
Une analyse discriminante relationnelle sur données évolutives

Rafik Abdesselam

Laboratoire GEMMA (UMR-CNRS 6154)
Université de Caen
Esplanade de la Paix
F-14032 CAEN Cedex
abdesselam@econ.unicaen.fr

RÉSUMÉ. Une méthode d'analyse de données évolutives dans un but de discrimination et de classement est présentée. L'Analyse factorielle Discriminante Multiple Relationnelle (ADMR) proposée est basée sur la recherche de moments principaux dans l'espace des individus muni d'une distance relationnelle. Il s'agit ici de mesurer, en terme d'inertie dans l'espace des individus, la liaison observée entre les ensembles évolutifs de variables explicatives et l'ensemble des variables indicatrices associées aux modalités d'une variable à expliquer. Nous sommes en présence d'un problème relevant d'une analyse discriminante dans laquelle la décision doit être prise au fur et à mesure de la connaissance de l'information "évolutive" des observations de l'unité statistique que l'on désire classer. La contrainte consiste à imposer aux facteurs discriminants d'appartenir au sous-espace engendré par l'ensemble des observations mesurées durant cette période partielle ou globale. L'ADMR est comparée à l'Analyse Discriminante Multiple (ADM), présentée comme une analyse factorielle multiple de tableaux de centres de gravité. Un exemple sur données réelles est présenté.

ABSTRACT. A method which analyses evolutionary data with a discrimination and classification aim is described. The Relational Multiple Discriminant Analysis (RMDA) proposed is based on the research of principal axes of a configuration of points in the individual-space with a relational distance. The aim is to measure in inertia term the link observed between the evolutionary sets of explanatory variables and the set of dummy-variables associated to the modalities of the explain variable. We are in presence of a discriminant analysis problem, in which the decision must be taken as the knowledge "evolutionary" information of the observations of the statistical unit, which we want to classify. The constraint consists in imposing to the discriminant factors to be part of the subspace associated to all the observations carried out during this partial or global period. The RMDA is compared to the Multiple Discriminant Analysis (MDA) set out like a multiple factorial analysis of centre of gravity tables. An analysis using real data is here described using these methods.

MOTS-CLÉS: Analyse discriminante, distance relationnelle, rapport de corrélation, données évolutives.

KEYWORDS: Discriminant analysis, relational distance, correlation ratio, evolutionary data.

1. Introduction

Une analyse factorielle discriminante sur données évolutives est proposée. Cette Analyse Discriminante Multiple Relationnelle (ADMR) est basée sur la notion de distance relationnelle dans l'espace des individus introduite par Y. Schektman. L'ADMR est ensuite comparée à une autre analyse proposée, L'Analyse Discriminante Multiple (ADM) qui est présentée comme une extension de l'Analyse Factorielle Multiple (AFM), introduite par B. Escofier et J. Pagès, sur tableaux de centres de gravité évolutifs. Ces analyses sont spécialement conçues pour étudier une population d'individus caractérisés par un certain nombre de groupes de variables, c'est-à-dire de tableaux de données évolutives. On distingue deux, voire plusieurs groupes d'individus définis a priori par une variable nominale et sur lesquels ont été mesurées les mêmes variables continues à différents instants. L'objectif est de caractériser l'évolution de ces groupes d'individus en fonction de ces mesures répétées dans le temps. Enfin, une relation est donnée pour montrer ce qui différencie l'ADMR de l'analyse discriminante classique. Un exemple d'application sur données réelles issues du domaine de l'agronomie est présenté.

2. Distance et Analyse dans un Modèle Euclidien Relationnel

Rappelons la notion de distance relationnelle introduite dans [SCH 78] et qui est à l'origine de nombreux travaux et des méthodes proposées notamment dans [SCH 87, ABD 96, SCH 00a, ABD 00].

2.1. Distance Relationnelle

On utilisera les notations suivantes pour construire la matrice M associée à la distance de référence dans l'espace E des individus. Le contexte est celui des relations observées entre deux ensembles de variables dans le cadre d'une Analyse Factorielle Discriminante (AFD) classique à l'instant t .

$E_t = E_{x(t)} = IR^p$ étant le sous-espace des individus, associé par dualité aux p variables continues centrées $\{x^j(t); j = 1, p\}$,

$E_y = IR^q$ étant le sous-espace des individus, associé par dualité aux variables indicatrices centrées des modalités de la variable nominale y , notées $\{y^k; k = 1, q\}$,

X_t est la matrice d'ordre (n, p) des données explicatives à l'instant t associée à l'ensemble des variables $\{x^j(t); j = 1, p\}$,

$Y_{(n,q)}$ est la matrice des données associée à l'ensemble des variables $\{y^k; k = 1, q\}$,

M_y [resp. M_t] est la matrice de la distance dans l'espace E_y [resp. E_t], isomorphe du sous-espace de même nom, via l'injection canonique notée Iny [resp. $Inx(t)$].

On pose $M_t = {}^t\mathbf{Inx}(t)M\mathbf{Inx}(t)$ et $M_y = {}^t\mathbf{Iny}M\mathbf{Iny}$, où (M_t, M_y) est un couple de distances euclidiennes, M est une distance relationnelle dans $E = E_t \oplus E_y$, relativement aux variables $\{x^j(t)\}$ et $\{y^k\}$, et est notée $R[M_t, M_y]$, si et seulement si :

$$M_{ty} = {}^t\mathbf{Inx}(t)M\mathbf{Iny} = M_t[(V_t M_t)^{\frac{1}{2}}] + V_{ty} M_y [(V_y M_y)^{\frac{1}{2}}] + \quad [1]$$

où, $V_{x(t)} = V_t = {}^t X_t D X_t$, $V_y = {}^t Y D Y$ et $V_{ty} = {}^t X_t D Y$ désignent les matrices de variances et covariances, $D = (1/n)I_n$ est la matrice diagonale des poids des n individus et I_n la matrice unité d'ordre n ,

$[(V_t M_t)^{\frac{1}{2}}]^+$ [resp. $[(V_y M_y)^{\frac{1}{2}}]^+$] est l'inverse généralisée de Moore-Penrose, pondérée par M_t [resp. M_y] de $(V_t M_t)^{\frac{1}{2}}$ [resp. $(V_y M_y)^{\frac{1}{2}}$]. L'introduction d'inverses généralisées est une conséquence de la singularité des matrices V_t et V_y , lorsque $\text{rang}[V_t] < p$ et puisque $\text{rang}[V_y] = q - 1$.

2.2. Analyse Discriminante Relationnelle

On utilisera aussi les notations suivantes pour présenter l'AFD classique à l'instant t dans un Modèle Euclidien Relationnel.

