'agit à la fois d'une analyse multi-blocs et d'une analyse multi-groupes, pour utiliser la terminologie de Tenenhaus and Tenenhaus (2014) ; comme le notaient déjà Louwerse et al. (1999), ce type de données se rencontre de plus en plus fréquemment, du fait du développement des bases de données.

Ainsi, pour caractériser différents terroirs de vins de Loire et mettre en évidence une structure commune aux différents ensembles de variables explicatives, Escofier and Pagès (1994) utilisent l'Analyse Factorielle Multiple et ensuite complètent les graphiques obtenus par la projection en supplémentaire de la variable qualitative décrivant la partition des terroirs en groupes.

C'est l'analyse discriminante, l'analyse canonique généralisée, puis une généralisation de l'analyse procustéenne qu'appliquent Gardner et al. (2006) à des données bi-partitionnées issues de l'industrie minière. Morand and Pagès (2006) utilisent aussi l'analyse procustéeenne, mais conjointement à l'Analyse Factorielle Multiple, sur des données d'analyse sensorielle, tandis que Krzysko, Smialowki et Wolinsky(2014) traitent des données agricoles bi-partitionnées à l'aide de MANOVA et de l'analyse en composantes principales. Les tableaux de données qui font l'objet de l'article de Shen et al. (2014) se présentent sous forme de matrices de dissimilarités et une généralisation de l'analyse canonique est utilisée pour expliquer la partition des individus en groupes.

Kang et al. (2015) proposent une technique de discrimination multi-blocs basée sur un algorithme itératif pour établir un lien entre des gènes et des maladies, prenant en compte à la fois les facteurs discriminants et les corrélations entre les blocs de variables caractérisant les gènes. Eslami et al. (2013), (2014) utilisent l'ACP multi-groupes dans une optique de description de ce même type de données bi-partitionnées. Vallejo-Arnadela et al. (2007), d'une part, dans le cas particulier où chaque bloc a le même nombre de colonnes, et Sabatier et al. (2013), d'autre part, dans un cadre général, développent des variantes de la méthode STATIS pour traiter des données bi-partitionnées, dans une optique de discrimination entre les individus.

La technique qui fait l'objet de cet article s'applique aussi à des données bi-partitionnées, dans une optique de discrimination entre les individus ; elle se situe dans la continuité des techniques d'analyse des tableaux multiples utilisant des variables auxiliaires et est basée à la fois sur le critère de discrimination de Fisher et sur le critère de Carroll de généralisation de l'analyse canonique.

Les notations sont définies dans la section 2 ; après avoir rappelé dans la section 3 les principes de l'analyse factorielle discriminante de Fisher et ceux de la généralisation de l'analyse canonique de Carroll (1968), une technique d'analyse de données bi-partitionnées, l'Analyse Factorielle Discriminante de Tableaux Multiples (AFD-TM), est proposée section 4 ; ses propriétés, exposées dans la section 5, permettent alors de définir des règles d'interprétation des résultats et il est montré, section 6, qu'en cas d'absence de partition des individus l'AFD-TM est une généralisation de l'analyse canonique et qu'en cas d'absence de partition des variables, l'AFD-TM est l'AFD. Une application sur des données simulées et bruitées est présentée section 7 ; le jeu de données utilisé par Escofier and Pagès (1994), puis ensuite par Sabatier et al. (2013) est utilisé section 8 pour illustrer cette technique, puis pour la comparer à STATIS-LDA, section 9. Enfin, une conclusion est donnée section 10.

 ISSN: 2102-6238

2. Notations

On considère K tableaux de données X_k pour $k = 1, \ldots, K$, dont les lignes sont les observations des mêmes n individus pour différents ensembles de variables ; la matrice X_k est de dimension $n \times m$. Chaque variable $X_{k,j}$ (la j-ème colonne de X_k) est centrée. m désigne la somme des m_k, soit $m = \sum_{k=1}^{K} m_k$. La matrice $X = \left[X_1, \ldots, X_K\right]$ de dimension $n \times m$ juxtapose les K tableaux de données.

W_k est l'espace engendré par les colonnes X_k et P_{W_k} est le projecteur orthogonal sur W_k : $P_{W_k} = X_k(X_k'X_k)^{-1}X_k'$. Pour simplifier les notations, nous supposerons que la dimension de chacun des espaces est égale à m_k et que les individus ont tous un même poids égal à $\frac{1}{n}$. Pour les mêmes raisons, A étant une matrice carrée, nous noterons A^{-1} l'inverse ou l'inverse généralisée de A, sans distinction.

Enfin, le tableau disjonctif complet G à n lignes et g colonnes décrit la partition des n individus en g groupes, W_G est l'espace engendré par les colonnes de G et P_{W_G} le projecteur sur cet espace.

Le tableau de données X est donc bi-partitionné : les lignes de ce tableaux décrivent n individus qui sont regroupés en g groupes et les colonnes de ce tableau comportent m variables qui constituent K ensembles de variables ; le tableau X se présente alors sous la forme suivante :

TABLE 1. *Partition des individus en groupes et des variables en blocs*

	m_1 variables	m_k variables	m_K variables
Groupe 1 n_1 individus			
	X_1	X_k	X_K
Groupe j n_j individus			
Groupe g n_g individus			

 ISSN: 2102-6238

3. L'analyse factorielle discriminante et l'analyse canonique généralisée

3.1. *Le pouvoir discriminant*

Le pouvoir discriminant d'une variable d_k mesure la capacité de d_k à «expliquer» une partition G des individus en groupes. Soit u_k un vecteur colonne à m_k lignes et $d_k = X_k u_k$ une variable synthétique définie pour les n individus. A partir de la partition de ces n individus en g groupes définie par G, il est possible de calculer la variance inter-groupes : les poids des groupes forment la diagonale de la matrice $D = \frac{1}{n} G'G$, les centres des groupes ont pour coordonnées $\frac{1}{n} D^{-1} G' d_k$ et par conséquent la variance inter-groupes est égale à $\frac{1}{n^2} d_k' G D^{-1} G' d_k$.

La variance totale est égale à $\frac{1}{n} d_k' d_k$ et donc le pouvoir discriminant μ_k est $\mu_k = \frac{d_k' G D^{-1} G' d_k}{n d_k' d_k}$, c'est-à-dire est égal au rapport entre la variance inter-groupes et la variance totale de la variable d_k.

Notons y la projection de d_k sur W_G, alors $y = P_{W_G} d_k = G(G'G)^{-1} G' d_k = G(nD)^{-1} G' d_k$ et il s'ensuit que $\mu_k = \frac{Var(y)}{Var(d_k)} = R^2(y, d_k)$.

