

# REVUE DE STATISTIQUE APPLIQUÉE

S. CHAH SLAOUI

H. CHAMLAL

## **Nouvelles approches pour la sélection de variables discriminantes**

*Revue de statistique appliquée*, tome 48, n° 4 (2000), p. 59-82

[http://www.numdam.org/item?id=RSA\\_2000\\_\\_48\\_4\\_59\\_0](http://www.numdam.org/item?id=RSA_2000__48_4_59_0)

© Société française de statistique, 2000, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

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

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

## NOUVELLES APPROCHES POUR LA SÉLECTION DE VARIABLES DISCRIMINANTES

S. Chah Slaoui, H. Chamlal

*Département de Mathématiques et Informatique, Faculté des Sciences, B.P. 1014, rue Ibn  
Batouta, Rabat-Maroc*

### RÉSUMÉ

L'objectif de cet article est d'introduire des algorithmes de sélection de variables **explicatives hétérogènes** dans le cadre de l'analyse discriminante. Les algorithmes proposés font appel aux techniques d'association entre deux variables hétérogènes introduites par [Chah S. (86b)] ( $\Psi_{\text{cov}}$ ,  $\Psi_{\text{cor}}$ ) et à des mesures d'association multiple entre variables hétérogènes ( $\Psi_{R_{\text{cor}}}$ ,  $\Psi_{R_{\text{cov}}}$ ) qu'on introduit ici. Après une brève présentation de ces mesures, on établit une relation entre le coefficient  $\tau$  de Kendall d'une part, et les coefficients  $\Psi_{\text{cor}}$  et  $\Psi_{\text{cov}}$  d'autre part, dont on donne la loi de distribution sous l'hypothèse d'indépendance. Une mesure d'association partielle entre variables hétérogènes est introduite. La dernière partie est consacrée à la présentation des nouveaux critères de discrimination, des algorithmes de sélection qui en découlent ainsi qu'à des applications.

**Mots-clés :** *Variables hétérogènes, préordonnances, triordonnances, sélection de variables, discrimination, association partielle.*

### ABSTRACT

This paper aims at introducing algorithms of **heterogeneous variables** selection in the problem of discrimination. The proposed algorithms call for association techniques between two heterogeneous variables ( $\Psi_{\text{cov}}$ ,  $\Psi_{\text{cor}}$ ) introduced by [Chah S. (86b)] and for multiple association coefficients between heterogeneous variables ( $\Psi_{R_{\text{cor}}}$ ,  $\Psi_{R_{\text{cov}}}$ ) that we introduce here. After a short survey of these measures, we establish a relation between the Kendall's coefficient  $\tau$  on one hand, and the coefficients  $\Psi_{\text{cor}}$  and  $\Psi_{\text{cov}}$  on the other hand, whose one calculate the distribution under the hypothesis of independence. A partial association measure between heterogeneous variables is introduced. The last part is devoted to new criteria of discrimination and algorithms of selection. To assess the performance of proposed algorithms, practical examples are provided.

**Keywords :** *Heterogeneous variables, preordonnances, triordonnances, partial association, discrimination, selection of variables.*

## 1. Introduction

Le problème de sélection de variables explicatives en analyse discriminante a été largement abordé dans le cas où les variables sont homogènes. On introduit ici un nouveau critère de discrimination basé sur les préordonnances et les triordonnances [Chah S. (86a, 86b)], dont l'avantage est sa validité pour une variable explicative qu'elle soit quantitative, qualitative ou ordinale. Dans le cas d'une variable explicative qualitative on retrouve le critère de discrimination basé sur le chi-2 de contingence [Nakache J.-P. & Celeux G., 94]. Le nouveau critère est «plus pertinent» – dans un sens que l'on explicitera – que le critère basé sur la maximisation du rapport de la variance interclasse à la variance intraclasse d'une variable explicative quantitative. Par la suite, on introduit une mesure d'association partielle et une mesure d'association multiple entre variables hétérogènes. Ces mesures d'association sont utilisées pour construire des algorithmes de sélection de **variables hétérogènes**, dans le cadre d'un problème de discrimination ou de régression.

Les deux premiers algorithmes sont basés sur la notion de préordonnance; ils consistent à fixer le nombre  $k$  ( $k < p$ ) de variables à retenir, sélectionner à chaque étape une variable par maximisation d'un critère que l'on déterminera, éliminer l'effet de la variable en question, et passer à l'étape suivante, la procédure s'arrête quand le nombre  $k$  est atteint. La même approche peut être appliquée en utilisant la théorie des triordonnances, [Chah S., 86a].

À l'inverse de cette procédure, on introduit des algorithmes pas à pas descendant, basés sur la notion de préordonnance, qui proposent d'éliminer successivement les variables les moins discriminantes.

Les algorithmes proposés sont une adaptation des procédures pas à pas ascendant et pas à pas descendant; on peut dans un tel cas, ne pas fixer le nombre de variables à sélectionner et suivre la stratégie souvent recommandée consistant à s'arrêter sur le nombre de variables qui minimisent le risque d'erreur par validation croisée [Nakache J.-P. & Celeux G., 94].

Dans la section 2, on rappelle les techniques d'association entre deux variables hétérogènes basées sur les préordonnances et triordonnances. Dans la section 3, on présente le test de signification des coefficients  $\Psi_{\text{cor}}$  et  $\Psi_{\text{cov}}$ . Dans la section 4, on introduit les techniques d'association partielle et multiple entre variables hétérogènes. Dans la section 5, on introduit les nouveaux critères de discrimination et on décrit les algorithmes de sélection proposés. Des applications sont traitées dans la dernière section. Une présentation des coefficients  $\tau$ , d'association partielle et multiple de Kendall est donnée en Annexe.

## 2. Présentation des mesures d'association

### 2.1. Techniques d'association entre deux variables hétérogènes

#### Notations

$E$  : un  $n$ -échantillon, identifié à l'ensemble  $\{1, 2, \dots, n\}$ .

$H = \{(i, j) \in E^2 / i < j\}$ .

$$G = \{((i, j), (j, k)) / (i, j, k) \in E^3, i \neq j, j \neq k\}.$$

$$M = |H| = \frac{n(n-1)}{2}.$$

$A \setminus B$  : complémentaire d'un ensemble  $B$  dans  $A$ .

**Définition 1.** On appelle *préordonnance définie sur  $E$* , une relation de préordre total, notée  $P$ , définie sur  $H$ .

**Définition 2.** Une *triordonnance sur  $E$*  est une relation de préordre partiel, définie sur  $H$ , comparant uniquement les éléments de  $G$ .

Une préordonnance  $P$ , peut être codée d'une façon binaire ou ternaire :

**Codage ternaire :**

**Codage binaire :**

$$T_P(i, j, k, l) = \begin{cases} 1 & \text{si } (i, j) >_P (k, l) \\ 0 & \text{si } (i, j) =_P (k, l) \\ -1 & \text{si } (i, j) <_P (k, l) \end{cases} \quad B_P(i, j, k, l) = \begin{cases} 1 & \text{si } (i, j) >_P (k, l) \\ 0 & \text{sinon} \end{cases}$$

$$T_P(i, j, k, l) = B_P(i, j, k, l) - B_P(k, l, i, j)$$

Le symbole  $>_P$  désigne le fait que la paire  $(i, j)$  précède la paire  $(k, l)$  dans l'ordre induit par  $P$  sur  $H$ ,  $=_P$  désigne le fait que les paires sont *ex aequo*,  $<_P$  désigne le fait que la paire  $(k, l)$  précède la paire  $(i, j)$ .

Dans le cas d'une triordonnance :

**Codage ternaire :**

$$T'_P(i, j, k) = T_P(i, j, j, k)$$

**Codage binaire :**

$$B'_P(i, j, k) = B_P(i, j, j, k)$$

Une préordonnance (respectivement triordonnance) peut être induite par une variable aléatoire  $V$  qu'elle soit qualitative, ordinale ou quantitative :

Si  $V$  est qualitative :  $(i, j)P_V(k, l) \iff (V(i) = V(j)) \text{ et } (V(k) \neq V(l))$ .

Si  $V$  est ordinale :  $(i, j)P_V(k, l) \iff |r(i) - r(j)| < |r(k) - r(l)|$ ;  $r$  est le rang induit par  $V$ .

Si  $V$  est quantitative :  $(i, j)P_V(k, l) \iff |V(i) - V(j)| < |V(k) - V(l)|$ .

Dans tout ce qui suit, si aucune confusion n'est à craindre,  $P_X$  désigne la préordonnance ou la triordonnance induite par une variable  $X$ .  $T_X$  et  $B_X$  désignent respectivement les codages ternaire et binaire associés à la préordonnance  $P_X$ .  $T'_X$  et  $B'_X$  désignent respectivement les codages ternaire et binaire associés à la triordonnance  $P_X$ .