$\{e_j(x(t)) \in E_t; j = 1, p\} \cup \{e_k(y) \in E_y; k = 1, q\}$ est la base canonique dans l'espace des individus $E = E_t \oplus E_y$,

$N_t = \{x_i(t) \in E_t; i = 1, n\}$ est le nuage des individus associé à X_t ,

$N_y = \{y_i \in E_y; i = 1, n\}$ est le nuage des individus associé à Y , il est constitué d'au plus q points distincts qui représentent aussi les modalités de y ,

D_y est la matrice diagonale des poids définie par $[D_y]_{kk} = n_k/n$ pour tout $k = 1, q$, où n_k est le nombre d'individus possédant la modalité k de y ,

$M_t = V_t^-$ (distance de Mahalanobis) est une inverse généralisée interne de plein rang de V_t et $M_y = \chi_y^2$ est une inverse généralisée interne de plein rang de V_y ,

$\chi_y^2 = D_y^{-1}$ est la matrice de la distance du khi-deux dans IR^q ,

$N_g(x(t)/y) = \{g(x(t)/y^k); k = 1, q\} \subset E_t$ est le nuage des q centres de gravité, associé au tableau des moyennes à l'instant t , noté $G_t = \chi_y^2 V_{yt}$,

$g(x(t)/y^k) = \sum_{\{j/j=1,p\}} [V_{ty} \chi_y^2]_{jk} e_j(x(t))$ est le k^{me} centre de gravité relativement à

la matrice des poids des groupes d'individus D_y ,

$P_t = \mathbf{Inx}(t) M_t^{-1} {}^t \mathbf{Inx}(t) M$ est l'opérateur de projection orthogonale sur E_t ,

$I_t[N_y] = I[P_t(N_y)]$ est l'inertie de la projection orthogonale, relativement à M , de N_y sur E_t .

Lemme 1

Si $M = R[V_t^-, \chi_y^2]$ alors $P_t(y_i) = P_t[e_k(y)] = g(x(t)/y^k)$.

On montre alors que les deux Analyses en Composantes Principales (ACP) suivantes sont équivalentes et traduisent l'AFD classique à l'instant t :

$$\begin{aligned} & ACP [N_g(x(t)/y); V_t^-; D_y] \\ & \quad \Downarrow \\ & ACP [P_t(N_y); M; D] \Leftrightarrow ACP [\{P_t[e_k(y)]; k = 1, q\}; M; D_y]. \end{aligned} \quad [2]$$

Preuve

Si $M = R[V_t^-, \chi_y^2]$ est relationnelle on a $M_{ty} = {}^t\mathbf{Inx}(\mathbf{t})M\mathbf{Iny} = V_t^- V_{ty} \chi_y^2$,

$$\begin{aligned} P_t \mathbf{Iny} &= \text{In}x(t) M_t^{-1} {}^t\mathbf{Inx}(\mathbf{t}) M \mathbf{Iny} = \mathbf{Inx}(\mathbf{t}) M_t^{-1} M_{ty} \\ &= \mathbf{Inx}(\mathbf{t}) V_{ty} \chi_y^2 = \mathbf{Inx}(\mathbf{t}) V_{\underline{t}y} \chi_y^2. \end{aligned}$$

On notera que $\forall i = 1, n$ si $y_i^k = 1$ alors $\underline{y}_i = e_k(y)$ et $y_i = e_k(y) - g(y)$, où \underline{y}_i^k est la valeur de l'indicatrice non centrée \underline{y}_i^k pour l'individu $\underline{y}_i = \sum_{\{k/k=1,q\}} \underline{y}_i^k e_k(y)$.

De plus, $g(y) = \mathbf{Iny} {}^t \underline{Y} D 1_n = \sum_{\{k/k=1,q\}} (n_k/n) e_k(y)$ et $Q_{\underline{y}} = \underline{Y} \chi_y^2 {}^t \underline{Y} D$ est une expression de l'opérateur de projection orthogonale sur $\mathbf{Im} \underline{Y} \subset IR^n$,

$$\begin{aligned} P_t[g(\underline{y})] &= P_t \mathbf{Iny} {}^t \underline{Y} D 1_n = \mathbf{Inx}(\mathbf{t}) V_{\underline{t}y} \chi_y^2 {}^t \underline{Y} D 1_n = \mathbf{Inx}(\mathbf{t}) {}^t X_t D \underline{Y} \chi_y^2 {}^t \underline{Y} D 1_n \\ &= \mathbf{Inx}(\mathbf{t}) {}^t X_t D Q_{\underline{y}} 1_n = \mathbf{Inx}(\mathbf{t}) {}^t X_t D 1_n = 0 \end{aligned}$$

car $1_n \in \mathbf{Im} \underline{Y}$ et que les variables $\{x^j(t); j = 1, p\}$ sont centrées, ce qui implique $P_t(y_i) = P_t[\underline{y}_i - g(\underline{y})] = P_t(\underline{y}_i)$.

Enfin, si $\underline{y}_i^k = 1$, on a,

$$\begin{aligned} P_t(y_i) &= P_t[e_k(y)] = \sum_{\{j/j=1,p\}} \langle \mathbf{Inx}(\mathbf{t}) V_{ty} \chi_y^2 \text{Pr}y e_k(y), e_j^*(x(t)) \rangle e_j(x(t)) \\ &= \sum_{\{j/j=1,p\}} [V_{ty} \chi_y^2]_{jk} e_j(x(t)) = g(x(t)/y^k). \end{aligned}$$

où, $\text{Pr}y$ désigne le projecteur cartésien de E sur l'espace E_y et $e_j^*(x(t))$ le j^{me} vecteur dual de la base canonique de $E = E_t \oplus E_y$.

REMARQUE. — Faire une AFD à l'instant t , revient à effectuer une décomposition en moments principaux de l'association symétrique entre l'ensemble des variables $\{x^j(t); j = 1, p\}$ et y , mesurée par le rapport de corrélation généralisé dont la valeur est égale à l'inertie $I_t[N_y] = I[N_g(x(t)/y)]$. Les représentations simultanées et barycentriques de l'AFD à l'instant t sont les projections orthogonales, sur les plans principaux de l'ACP [2], du nuage $\{g(x(t)/y^k) = P_t[e_k(y)]; k = 1, q\} \cup \{x_i(t); i = 1, n\}$.

Lemme Ibis

Si $M = R[V_x^-, \chi_y^2]$ alors $P_x^M[e_k(y)] = g(x/y^k)$.

L'AFD globale est équivalente dans $E = E_x \oplus E_y$, à l'ACP suivante :

$$\text{ACP} [\{P_x[e_k(y)]; k = 1, q\}; M = R[V_x^-; \chi_y^2]; D_y]. \quad [3]$$

3. Analyses Multiples

On utilisera les notations suivantes pour construire la matrice associée à la distance de référence dans l'espace E des individus.