3.2. *L'analyse factorielle discriminante de Fisher*

Considérons l'espace W engendré par les colonnes d'un tableau X et l'espace W_G engendré par les colonnes de G, alors l'analyse factorielle discriminante de Fisher (1936) détermine la variable $z^1 \in W$ ayant le pouvoir discriminant le plus élevé possible, puis la variable $z^2 \in W$ non corrélée à z^1 ayant le pouvoir discriminant le plus élevé, etc ... l'analyse s'arrêtant après $\inf(g-1, m)$ étapes. Les variables discriminantes sont obtenues de la manière suivante :

- à l'étape 1, on détermine $z^1 \in W$ de telle manière que le pouvoir discriminant de z^1, noté μ^1, soit maximal. Autrement dit, il s'agit de résoudre :

$$\begin{cases} \text{Max } \mu^1 = f(z^1) = \frac{z^{1'} G D^{-1} G' z^1}{n z^{1'} z^1} \\ \text{avec } z^1 \in W. \end{cases}$$

- à l'étape j, y^j désignant la projection de z^j sur W_G, on détermine $z^j \in W$ de telle manière que μ^j le pouvoir discriminant de z^j soit maximal :

$$\begin{cases} \text{Max } \mu^j = f(z^j) = \frac{z^{j'} G D^{-1} G' z^j}{n z^{j'} z^j} \\ \text{avec pours} < j, R(z^s, z^j) = 0 \\ \text{avec pour s} < j, R(y^s, y^j) = 0. \end{cases}$$

Comme $\mu^j = R^2(y^j, z^j)$ (section 3.1), ce problème de maximisation est équivalent à

$$\begin{cases} \text{Max } R^2(z^j, y^j) \\ \text{avec pour s} < \text{j }, R(z^s, z^j) = 0 \\ \text{avec pour s} < \text{j }, R(y^s, y^j) = 0. \end{cases}$$

Par conséquent, le couple (z^j, y^j) est le j-ième couple de variables canoniques issu de l'analyse canonique entre W et W_G.

 ISSN: 2102-6238

3.3. *La généralisation de l'analyse canonique de Carroll*

Si on considère deux tableaux de variables X_k et X_l engendrant respectivement les espaces W_k etW_l, l'analyse canonique entre ces deux tableaux consiste à déterminer à l'étape j le couple $R(z_k^j, z_l^j)$ solution de :

$$\begin{cases} \text{Max } R^2(z_k^j, z_l^j) \\ \text{avec } z_k^j \in W_k \text{ et pour } s < j, R(z_k^s, z_k^j) = 0 \\ \text{avec } z_k^l \in W_l \text{ et pour } s < j, R(z_l^s, z_l^j) = 0. \end{cases}$$

Carroll (1968) généralise l'analyse canonique à K tableaux en introduisant à chaque étape j une variable auxiliaire z^j solution de

$$\begin{cases} \text{Max } G(z^j, z_1^j, ..., z_K^j) = \sum_{k=1}^{K} R^2(z^j, z_k^j) \\ \text{avec pour s < j }, R(z^s, z^j) = 0. \end{cases}$$

z^j est alors le j-ième vecteur propre de $\sum_{k=1}^{K} P_{W_k}$ et z_k^j est la projection de z^j sur W_k (Carroll, 1968).

4. L'analyse factorielle discriminante de tableaux multiples

4.1. *Le principe*

On considère à la fois une partition des n individus en g classes et K tableaux de données. La technique proposée, l'analyse factorielle discriminante de tableaux multiples (AFD-TM) est dans la continuité des techniques d'analyse de tableaux multiples utilisant des variables auxiliaires (Carroll, 1968 ; Saporta, 1976 ; Casin, 1996 ; Casin, 2001 ; Jollife, 2002 ; Cazes, 2004) ; le critère utilisé pour mesurer l'homogénéité des groupes d'individus est celui de Fisher et celui de Carroll est utilisé pour mesurer l'homogénéité des blocs de variables.

A l'étape j, il s'agit de déterminer la variable auxiliaire normée z^j, $z^j \in R^n$, telle que ses projections orthogonales z_k^j sur ces sous-espaces $W_k, k = 1, \ldots, K$ soient à la fois :
-les plus liées linéairement entre elles, au sens du critère de Carroll,
-les plus discriminantes possible, au sens du critère de Fisher.

Afin d'obtenir à chaque étape une description du tableau X_k qui soit complémentaire de celle des étapes différentes, on impose que pour $j \neq l$, $R(z_k^j, z_k^l) = 0$ pour chacun des tableaux X_k, $k = 1, \ldots, K$.

Et y_k^j désignant la projection orthogonale de z_k^j sur W_G, pour que la capacité à discriminer des variables z_k^j soit complémentaire à l'étape j de celles des autres étapes, on impose aussi que si $j \neq l$, alors $R(y_k^j, y_k^l) = 0$.

 ISSN: 2102-6238

4.2. *La technique proposée*

L'AFD-TM procède par étapes successives.

A l'étape 1, il s'agit donc de déterminer la variable auxiliaire normée z^1, $z^1 \in R^n$, à la fois la plus corrélée possible avec les espaces $W_k, k = 1, \ldots, K$ et telle que ses projections orthogonales z_k^1 sur ces sous-espaces W_k, soient les plus discriminantes possible.

Autrement dit, à l'étape 1, il s'agit de déterminer z^1 solution de

$$\left\{ \begin{array}{c} \text{Max } H(z^1) = \sum_{k=1}^{K} \|P_{W_G}(P_{W_k}(z^1))\|^2 = \sum_{k=1}^{K} \|P_{W_G}(z_k^1)\|^2 = \sum_{k=1}^{K} \|y_k^1\|^2 \\ \text{avec } z^1 \in R^n \text{ et } \|z^1\|^2 = 1. \end{array} \right.$$

A l'étape 2, le problème est le même qu'à l'étape 1, mais on impose de plus des contraintes d'orthogonalisation, z_k^2 désignant la projection orthogonale de z^2 sur le sous-espace W_k et y_k^2 étant la projection orthogonale de z_k^2 sur W_G.

$$\left\{ \begin{array}{c} \text{Max } H(z^2) = \sum_{k=1}^{K} \|P_{W_G}(P_{W_k}(z^2))\|^2 = \sum_{k=1}^{K} \|P_{W_G}(z_k^2)\|^2 = \sum_{k=1}^{K} \|y_k^2\|^2 \\ \text{avec } \|z^2\|^2 = 1 \\ \text{avec } R(z_k^1, z_k^2) = 0 \text{ pour } k = 1, \ldots, K \\ \text{avec } R(y_k^1, y_k^2) = 0 \text{ pour } k = 1, \ldots, K. \end{array} \right.$$

Plus généralement, on construit, étape après étape, une base orthogonale de chaque espace W_k et K bases orthogonales de W_G.
z_k^j étant la projection orthogonale de z^j sur le sous-espace W_k et y_k^j la projection orthogonale de z_k^j sur W_G, à l'étape j, le problème est le suivant :

$$\left\{ \begin{array}{c} \text{Max } H(z^j) = \sum_{k=1}^{K} \|P_{W_G}(P_{W_k}(z^j))\|^2 = \sum_{k=1}^{K} \|P_{W_G}(z_k^j)\|^2 = \sum_{k=1}^{K} \|y_k^j\|^2 \\ \text{avec } \|z^j\|^2 = 1 \\ \text{avec } R(z_k^j, z_k^l) = 0 \text{ pour } k = 1, \ldots, K \text{ et pour } l = 1, \ldots, j-1 \\ \text{avec } R(y_k^j, y_k^l) = 0 \text{ pour } k = 1, \ldots, K \text{ et pour } l = 1, \ldots, j-1. \end{array} \right.$$

4.3. *La solution*

A l'étape 1, z^1 est le premier vecteur propre normé de XMX' avec

$$M = \begin{bmatrix} M_1 & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & M_K \end{bmatrix}$$

et $M_k = (X_k'X_k)^{-1}X_k'GD^{-1}G'X_k(X_k'X_k)^{-1}$.