Étant donné deux préordonnances  $P$  et  $Q$ ,  $T_P$  et  $T_Q$  les codages associés, le coefficient de covariance (respectivement de corrélation) induit sur l'ensemble des préordonnances une mesure d'adéquation, notée  $\Psi_{\text{cov}}$  (respectivement  $\Psi_{\text{cor}}$ ) :

$$\begin{aligned}\Psi_{\text{cov}}(P, Q) &= \text{cov}(T_P, T_Q) \\ &= \frac{\sum T_P(i, j, k, l)T_Q(i, j, k, l)}{M(M-1)} \\ \Psi_{\text{cor}}(P, Q) &= \text{cor}(T_P, T_Q) \\ &= \frac{\sum T_P(i, j, k, l)T_Q(i, j, k, l)}{\sqrt{(\sum T_P^2(i, j, k, l) \sum T_Q^2(i, j, k, l))}}\end{aligned}$$

En effet, les variables  $T_P$  et  $T_Q$  sont centrées car

$$T_P(i, j, k, l) = -T_P(k, l, i, j) \quad T_Q(i, j, k, l) = -T_Q(k, l, i, j)$$

Les signes « $\sum$ » portent sur l'ensemble  $\{(i, j), (k, l) \in H \times (H \setminus \{(i, j)\})\}$ .

Dans le cas d'une triordonnance on obtient :

$$\begin{aligned}\Psi'_{\text{cov}}(P, Q) &= \text{cov}(T'_P, T'_Q) \\ &= \frac{\sum T'_P(i, j, k)T'_Q(i, j, k)}{n(n-1)^2} \\ \Psi'_{\text{cor}}(P, Q) &= \text{cor}(T'_P, T'_Q) \\ &= \frac{\sum T'_P(i, j, k)T'_Q(i, j, k)}{\sqrt{(\sum T'^2_P(i, j, k) \sum T'^2_Q(i, j, k))}}\end{aligned}$$

En effet, les variables  $T'_P$  et  $T'_Q$  sont centrées et  $|G| = n(n-1)^2$ . Les signes « $\sum$ » portent sur  $\{(i, j, k) \in G\}$ .

Dans le cas d'absence d'*ex aequo* :

$$\sum T^2(i, j, k, l) = M(M-1) \implies \Psi_{\text{cov}} = \Psi_{\text{cor}}$$

puisque,  $T^2(i, j, k, l) = 1, \forall (i, j) \in H, \forall (k, l) \in H \setminus \{(i, j)\}$ . De même  $\Psi'_{\text{cor}} = \Psi'_{\text{cov}}$ .

### 3. Relation entre le $\tau$ de Kendall et les coefficients $\Psi_{\text{cor}}$ , $\Psi_{\text{cov}}$ Application au test de signification de $\Psi_{\text{cor}}$ et $\Psi_{\text{cov}}$

Les nouveaux critères de discrimination seront construits sur la base des coefficients  $\Psi_{\text{cor}}$  et  $\Psi_{\text{cov}}$ . L'objectif de ce paragraphe est de calculer la distribution sous l'hypothèse d'indépendance de ces coefficients.

Étant donné deux variables aléatoires (non nécessairement homogènes)  $X_1$  et  $X_2$ ,  $P_{X_1}$  et  $P_{X_2}$  désignent les préordonnances induites respectivement par  $X_1$  et  $X_2$ , et  $T_{X_1}, T_{X_2}$  les codages associés. Notons  $r_1, r_2$  les deux variables rangs définies sur  $H$  associées respectivement à  $P_{X_1}, P_{X_2}$  :

$$\forall (i, j) \in H, \forall (k, l) \in H \begin{cases} r_m(i, j) < r_m(k, l) \iff (i, j) >_{P_{X_m}} (k, l) \\ r_m(i, j) = r_m(k, l) \iff (i, j) =_{P_{X_m}} (k, l), (m = 1, 2) \\ r_m(i, j) > r_m(k, l) \iff (i, j) <_{P_{X_m}} (k, l) \end{cases}$$

**Proposition 1.** *Les expressions  $\tau_1$  et  $\tau_2$  du coefficient  $\tau$  de Kendall (cf. annexe) induisent les mesures d'association entre deux variables hétérogènes  $\Psi_{\text{cor}}$  et  $\Psi_{\text{cov}}$  de la manière suivante :*

$$\Psi_{\text{cor}}(P_{X_1}, P_{X_2}) = \tau_1(r_1, r_2) \quad (3.1)$$

$$\Psi_{\text{cov}}(P_{X_1}, P_{X_2}) = \tau_2(r_1, r_2) \quad (3.2)$$

En effet, (3.1) et (3.2) découlent des propositions (4) et (5) de l'annexe et du fait que

$$\forall (i, j) \in H, \forall (k, l) \in H \begin{cases} r_m(i, j) < r_m(k, l) \iff T_{X_m}(i, j, k, l) = 1 \\ r_m(i, j) = r_m(k, l) \iff T_{X_m}(i, j, k, l) = 0 \\ r_m(i, j) > r_m(k, l) \iff T_{X_m}(i, j, k, l) = -1 \end{cases} \quad (m = 1, 2)$$

Ce résultat permet de construire un test d'indépendance entre deux variables hétérogènes de nature quelconque, ce qui généralise le test classique entre deux variables rangs, basé sur le coefficient  $\tau$  de Kendall (cf. annexe).

Si on note  $S$  la quantité  $\frac{1}{2} \sum T_{X_1}(i, j, k, l) T_{X_2}(i, j, k, l)$ , les résultats asymptotiques concernant la distribution de  $\tau$  (cf. annexe; où il faut remplacer  $n$  par  $M$ ) s'adaptent immédiatement à  $S$ . Dans notre cas les approximations sont bonnes dès que la taille de l'échantillon  $n$  est supérieure ou égale à 5 (ce qui correspond à  $M = \frac{5 \times 4}{2} = 10$ ).

### 4. Techniques d'association entre variables hétérogènes

Les algorithmes de sélection pas à pas ascendant consistent à ajouter à chaque étape, l'effet de la variable préalablement retenue. Les algorithmes pas à pas descendant consistent à éliminer à chaque étape la variable la moins discriminante.

L'objectif de ce paragraphe est d'introduire une mesure d'association partielle et multiple entre variables hétérogènes.

Soient  $X_1, X_2, \dots, X_p$ ,  $p$  variables hétérogènes;  $P_{X_1}, P_{X_2}, \dots, P_{X_p}$  les préordonnances correspondantes;  $T_{X_1}, T_{X_2}, \dots, T_{X_p}$  les codages ternaires associés.  $r_1, r_2, \dots, r_p$  désignent les  $p$  variables rangs induites sur  $H$  par  $P_{X_1}, P_{X_2}, \dots, P_{X_p}$  où

$$\forall (i, j) \in, \forall (k, l) \in H \begin{cases} r_m(i, j) < r_m(k, l) \iff (i, j) >_{P_{X_m}} (k, l) \\ r_m(i, j) = r_m(k, l) \iff (i, j) =_{P_{X_m}} (k, l), (m = 1, 2, \dots, p) \\ r_m(i, j) > r_m(k, l) \iff (i, j) <_{P_{X_m}} (k, l) \end{cases}$$

#### 4.1. Association partielle entre variables hétérogènes

Les coefficients d'association partielle des rangs [Kendall M.G., 62], (cf. annexe) induisent sur l'ensemble des préordonnances deux mesures d'association partielle, notées  $(\Psi_{\text{cor}})$ . et  $(\Psi_{\text{cov}})$ . :

$$\begin{aligned} \Psi_{\text{cor}}(P_{X_1}, P_{X_2})_{.P_{X_3}} &= \tau_1(r_1, r_2)_{.r_3} \\ \Psi_{\text{cor}}(P_{X_1}, P_{X_2})_{.P_{X_3} \dots P_{X_h}} &= \tau_1(r_1, r_2)_{.r_3 \dots r_h} \\ \Psi_{\text{cov}}(P_{X_1}, P_{X_2})_{.P_{X_3}} &= \tau_2(r_1, r_2)_{.r_3} \\ \Psi_{\text{cov}}(P_{X_1}, P_{X_2})_{.P_{X_3} \dots P_{X_h}} &= \tau_2(r_1, r_2)_{.r_3 \dots r_h} \end{aligned}$$

De même, on définit sur l'ensemble des triordonnances deux mesures d'association partielle, notées :  $(\Psi'_{\text{cor}})$ . et  $(\Psi'_{\text{cov}})$ .