$X_{(n,Tp)} = [X_1, X_2, \dots, X_t, \dots, X_T]$ est la matrice globale des données évolutives : juxtaposition des T tableaux X_t à n individus et p variables,

$Y_{(n,q)}$ est la matrice des données associée à l'ensemble des variables $\{y^k; k = 1, q\}$,
 $E_x = E_1 \oplus \dots \oplus E_t \oplus \dots \oplus E_T = IR^{Tp}$ étant le sous-espace des individus, associé par dualité aux T ensembles de variables explicatives centrées $\{x^j(t); j = 1, p\}_{\{t=1, T\}}$,
 $E_y = IR^q$ étant le sous-espace des individus, associé par dualité aux variables indicatrices centrées des modalités de la variable nominale y , notées $\{y^k; k = 1, q\}$.

Pour mesurer l'association, en terme d'inertie d'un nuage de points dans E , entre les ensembles de variables, et comme il s'agit d'analyse discriminante, on choisit des distances de Mahalanobis $M_t = V_t^-$ comme distances intra dans tous les sous-espaces explicatifs E_t , on élimine ainsi les effets variances-corrélations à l'intérieur de chaque sous-ensemble de variables, et $M_y = \chi_y^2$ la distance du khi-deux dans E_y .

Les deux analyses discriminantes multiples proposées diffèrent principalement par le choix des distances inter dans les sous-espaces explicatifs E_t : à effet relationnel pour l'ADMR et à effet relationnel nul pour l'ADM.

3.1. Analyse Discriminante Multiple Relationnelle

L'ADMR consiste à utiliser l'expression relationnelle [1] entre tous les couples d'ensembles de variables $\{x^j(t); j = 1, p\}_{t=1, T}$ et $\{y^k; k = 1, q\}$,

$$\begin{cases} R_t = V_t^- & \text{pour } t = 1, T \\ R_{tt'} = R[V_t^-, V_{t'}^-] = V_t^- V_{tt'} V_{t'}^- & \text{pour } t \neq t' \\ R_{ty} = R[V_t^-, \chi_y^2] = V_t^- V_{ty} \chi_y^2 \end{cases}$$

on construit la matrice $R = R[V_1^-, \dots, V_t^-, \dots, V_T^-, \chi_y^2] = R[R_x, \chi_y^2]$ relationnelle équilibrée, partitionnée d'ordre $Tp + q$,

$$R = \begin{pmatrix} V_1^- & \dots & R_{1t} & \dots & R_{1T} & R_{1y} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & V_t^- & \dots & R_{tT} & R_{ty} \\ & & & \dots & \dots & \dots \\ & & & & V_T^- & R_{Ty} \\ & & & & & \chi_y^2 \end{pmatrix} = \begin{pmatrix} R_x & R_{xy} \\ & \chi_y^2 \end{pmatrix}$$

associée à la distance dans l'espace des individus $E = E_x \oplus E_y$.

On note P_x^R l'opérateur de projection R -orthogonale sur $E_x = \oplus \{E_t\}_{\{t=1, T\}}$,
 $N_g^R(x/y) = \{P_x^R(e_k(y); k = 1, q)\} \subset E$ est le nuage des q points centres de gravité,
 $N_{\tilde{x}} = \{\tilde{x}_i \subset E_x; i = 1, n\}$ est le nuage points-individus actifs associé au tableau $\tilde{X} = X \text{Diag}[V_t^-] R_x^-$ et dont $N_g^R(x/y)$ est le nuage points-centres de gravité dans le sous-espace explicatif $E_x = \oplus \{E_t\}_{\{t=1, T\}}$.

Définition 1

L'ADMR proposée consiste à effectuer, dans le MER, l'ACP suivante :

$$ACP [\{P_x^R(e_k(y)); k = 1, q\}; R; D_y]. \tag{4}$$

REMARQUE. — Les représentations simultanées et barycentriques de l'ADMR sont les projections orthogonales, sur les plans principaux de l'ACP [4], du nuage $\{P_x^R[e_k(y)], k = 1, q\} \cup \{[\bar{x}_i, 0] \subset E; i = 1, n\}$.

3.2. Analyse Discriminante Multiple

Pour l'ADM, la relation [1] est utilisée uniquement entre les T couples d'ensembles de variables explicatives $\{x^j(t); j = 1, p\}_{t=1, T}$ et à expliquer $\{y^k; k = 1, q\}$,

$$\begin{cases} M_t = V_t^- & \text{pour } t = 1, T \\ M_{tt'} = R[V_t^-, V_{t'}^-] = 0 & \text{pour } t \neq t' \\ M_{ty} = R[V_t^-, \chi_y^2] = V_t^- V_{ty} \chi_y^2 \end{cases}$$

on construit la matrice $M = [R(V_1^-, \chi_y^2), \dots, R(V_t^-, \chi_y^2), \dots, R(V_T^-, \chi_y^2)]$ semi-relationnelle, partitionnée d'ordre $Tp + q$,

$$M = \begin{pmatrix} V_1^- & \dots & 0 & \dots & 0 & M_{1y} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & V_t^- & \dots & 0 & M_{ty} \\ & & \dots & \dots & \dots & \dots \\ & & & & V_T^- & M_{Ty} \\ & & & & & \chi_y^2 \end{pmatrix} = \begin{pmatrix} M_x & M_{xy} \\ & \chi_y^2 \end{pmatrix}$$

associée à la distance dans l'espace des individus $E = E_x \oplus E_y$ où, la distance $M_x = \text{Diag}[V_t^-]$ dans $E_x = \oplus\{E_t\}_{t=1, T}$ est à effet relationnel nul.

REMARQUE. — On appliquant le lemme 1 pour $t = 1, T$, c'est-à-dire en projetant le nuage N_y sur chacun des T sous-espaces E_t , on obtient ainsi les T AFD instantanées séparées avec leurs éléments propres $\{(\lambda_r(t), u_r(t))\}_{t=1, T}$: les moments discriminants non nuls et les vecteurs axiaux correspondants.

On note P_x^M l'opérateur de projection M -orthogonale sur $E_x = \oplus\{E_t\}_{t=1, T}$, $N_g^M(x/y) = \{P_x^M(e_k(y)); k = 1, q\} \subset E$ est le nuage des q centres de gravité, associé au tableau de moyennes évolutives $G = [G_1, \dots, G_t, \dots, G_T, 0]$: juxtaposition des T tableaux $G_t = \chi_y^2 V_{yt}$.

On démontrerait de même dans E le lemme suivant.