 ISSN: 2102-6238

A l'étape j, $X_k^{(j)}$ est le tableau dont les colonnes sont les résidus de la régression des colonnes de X_k par les variables z_k^l pour $l = 1, \ldots, j-1$ et $X^{(j)}$ le tableau juxtaposant les $X_k^{(j)}$; $G_k^{(j)}$ désigne le tableau des résidus des colonnes du tableau G_k par les variables y_k^j, $k = 1, \ldots, K$.

Alors z^j est le premier vecteur propre normé de $X^{(j)}M^{(j)}X^{(j)'}$, où :

$$M^{(j)} = \begin{bmatrix} M_1^{(j)} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & M_K^{(j)} \end{bmatrix}$$

avec $M_k^{(j)} = (X_k^{(j)'}X_k^{(j)})^{-1}X_k^{(j)'}P_{G_k}^{(j)}X_k^{(j)}(X_k^{(j)'}X_k^{(j)})^{-1}$, $P_{G_k}^{(j)}$ étant le projecteur sur l'espace engendré par les colonnes de $G_k^{(j)}$.

4.4. Démonstration

A l'étape 1, il s'agit donc de déterminer z^1 maximisant $\sum_{k=1}^{K} \|P_{W_G}(P_{W_k}(z^1))\|^2$. En notant μ_k^1 le pouvoir discriminant de la variable z_k^1 :

$$\|P_{W_G}(P_{W_k}(z^1)\|^2 = \|P_{W_G}(z_k^1)\|^2 = \mu_k^1\|z_k^1\|^2 = \mu_k^1 R^2(z^1, z_k^1)$$

En effet, z_k^1 est la projection orthogonale de z^1 sur W_k et comme z^1 est normée, le cosinus carré de l'angle entre z_k^1 et z^1 est à la fois égal à $\text{var}(z_k^1)$ et à $R^2(z^1, z_k^1)$ et donc $R^2(z^1, z_k^1) = \text{var}(z_k^1) = \frac{1}{n}z_k^{1'}z_k^1$.

Et finalement, il s'agit donc de déterminer z^1 maximisant $\sum_{k=1}^{K} R^2(z^1, z_k^1)\mu_k^1$.

Or $z_k^1 = P_k z^1 = X_k(X_k'X_k)^{-1}X_k'z^1$.

Comme $\mu_k^1 = \frac{z_k^{1'}GD^{-1}G'z_k^1}{nz_k^{1'}z_k^1}$

alors $\mu_k^1 = \frac{z^{1'}X_k(X_k'X_k)^{-1}X_k'GD^{-1}G'X_k(X_k'X_k)^{-1}X_k'z^1}{nz_k^{1'}z_k^1}$ et donc :

$$\mu_k^1 R^2(z^1, z_k^1) = \frac{1}{n^2}z^{1'}X_k(X_k'X_k)^{-1}X_k'GD^{-1}G'X_k(X_k'X_k)^{-1}X_k'z^1$$

avec $M_k = (X_k'X_k)^{-1}X_k'GD^{-1}G'X_k(X_k'X_k)^{-1}$.

Alors $\sum_{k=1}^{K} \mu_k^1 R^2(z^1, z_k^1) = \frac{1}{n^2}z^{1'}\left(\sum_{k=1}^{K} X_k M_k X_k'\right)z^1$

Soit

$$M = \begin{bmatrix} M_1 & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & M_K \end{bmatrix}$$

Alors $\left(\sum_{k=1}^{K}(X_k M_k X_k')\right) = XMX'$ et par conséquent z^1 est le premier vecteur propre normé de XMX'. La technique proposée est donc la première étape d'une ACP dans la métrique M décrite ci-dessus.

 ISSN: 2102-6238

A l'étape 2, écrire la contrainte $R(z_k^2, z_k^1) = 0$ est équivalent à écrire que les variables z_k^2 sont des combinaisons linéaires des colonnes du tableau $X_k^{(2)}$, dont les colonnes sont les résidus de la régression des colonnes du tableau X_k par la variable z_k^1. En effet, $R(z_k^2, z_k^1) = 0$ signifie que z_k^2 appartient au sous-espace de W_k orthogonal à z_k^1. Ce sous-espace est engendré par les résidus de la régression des variables du tableau X_k par z_k^1, et donc z_k^2 est une combinaison linéaire de ces résidus.

De la même manière la contrainte $R(y_k^2, y_k^1) = 0$ signifie que les variables y_k^2 sont des combinaisons linéaires des colonnes des tableaux $G_k^{(2)}$ pour $k = 1, \ldots, K$ obtenus en régressant les colonnes de G par y_k^1.

Dès lors, on procède de la même manière qu'à l'étape 1, en substituant aux tableaux X_k les tableaux $X_k^{(2)}$ pour $k = 1, \ldots, K$ et au tableau G les tableaux $G_k^{(2)}$ pour $k = 1, \ldots, K$. μ_k^2 désigne le pouvoir discriminant de la variable z_k^2.

La maximisation de $\sum_{k=1}^{K} R^2(z^2, z_k^2)\mu_k^2$ conduit donc à déterminer z^2, qui est le premier vecteur propre normé de $X^{(2)}M^{(2)}X^{(2)'}$, expression dans laquelle $X^{(2)}$ et $M^{(2)}$ sont obtenus en remplaçant les tableaux X_k et G respectivement par les tableaux $X_k^{(2)}$ et par $G_k^{(2)}$ pour $k = 1, \ldots, K$ dans X et M :

Soit

$$M^{(2)} = \begin{bmatrix} M_1^{(2)} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & M_K^{(2)} \end{bmatrix}$$

avec $M_k^{(2)} = (X_k^{(2)'}X_k^{(2)})^{-1}X_k^{(2)'}\, P_{G_k}^{(2)} X_k^{(2)} (X_k^{(2)'}X_k^{(2)})^{-1}$, $P_{G_k}^{(2)}$ étant le projecteur sur l'espace engendré par les colonnes de $G_k^{(2)}$ pour $k = 1, \ldots, K$.

A l'étape j, z_k^j est orthogonale à $z_k^1, \ldots, z_k^{j-1}$. De façon itérative, on détermine ainsi une base orthogonale pour chacun des espaces W_k, $k = 1, \ldots, K$ et pour chacun des espaces $G_k^{(j)}$ pour $k = 1, \ldots, K$. Soit donc $X_k^{(j)}$ le tableau dont les colonnes sont les résidus des colonnes de X_k par $z_k^1, \ldots, z_k^{j-1}$ et $G_k^{(j)}$ le tableau dont les colonnes sont les résidus des colonnes de G par $y_k^1, \ldots, y_k^{j-1}$. A l'étape j, z^j est donc le premier vecteur propre de $X^{(j)}M^{(j)}X^{(j)'}$, $X^{(j)}$ étant le tableau juxtaposant les $X_k^{(j)}$ et

$$M^{(j)} = \begin{bmatrix} M_1^{(j)} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & M_K^{(j)} \end{bmatrix}$$

avec $M_k^{(j)} = (X_k^{(j)'}X_k^{(j)})^{-1}X_k^{(j)'} P_{G_k}^{(j)} X_k^{(j)} (X_k^{(j)'}X_k^{(j)})^{-1}$, $P_{G_k}^{(j)}$ étant le projecteur sur l'espace engendré par les colonnes de $G_k^{(j)}$.