#### 4.2. Association multiple entre variables hétérogènes

Le coefficient de corrélation multiple des rangs [Kendall M.G., 62], (cf. annexe) induit sur l'ensemble des préordonnances deux mesures d'association multiple, notées  $\Psi_{R_{\text{cor}}}$ ,  $\Psi_{R_{\text{cov}}}$ . Dans le cas  $p = 3$ , elles sont définies par :

$$\begin{aligned} 1 - \Psi_{R_{\text{cor}}}^2(P_{X_1}, (P_{X_2}, P_{X_3})) &= (1 - \Psi_{\text{cor}}^2(P_{X_1}, P_{X_2}))(1 - \Psi_{\text{cor}}^2(P_{X_1}, P_{X_3})_{.P_{X_2}}) \\ 1 - \Psi_{R_{\text{cov}}}^2(P_{X_1}, (P_{X_2}, P_{X_3})) &= (1 - \Psi_{\text{cov}}^2(P_{X_1}, P_{X_2}))(1 - \Psi_{\text{cov}}^2(P_{X_1}, P_{X_3})_{.P_{X_2}}) \end{aligned}$$

expressions qui se généralisent au cas  $p > 3$ .

Notons qu'en l'absence d'*ex aequo*, on a :

$$(\Psi_{\text{cor}})_{.} = (\Psi_{\text{cov}})_{.}; (\Psi'_{\text{cor}})_{.} = (\Psi'_{\text{cov}})_{.}; \Psi_{R_{\text{cor}}} = \Psi_{R_{\text{cov}}}$$

### 5. Nouveaux algorithmes de sélection

#### 5.1. Critères de discrimination

Les deux expressions du  $\tau$  de Kendall,  $\tau_1$  et  $\tau_2$ , induisent les mesures d'association entre variables hétérogènes  $\Psi_{\text{cor}}$  et  $\Psi_{\text{cov}}$ . S'agissant des préordonnances et

triordonnances, les coefficients  $\Psi_{\text{cov}}$ ,  $\Psi'_{\text{cov}}$  et  $\Psi_{\text{cov}}$ ,  $\Psi'_{\text{cov}}$  induisent quatre algorithmes de sélection, un par critère de discrimination. Cependant, dans la littérature statistique [Kendall M. G, 62], le test de signification des deux expressions du coefficient  $\tau$  de Kendall se base sur le numérateur (quantité  $S$  (cf. annexe)), ce qui justifie, par la suite, le raisonnement effectué en terme des coefficients  $\Psi_{\text{cov}}$  et  $\Psi'_{\text{cov}}$  uniquement.

**Position du problème :** Étant donné un ensemble de variables explicatives (non nécessairement homogènes) et une variable qualitative à expliquer  $Q$ , le problème consiste à chercher la variable explicative  $X$  qui maximise l'association, au sens de  $\Psi_{\text{cov}}$ , avec la variable à expliquer.

Si  $X$  est qualitative, il a été démontré que  $\Psi_{\text{cov}}(P_Q, P_X)$  coïncide avec le chi-2 de contingence [Chah S., 86b]; on retrouve donc un critère de discrimination dans le cas de variables explicatives qualitatives.

Si  $X$  est quantitative :

$$T_X(i, j, k, l) = B_X(i, j, k, l) - B_X(k, l, i, j)$$

$$T_Q(i, j, k, l) = B_Q(i, j, k, l) - B_Q(k, l, i, j)$$

$$B_X(i, j, k, l) = \begin{cases} 1 & \text{si } |X(i) - X(j)| < |X(k) - X(l)| \\ 0 & \text{sinon} \end{cases}$$

$$B_Q(i, j, k, l) = \begin{cases} 1 & \text{si } Q(i) = Q(j), Q(k) \neq Q(l) \\ 0 & \text{sinon} \end{cases}$$

de sorte que :

$$M(M - 1)\Psi_{\text{cov}}(P_Q, P_X) = 2 \sum_{(i,j) \in H(k,l)} \sum_{(i,j) \in H \setminus (i,j)} (B_X(i, j, k, l) - B_X(k, l, i, j)) B_Q(i, j, k, l) \quad (5.3)$$

en effet :

$$\begin{aligned} & \sum T_X(i, j, k, l) T_Q(i, j, k, l) \\ &= \sum (B_X(i, j, k, l) - B_X(k, l, i, j)) (B_Q(i, j, k, l) - B_Q(k, l, i, j)) \\ &= \sum (B_X(i, j, k, l) - B_X(k, l, i, j)) B_Q(i, j, k, l) \\ &+ \sum (B_X(k, l, i, j) - B_X(i, j, k, l)) B_Q(k, l, i, j) \\ &= 2 \sum (B_X(i, j, k, l) - B_X(k, l, i, j)) B_Q(i, j, k, l) \end{aligned}$$

Les sommations portent sur l'ensemble  $\{((i, j), (k, l)) \in H \times (H \setminus (i, j))\}$ .



Une variable explicative maximisant ce critère doit «rapprocher» deux individus réunis par la variable à expliquer, et «éloigner» deux individus séparés par la variable dépendante; en d'autres termes, la variable en question doit satisfaire au mieux, la condition suivante :

$$\max_{Q(i)=Q(j)} |X(i) - X(j)| < \min_{Q(k) \neq Q(l)} |X(k) - X(l)| \quad (5.4)$$

Ce qui assure la compacité et la séparabilité des groupes décrits par la variable explicative. Notons qu'en général, il n'existe pas de variable  $X$  vérifiant (5.4) pour tout quadruplet  $(i, j, k, l)$ . Si une telle variable existe, on peut la qualifier d'idéale pour expliquer  $Q$ .

Le nouveau critère est «plus pertinent» que celui basé sur la maximisation du rapport de Fisher ou encore du rapport de la variance interclasse à la variance intraclasse. En effet l'optimisation du critère  $\Psi_{\text{cov}}$  exige que chaque paire d'individus de groupes différents soit plus dispersée que toute autre paire d'individus provenant d'un même groupe, alors que le rapport de Fisher consiste à maximiser la variance des centres de gravité des groupes (variance interclasse) et à minimiser la variance à l'intérieur des groupes en moyenne (variance intraclasse). En outre dans le cas gaussien (cas le plus fréquent en pratique), on se place sous l'hypothèse d'homoscédasticité, qui fait rarement l'objet d'un test statistique, les tests disponibles étant peu fiables. Le nouveau critère basé sur  $\Psi_{\text{cov}}$  est plus «fin» en ce sens qu'il mesure le degré «d'accord» ( $\Psi_{\text{cov}}$  positif) ou de «désaccord» ( $\Psi_{\text{cov}}$  négatif) entre préordonnances. Illustrons cet aspect du critère à travers l'exemple suivant :

$X_1$	1	2	3	4
$X_2$	2	2	4	4
$Q$	non	oui	oui	non

$$\Psi_{\text{cov}}(P_Q, P_{X_1}) = \frac{-4}{30} = -0.13 \text{ et } \Psi_{\text{cov}}(P_Q, P_{X_2}) = \frac{-8}{30} = -0.27.$$

Les deux variables aléatoires  $X_1$  et  $X_2$  ont chacune une variance interclasse nulle, et un même rapport de Fisher nul. Relativement au critère classique, les deux variables sont à «égalité». Par contre,  $\Psi_{\text{cov}}(P_{X_2}, P_Q) < \Psi_{\text{cov}}(P_{X_1}, P_Q) < 0$  : soit un «désaccord» entre préordonnances induites par chacune des variables explicatives et celle induite par la variable à expliquer et une mise en évidence du caractère «ambigu» de la variable  $X_2$ .

Dans le cas de la théorie des triordonnances, si  $X$  est qualitative, le critère  $\Psi_{\text{cov}}$  équivalait au critère de Jordan (résultat établi par [Chah S., (86b)]); on retrouve un critère de discrimination dans le cas de variables explicatives qualitatives.

Si  $X$  est quantitative, on a par un raisonnement analogue à celui ayant abouti à l'équation (5.3)

$$|G|\Psi'_{\text{cov}}(P_Q, P_X) = 2 \left[ \sum_{(i,j,k) \in G} (B'_X(i, j, k) - B'_X(k, j, i))B'_Q(i, j, k) \right]$$

Une variable maximisant ce critère doit satisfaire au mieux, pour tout  $j$ , la condition suivante :

$$\max_{Q(i)=Q(j)} |X(i) - X(j)| < \min_{Q(j) \neq Q(k)} |X(j) - X(k)| \quad (5.5)$$

En d'autres termes, chaque individu doit être plus proche de l'élément le plus éloigné de son groupe que de l'élément le plus proche issu d'un groupe différent. Le critère  $\Psi'_{\text{cov}}$  assure la séparabilité des groupes et leur compacité, mais avec un degré moindre que le critère basé sur les préordonnances.