Lemme Iter

Si $M = [R(V_1^-, \chi_y^2), \dots, R(V_t^-, \chi_y^2), \dots, R(V_T^-, \chi_y^2)]$ alors $P_x^M[e_k(y)] = g(x/y^k)$.

Afin d'équilibrer l'influence des différents groupes de variables du tableau G des centres de gravité, on introduit la pondération proposée dans [ESC 88] sur laquelle est basée la méthode AFM. Elle consiste ici à donner à chacune des variables d'un même groupe un poids égal à l'inverse du premier moment ou pouvoir discriminant de l'AFD instantanée : le poids ou encore le coefficient de pondération des variables à l'instant t est égal à $1/\lambda_1(t)$.

Définition 2

L'ADM revient à effectuer l'ACP pondérée suivante :

$$ACP [\{P_x^M(e_k(y)); k = 1, q\}; M_p; D_y]. \quad [5]$$

$$\text{avec } M_p = [R(\frac{V_1^-}{\lambda_1(1)}, \chi_y^2), \dots, R(\frac{V_t^-}{\lambda_1(t)}, \chi_y^2), \dots, R(\frac{V_T^-}{\lambda_1(T)}, \chi_y^2)]$$

REMARQUE. — Les représentations simultanées et barycentriques de l'ADM sont les projections orthogonales, sur les plans principaux de l'ACP [5], du nuage $\{P_x^M(e_k(y)); k = 1, q\} \cup \{x_i = [x_i(1), \dots, x_i(t), \dots, x_i(T), 0]; i = 1, n\}$.

REMARQUE. — L'ADM est équivalente à l'AFM du tableau G des moyennes évolutives : centres de gravité dans $E_x = \oplus\{E_t\}_{t=1,T} = IR^{pT}$.

NOTE. — L'ADMR ne nécessite aucune pondération ; la distance relationnelle positionne les sous-espaces évolutifs E_t tel que l'on puisse traduire en terme d'inertie dans E , la structure des corrélations observées entre les sous-espaces de variables. L'ADMR est identique à l'AFD instantanée pour la première période, c'est-à-dire, au temps $t = 1$.

REMARQUE. — Dans [4] et [5], les nuages de points centres de gravité $N_g^R(x/y)$ et $N_g^M(x/y)$ sont dans le même sous-espace explicatif $E_x = \oplus\{E_t\}_{t=1,T} = IR^{pT}$ de l'espace euclidien des individus $E = E_x \oplus E_y$. Le critère de comparaison de ces deux analyses est celui du rapport maximum de l'inertie inter-classe sur l'inertie totale. Dans le cas de l'AMDR, l'inertie inter-classe $I[N_g^R(x/y)]$ est égale au rapport de corrélation généralisé.

4. Relation entre l'ADMR et l'AFD classique

En quoi l'ADMR est-elle une AFD ? Qu'est-ce qui différencie l'analyse multiple de l'analyse classique ?

A partir de l'expression [1] et vu que pour tout t allant de 1 à T , $M_t = V_t^-$, on a la relation : $R_x = \text{Diag}[V_t^-] V_x \text{Diag}[V_t^-]$ entre les distances R_x et V_x^- choisies dans le sous-espace explicatif, respectivement $E_x = \oplus\{E_t\}_{\{t=1,T\}}$, pour l'ADMR (ACP [4]) et E_x pour l'AFD (ACP [3]).

Soit $\{\gamma_r, u_r, U^r\}$ [resp. $\{\beta_r, v_r, V^r\}$] les moments principaux non nuls, les vecteurs axiaux principaux correspondants et les composantes principales de l'AFD [resp. l'ADMR].

Propriété 1

$$\forall r \text{ on a : } \gamma_r = \beta_r \quad ; \quad \begin{cases} v_r = (1/\sqrt{\gamma_r}) \text{Diag}[V_t] V_x^- u_r \\ u_r = (1/\sqrt{\gamma_r}) \text{Diag}[V_t] R_x v_r \end{cases} \quad \text{et} \quad V^r = U^r.$$

Preuve

Notons par $G = \chi_y^2 V_{yx}$ le tableau des centres de gravité (moyennes) associé à la restriction dans E_x du nuage $N_g^M(x/y)$ de l'ACP [3] et par \tilde{G} le tableau des centres de gravité associé à la restriction dans E_x du nuage $N_g^R(x/y)$ de l'ACP [4]. On a la relation suivante :

$$\begin{aligned} \tilde{G} &= [R_{y1}, \dots, R_{yt}, \dots, R_{yT}] R_x^- = [\chi_y^2 V_{y1} V_1^-, \dots, \chi_y^2 V_{yt} V_t^-, \dots, \chi_y^2 V_{yT} V_T^-] R_x^- \\ &= \chi_y^2 [V_{y1}, \dots, V_{yt}, \dots, V_{yT}] \text{Diag}[V_t^-] R_x^- = \chi_y^2 V_{yx} V_x^- \text{Diag}[V_t] \\ &= G V_x^- \text{Diag}[V_t]. \end{aligned}$$

Sachant que $\text{Diag}[V_t] R_x \text{Diag}[V_t] = V_x$, on montre que les deux analyses ont d'une part la même inertie $I_x(N_y)$ égale au rapport de corrélation généralisé, en effet,

$$\begin{aligned} I_x(N_y) &= \text{trace}({}^t \tilde{G} D_y \tilde{G} R_x) = \text{trace}(\text{Diag}[V_t] V_x^- {}^t G D_y G V_x^- \text{Diag}[V_t] R_x) \\ &= \text{trace}({}^t G D_y G V_x^- \text{Diag}[V_t] R_x \text{Diag}[V_t] V_x^-) = \text{trace}({}^t G D_y G V_x^-), \end{aligned}$$

et d'autre part les mêmes valeurs propres non nulles, puisque les moments et vecteurs axiaux discriminants de l'ACP [4] sont les éléments propres (normés) de l'opérateur : $\text{Diag}[V_t] V_x^- V_{xy} \chi_y^2 V_{yx} V_x^- \text{Diag}[V_t] R_x$ et ceux de l'ACP [3] sont les éléments propres (normés) de l'opérateur : $V_{xy} \chi_y^2 V_{yx} V_x^-$.

Enfin, on déduit aisément les assertions suivantes, compte tenu de ce que les vecteurs u_r [resp. v_r] sont normés pour V_x^- [resp. R_x].