 ISSN: 2102-6238

4.5. Le nombre maximal d'étapes

Lorsque l'on considère un seul tableau X_k de rang m_k, le nombre maximal d'étapes de l'AFD est $\inf(g-1, m_k)$. Le nombre maximal d'étapes de l'AFD-TM est donc égal à $netap = \inf(g-1, max_k(m_k))$. Mais pour les tableaux pour lesquels $m_k < max_k(m_k)$, les calculs s'arrêtent à l'étape m_k, puisque l'espace W_k est alors complètement décrit par la base $z_k^1, \ldots, z_k^{m_k}$.

4.6. L'algorithme de calcul

L'algorithme de calcul est alors le suivant :

1. Calcul de z^1, premier vecteur propre de XMX'
2. Détermination de z_k^1 , pour $k = 1, \ldots, K$ par projection de z^1 sur W_k ; on en déduit les valeurs des $R^2(z^1, z_k^1)$
3. Détermination de y_k^1 pour $k = 1, \ldots, K$ par projection de z_k^1 sur l'espace W_G ; on en déduit les valeurs des μ_k^1
4. L'algorithme s'arrête pour les tableaux dont le rang est égal à 1 ; en effet, si un tableau est de rang 1, le vecteur z_k^2 est le vecteur nul puisque la dimension de W_k est égale à 1. Le tableau concerné est alors ôté de l'algorithme de calcul et les calculs se poursuivent sans tenir compte de ce tableau. Si G ne décrit que 2 groupes, l'algorithme s'arrête et les calculs sont terminés.
5. Détermination des tableaux $X_k^{(2)}$ (tableau des résidus des régressions des colonnes de X_k par z_k^1) et $G_k^{(2)}$ (tableau des résidus des régressions des colonnes du tableau G par y_k^1).
6. Calcul de z^2, premier vecteur propre de $X^{(2)}M^{(2)}X^{(2)'}$, $X^{(2)}$ étant le tableau juxtaposant les $X_k^{(2)}$ et
$$M^{(2)} = \begin{bmatrix} M_1^{(2)} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & M_K^{(2)} \end{bmatrix}$$
avec $M_k^{(2)} = (X_k^{(2)'}X_k^{(2)})^{-1}X_k^{(2)'}P_{G_k}^{(2)}X_k^{(2)'}(X_k^{(2)'}X_k^{(2)})^{-1}$, $P_{G_k}^{(2)}$ étant le projecteur sur l'espace engendré par les colonnes de $G_k^{(2)}$
7. L'algorithme se poursuit de la même manière pour $\inf(g-1, max_k(m_k))$ étapes.

5. L'interprétation des résultats de l'AFD-TM

Cette interprétation se fait à partir de deux séries d'indicateurs et de deux séries de graphiques, comme dans le cas d'autres techniques de description de tableaux multiples bâties sur le même principe que l'AFD-TM, à savoir une suite de premières étapes d'ACP dans des métriques variant à chaque étape (voir Casin, 1996 et Casin, 2001).

 ISSN: 2102-6238

5.1. Les indicateurs

On considère, pour chaque étape j, les deux séries d'indicateurs suivantes :

- -la corrélation de chacune des variables z_k^j avec la variable z^j exprimée par $R^2(z^j, z_k^j)$, indicateur du lien entre la variable calculée à l'étape j pour le tableau k et les autres tableaux : c'est le critère de Carroll ;
- -le pouvoir discriminant de chaque variable z_k^j , soit μ_k^j , qui mesure la qualité de la discrimination à l'étape j pour le tableau k : c'est le critère de Fisher.

5.2. Les propriétés de l'AFD-TM

Les propriétés suivantes de l'analyse en composantes principales généralisée (Casin, 2001) sont aussi vérifiées pour l'AFD-TM. Elles sont rappelées ici car elles justifient les représentations graphiques utilisées.

Propriété 1. *Pour $r > j$ la v-ième colonne de X_k^r, $X_{k,v}^r$ est orthogonale à z^j .*

En effet, z_k^j est la projection de z^j sur le sous espace de W_k orthogonal à z_k^1,...,z_k^{j-1} , $z^j = z_k^j + e$, e étant orthogonal à tout vecteur de ce sous-espace et donc orthogonal à $X_{k,v}^r$. Comme du fait de la contrainte d'orthogonalisation, $X_{k,v}^r$ est orthogonale à z_k^j , $X_{k,v}^r$ est orthogonale à z^j .

Propriété 2. *Pour $r > j$, z_k^r est orthogonale à z^j .*

Comme z_k^r est une combinaison linéaire des variables $X_{k,v}^r$, z_k^r est orthogonale à z^j en vertu de la propriété 1.

Propriété 3. *Les variables z^j sont deux à deux orthogonales.*

z^r est une combinaison linéaire des $X_{k,v}^r$, qui sont orthogonaux à z^j (propriété 1) , et si $r > j$,z^r est donc orthogonal à z^j .

5.3. Les graphiques de l'AFD-TM

Ces propriétés permettent donc de disposer de deux séries de graphiques :

- -la première série obtenue, pour les étapes j et j', en croisant les deux variables non corrélées entre elles z^j et $z^{j'}$, qui permet de représenter l'ensemble des variables sur un même graphique d'une part, l'ensemble des individus d'autre part sur un même graphique ;
- -la seconde série obtenue, pour les étapes j et j', et le tableau k, en croisant les deux variables non corrélées entre elles z_k^j et $z_k^{j'}$, et qui permet de retrouver les sorties classiques de l'AFD de Fisher.

6. Remarques

6.1. Les cas particuliers de l'AFD-TM

Lorsqu'il n'y a qu'un seul individu dans chaque groupe d'individus, alors la variance intra-groupes est nulle et le pouvoir discriminant de toute variable est égal à 1 ; dès lors, maximiser à l'étape j

 ISSN: 2102-6238

$\sum_{k=1}^{K} R^2(z^j, z_k^j)\mu_k^j$ revient à maximiser $\sum_{k=1}^{K} R^2(z^j, z_k^j)$; compte tenu des contraintes d'orthogonalisation entre les variables z_k^j, la technique obtenue est alors la généralisation de l'analyse canonique décrite dans Casin (1996). Si de surcroit, il n'y a que K=2 tableaux, alors l'AFD-TM est l'analyse canonique entre ces deux tableaux.

Lorsqu'il n' y a qu'un seul tableau, alors K=1 et $R^2(z^j, z_1^j)$ =1, car z^j se confond avec z_1^j, et il s'agit donc à l'étape j de maximiser μ_1^j, c'est- à- dire d'effectuer l'analyse discriminante du tableau X_1, la partition des individus étant donnée par G.