Bien que la condition (5.5) soit moins forte que (5.4), il n'existe pas en général de variable  $X$  vérifiant (5.5) pour tout triplet  $(i, j, k)$  de  $G$

**Proposition 2.** *Supposons que la variable à expliquer  $Q$  possède  $nm$  modalités. On note  $G_l (l = 1, \dots, nm)$  le groupe d'individus possédant la  $l^{\text{ème}}$  modalité et  $n_l$  désigne son cardinal. Alors, dans le cadre de la théorie des préordonnances, la valeur maximale possible du critère  $\Psi_{\text{cov}}$  est égale à :*

$$\frac{2 \left[ M \sum_{l=1}^{nm} C_{n_l}^2 - \left( \sum_{l=1}^{nm} C_{n_l}^2 \right)^2 \right]}{M(M-1)}$$

Dans le cadre de la théorie des triordonnances, elle est égale à :

$$\frac{2 \left[ (n-1) \sum_{l=1}^{nm} n_l(n_l-1) - \sum_{l=1}^{nm} n_l(n_l-1)^2 \right]}{n(n-1)^2}$$

En effet :

**1. Cas des préordonnances :** La valeur maximale possible du critère  $\Psi_{\text{cov}}$  est atteinte si pour une variable  $X$ ,

$$T_X(i, j, k, l) = T_Q(i, j, k, l), \forall [(i, j), (k, l)] \in H \times H \setminus \{(i, j)\}$$

Dans ce cas :

$$\sum_{(i,j) \in H, (k,l) \in H \setminus \{(i,j)\}} T_Q(i, j, k, l) T_X(i, j, k, l) = \sum_{(i,j) \in H, (k,l) \in H} T_Q^2(i, j, k, l), (T_Q(i, j, i, j) = 0)$$

or  $T_Q(i, j, k, l) = Y_Q(i, j) - Y_Q(k, l)$  [Chah. S, 86b], où :

$$Y_Q(i, j) = \begin{cases} 1 & \text{si } Q(i) = Q(j) \\ 0 & \text{sinon} \end{cases}$$

posons  $A = \sum_{(i,j) \in H, (k,l) \in H} T_Q^2(i, j, k, l)$ , alors, puisque  $Y_Q^2 = Y_Q$ , on a :

$$\begin{aligned} A &= \sum_{(i,j) \in H} \sum_{(k,l) \in H} (Y_Q(i, j) - Y_Q(k, l))^2 \\ &= \sum_{(i,j) \in H} \sum_{(k,l) \in H} Y_Q^2(i, j) + \sum_{(i,j) \in H} \sum_{(k,l) \in H} Y_Q^2(k, l) - 2 \sum_{(i,j) \in H} \sum_{(k,l) \in H} Y_Q(i, j) Y_Q(k, l) \\ &= \sum_{(i,j) \in H} \sum_{(k,l) \in H} Y_Q(i, j) + \sum_{(i,j) \in H} \sum_{(k,l) \in H} Y_Q(k, l) - 2 \sum_{(i,j) \in H} Y_Q(i, j) \sum_{(k,l) \in H} Y_Q(k, l) \\ &= 2 \left[ M \sum_{(i,j) \in H} Y_Q(i, j) - \left( \sum_{(i,j) \in H} Y_Q(i, j) \right)^2 \right] \\ &= 2 \left[ M \sum_{l=1}^{nm} C_{n_l}^2 - \left( \sum_{l=1}^{nm} C_{n_l} \right)^2 \right] \end{aligned}$$

**2. Cas des triordonnances :** De même, la valeur maximale possible du critère  $\Psi'_{cov}$  est atteinte, si pour une variable  $X$ ,  $T'_X(i, j, k) = T'_Q(i, j, k), \forall (i, j, k) \in G$ . Dans ce cas :

$$\begin{aligned}
 \sum_{i \neq j, k \neq j} T_Q^2(i, j, k) &= \sum_{i \neq j, k \neq j} (Y_Q(i, j) - Y_Q(j, k))^2 \\
 &= \sum_{i \neq j, k \neq j} Y_Q^2(i, j) + \sum_{i \neq j, k \neq j} Y_Q^2(j, k) - 2 \sum_{i \neq j, k \neq j} Y_Q(i, j) Y_Q(j, k) \\
 &= \sum_{i \neq j, k \neq j} Y_Q(i, j) + \sum_{i \neq j, k \neq j} Y_Q(j, k) - 2 \sum_{i \neq j, k \neq j, Q(i)=Q(j)=Q(k)} 1 \\
 &= 2 \left[ (n-1) \sum_{l=1}^{nm} n_l(n_l-1) - \sum_{l=1}^{nm} n_l(n_l-1)^2 \right]
 \end{aligned}$$

**Proposition 3.** *S'il existe une variable  $X$  pour laquelle le critère  $\Psi_{\text{cov}}$  prend sa valeur maximale possible, alors le critère  $\Psi'_{\text{cov}}$  prend sa valeur maximale possible pour cette même variable.*

En effet : dans le cas des préordonnances, supposons qu'une telle variable  $X$  existe, on aura :

$$T_X(i, j, k, l) = T_Q(i, j, k, l), \forall [(i, j), (k, l)] \in H \times H \setminus \{(i, j)\}$$

et donc :

$$T'_X(i, j, k) = T_X(i, j, j, k) = T_Q(i, j, j, k) = T_Q^l(i, j, k), \forall (i, j, k) \in G$$

ce qui entraîne que  $\Psi'_{\text{cov}}$  prend aussi sa valeur maximale possible pour la variable  $X$ .

## 5.2. Algorithmes de sélection

### Données du problème

$X_1, X_2, \dots, X_p$  :  $p$  variables aléatoires explicatives (non nécessairement homogènes).

$Q$  : variable qualitative à expliquer.

### Notations

$$V = \{X_1, \dots, X_p\}.$$

$$\Psi_{R_{\text{cor}}}(P_Q, P_{X_m})_V = \Psi_{R_{\text{cor}}}(P_Q, (P_{X_1}, \dots, P_{X_p})).$$

$$\Psi_{R_{\text{cor}\cdot j}}(P_Q, P_{X_m})_V = \Psi_{R_{\text{cor}}}(P_Q, (P_{X_1}, \dots, P_{X_{j-1}}, P_{X_{j+1}}, \dots, P_{X_p})), X_j \in V.$$

### Algorithmes de sélection pas à pas ascendant

Il s'agit d'une adaptation de l'algorithme pas à pas ascendant appliqué au cas de variables hétérogènes. Pour ce faire, on fait appel aux notions introduites précédemment, concernant les mesures d'association simple et partielle entre variables hétérogènes.

#### Algorithme de sélection pas à pas ascendant (Cas des préordonnances)

Le choix du critère  $|\Psi_{\text{cov}}|$  (resp.  $|\Psi_{\text{cor}}|$ ) au lieu de  $\Psi_{\text{cov}}$  (resp.  $\Psi_{\text{cor}}$ ), permet de sélectionner aussi bien des variables en accord (une valeur de  $\Psi_{\text{cov}}$  ou  $\Psi_{\text{cor}}$  fortement positive) que des variables en désaccord (une valeur de  $\Psi_{\text{cov}}$  ou  $\Psi_{\text{cor}}$  fortement négative) avec la variable à expliquer. Cependant, dans le cadre d'un problème de discrimination, la sélection concerne uniquement les variables en accord avec la variable dépendante.

#### Algorithme

Introduire  $k$ ; le nombre de variables à sélectionner ( $k < p$ ),

**Pas 1 :** La première variable sélectionnée est  $X_{i_1}$  solution du problème :

$$\max_{i=1,2,\dots,p} \Psi_{\text{cov}}(P_{X_i}, P_Q),$$

**Pas 2 :** La deuxième variable sélectionnée est  $X_{i_2}$  solution du problème :

$$\max_{i=1,2,\dots,p, i \neq i_1} \Psi_{\text{cov}}(P_{X_i}, P_Q)_{.P_{X_{i_1}}},$$

...

**Pas  $k$  :** La dernière variable sélectionnée est  $X_{i_k}$  solution du problème :

$$\max_{i=1,2,\dots,p, i \notin \{i_1, i_2, \dots, i_{k-1}\}} \Psi_{\text{cov}}(P_{X_i}, P_Q)_{.P_{X_{i_1}}, P_{X_{i_2}}, \dots, P_{X_{i_{k-1}}}} \quad \blacksquare$$

le même algorithme peut être appliqué, en remplaçant le critère  $\Psi_{\text{cov}}$  par  $\Psi_{\text{cor}}$

En appliquant la théorie des triordonnances, on obtient deux autres algorithmes de sélection pas à pas ascendant, opérant sur variables hétérogènes. Pour ce faire, il suffit de remplacer dans les algorithmes précédents, le critère de sélection  $\Psi_{\text{cor}}$  par  $\Psi'_{\text{cor}}$  et  $\Psi_{\text{cov}}$  par  $\Psi'_{\text{cov}}$ .