REMARQUE. — Les représentations simultanées et barycentriques des individus de l'ADMR et de l'AFD sont identiques ainsi que les corrélations entre variables initiales et facteurs discriminants ; calculées au niveau T-Totale, c'est-à-dire sur l'ensemble des observations, sans tenir compte de la structure du tableau engendré par l'existence de groupes. Par contre, les représentations des variables ainsi que les corrélations entre variables et facteurs discriminants de ces deux analyses, calculées au niveau B-Between-Inter, c'est-à-dire sur les moyennes estimées des populations, sont différentes. A priori, ces corrélations ont un intérêt pour l'interprétation de l'analyse discriminante, elles permettent de juger du pouvoir discriminant des variables initiales et donc de caractériser ce qui différencie les groupes d'individus que l'on cherche à séparer "au mieux" et à reconnaître.

5. Exemple d'application - Variétés de pommes

Les données analysées ici m'ont été obligeamment fournies par l'Institut National de la Recherche Agronomique (INRA) - Centre de recherche d'Angers. L'objectif de cette étude est de caractériser l'évolution de trois variétés de pommes en fonction de deux caractéristiques : la teneur en sucre et en acidité de pommes récoltées avant, pendant et après la date de maturité optimale. Les mesures ont été obtenues à partir de lots de trois variétés de pommes (Golden, Fuji et Braeburn). Les mêmes variables (teneurs en sucre et en acidité des trois variétés de pommes) ont été mesurées à environ 15 jours d'intervalle avant et après maturité. Les variables mesurées à trois instants différents (prématurité, maturité et postmaturité) sont : *TS1* - Teneur globale en sucre mesurée sur un jus de pommes en prématurité (*TS2* : maturité et *TS3* : postmaturité) et *TA1* - Teneur en acidité, exprimée en équivalent *g/l* d'acide malique, également dosée sur un jus de pommes en prématurité (*TA2* : maturité et *TA3* : postmaturité).

Les données étudiées sont présentées en annexe dans la table A1. Le tableau 1. récapitule les statistiques élémentaires des variables considérées.

Période	prématurité		Maturité		Postmaturité	
	TS1	TA1	TS2	TA2	TS3	TA3
Moyenne	10.42	5.51	12.88	5.02	13.16	4.41
Ecart-type	1.07	1.46	0.97	1.47	0.99	1.58
CV(%)	8.64	26.46	7.54	29.35	7.55	35.76
Minimum	10.40	3.83	11.50	3.54	10.80	2.96
Maximum	13.80	8.42	14.10	8.29	15.40	7.12

Tableau 1. Statistiques sommaires

Les résultats des analyses multiples proposées, l'ADMR et l'ADM sur variables réduites, sont présentés en parallèle. Ces résultats n'ont pour but que de comparer ces deux méthodes. Notons qu'il serait possible de remettre en cause la supposition de normalité des variables et l'égalité des matrices de covariances de chaque groupe.

Analyses jusqu'à maturité : temps t_{12}

ADMR - Variétés	TS1	TA1	TS2	TA2
Golden	1.312	-1.378	-0.876	-1.532
Fuji	-0.168	0.136	1.000	1.435
Braeburn	-1.144	1.241	-0.124	0.097
ADM - Variétés	TS1	TA1	TS2	TA2
Golden	0.460	-0.265	0.385	-0.454
Fuji	0.768	-1.015	0.859	1.331
Braeburn	-1.227	1.280	-1.243	-0.877

Tableau 2.1. Coordonnées des centres de gravité dans $E_x = E_1 \oplus E_2$

Inertie	ADMR	ADM
B-Inter-classe	1.6564	2.6566
W-Intra-classe	0.3436	0.7082
T-Totale	2.0000	3.3274
Expliquée	82.82%	79.84%

Tableau 2.2. Critère d'inertie expliquée

ADMR	Golden	Fuji	Braeburn	Inertie intra	Poids	Effectif
Golden	0.000			0.153	0.333	10
Fuji	2.137	0.000		0.624	0.333	10
Braeburn	2.142	2.398	0.000	0.254	0.333	10
ADM	Golden	Fuji	Braeburn	Inertie intra	Poids	Effectif
Golden	0.000			0.772	0.333	10
Fuji	2.327	0.000		0.970	0.333	10
Braeburn	2.622	3.409	0.000	0.271	0.333	10

Tableau 2.3. Distances et inertie intra-classe

F1	F2	ADMR ← Facteurs → ADM	F1	F2
0.96	0.70	Pouvoirs discriminants	1.95	0.70
57.85	42.15	Pourcentages	73.52	26.48
F1	F2	ADMR ← Contributions → ADM	F1	F2
0.01	66.66	Golden	0.85	65.81
49.53	17.13	Fuji	43.08	23.59
50.46	16.21	Braeburn	56.07	10.60

Tableau 2.4. Facteurs discriminants - Contributions (%) des points moyens aux facteurs

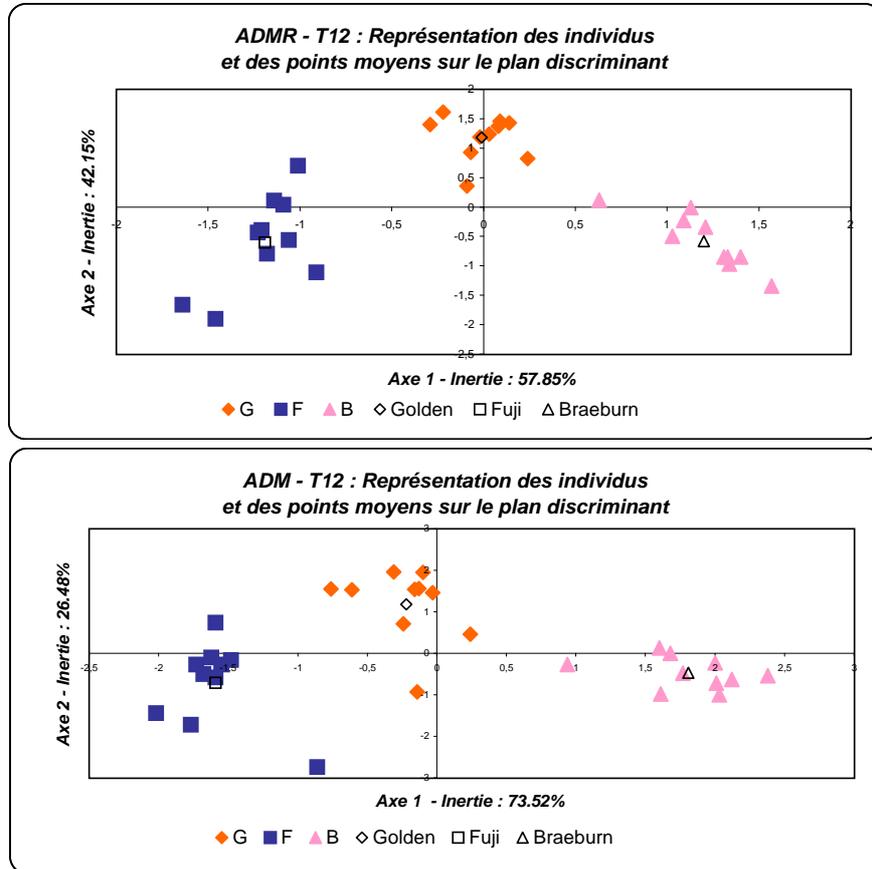


Figure 1.1. Représentations simultanées et barycentriques sur le plan discriminant

Pour mieux interpréter les variables discriminantes, le tableau 2.5. et la figure 1.2. présentent les coefficients de corrélation entre les facteurs et les variables initiales au niveau T, c'est-à-dire sur l'ensemble des observations, et entre les facteurs et les variables au niveau B, c'est-à-dire sur les moyennes estimées des populations.