Ainsi, l'AFD-TM est une technique d'analyse de tableaux partitionnés qui se ramène à une analyse canonique généralisée lorsqu'il existe une partition des variables et pas de partition des individus, et à une analyse discriminante dans le cas où existe une partition des individus et pas de partition des variables.

6.2. Chaque tableau est composé d'une seule variable qualitative

Lorsque chacun des K tableaux de données X_k pour $k = 1, \ldots, K$ est constitué des indicatrices d'une seule variable qualitative, alors l'AFD-TM est une technique de comparaison des tableaux de contingence $X_k^{'}G$ dont les colonnes décrivent une même variable qualitative G (voir pour ce type de technique (Zarraga and Goitisolo, 2009). La variable en ligne peut être différente pour chaque tableau k, ou il peut s'agir de la même variable observée dans des situations différentes.

6.3. Les contraintes d'orthogonalisation

Les contraintes d'orthogonalisation sont doubles puisque l'on recherche une base orthogonale pour chaque espace W_k d'une part, et K bases orthogonales pour W_G, chaque base étant associée à un tableau K. Il est évidemment possible de construire des variantes de l'AFD-TM, soit en abandonnant les contraintes d'orthogonalisation dans W_G, ou encore en ne posant qu'une contrainte d'orthogonalisation globale pour les tableaux (c'est-à-dire, une contrainte portant sur les variables z^j, comme dans l'Analyse Canonique Généralisée de Carroll, 1968) ; le problème devient alors :

$$\begin{cases} \text{Max } I(z^j) = \sum_{k=1}^{K} \|P_{W_G}(P_{W_k}(z^j))\|^2 = \sum_{k=1}^{K} \|P_{W_G}(z_k^j)\|^2 = \sum_{k=1}^{K} \|y_k^j\|^2 \\ \text{avec } \|z^j\|^2 = 1 \\ \text{avec } R(z^j, z^r) = 0 \text{ pour } r = 1, \ldots, j-1. \end{cases}$$

Les variables z^j successives sont alors les vecteurs propres successifs de XMX'. Cette contrainte est moins forte que celle de l'AFD-TM (puisque l'AFD-TM fournit aussi des variables z^j deux à deux orthogonales, propriété 3, section 5.2) ; on perd cependant la représentation des variables et des individus pour chacun des tableaux qui était obtenue avec l'AFD-TM.

 ISSN: 2102-6238

7. Un exemple d'application sur données simulées

On considère 5 tableaux de tailles différentes comportant respectivement 6, 2, 8, 4 et 2 variables centrées et normées observés sur les mêmes 100 individus. Le tableau 1 caractérise 4 groupes bien distincts (groupe 1 : individus 1 à 25, groupe 2 : individus : 26 à 50, groupe 3 : individus 51 à 75, et groupe 4 : individus 76 à 100), le tableau 2 caractérise le groupe 4, les autres groupes étant confondus. En ce qui concerne le tableau 3, les groupes 1 et 2 d'une part, et les groupes 3 et 4 d'autre part sont confondus. Pour le tableau 4, 3 groupes sont constitués, qui chevauchent les 4 groupes du tableau 1 : le premier groupe comporte les individus 1 à 33, le deuxième les individus 34 à 67, et le troisième les individus 68 à 100. Le tableau 5, enfin, ne met en évidence aucun groupe.

La validation croisée est obtenue en formant aléatoirement 10 groupes d'observations de taille identique ; chacun de ces 10 groupes est utilisé successivement comme échantillon test, la règle de décision étant établie à partir de l'échantillon d'apprentissage formé par les 9 autres groupes. Le taux de de bien classés est alors de 71 %. Après bruitage des données par une loi normale d'espérance nulle et d'espérance 0.2, ce taux de bien classés est de 69,9 % pour cent en moyenne pour 100 échantillons bruités.

Les tableaux ci-dessous donnent la valeur des différents paramètres de l'AFD-TM pour les étapes 1 et 2 ; pour chaque paramètre sont donnés la valeur sur le jeu de données initial, puis le maximum et le minimum calculés pour les 100 simulations.

TABLE 2. *Données simulées, résultats pour l'étape 1*

	Tableau 1	Tableau 2	Tableau 3	Tableau 4	Tableau 5
$R^2(z^1, z_k^1)$	.88 (.93 .82)	.26 (.42 .07)	.86 (.90 .80)	.28 (.44 .11)	.00 (.18 .00)
μ_k^1	.84 (.91 .78)	.34 (.41 .19)	.72 (.76 .65)	.23 (.38 .14)	.15 (.22 .06)

TABLE 3. *Données simulées, résultats pour l'étape 2*

	Tableau 1	Tableau 2	Tableau 3	Tableau 4	Tableau 5
$R^2(z^2, z_k^2)$	.13 (.21 .00)	.00 (.04 .00)	.99 (.99 .98)	.04 (.11 .00)	.04 (.05 .00)
μ_k^2	.03 (.08 .00)	.03 (.05 .00)	.27 (.38 .18)	.00 (.04 .00)	.00 (.04 .00)

L'AFD-TM met en évidence, à l'étape 1, la structure en 4 groupes du tableau 1, strucure que l'AFD-TM retrouve en partie dans le tableau 3 et dans une moindre mesure dans les tableaux 2 et 4. En effet, la corrélation avec les tableaux 1 et 3 est forte, moyenne pour le tableau 4 et quasi-nulle pour les deux autres tableaux ; la discrimination est forte pour le tableau 1 (il y a effectivement 4 groupes bien distincts) et pour le tableau 3 (deux groupes bien distincts, regroupant deux à deux les groupes du tableau 1), plus faible pour le tableau 2 (où seul le groupe 4 est caractérisé) et le tableau 4 (les 3 groupes sont une version "floutée" des 4 groupes du tableau 1), faible voire nulle pour les tableau 5. L'étape 2 met en évidence l'opposition entre les groupes 1 et 2, et entre les groupes 3 et 4, qui est décrite par le tableau 3.

Les résultats sont stables lorsque les données sont bruitées. Les plus fortes variations des paramètres (corrélation et discrimination) sont enregistrées pour des valeurs de ces paramètres

 ISSN: 2102-6238

faibles ou moyennes, mais ne remettent pas en cause les conclusions de l'AFD-TM, c'est-à-dire la détection des 4 groupes à l'étape 1 et le regroupement en 2 groupes à l'étape 2. De la même façon, le bruitage a peu de conséquences sur le taux de bien classés obtenu après validation croisée.

8. Un exemple d'application sur données réelles

8.1. Les données

Le jeu de données provient du centre de recherches de l'INRA d'Angers et a été utilisé plusieurs fois par Escofier and Pagès (1994) ; il est constitué par 5 tableaux de données ($K = 5$) observés sur 21 observations ($n = 21$). Ces observations caractérisent des vins de trois origines différentes ($g = 3$) : Saumur (11 observations), Chinon (4 observations) et Bourgueil (6 observations).

Les 5 tableaux comportent respectivement 5 variables (Olfaction avant agitation), 3 variables (Vision), 10 variables (Olfaction après agitation), 9 variables (Goût) et 2 variables (Jugement global). Par conséquent, $m = 29$.