### Algorithmes de sélection pas à pas descendant

Dans ce paragraphe, nous proposons une adaptation de l'algorithme de sélection pas à pas descendant, au cas de variables hétérogènes. En effet chacune des mesures d'association multiple introduite au paragraphe (4.2), induit un critère de discrimination pouvant servir à l'élaboration d'un algorithme de sélection pas à pas descendant. L'algorithme procède par éliminations successives : à chaque étape, l'algorithme

détermine la variable dont l'élimination maximise le critère de discrimination induit par l'une des mesures d'association  $\Psi_{R_{cor}}$  ou  $\Psi_{R_{cov}}$ .

**Algorithme**

**Étape 0 :** Introduire  $k$ /nombre de variables à sélectionner/,  $V = \{X_1, \dots, X_p\}$

**Étape 1 :** Chercher  $X_{i_1^*} \in V$ , vérifiant :

$$\Psi_{R_{cor}.i_1^*}(P_Q, P_{X_m})_V = \max_{X_j \in V} \Psi_{R_{cor}.j}(P_Q, P_{X_m})_V, \text{ (cf. notations début du § 5.2)}$$

$$V := V \setminus \{X_{i_1^*}\}$$

**Étape 2 :** Chercher  $X_{i_2^*} \in V$ , vérifiant :

$$\Psi_{R_{cor}.i_2^*}(P_Q, P_{X_m})_V = \max_{X_j \in V} \Psi_{R_{cor}.j}(P_Q, P_{X_m})_V,$$

$$V := V \setminus \{X_{i_2^*}\}$$

...

**Étape  $p - k$  :** Chercher  $X_{i_{p-k}^*} \in V$ , vérifiant :

$$\Psi_{R_{cor}.i_{p-k}^*}(P_Q, P_{X_m})_V = \max_{X_j \in V} \Psi_{R_{cor}.j}(P_Q, P_{X_m})_V,$$

$$V := V \setminus \{X_{i_{p-k}^*}\} \quad \blacksquare$$

**6. Exemples d'application**

**6.1. Exemple 1**

TABLEAU 1

$X_1$	38	40	33	30	50	55	35	31	30	25	37	41	32	47	33	36	29
$X_2$	85	85	85	85	85	85	85	85	85	85	85	85	75	85	85	85	85
$X_3$	i n	f o	f n	m n	i o	m n	m o	m n	m o	i n	i o	m o	m n	m o	m n	f n	m n
$X_4$	20	17.2	20.2	17.5	20.5	19.5	18	16.7	18.5	15.8	21	23	19.7	19	17.6	16.7	18
$X_5$	o o	o o	n n	n n	n n	o o	o o	o n	o n	o o	o o	o o	o n	o o	o o	o n	n n
$Q$	2	4	4	2	4	2	4	2	2	2	4	4	2	4	2	2	2

i : important, m : moyen, f : faible, o : oui, n : non. Les valeurs données de  $X_1$ , sont les prix multipliés par  $10^{-2}$ .

Le tableau de données est extrait de la revue INFO PC (n° 146 Avril 98). 17 modèles de moniteurs, décrits par cinq variables explicatives hétérogènes : trois variables quantitatives,  $X_1$  : Prix (F),  $X_2$  : Fréquence maximale (Hz),  $X_4$  : Poids (Kg), et deux variables qualitatives,  $X_3$  : Moirage|correction possible, possédant 6 modalités (i|o,i|n,m|o,m|n,f|o,f|n),  $X_5$  : TCO92|TCO95, possédant 3 modalités (o|o, o|n, n|n). La variable à expliquer  $Q$  concerne la Configuration et les possibilités de réglages, elle possède deux modalités, 2 : moyen et 4 : bon.

### Étude des associations simples :

TABLEAU 2

Variable $X$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
$\frac{S_X}{\sqrt{(\text{var}(S_X))}}$	2.089	-0.651	3.742	1.279	0.25

$S_X$  désigne la quantité  $\frac{\sum T_X(i, j, k, l)T_Q(i, j, k, l)}{2}$  (cf. section (3.1));  $T_X, T_Q$  codage associé respectivement à la préordonnance induite par la variable explicative  $X$  et la variable à expliquer  $Q$ .  $\text{var}(S_X)$  désigne la variance de  $S_X$  (cf. annexe). Les valeurs observées de  $\frac{S_X}{\sqrt{(\text{var}(S_X))}}$  sont comparées au quantile d'ordre 0.95 de la loi normale centrée réduite.

Les variables  $X_3$  (Moirage|Correction possible) et  $X_1$  (Prix), sont les plus liées à la variable à expliquer, la concordance est significativement forte. L'examen des valeurs observées de la variable  $X_2$  (Fréquence maximale), montre que c'est la variable la moins discriminante (relativement à cette variable, les deux groupes sont presque identiques, (cf. tableau 1)), ce qui est aussi confirmé par la discordance entre  $P_{X_2}$  et  $P_Q$  ( $\frac{S_{X_2}}{\sqrt{(\text{var}(S_{X_2}))}} < 0$ ).

### Résultats des algorithmes pas à pas ascendant (cas des préordonnances)

TABLEAU 3

Pas	Critère à maximiser	Valeur maximale du critère	Variable sélectionnée
1	$(\Psi_{\text{cor}})$	0.322	$X_3$
2	$(\Psi_{\text{cor}}).P_{X_3}$	0.147	$X_1$
3	$(\Psi_{\text{cor}}).P_{X_3}, P_{X_1}$	0.064	$X_4$
4	$(\Psi_{\text{cor}}).P_{X_3}, P_{X_1}, P_{X_4}$	0.051	$X_5$
5	$(\Psi_{\text{cor}}).P_{X_3}, P_{X_1}, P_{X_4}, P_{X_5}$	-0.002	$X_2$

TABLEAU 4

Pas	Critère à maximiser	Valeur maximale du critère	Variable sélectionnée
1	$(\Psi_{cov})$	0.123	$X_3$
2	$(\Psi_{cov}).P_{X_3}$	0.102	$X_1$
3	$(\Psi_{cov}).P_{X_3},P_{X_1}$	0.054	$X_4$
4	$(\Psi_{cov}).P_{X_3},P_{X_1},P_{X_4}$	0.017	$X_5$
5	$(\Psi_{cov}).P_{X_3},P_{X_1},P_{X_4},P_{X_5}$	-0.009	$X_2$

Résultats des algorithmes pas à pas ascendant (cas des triordonnances)

TABLEAU 5

Pas	Critère à maximiser	Valeur maximale du critère	Variable sélectionnée
1	$(\Psi'_{cov})$	0.314	$X_3$
2	$(\Psi'_{cov}).P_{X_3}$	0.166	$X_1$
3	$(\Psi'_{cov}).P_{X_3},P_{X_1}$	0.059	$X_4$
4	$(\Psi'_{cov}).P_{X_3},P_{X_1},P_{X_4}$	0.056	$X_5$
5	$(\Psi'_{cov}).P_{X_3},P_{X_1},P_{X_4},P_{X_5}$	0.002	$X_2$

TABLEAU 6

Pas	Critère à maximiser	Valeur maximale du critère	Variable sélectionnée
1	$(\Psi'_{cov})$	0.112	$X_3$
2	$(\Psi'_{cov}).P_{X_3}$	0.109	$X_1$
3	$(\Psi'_{cov}).P_{X_3},P_{X_1}$	0.049	$X_4$
4	$(\Psi'_{cov}).P_{X_3},P_{X_1},P_{X_4}$	0.02	$X_5$
5	$(\Psi'_{cov}).P_{X_3},P_{X_1},P_{X_4},P_{X_5}$	-0.004	$X_2$

Résultats des algorithmes pas à pas descendant

TABLEAU 7

Étape	0	1	2	3	4
Valeur maximale du critère $\Psi_{R_{cor}}$	0.129	0.129	0.127	0.123	0.104
Variable à éliminer		$X_2$	$X_5$	$X_4$	$X_1$



TABLEAU 8

Étape	0	1	2	3	4
Valeur maximale du critère $\Psi_{R_{cov}}$	0.029	0.028	0.028	0.025	0.015
Variable à éliminer		$X_2$	$X_5$	$X_4$	$X_1$

Les deux approches (préordonnance et triordonnance) donnent le même résultat (cf. tableaux 3, 4, 5 et 6). La variable  $X_2$ , à association négative avec  $Q$  (cf. tableau 2), est la dernière variable à être retenue, par application des algorithmes de sélection pas à pas ascendant; elle est aussi la première variable éliminée par application des algorithmes pas à pas descendant (cf. tableaux 7 et 8); remarquons que cette variable quantitative prend deux valeurs, l'une observée sur un seul individu, l'autre observée sur tout le reste de l'échantillon; elle ne présente donc pas d'intérêt.