	F1	F2	ADMR ← Corrélations → ADM	F1	F2
T	-0.875	0.381	TS1	-0.870	0.356
	0.958	-0.215	TA1	0.972	-0.030
	-0.879	0.317	TS2	-0.926	0.351
	-0.918	-0.393	TA2	-0.861	-0.414
B	-0.402	0.916	TS1	-0.965	0.263
	0.427	-0.904	TA1	0.996	-0.084
	-0.589	-0.808	TS2	-0.981	0.192
	-0.443	-0.896	TA2	-0.898	-0.440

Tableau 2.5. Différents coefficients de corrélation variables-facteurs discriminants

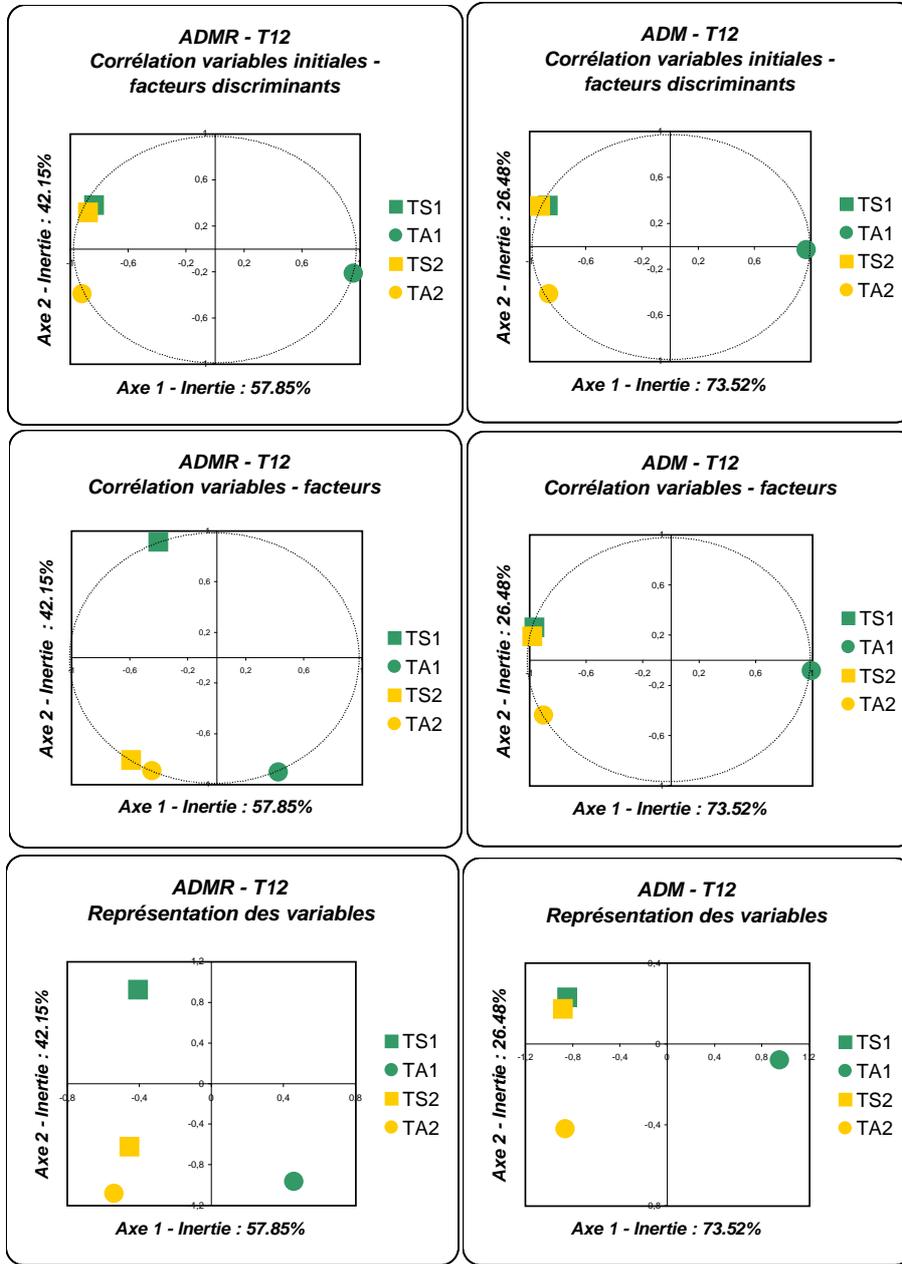


Figure 1.2. Représentations des corrélations variables-facteurs sur le plan discriminant

Analyses jusqu'à postmaturité : temps t_{13}

ADMR - Variétés	TS1	TA1	TS2	TA2	TS3	TA3
Golden	0.041	-0.023	-0.895	-1.328	1.264	-1.456
Fuji	0.155	-0.211	1.012	1.385	-0.339	0.379
Braeburn	-0.195	0.234	-0.117	-0.056	-0.925	1.077
ADM - Variétés	TS1	TA1	TS2	TA2	TS3	TA3
Golden	0.460	-0.265	0.385	-0.454	0.698	-0.699
Fuji	0.768	-1.015	0.859	1.331	0.486	-0.683
Braeburn	-1.227	1.280	-1.243	-0.877	-1.184	1.382

Tableau 3.1. Coordonnées des centres de gravité dans $E_x = E_1 \oplus E_2 \oplus E_3$

Inertie	ADMR	ADM
B-Inter-classe	1.8133	3.6749
W-Intra-classe	0.1867	0.5263
T-Totale	2.0000	4.2275
Expliquée	90.67%	86.93%

Tableau 3.2. Critère d'inertie expliquée

	Golden	Fuji	Braeburn	Inertie intra	Poids	Effectif
Golden	0.000			0.056	0.333	10
Fuji	2.245	0.000		0.341	0.333	10
Braeburn	2.338	2.411	0.000	0.162	0.333	10
ADM	Golden	Fuji	Braeburn	Inertie intra	Poids	Effectif
Golden	0.000			0.637	0.333	10
Fuji	2.351	0.000		0.787	0.333	10
Braeburn	3.392	4.006	0.000	0.234	0.333	10