L'objectif d'Escofier and Pagès (1994) est de " mettre en évidence les principales dimensions de la variabilité sensorielle des vins et de relier ces dimensions avec le type de terroir" et pour ce faire, une Analyse Factorielle Multiple (AFM) des 5 tableaux de données est effectuée, et ensuite la variable qualitative caractérisant les différents terroirs est projetée sur l'espace des premiers facteurs. L'AFM étant une technique d'analyse de données multi-blocs, l'aspect multi-groupes (c'est-à-dire la partition des individus en groupes) n'est pas pris en compte pour la détermination des variables synthétiques que l'AFM calcule et n'est introduit qu'à titre illustratif.

Ce jeu de données est disponible notamment dans le package R consacré à l'analyse des données Factominer (Husson et al., 2010) et est traité, à titre d'exemple d'application, dans la documentation de Factominer. Il est aussi utilisé par Sabatier et al. (2013) pour illustrer la technique d'analyse de tableaux bi-partionnés, appelée STATIS-LDA, qu'ils proposent. Utiliser ce jeu de données permet donc de comparer STATIS-LDA et l'AFD-TM sur un jeu de données.

Ici, l'AFD-TM détermine $netap = 2$ variables synthétiques par tableau (car il y a 3 groupes, et chaque groupe comporte plus d'une seule variable) le plus liées linéairement entre elles (autrement dit "mettre en évidence les principales dimensions communes de ces tableaux"), tout en tenant compte le mieux possible de la partition des variables en groupes (autrement dit, "reliant les principales dimensions aux différents types de terroirs").

8.2. Les résultats numériques

Ces résultats, comme ceux effectués pour données simulées de la section précédente, ont été obtenus grâce au logiciel d'économétrie Eviews.

A l'étape 1, la première valeur propre est $\lambda_1 = \sum_{k=1}^{K} R^2(z^1, z_k^1)\mu_k^1 = 1.526$.

Plus précisément, on obtient les résultats suivants :

 ISSN: 2102-6238

TABLE 4. *Données réelles, résultats pour l'étape 1*

	Tableau 1	Tableau 2	Tableau 3	Tableau 4	Tableau 5
$R^2(z^1, z_k^1)$	0.841	0.349	0.907	0.598	0.058
μ_k^1	0.702	0.313	0.706	0.300	0.107

La première étape met en évidence une proximité entre le tableau 1 et le tableau 3 (valeurs élevées du coefficient de détermination), et dans une moindre mesure avec le tableau 4, associée à un pouvoir discriminant assez fort des variables z_1^1 et z_3^1 . On note ici que le pouvoir discriminant de z_4^1 est peu élevé. Les deux tableaux 1 et 3 ont donc en évidence une caractéristique commune associée à un pouvoir discriminant élevé, qui ne se retrouve pas dans les autres tableaux.

La recherche peut ici être complétée en effectuant l'AFD-TM des seuls tableaux 1 et 3 : la première valeur propre λ_1 est alors égale à 1.29, les coefficients de corrélation $R^2(z^1, z_1^1)$ et $R^2(z^1, z_2^1)$ sont égaux respectivement à 0.871 et 0.923, tandis que $\mu_1^1 = 0.710$ et $\mu_2^1 = 0.733$. L'AFD-TM a ainsi mis en évidence une variable synthétique très liée aux deux tableaux 1 et 3 et ayant un fort pouvoir discriminant.

A l'étape 2, on a $\lambda_2 = \sum_{k=1}^{K} R^2(z^2, z_k^2)\mu_k^2 = 0.8095$.

Plus précisément, les résultats de l'étape deux sont les suivants :

TABLE 5. *Données réelles, résultats pour l'étape 2*

	Tableau 1	Tableau 2	Tableau 3	Tableau 4	Tableau 5
$R^2(z^2, z_k^2)$	0.724	0.009	0.539	0.762	0.164
μ_k^2	0.484	0.002	0.215	0.448	0.013

C'est ici une relation entre deux variables synthétiques, l'une issue du tableau 4 et l'autre du tableau 1, qui est mise en évidence. Le pouvoir discriminant de ces deux variables est cependant assez peu élevé. Les tableaux 2 et 5 sont peu utiles dans l'analyse ; ceci est cohérent avec leurs faibles pouvoirs de discrimination (calculés dans Sabatier et al., 2013). Les résultats de cette seconde étape suggèrent d'effectuer une AFD-TM à partir des seuls tableaux 1 et 4. La première valeur propre de cette AFD-TM est $\lambda_1 = 0.901$. $R^2(z^1, z_1^1)$ et $R^2(z^1, z_4^1)$ sont égaux respectivement à 0.928 et 0.646, et $\mu_1^1 = 0.706$ et $\mu_4^1 = 0.381$. Le pouvoir discriminant associé au tableau 4 est donc assez faible et la variable est donc peu pertinente.

On forme aléatoirement, et 100 fois, 10 groupes d'observations de taille identique ; chacun de ces 10 groupes est utilisé successivement comme échantillon test, la règle de classement étant établie à partir de l'échantillon formé par les 9 autres groupes. Le taux de de bien classés est alors de 48,3 %.

8.3. Les graphiques

Plusieurs graphiques complètent l'analyse.

Le premier graphique permet de visualiser les corrélations entre les variables z^1 et z^2 et les variables z_k^j, pour $j = 1, 2$ et $k = 1, \ldots 5$ et confirme la forte corrélation de la variable z^1 avec ses

 ISSN: 2102-6238

projections sur les espaces engendrés par les tableaux 1 et 3, et dans une moindre mesure avec le tableau 4. Sur ce graphique on notera aussi que toutes les variables z_k^2 sont alignées dans l'axe de la variable z^2 : en effet, par construction les variables z_k^2 sont non corrélées avec la variable z^1 et leur position sur ce graphique ne renseigne donc que sur l'intensité de leur corrélation avec z^2.

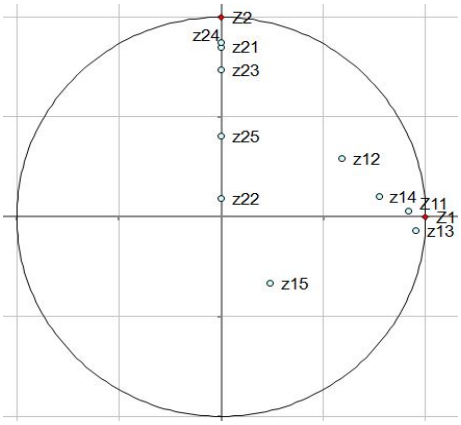

FIGURE 1. *Corrélations entre les variables z^1 et z^2 et les variables z_k^j, (notées sur le graphique 1 zkj)*

Le deuxième graphique croise les individus dans un repère dont l'abscisse est z^1 et l'ordonnée z^2 et décrit une bonne séparation des trois groupes.