Les deux algorithmes pas à pas descendant donnent le même résultat. Elles concordent avec les algorithmes pas à pas ascendant

### 6.2. Exemple 2

Les données suivantes sont extraites de la thèse présentée par Lazraq A. (1988).

TABLEAU 9

$X_1$	2.02	2.62	2.08	2.2	2.38	2.87	1.88	1.93	2.57	1.95	2.03	2.5
$X_2$	2.9	2.78	2.68	3.17	2.52	2.67	2.58	2.26	1.74	2.15	2	2.07
$X_3$	2.17	1.72	2.4	2.06	2.18	2.64	2.22	2.15	1.64	2.48	2.38	2.32
$X_4$	0.51	0.5	0.43	0.52	0.42	0.5	0.49	0.56	0.51	0.48	0.5	0.48
$X_5$	3.47	4.57	3.52	3.69	4.01	3.92	3.58	3.57	4.38	3.28	3.31	3.72
$X_6$	0.91	1.25	0.82	0.97	1.12	1.06	1.01	0.92	1.22	0.81	0.98	1.04
$Q$	1	2	1	1	2	2	1	1	2	1	1	2
$X_1$	1.72	2.53	1.9	2.18	2.16	2.14	1.98	1.89	2.07	2.08	2.21	2
$X_2$	2.24	1.74	1.46	0.74	2.84	3.3	2.9	2.82	2.79	3.14	2.81	3.16
$X_3$	2.25	2.64	1.97	2.46	2.36	2.04	2.16	2.04	2.15	2.6	2.18	2.22
$X_4$	0.52	0.5	0.46	0.48	0.49	0.48	0.48	0.53	0.52	0.5	0.44	0.51
$X_5$	3.1	3.48	3.48	3.16	3.68	3.28	3.56	3.56	3.49	3.3	4.16	3.73
$X_6$	0.78	0.93	0.9	0.86	0.95	1.06	0.84	1.02	1.04	0.8	0.92	1.07
$Q$	1	2	2	2	1	1	1	1	2	1	2	1

L'échantillon est constitué de 24 individus, six variables explicatives, qui mesurent respectivement le pourcentage d'azote, de chlore, de potassium, de phosphore, de calcium et de magnésium dans la feuille de tabac, et une variable à expliquer  $Q$ , qui représente le pourcentage de sucre dans la feuille de tabac.  $Q$  a deux modalités, 1 : pourcentage élevé, 2 : pourcentage faible.

On cherche à expliquer la variable  $Q$  à partir d'un sous ensemble de variables inorganiques de la feuille de tabac, par application des nouveaux algorithmes introduits dans cet article et de la méthode classique de sélection basée sur la statistique de Fisher. L'analyse discriminante est effectuée selon l'approche bayésienne, au moyen de la troisième version du logiciel d'analyse de données : **Chadoc**, sous les hypothèses de normalité et d'homoscédasticité.

### Résultats de l'approche classique

La méthode nécessite de préciser le nombre  $k$  de variables à sélectionner. Pour chaque valeur de  $k$ , on indique les variables sélectionnées en pas à pas et le pourcentage de bien classés, qui correspond à la moyenne arithmétique des pourcentages de bien classés à l'intérieur des différentes classes :

Pour  $k = 6$ ; les variables sélectionnées en pas à pas sont :  $X_1, X_2, X_5, X_4, X_6, X_3$ .

Pour  $k = 6, k = 5, k = 4, k = 3$  : le pourcentage de bien classés est 95%.

Pour  $k = 2$  : le pourcentage de bien classés est 85%.

### Résultats des algorithmes pas à pas ascendant (cas des préordonnances)

TABLEAU 10

Pas	Critère à maximiser	Valeur maximale du critère	Variable sélectionnée
1	$(\Psi_{\text{cor}})$	0.311	$X_1$
2	$(\Psi_{\text{cor}}).P_{X_1}$	0.103	$X_2$
3	$(\Psi_{\text{cor}}).P_{X_1}, P_{X_2}$	0.088	$X_5$
4	$(\Psi_{\text{cor}}).P_{X_1}, P_{X_2}, P_{X_5}$	0.021	$X_4$
5	$(\Psi_{\text{cor}}).P_{X_1}, P_{X_2}, P_{X_5}, P_{X_4}$	-0.005	$X_6$
6	$(\Psi_{\text{cor}}).P_{X_1}, P_{X_2}, P_{X_5}, P_{X_4}, P_{X_6}$	-0.015	$X_3$

TABLEAU 11

Pas	Critère à maximiser	Valeur maximale du critère	Variable sélectionnée
1	$(\Psi_{\text{cov}})$	0.218	$X_1$
2	$(\Psi_{\text{cov}}).P_{X_1}$	0.071	$X_2$
3	$(\Psi_{\text{cov}}).P_{X_1}, P_{X_2}$	0.061	$X_5$
4	$(\Psi_{\text{cov}}).P_{X_1}, P_{X_2}, P_{X_5}$	0.013	$X_4$
5	$(\Psi_{\text{cov}}).P_{X_1}, P_{X_2}, P_{X_5}, P_{X_4}$	-0.003	$X_6$
6	$(\Psi_{\text{cov}}).P_{X_1}, P_{X_2}, P_{X_5}, P_{X_4}, P_{X_6}$	-0.01	$X_3$

**Résultats des algorithmes pas à pas ascendant (cas des triordonnances)**

TABLEAU 12

Pas	Critère à maximiser	Valeur maximale du critère	Variable sélectionnée
1	$(\Psi'_{cor})$	0.306	$X_1$
2	$(\Psi'_{cor}).P_{X_1}$	0.111	$X_2$
3	$(\Psi'_{cor}).P_{X_1}, P_{X_2}$	0.078	$X_5$
4	$(\Psi'_{cor}).P_{X_1}, P_{X_2}, P_{X_5}$	0.018	$X_4$
5	$(\Psi'_{cor}).P_{X_1}, P_{X_2}, P_{X_5}, P_{X_4}$	0.013	$X_6$
6	$(\Psi'_{cor}).P_{X_1}, P_{X_2}, P_{X_5}, P_{X_4}, P_{X_6}$	-0.022	$X_3$

TABLEAU 13

Pas	Critère à maximiser	Valeur maximale du critère	Variable sélectionnée
1	$(\Psi'_{cov})$	0.208	$X_1$
2	$(\Psi'_{cov}).P_{X_1}$	0.073	$X_2$
3	$(\Psi'_{cov}).P_{X_1}, P_{X_2}$	0.053	$X_5$
4	$(\Psi'_{cov}).P_{X_1}, P_{X_2}, P_{X_5}$	0.011	$X_4$
5	$(\Psi'_{cov}).P_{X_1}, P_{X_2}, P_{X_5}, P_{X_4}$	0.01	$X_6$
6	$(\Psi'_{cov}).P_{X_1}, P_{X_2}, P_{X_5}, P_{X_4}, P_{X_6}$	-0.013	$X_3$

**Résultats des algorithmes pas à pas descendant**

TABLEAU 14

Étape	0	1	2	3	4	5
Valeur maximale du critère $\Psi_{R_{cor}}$	0.114	0.114	0.113	0.113	0.106	0.096
Variable à éliminer		$X_6$	$X_3$	$X_4$	$X_5$	$X_2$

TABLEAU 15

Étape	0	1	2	3	4	5
Valeur maximale du critère $\Psi_{R_{cov}}$	0.056	0.056	0.056	0.056	0.052	0.048
Variable à éliminer		$X_6$	$X_3$	$X_4$	$X_5$	$X_2$

Les deux approches (préordonnances et triordonnances) (cf. tableaux 10,11,12, et 13) donnent le même résultat : ainsi pour  $k = 6$ , les variables sélectionnées en pas

à pas sont :  $X_1, X_2, X_5, X_4, X_6, X_3$ , le pourcentage de bien classés est 95%. Les résultats des algorithmes pas à pas ascendant concordent avec ceux de l'approche classique. Concernant les variables explicatives à association significative avec la variable à expliquer, à savoir les variables  $X_1, X_2, X_5$ , les rapports de Fisher sont les plus significatifs : (cf. tableau 16)

TABLEAU 16

Variable $X$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
$\frac{S_X}{\sqrt{\text{var}(S_X)}}$	6.273	2.018	1.373	-0.175	3.316	1.32
Rapport de Fisher	18.82	6.66	0.34	2.2	9.55	5.31

Les valeurs observées du rapport de Fisher sont comparées au quantile d'ordre 0.95 de la loi Fisher à (1, 22) degré de liberté (= 4.3009). Remarquons que la valeur observée  $\Psi_{\text{cor}}(P_{X_6}, P_Q)$  est non significative ( $\frac{S_{X_6}}{\sqrt{\text{var}(S_{X_6})}} = 1.32$ ), alors que le rapport de Fisher correspondant à  $X_6$  est significatif; pourtant la variable  $X_6$  est parmi les variables les moins dispersées (variance totale égale à 0.36). Sa variance intraclasse est aussi parmi les plus faibles (égale à 0.29), mais elle reste voisine de la variance totale au profit d'une variance interclasse très faible (égale à 0.07). Ce cas montre que le rapport de Fisher associé à une variable explicative peut être significativement élevé sans que les groupes décrits par la variable en question ne soient homogènes.