Tableau 3.3. Distances et inerties intra-classe

F1	F2	ADMR ← Facteurs → ADM	F1	F2
0.98	0.83	Pouvoirs discriminants	2.85	0.82
54.12	45.88	Pourcentages	73.63	22.37
F1	F2	ADMR ← Contributions → ADM	F1	F2
5.57	61.10	Golden	5.87	60.80
31.24	35.43	Fuji	30.71	35.96
63.19	3.48	Braeburn	63.42	3.24

Tableau 3.4. Facteurs discriminants - Contributions (%) des points moyens aux facteurs

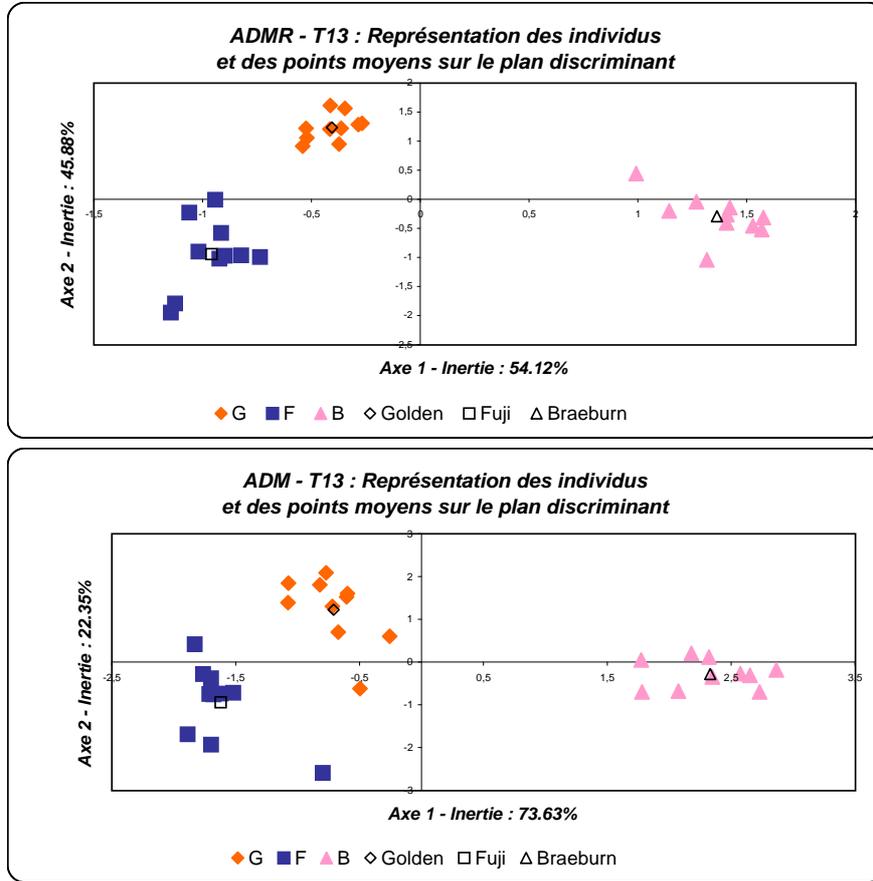


Figure 2.1. Représentations simultanées et barycentriques sur le plan discriminant

	F1	F2	ADMR ← Corrélations → ADM	F1	F2
T	-0.882	-0.083	TS1	-0.909	0.188
	0.960	0.100	TA1	0.962	0.122
	-0.909	0.014	TS2	-0.941	0.187
	-0.777	-0.623	TA2	-0.760	-0.563
	-0.803	0.302	TS3	-0.785	0.279
	0.959	-0.252	TA3	0.949	-0.189
B	-0.996	-0.094	TS1	-0.997	0.079
	0.979	0.204	TA1	0.995	0.104
	-0.330	-0.944	TS2	-1.000	0.006
	-0.263	-0.965	TA2	-0.800	-0.600
	-0.526	0.850	TS3	-0.948	0.319
	0.534	-0.845	TA3	0.974	-0.227