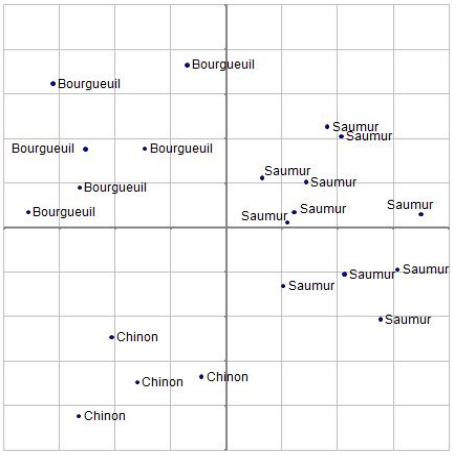

FIGURE 2. *Individus dans le repère de variables z^1 (horizontal) et z^2 (vertical)*

Les variables sont alors représentées classiquement à l'intérieur d'un cercle des corrélations (graphique 3), qui met notamment en évidence les variables des tableaux 1, 3 et 4 les plus liées à z^1 et –mais les corrélations sont beaucoup plus faibles- à z^2.

Enfin, l'AFD-TM permet de représenter les individus et les variables du tableau k dans un

 ISSN: 2102-6238

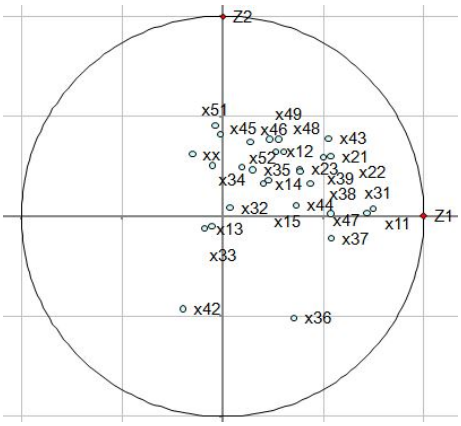

FIGURE 3. *Corrélations entre les variables z^1 et z^2 et les variables de départ (xkj désigne la variable j du tableau k)*

repère constitué des deux variables non corrélées entre elles z_k^1 et z_k^2. Ces deux variables étant dans l'espace W_k permettent une meilleure visualisation des individus et des variables que par projection sur l'espace engendré par z^1 et z^2.

Pour $k = 2$, le pouvoir discriminant de 0.313 obtenu à l'axe 1 semble ici correspondre à une assez bonne séparation de Chinon et Saumur sur l'axe 1.

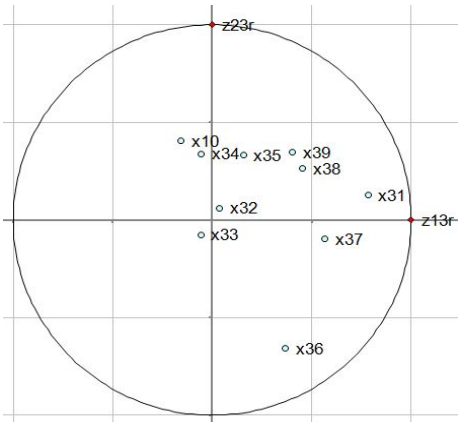

FIGURE 4. *Individus dans le repère de variables z_3^1 (horizontal) et z_3^2 (vertical)*

En ce qui concerne les variables, l'axe 1 est bien corrélé avec la variable 1, et dans une moindre mesure avec la variable 7 du tableau 3.

9. Comparaison avec STATIS-LDA

9.1. La méthode STATIS-LDA

Il s'agit de la technique STATIS (Lavit, 1988 ; Lavit et al., 1994), qui est appliquée ici aux K tableaux de centres de gravité des groupes $C_k = (\frac{1}{n}G'G)^{-1}G'X_k$ (Sabatier et al., 2013), chacun de

 ISSN: 2102-6238

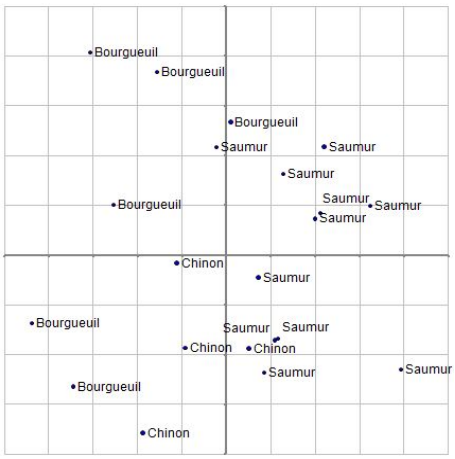

FIGURE 5. *Corrélations entre les variables du tableau 3 et les variables* z_3^1 *et* z_3^2

ces tableaux étant doté de la métrique $D_k = (\frac{1}{n}X_k'X_k)^{-1}$. Autrement dit, à chaque tableau X_k est associé un opérateur

$$O_k = \frac{1}{n}C_k D_k C_k'$$

soit

$$O_k = \frac{1}{n}(\frac{1}{n}G'G)^{-1}G'X_k(\frac{1}{n}X_k'X_k)^{-1}X_k'G(\frac{1}{n}G'G)^{-1}$$

.

Après avoir défini le produit scalaire entre opérateurs

$$\langle O_k, O_{k^*}\rangle = trace(O_k O_{k^*})$$

STATIS-LDA considère la matrice carrée de dimension K des produits scalaires des K tableaux entre eux, notée C ; le calcul des premiers vecteurs propres de cette matrice C permet la représentation des opérateurs dans un espace de dimension faible. Cette étape est celle de l'interstructure. La seconde étape de STATIS-LDA est celle du compromis ; les produits scalaires $\langle O_k, O_{k^*}\rangle$ étant tous positifs, la matrice C a tous ses éléments positifs, donc son premier vecteur propre $v = (v_1, \ldots, v_k, \ldots, v_K)$ a tous ses éléments positifs (théorème de Frobénius) ; STATIS-LDA procède alors à la diagonalisation de la matrice compromis $T = \sum_{k=1}^{K} v_k O_k$.

Le traitement des données de la section 6 a aussi été réalisé par Sabatier et al. (2013) et donc des comparaisons peuvent être établies avec le traitement effectué avec l'AFD-TM.

STATIS-LDA fournit deux types de représentation :
- une représentation des opérateurs lors de l'étape de l'interstructure,
- une représentation des individus et des variables lors de l'étape du compromis.

 ISSN: 2102-6238

9.2. *Les représentations des tableaux*

Contrairement à STATIS-LDA, l'AFD-TM ne fournit pas de représentation globale des proximités entre les opérateurs, c'est-à-dire entre les tableaux ; mais, à chaque étape de calcul j, l'AFD-TM mesure la proximité entre les phénomènes mis en avant à cette étape, en calculant les paramètres $R^2(z^k, z_j^k)$ et mesure la capacité à discriminer la variable G en calculant les valeurs μ_k^j.

Ainsi, STATIS-LDA met en évidence lors de sa première étape l'importance des tableaux 1, 3 et 4, et la proximité entre les tableaux 1 et 4 ; l'AFD-TM met en évidence lors de la première étape une forte proximité entre z_1^1 et z_3^1, donc entre les tableaux 1 et 3, associée à un pouvoir discriminant assez élevé, et, à la deuxième étape, une proximité entre les tableaux 1 et 4, mais associée à un pouvoir de discrimination moins élevé.
Dans le deux cas, ce sont les proximités entre les mêmes tableaux 1, 3 et 4 que les analyses mettent en évidence, mais de façon globale pour STATIS-LDA, et étape après étape pour l'AFD-TM. L'AFD-TM montre aussi que dans le cas des tableaux 1 et 4, cette proximité n'est pas associée à un pouvoir discriminant important du tableau 4 et n'est donc pas pertinente au regard des objectifs poursuivis.