Les résultats obtenus par les deux algorithmes pas à pas descendant (cf. tableaux 14, 15) concordent. Pour  $k = 4$ , les deux approches pas à pas descendant et pas à pas ascendant conduisent au même résultat. Par contre pour  $k = 5$  les résultats diffèrent légèrement.

### 6.3. Discussion

Dans les algorithmes proposés, on s'intéresse à des associations positives. Une association négative correspond à un désaccord entre variable dépendante et variable explicative, et par conséquent, cette dernière ne peut être sélectionnée. Lors du déroulement des algorithmes, le processus de sélection est interrompu, si à un pas  $k_0$ , l'association calculée est négative, même si  $k_0 < k$ . Dans le cas où, à chaque pas, les associations calculées sont positives, soit on se fixe  $k$  (le nombre de variables à sélectionner), soit on s'arrête lorsque le nombre de variables qui minimisent le risque d'erreur par validation croisée est atteint.

Examinons les tableaux 3, 4, 5, 6, 10, 11, 12 et 13 :

Dans le cas des préordonnances (cf. tableaux 3 et 4), la valeur du critère, au pas 5, calculée pour la variable  $X_2$  est négative. Donc, au plus, le processus de sélection doit s'arrêter au pas 5. Dans le cas des triordonnances (cf. tableaux 5 et 6), la valeur du critère ( $\Psi'_{\text{cor}}$ ), obtenue au cinquième pas, pour la variable  $X_2$  est positive faible

(cf. tableau 5). Celle du critère  $(\Psi'_{\text{cov}})$ , obtenue au même pas, pour la même variable est négative (cf. tableau 6). Le processus de sélection s'arrête, au plus, au cinquième pas. Les résultats consignés dans les tableaux 10 et 11 (cas des préordonnances), montrent que les valeurs des critères obtenues au cinquième et sixième pas sont négatives. Le processus de sélection s'arrête, au plus, au cinquième pas. Dans le cas des triordonnances (cf. tableaux 12 et 13), le processus de sélection s'arrête, au plus, au sixième pas.

Appliqués sur les deux exemples, les algorithmes basés sur les notions de préordonnance et triordonnance concordent. Quoiqu'on relève une différence entre les résultats du test d'arrêt basé sur le signe des valeurs des critères. En effet, une préordonnance est une relation de préordre total alors qu'une triordonnance est une relation de préordre partiel. De sorte que, si désaccord il y'a, celui ci est mieux mis en évidence par le critère basé sur les préordonnances que celui basé sur les triordonnances.

Notons, en outre, qu'il n'existe aucune relation entre  $\Psi_{\text{cor}}$  et  $\Psi'_{\text{cor}}$ , ni entre  $(\Psi_{\text{cor}})$  et  $(\Psi'_{\text{cor}})$ , de sorte qu'aucune comparaison entre les valeurs de ces critères n'est possible : dans le cas de l'exemple 1

$$\Psi_{\text{cor}}(P_{X_1}, P_Q)_{.P_{X_3}} < \Psi'_{\text{cor}}(P_{X_1}, P_Q)_{.P_{X_3}}$$

alors que

$$\Psi_{\text{cor}}(P_{X_4}, P_Q)_{.P_{X_1}, P_{X_3}} > \Psi'_{\text{cor}}(P_{X_4}, P_Q)_{.P_{X_1}, P_{X_3}}$$

dans le cas de l'exemple 2

$$\Psi_{\text{cor}}(P_{X_2}, P_Q)_{.P_{X_1}} < \Psi'_{\text{cor}}(P_{X_2}, P_Q)_{.P_{X_1}}$$

alors que

$$\Psi_{\text{cor}}(P_{X_5}, P_Q)_{.P_{X_1}, P_{X_2}} > \Psi'_{\text{cor}}(P_{X_5}, P_Q)_{.P_{X_1}, P_{X_2}}$$

Finalemnt, bien que pour les deux exemples, les valeurs du  $\Psi_{\text{cor}}$  et  $\Psi'_{\text{cor}}$  décroissent d'un pas au suivant, il n'en est pas de même en général. En effet : considérons deux variables explicatives  $X_1$  et  $X_2$  pour lesquelles,

$$\Psi_{\text{cor}}(P_{X_1}, P_Q) = 0.06, \Psi_{\text{cor}}(P_{X_2}, P_Q) = 0.05, \Psi_{\text{cor}}(P_{X_1}, P_{X_2}) = -0.2$$

$$\Psi_{\text{cor}}(P_{X_2}, P_Q)_{.P_{X_1}} = 0.062$$

C'est que alors

$$\Psi_{\text{cor}}(P_{X_2}, P_Q)_{.P_{X_1}} > \Psi_{\text{cor}}(P_{X_1}, P_Q)$$

Concernant les lois limites des coefficients  $(\Psi_{\text{cor}})$ ,  $\Psi_{R_{\text{cor}}}$ , dans la mesure où même pour les petites valeurs de  $n$ , les lois de distribution des coefficients de

corrélation partielle et multiple des rangs sont difficiles à calculer [Kendall M.G., 1962], il en est de même pour  $(\Psi_{cor}), \Psi_{R_{cor}}$ .

La variable  $X_2$  (cf. Exemple 1), qui prend la même valeur sur tous les individus sauf un (cf. tableau 1), ne présente aucun intérêt pour la discrimination; on pourrait initialement ne pas la garder. Cependant notre objectif étant de montrer que les algorithmes conduisent bien à l'élimination de cette variable, nous l'avons conservée.

### Remerciement

Les auteurs ont bénéficié des critiques et suggestions du Professeur Pierre Cazes.

### Annexe [Kendall M. G., 1962]

#### Coefficient $\tau$ de Kendall

Soient  $r_1$  et  $r_2$  les variables rangs induites par deux caractères  $X$  et  $Y$ .

#### Cas d'absence d'*ex aequo*

Soient  $P$  le nombre de paires  $(i, j)$  appartenant à  $H$  vérifiant :

$(r_1(i) - r_1(j))(r_2(i) - r_2(j)) > 0$  et  $Q$  le nombre de paires  $(i, j)$  appartenant à  $H$  vérifiant :

$(r_1(i) - r_1(j))(r_2(i) - r_2(j)) < 0$ . Posons  $S = P - Q$ , le nombre maximum que peut atteindre  $P$  est  $\frac{n(n-1)}{2}$  et  $\tau(X, Y)$  est défini par :

$$\tau(X, Y) = \frac{S}{M}$$

**Proposition 4.** Soient  $A$  et  $B$  le codage des comparaisons par paires des variables rangs  $r_1$  et  $r_2$  :

$$A_{ij} = \begin{cases} 1 & \text{si } r_1(i) < r_1(j) \\ -1 & \text{si } r_1(i) > r_1(j) \end{cases} \quad B_{ij} = \begin{cases} 1 & \text{si } r_2(i) < r_2(j) \\ -1 & \text{si } r_2(i) > r_2(j) \end{cases} \quad A_{ii} = B_{ii} = 0$$

alors  $\tau(X, Y) = \text{cor}(A, B) = \text{cov}(A, B)$

en effet :

$$\sum_{i \neq j} A_{ij} B_{ij} = 2S = 2(P - Q) \text{ et } \sum_{i \neq j} A_{ij}^2 = \sum_{i \neq j} B_{ij}^2 = n(n - 1).$$

#### Distribution sous l'hypothèse d'indépendance

La distribution de  $\tau$  est tabulée pour les petites valeurs de  $n$ , mais elle peut être approchée par une loi de Laplace Gauss : de moyenne 0 et de variance  $\frac{2(2n + 5)}{9n(n - 1)}$ .

L'approximation est valable dès que  $n \geq 10$ . Lors du calcul de la distribution de  $\tau$  sous l'hypothèse nulle, on s'intéresse à la quantité  $S$  qui est discrète et approchée par une variable continue, on procède à une correction de continuité qui consiste à retrancher 1 de la valeur observée de  $S$  si elle est positive, d'ajouter 1 si elle est négative.