Tableau 3.5. Différents coefficients de corrélation variables-facteurs

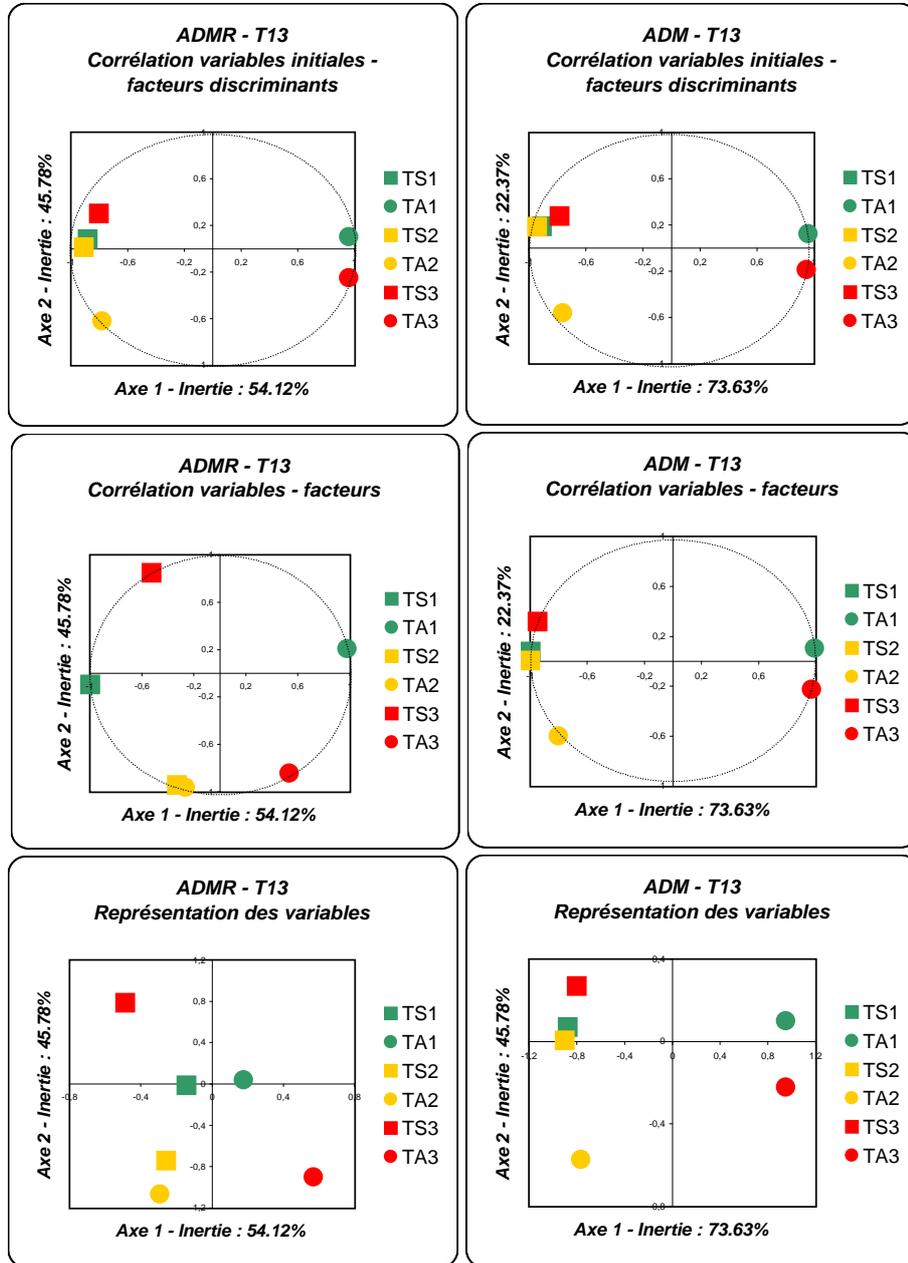


Figure 2.2. Représentations des corrélations variables-facteurs sur le plan discriminant

Remerciements

Je tiens à remercier Monsieur François Laurens, du centre de recherche de l'Institut National de la Recherche Agronomique d'Angers, pour m'avoir aimablement fourni les données de l'exemple d'application.

Annexe

N°	TS1	TA1	TS2	TA2	TS3	TA3	Variété
1	11.6	4.52	12.6	4.78	14.0	3.17	1
2	12.3	4.79	13.2	4.25	13.5	3.27	1
3	13.1	5.32	13.1	4.05	14.1	3.28	1
4	13.7	5.73	13.6	4.71	13.5	3.33	1
5	12.4	5.41	12.7	4.20	13.5	3.27	Golden
6	12.6	4.62	14.0	4.47	15.4	3.92	1
7	12.8	5.25	13.4	4.09	13.7	3.53	1
8	13.8	5.55	13.2	4.21	14.3	3.12	1
9	13.1	4.64	13.6	4.36	13.7	3.21	1
10	13.7	5.46	13.1	4.42	12.8	2.96	1
11	13.1	4.41	14.0	6.76	12.6	3.32	2
12	13.5	3.84	13.8	6.51	13.7	3.37	2
13	13.5	3.91	14.1	5.96	13.8	3.14	2
14	13.1	4.23	13.7	8.13	13.6	3.25	2
15	13.7	4.01	13.8	8.29	13.8	3.42	Fuji
16	13.1	3.83	13.8	6.42	14.5	3.11	2
17	13.1	4.24	13.9	7.06	13.9	3.24	2
18	13.1	3.89	13.9	6.84	13.8	3.63	2
19	13.1	3.97	12.0	7.02	13.6	3.44	2
20	13.1	4.00	14.1	6.87	13.1	3.40	2
21	12.0	6.18	12.0	4.15	10.8	6.88	3
22	10.9	7.62	11.5	3.74	12.5	7.10	3
23	11.1	6.83	11.5	3.90	12.4	6.26	3
24	11.7	7.30	11.9	3.74	12.0	6.83	3
25	11.9	7.16	11.6	3.60	11.7	5.33	Braehurn
26	10.5	8.42	11.6	3.66	12.0	6.80	3
27	10.4	7.43	11.5	3.63	12.0	7.12	3
28	10.9	7.14	11.7	3.54	11.8	6.70	3
29	10.8	7.77	11.5	3.62	12.2	6.57	3
30	10.8	7.96	11.9	3.74	12.4	6.27	3

Table A1. Données - Variétés de pommes (INRA-Angers)

6. Bibliographie

- [ABD 96] ABDESSELAM R., SCHEKTMAN Y., « Une analyse factorielle de l'association dissymétrique entre deux variables qualitatives », *Revue de Statistique Appliquée*, vol. XLIV(2), 1996, p. 5-34.
- [ABD 00] ABDESSELAM R., SCHEKTMAN Y., « A relational model for data structured by factors », *7th Conference of the IFCS'2000*, Namur, 11-14 July 2000.
- [BOU 75] BOUROCHE J., *Analyse des données ternaires : la double analyse en composantes principales*, Thèse, 1975.
- [BRA 80] BRAGARD L., RICHARD J., SIMAR L., *Analyse discriminante*, Comité de statistique, Université Catholique de Louvain, N°4, 1980.
- [CAI 76] CAILLEZ F., PAGÈS J., *Introduction à l'analyse des données*, SMASH, 1976.
- [CAS 95] CASIN PH., « L'analyse discriminante de tableaux évolutifs », *Revue de Statistique Appliquée*, vol. XLIII(3), 1995, p. 73-91.
- [CEL 94] CELLEUX G., NAKACHE J-P., *Analyse discriminante sur variables qualitatives*, Polytechnica, 1994.
- [DAZ 96] DAZY F., LEBARZIC J-F., *L'analyse des données évolutives*, Edts technip, 1996.
- [ESC 88] ESCOFIER B., PAGÈS J., *Analyses factorielles simples et multiples*, Dunod, 1988.
- [LAV 88] LAVIT C., *Analyse Conjointe de Tableaux Quantitatifs*, Masson, Paris, 1988.
- [ROM 73] ROMEDER J., *Méthodes et programmes d'analyse discriminante*, Dunod, 1973.
- [RTO 88] TOMASSONE R., DANZAET M., DAUDIN J-J., MASSON J., *Discrimination et classement*, Techniques stochastiques, Masson, 1988.
- [SCH 78] SCHEKTMAN Y., « Contribution à la mesure en facteurs dans les sciences expérimentales et à la mise en oeuvre des calculs statistiques », Thèse de Doctorat d'Etat, Université de Toulouse, UPS, 1978.
- [SCH 87] SCHEKTMAN Y., « A general euclidean approach for measuring and describing associations between several sets of variables », p. 37-48, Academic press, Tokyo, 1987.
- [SCH 00a] SCHEKTMAN Y., ABDESSELAM R., « A Geometrical Relational Model for Data Analyses », p. 359-368, W.Gaul, O.Opitz, M.Schader Editors, by Springer, Berlin, 2000.
- [SCH 00b] SCHEKTMAN Y., ABDESSELAM R., « Semi-produits scalaires relationnels », Note interne, 2000, Laboratoire GEMMA (UMR-CNRS 6154), Université de Caen.