REVUE DE STATISTIQUE APPLIQUÉE

A. BACCINI
H. CAUSSINUS
A. RUIZ-GAZEN

Apprentissage progressif en analyse discriminante

Revue de statistique appliquée, tome 49, nº 4 (2001), p. 87-99
<http://www.numdam.org/item?id=RSA_2001__49_4_87_0>

© Société française de statistique, 2001, 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/

Rev. Statistique Appliquée, 2001, **XLIX** (4), 87-99

APPRENTISSAGE PROGRESSIF EN ANALYSE DISCRIMINANTE

A. Baccini, H. Caussinus et A. Ruiz-Gazen

Laboratoire de Statistique et Probabilités
U.M.R. - C.N.R.S. C 5583
Université Paul Sabatier — 118, route de Narbonne
31062 Toulouse cedex 4

RÉSUMÉ

L'analyse discriminante affecte une unité statistique à un groupe selon une règle de décision obtenue à partir d'un échantillon d'apprentissage. Cette procédure suppose que les conditions d'utilisation restent celles de l'apprentissage, ce qui n'est pas toujours garanti. Sous des hypothèses assez larges, il est cependant possible de faire face à une dérive inconnue des conditions expérimentales. Nous introduisons pour cela une analyse convenablement modifiée qui complète l'apprentissage initial par un contrôle progressif des conditions d'utilisation. Un exemple réel permet de montrer la mise en œuvre concrète de la méthode ainsi que son efficacité.

Mots-clés : *analyse discriminante, classification supervisée, conditionnement, apprentissage progressif.*

ABSTRACT

Discriminant analysis assigns a statistical unit to a group according to a decision rule obtained from a learning sample. This needs that the background remains stable along the use of the procedure after the learning step, which is not always warranted. Under fairly mild assumptions, however, one can face unknown trends in experimental conditions. For that, we introduce a modified analysis which completes the initial learning by a suitable control of the possible trend. The procedure is implemented on a real example which serves also to evidence its efficiency.

Keywords : *discriminant analysis, supervised classification, conditioning, progressive learning.*

1. Introduction

Dans cet article, nous considérons le problème de l'analyse discriminante décisionnelle dans lequel il s'agit d'affecter une unité statistique à un parmi g groupes, sur la base de p mesures réelles. On suppose que, pour chaque groupe, les mesures se dispersent autour d'un vecteur moyen caractérisant le groupe, la dispersion à l'intérieur d'un groupe étant sensiblement de même nature pour chacun d'eux. Dans la situation usuelle, la moyenne de chaque groupe ainsi que la variabilité à l'intérieur

du groupe sont estimées au moyen d'un échantillon d'apprentissage sur lequel les appartenances aux groupes sont connues, ce qui permet de définir la combinaison des mesures qui différencie le mieux ces groupes, c'est-à-dire qui fait la plus grande part possible à la dispersion des mesures d'un groupe à l'autre (intergroupe) par rapport à la variabilité à l'intérieur de chacun d'entre eux (intragroupe). Après apprentissage, cette combinaison discriminante sert à affecter toute nouvelle unité statistique dont l'appartenance à un groupe est désormais inconnue.

Bien entendu, une telle procédure n'a de sens (ou, du moins, n'est pleinement efficace) que si l'échantillon d'apprentissage et les données ultérieures sont observés dans les mêmes conditions. Or, cette hypothèse est très douteuse dans de nombreuses situations réelles : par exemple, un apprentissage réalisé sur une courte période ne peut pas intégrer les facteurs de variabilité susceptibles d'apparaître seulement sur une longue période; en ce sens, l'apprentissage n'est que *partiel*. Ces facteurs, que nous appellerons de *dérive*, peuvent s'avérer importants au point de compromettre très sérieusement, et souvent *insidieusement*, l'efficacité, voire la pertinence, de la méthode de discrimination. L'objet de cet article est de proposer quelques pistes pour prendre en compte ces facteurs dans la procédure d'affectation, problème négligé par les ouvrages classiques sur le sujet (par exemple, McLachlan, 1992). Notons cependant qu'une problématique analogue a parfois été abordée dans d'autres domaines de la statistique : voir, par exemple, Goupy (1989) pour la prise en compte, dans le contexte de l'expérimentation planifiée, d'une dérive *systématique* (hypothèse nettement plus restrictive que celle envisagée ici).