### Cas de présence d'*ex aequo*

Dans le cas de présence d'*ex aequo*, on retient le codage suivant pour les variables rangs  $r_1$  et  $r_2$  :

$$A_{ij} = \begin{cases} 1 & \text{si } r_1(i) < r_1(j) \\ 0 & \text{si } r_1(i) = r_1(j) \\ -1 & \text{si } r_1(i) > r_1(j) \end{cases} \quad B_{ij} = \begin{cases} 1 & \text{si } r_2(i) < r_2(j) \\ 0 & \text{si } r_2(i) = r_2(j) \\ -1 & \text{si } r_2(i) > r_2(j) \end{cases}$$

Au préalable, on affecte aux paires *ex aequo* la moyenne arithmétique des rangs qu'ils auraient s'il n'y avait pas eu d'*ex aequo*.

De façon précise, supposons qu'il y ait  $n_1$  (resp.  $n_2$ ) groupes d'*ex aequo* pour  $r_1$  (resp.  $r_2$ ) et désignons par  $u_i$  ( $1 \leq i \leq n_1$ ) (resp.  $v_j$ ,  $1 \leq j \leq n_2$ ) le nombre d'individus *ex aequo* du  $i^{\text{ème}}$  (resp.  $j^{\text{ème}}$ ) groupe.

Deux expressions  $\tau_1$  et  $\tau_2$  sont proposées pour le coefficient  $\tau$  de Kendall, la première,  $\tau_1$ , s'écrit sous la forme :

$$\tau_1(X, Y) = \frac{S}{\sqrt{\left(\left(\frac{n(n-1)}{2} - V\right)\left(\frac{n(n-1)}{2} - U\right)\right)},$$

$$V = \frac{\sum_j v_j(v_j - 1)}{2}, \quad U = \frac{\sum_i u_i(u_i - 1)}{2}$$

l'autre expression,  $\tau_2$ , du coefficient  $\tau$  s'écrit sous la forme :

$$\tau_2(X, Y) = \frac{S}{\frac{n(n-1)}{2}}$$

**Proposition 5.**  $\tau_1(X, Y) = \text{cor}(A, B)$ ,  $\tau_2(X, Y) = \text{cov}(A, B)$

En effet :

$$\sum_{i \neq j} A_{ij} B_{ij} = 2S, \quad \sum_{i \neq j} A_{ij}^2 = n(n-1) - 2U \quad \text{et} \quad \sum_{i \neq j} B_{ij}^2 = n(n-1) - 2V$$

Dans le cas d'absence d'*ex aequo*,  $\tau_1(X, Y) = \tau_2(X, Y)$

### Distribution sous l'hypothèse nulle

La distribution de  $S$  sous l'hypothèse nulle est tabulée pour les petites valeurs de  $n$  [Sillitto, 1947]; dès que  $n \geq 10$ , elle peut être approchée par une loi normale de moyenne 0, de variance :

$$\begin{aligned} & \frac{1}{18} \left[ n(n-1)(2n+5) - \sum_j (v_j(v_j-1)(2v_j+5)) - \sum_i (u_i(u_i-1)(2u_i+5)) \right] \\ & + \frac{1}{9n(n-1)(n-2)} \left[ \sum_j (v_j(v_j-1)(v_j-2)) \sum_i (u_i(u_i-1)(u_i-2)) \right] \\ & + \frac{1}{2n(n-1)} \left[ \sum_j (v_j(v_j-1)) \sum_i (u_i(u_i-1)) \right] \end{aligned}$$

### Corrélation partielle des rangs

Supposons que  $n$  individus ont été classés selon  $p$  critères :  $r'_1, r'_2, \dots, r'_p$ . Pour étudier l'association entre  $r'_1, r'_2$ , l'effet de  $r'_3$  étant éliminé, on utilise le coefficient de corrélation partielle des rangs, ( $\tau$  peut prendre l'une ou l'autre des expressions  $\tau_1, \tau_2$ ) défini par :

$$\tau(r'_1, r'_2)_{.r'_3} = \frac{\tau(r'_1, r'_2) - \tau(r'_1, r'_3)\tau(r'_2, r'_3)}{\sqrt{(1 - \tau^2(r'_1, r'_3))(1 - \tau^2(r'_2, r'_3))}}$$

Plus généralement, on définit le coefficient d'association partielle d'ordre  $h$  entre rangs :

$$\tau(r'_1, r'_2)_{.r'_h, r'_{h-1}, \dots, r'_3} = \frac{\tau(r'_1, r'_2)_{.r'_{h-1}, \dots, r'_3} - \tau(r'_1, r'_h)_{.r'_{h-1}, \dots, r'_3} \tau(r'_2, r'_h)_{.r'_{h-1}, \dots, r'_3}}{\sqrt{(1 - \tau^2(r'_1, r'_h)_{.r'_{h-1}, \dots, r'_3})(1 - \tau^2(r'_2, r'_h)_{.r'_{h-1}, \dots, r'_3})}}$$

### Corrélation multiple des rangs

Le carré du coefficient de corrélation multiple des rangs entre  $r_1$  et les variables rangs  $r_2, r_3$  est défini par :

$$1 - R^2(r_1, (r_2, r_3)) = (1 - \tau^2(r_1, r_2))(1 - \tau^2(r_1, r_3)_{.r_2})$$

De façon plus générale, on a

$$1 - R^2(r_1, (r_2, r_3, \dots, r_h)) = (1 - R^2(r_1, (r_2, r_3, \dots, r_{h-1}))) (1 - \tau^2(r_1, r_h)_{.r_2 \dots r_{h-1}})$$



### Références

- [1] CHAH S., (1985), *Critères de classification sur données hétérogènes*, Rev. Statistique Appliquée, vol. 33, n° 2, pp. 19-36.
- [2] CHAH S. (1986a), *Comparaison par triplets en classification automatique*, Rev. Statistique Appliquée, vol. 34, n° 1, pp. 61-79.
- [3] CHAH S. (1986b), *Nouvelles techniques de codage d'association et de classification*, Thèse de Doctorat d'état ès sciences, Université Pierre et Marie Curie Paris VI.
- [4] DIDAY E., LEMAIRE J., POUGET P., TESTU F. (1983), *Eléments d'analyse des données*, Dunod-Paris.
- [5] GUÉGUEN A., NAKACHE J.-P. (1988), *Méthode de discrimination basée sur la construction d'un arbre de décision binaire*, Rev. Statistique Appliquée 36, 1, pp. 19-38
- [6] KENDALL M.G. (1962), *Rank correlation methods*, Griffin, Londres.
- [7] KRUSŮRSKA E. M. (1991), *Discriminant Analysis in graphical and hierarchical interaction models*, Proceeding of the 6<sup>th</sup> international workshop on statistical modelling, Utrecht, W. Jansen et P. Van der heijden (éd).
- [8] LAZRAQ A. (1988), *Inférences sur plusieurs mesures de liaison entre deux vecteurs aléatoires et Algorithmes de sélection de variables*, Thèse présentée à la faculté des études supérieures en vue de l'obtention du grade de philosophiae doctor (PH.D) en Informatique, Université de Montréal.
- [9] LEBART L. (1992), *Discrimination through the regularized nearest cluster method*, *Compstat*, Y. Dodge et J. Whittaker (éd)., Neuchatel : Springer-Verlag, pp. 103-118.
- [10] LEBART L., MORINEAU A., FENELON J.-P., *Traitement des données statistiques, méthodes et programmes*, Dunod.
- [11] LEMAIRE J., *Chadoc vs-version 3-*, Département Informatique de l'IUT de Nice.
- [12] NAKACHE J.-P., CELEUX G. (1994), *Analyse discriminante sur variables qualitatives*, Polytechnica, chapitre 5, pp. 134-135.
- [13] NAKACHE J.-P., LORENTE P., BENZÉKRI J.-P., CHASTANG J. F. (1977), *Aspects pronostics et thérapeutiques de l'infarctus myocardique aigu[Aorte]*, Les cahiers de l'Analyse des données 2, pp. 415-434.
- [14] NAKACHE J.-P. (1980), *Méthodes de discrimination pour variables de nature quelconque, théorie et pratique*, Thèse présentée pour obtenir le grade de Docteur ès sciences, Université Pierre et Marie Curie.
- [15] NIEROP A. F. M. (1993), *Multidimensional analysis of grouped variables : an integrated approach*, DSWO Press, pp. 67-89.
- [16] SAPORTA G. (1975), *Liaison entre plusieurs ensembles de variables et codages de données qualitatives*, Thèse de troisième cycle, Université Paris VI.
- [17] SAPORTA G. (1990), *Probabilités, analyse des données et statistique*, Editions Technip-France.
- [18] SILLITO G. P. (1947), *The distribution of Kendall's  $\tau$  coefficient of rank correlation in ranking containing ties*, *Biometrika*, pp. 34-36.