En conclusion de ce paragraphe, la représentation des tableaux est basée sur un indicateur global de proximité entre tableaux en ce qui concerne STATI-LDA (le produit scalaire entre deux tableaux, égal à la trace du produit des opérateurs associés à ces tableaux) alors que pour l'AFD-TM, la proximité est mesurée, étape après étape, à la fois par la corrélation entre les variables synthétiques z_k^j et par le pouvoir discriminant de ces variables z_k^j.

9.3. *Les représentations des individus et des variables*

Les représentations des individus et des variables fournies par STATIS-LDA sont obtenues à partir des premiers axes de l'opérateur compromis T (Sabatier et al., 2013). L'AFD-TM fournit deux types de représentations pour les individus et les variables.
- La première de ces représentations permet de mettre en évidence les aspects communs à plusieurs tableaux, dans un espace commun à l'ensemble de ces tableaux, (à partir des variables z^j ; section 6.3 : graphiques 2 et 3) et est donc du même type que celle fournie par STATIS-LDA.
- La seconde de ces représentations permet la description complète de chaque tableau k (section 6.3 : graphiques 4 et 5), et donc des particularités de chacun des tableaux (à partir des variables z_k^j).

L'AFD-TM fournit une base complète pour chacun des espaces W_k décrits par les tableaux X_k, ce que ne fait pas STATIS-LDA.

9.4. *Le taux de bien classés*

En ce qui concerne STATIS-LDA, Sabatier et al. (2013) obtiennent un taux de bien classés de 52,4 %, alors que le taux de bien classés de l'AFD-TM est de 48,3 %. Mais si on utilise une variante de l'AFD-TM (section 6.3) qui ne considère qu'une contrainte d'orthogonalisation globale portant sur les variables z^j, comme c'est le cas pour STATIS-LDA, ce taux de bien classés obtenu par validation croisée augmente et passe à 54,2 % ; on perd cependant la possibilité d'avoir une représentation des individus et des variables de chaque tableau à partir des variables z_k^j.

 ISSN: 2102-6238

10. Conclusion

L'AFD-TM cherche à mettre en évidence un même phénomène, retrouvé dans différents tableaux, et ayant un bon pouvoir discriminant relativement à une partition des individus donnée. A contrario, il sera alors possible de repérer les tableaux dans lesquels ce phénomène n'apparait pas, et qui décrivent donc d'autres comportements que l'on ne retrouve pas dans les autres tableaux. La technique proposée apparait donc comme étant à mi-chemin entre une généralisation de l'analyse canonique (on recherche des proximités entre des combinaisons linéaires de variables issues de tableaux différents) et des analyses factorielles discriminantes effectuées indépendamment sur chacun des tableaux. L'AFD-TM peut s'étendre à des cas où les tableaux sont des indicatrices de variables qualitatives ou des ensembles mélangeant variables qualitatives et quantitatives.

Remerciements

Je tiens à remercier tout particulièrement Monsieur Pierre CAZES pour les conseils et suggestions qu'il m'a prodigués et qui ont permis d'améliorer une première version de ce travail, les rapporteurs anonymes pour leurs critiques constructives, et ma fille Marion CASIN, étudiante à l'ISUP-UPMC pour son aide technique.

 ISSN: 2102-6238

Références

Carroll, J.-D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th annual convention of the Americain Psychological Association*, 3 :227–228.

Casin, P. (1996). L'analyse en composantes principales généralisée. *Revue de statistique appliquée*, XLIV (3) :63–81.

Casin, P. (2001). A generalization of principal components analysis to k sets of variables. *Computational Statistics and Data Analysis*, 35 :417–428.

Cazes, P. (2004). Quelques méthodes d'analyse factorielle d'une série de tableaux de données. *Modulad*, 31.2004 :1–31.

Escofier, B. and Pagès, J. (1994). Multiple factor analysis (afmult package). *Computational Statistics and Data Analysis*, 18.1994 :121–140.

Eslami, A., Qannari, E., Kohler, A., and Bougeard, S. (2013). Analyse factorielle de données structurées en groupes d'individus. *Journal de la Société Française de Statistique*, 154(3) :44–57.

Eslami, A., Qannari, E., Kohler, A., and Bougeard, S. (2014). Multivariate analysis of multiblock and multigroup data. *Chemometrics and Intelligent Laboratory Systems*, 133 :63–69.

Fisher, R.-A. (1936). The use of multiple measurements in taxonomics problems. *Annals of eugenics*, 7-(2) :179–188.

Gardner, S., Gower, J., and Le Roux, N.-C. (2006). A synthesis of canonical variate analysis, generalised canonical correlation and procustes analysis. *Computational Statistics and Data Analysis*, 50 :107–134.

Hotelling, H. (1936). Relations between two sets of variants. *Biometrika*, 28 :321–337.

Husson, F., Josse, J., Le, S., and Mazet, J. (2010). *FactorMineR : Mutivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.14.

Jollife, I.-T. (2002). *Principal Component Analysis*. Springer.

Kang, M., Kim, D.-C., Liu, C., and Gao, J. (2015). Multiblock discriminant analysis for integrative genomic study. *Biomed Research International*, pages 1–10.

Kettenring, J.-R. (1971). Canonical analysis of several sets of variables. *Biometrika*, 58 (3) :333–351.

Krzysko, M., Smialowki, T., and Wolinsky, W. (2014). Analysis of multivariate repeated measures data using a manova model and principal components. *Biometrical Letters*, 51 :103–124.

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

Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (1994). The act (statis method). *Computational Statistics and Data Analysis*, 18 :97–119.

Louwerse, D., Tates, A., Smilde, A., Koot, G., and Berndt, H. (1999). Pls discriminant analysis with contribution plots to determine differences between parallel batch reactorsin the process industry. *Chemometrics and Intelligent Laboratory Systems*, 46 :197–206.

Morand, E. and Pagès, J. (2006). Procustes multiple factor analysis to analyse the overall perception of food products. *Food quality and preference*, 17 :36–42.

Sabatier, R., Vivien, M., and Reynès, C. (2013). Une nouvelle proposition, l'analyse discriminante multitableaux : Statis-lda. *Journal de la Société Française de Statistique*, 154 :31–43.

Saporta, G. (1976). Liaison entre plusieurs ensembles de variables et codage de variables qualitatives. *Thèse, Université de Paris VI.*

Shen, C., Sun, M., Tang, M., and Priebe, C. (2014). Generalized canonical correlation analysis for classification. *Journal of Multivariate Analysis*, 130 :310–322.

Tenenhaus, A. and Tenenhaus, M. (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of operational research*, 238 :391–403.

Vallejo-Arnadela, A., Vincente-Villardon, J., and Gamindo-Villaedon, M. (2007). Canonical-statis : Biplot analysis of multi-group structured data based on statis-act methodology. *Computational Statistics and Data Analysis*, 46 :4193–4205.

Zarraga, A. and Goitisolo, B. (2009). Simultaneous analysis and multiple factor analysis for contingency tables : Two methods for the joint study of contingency tables. *Computational Statistics and Data Analysis*, 53 :3171 – 3182.

 ISSN: 2102-6238