Une remarque préalable s'impose : une fois l'apprentissage effectué, on ne connait pas a priori le groupe auquel chaque produit contrôlé appartient réellement; dès lors, on voit mal comment identifier un groupe si la valeur des mesures effectuées peut être amenée à tel ou tel niveau par un phénomène externe, non contrôlé, dépendant de ce groupe inconnu. Il semble par contre envisageable de tenir compte de modifications qui toucheraient tous les groupes de la même manière (d'une façon précisée plus loin) : c'est le problème étudié ici.

2. Rappels sur l'Analyse Discriminante

L'objectif de ce paragraphe est de donner quelques rappels sur l'Analyse Discriminante (que l'on pourra trouver plus détaillée, par exemple, dans l'ouvrage de Saporta, 1990), en adoptant une présentation appropriée aux modifications proposées au paragraphe 3 en vue de prendre en compte d'éventuels phénomènes de dérive.

Soit g groupes de données vectorielles (à k dimensions), le j-ième groupe ayant une distribution $N_k(\mu_j, W)$. Si, de plus, la probabilité a priori qu'une observation x appartienne au j-ième groupe est p_j, la probabilité a posteriori du groupe j, cette observation étant faite, est

$$\frac{p_j \exp\left(-\frac{1}{2}\|x-\mu_j\|^2_{W^{-1}}\right)}{\sum_{i=1}^{g} p_i \exp\left(-\frac{1}{2}\|x-\mu_i\|^2_{W^{-1}}\right)},$$

avec, de façon usuelle,

$$\|x - \mu_j\|^2_{W^{-1}} = {}^t(x - \mu_j)W^{-1}(x - \mu_j),$$

x et μ_j étant ici des vecteurs $k \times 1$ et W une matrice $k \times k$ symétrique qu'on supposera régulière.

Le groupe de plus grande probabilité a posteriori est donc celui pour lequel est maximum $p_j \exp\left(-\frac{1}{2}\|x - \mu_j\|^2_{W^{-1}}\right)$, ou est minimum

$$\|x - \mu_j\|^2_{W^{-1}} - 2 \ln p_j. \tag{1}$$

Dans la pratique, les μ_j et W sont inconnus et estimés à partir d'un échantillon dit d'apprentissage dans lequel l'affectation au groupe est connue; l'estimation de W est la matrice des covariances intragroupes empiriques, tandis que les estimations des μ_j sont soit les moyennes empiriques des groupes, soit les projections (orthogonales au sens de W^{-1}) de ces moyennes empiriques sur un sous-espace de dimension q obtenu par Analyse Factorielle Discriminante (AFD) (comme $q \leqslant g - 1$, on obtient $q = 1$ si $g = 2$; dans les autres cas, on choisit le plus souvent $q = 2$, les deux procédures étant équivalentes pour $g = 3$).

Rappelons que l'AFD repose essentiellement sur la diagonalisation de BW^{-1} où $B = \dfrac{1}{g-1}\displaystyle\sum_{j=1}^{g}(\mu_j - \bar{\mu})\,{}^t(\mu_j - \bar{\mu})$ est la matrice de dispersion des μ_j, ou variance intergroupes, $\bar{\mu} = \dfrac{1}{g}\displaystyle\sum_{j=1}^{g}\mu_j$ étant la moyenne des μ_j (cette formule pour B est naturelle si les p_j sont égaux; sinon, il convient de pondérer convenablement).

L'AFD est utilisée de longue date à partir d'arguments empiriques variés, parmi lesquels l'utilité pratique d'une représentation en dimension 2 (pour $q = 2$); on peut trouver une justification théorique d'une réduction de dimension dans Ferré (1995), où est discutée l'estimation optimale des μ_j.

Par la suite, on remplacera dans (1) W et les μ_j par leurs estimations, sans changer les notations pour ne pas alourdir. Par ailleurs, on supposera les p_j égaux pour simplifier, ce qui ramène le problème (1) à choisir j qui minimise

$$\|x - \mu_j\|^2_{W^{-1}}. \tag{2}$$

Il est facile d'introduire la correction $-2 \ln p_j$ dans le cas contraire, à condition que les p_j soient connus ou correctement estimés.

Remarque 1

Transformons provisoirement en $W^{-1/2}X$ le vecteur aléatoire d'observations X; sa loi de probabilité pour le groupe j devient $N_k(W^{-1/2}\mu_j, I_k)$. Il est alors clair que les groupes sont d'autant mieux différenciés que les $W^{-1/2}\mu_j$ sont plus dispersés. Si

cette dispersion est mesurée de façon quadratique, elle peut s'écrire :

$$\sum_{j=1}^{g} {}^t(\mu_j - \bar{\mu})W^{-1/2}W^{-1/2}(\mu_j - \bar{\mu}) = \sum_{j=1}^{g} {}^t(\mu_j - \bar{\mu})W^{-1}(\mu_j - \bar{\mu}). \qquad (3)$$

En se basant sur (3), on voit que les groupes sont d'autant mieux différenciés que $\sum_{j=1}^{g} {}^t(\mu_j - \bar{\mu})W^{-1}(\mu_j - \bar{\mu})$ est plus grand. Donc, pour des μ_j donnés, W^{-1} doit être le plus grand possible au sens des formes quadratiques (on rappelle que, pour des matrices $k \times k$ symétriques A_1 et A_2, on pose $A_1 \geqslant A_2$ si et seulement si, pour tout vecteur colonne x, ${}^t x A_1 x \geqslant {}^t x A_2 x$), ce qui est équivalent à W le plus petit possible au sens des formes quadratiques (on vérifie en effet que, si A_1 et A_2 sont définies positives et $A_1 \geqslant A_2$, alors $A_1^{-1} \leqslant A_2^{-1}$).

Remarque 2

Dans la présentation ci-dessus, la loi normale des observations n'intervient que pour affirmer le caractère optimal de la technique de discrimination linéaire. Pour des observations non normales, ce caractère optimal n'est pas assuré, mais la méthode peut encore être justifiée si l'on s'en tient aux propriétés du second ordre, c'est-à-dire portant seulement sur les moyennes et les variances. Cette remarque reste vraie pour les raisonnements du paragraphe suivant.

3. Analyse Discriminante Conditionnelle

3.1. Modèle

Supposons maintenant que, les $\mu_j (j = 1, \ldots, g)$ et W ayant été préalablement estimés, s'ajoute à chaque observation nouvelle X l'observation complémentaire d'un vecteur aléatoire Y (à h dimensions) dans les conditions suivantes.

Pour un groupe j donné :

(a) la moyenne de Y est μ_Y, quel que soit j;

(b) la variance de $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$ s'écrit $W_Z = \begin{bmatrix} W_X & W_{XY} \\ W_{YX} & W_Y \end{bmatrix}$

où $W_X = W$, $W_{XY} = {}^t W_{YX}$ est la matrice des covariances de X et Y et W_Y est la matrice des covariances de Y supposée régulière (on notera que cette variance de Z est une variance intragroupe, supposée identique pour chaque groupe, comme au paragraphe 2);

(c) la loi de Z est une loi normale à $k+h$ dimensions (mais la remarque 2 s'applique encore).

Dans la pratique, Y est introduit pour chercher à prendre en compte des dérives éventuelles; on verra plus bas (paragraphe 4.2) quelques propositions pour le choix de

Y. On notera que la propriété (a) explicite les commentaires en fin du paragraphe 1. Le paragraphe 4.1 confirmera sa nécessité pour obtenir les estimateurs indispensables de W_Y et W_{XY}.

3.2. Nouvelle Analyse Discriminante

Si les valeurs des paramètres en jeu, ou des estimations convenables, sont disponibles, ce que nous supposons, la prise en compte de l'ensemble des données conduit à effectuer l'Analyse Discriminante non plus sur X mais sur $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$, un vecteur à $k + h$ dimensions.

On est donc amené à changer W en W_Z, les μ_j en $\begin{bmatrix} \mu_j \\ \mu_Y \end{bmatrix}$ et, le cas échéant, B en $B_Z = \dfrac{1}{g-1} \displaystyle\sum_{j=1}^{g} \begin{bmatrix} \mu_j - \bar{\mu} \\ 0 \end{bmatrix} [\, {}^t(\mu_j - \bar{\mu}) \quad 0] = \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix}$.

Posons $C = W_X - W_{XY} W_Y^{-1} W_{YX}$ et supposons C régulière. On a :

$$W_Z^{-1} = \begin{bmatrix} C^{-1} & -C^{-1} W_{XY} W_Y^{-1} \\ -W_Y^{-1} W_{YX} C^{-1} & (*) \end{bmatrix}$$

(le terme noté $(*)$ est inutile à expliciter).

Conformément au principe de l'Analyse Discriminante usuelle rappelé au paragraphe 2, on décide d'affecter l'observation $\begin{bmatrix} x \\ y \end{bmatrix}$ au groupe j qui minimise

$$[\, {}^t(x - \mu_j) \;\; {}^t(y - \mu_Y)] W_Z^{-1} \begin{bmatrix} x - \mu_j \\ y - \mu_Y \end{bmatrix}$$

$$= {}^t(x - \mu_j) C^{-1} (x - \mu_j) - 2\, {}^t(x - \mu_j) C^{-1} W_{XY} W_Y^{-1} (y - \mu_Y) + (**)$$

(le terme noté $(**)$ est inutile à expliciter car indépendant de j), soit encore d'affecter cette observation au groupe j qui minimise

$$\|(x - \mu_j) - W_{XY} W_Y^{-1} (y - \mu_Y)\|^2_{C^{-1}}.$$

Pour décider du groupe d'une nouvelle donnée x, à laquelle est associé y, il convient donc, pour tenir compte de la dérive, de remplacer $x - \mu_j$ dans (2) par $x - \mu_j - W_{XY} W_Y^{-1} (y - \mu_Y)$ et la métrique W^{-1} par $C^{-1} = (W_X - W_{XY} W_Y^{-1} W_{YX})^{-1}$ inverse de la variance de X *conditionnelle à* Y, selon un résultat classique. Notons que, en dehors de la correction sur x qui est le point majeur, l'introduction de C^{-1} améliore théoriquement l'analyse selon la remarque 1 (puisque $C \leqslant W$, soit $C^{-1} \geqslant W^{-1}$) mais qu'elle pose cependant des difficultés pratiques : voir la remarque 3 au paragraphe 4.

Pour des raisons évidentes, nous proposons d'appeler *Analyse Discriminante Conditionnelle* cette nouvelle analyse.

Considérons enfin le cas où l'on préfère travailler sur une projection obtenue par AFD. La nouvelle analyse factorielle est fondée sur la diagonalisation de

$$B_Z W_Z^{-1} = \begin{bmatrix} BC^{-1} & -BC^{-1}W_{XY}W_Y^{-1} \\ 0 & 0 \end{bmatrix}.$$

Or, si $v = \begin{bmatrix} v_X \\ v_Y \end{bmatrix}$ est vecteur propre de $B_Z W_Z^{-1}$ associé à la valeur propre λ, on a

$BC^{-1}(v_X - W_{XY}W_Y^{-1}v_Y) = \lambda v_X \qquad \text{et} \qquad \lambda v_Y = 0$

donc, pour $\lambda \neq 0$, v_Y est nul et v_X est vecteur propre de BC^{-1} associé à la valeur propre λ de cette dernière matrice. On sait que les vecteurs propres de BC^{-1} sont C^{-1}-orthogonaux et sont choisis en général C^{-1}-normés, ce que nous ferons. Nous avons donc ${}^t v_X C^{-1} v_X = 1$, ce qui est encore équivalent à $[\, {}^t v_X \; 0 \,]\, W_Z^{-1} \begin{bmatrix} v_X \\ 0 \end{bmatrix} = 1$.

Si l'on projette le vecteur $\begin{bmatrix} X \\ Y \end{bmatrix}$ sur $\begin{bmatrix} v_X \\ 0 \end{bmatrix}$ au sens de W_Z^{-1}, on obtient pour abscisse ${}^t v_X C^{-1}(X - W_{XY}W_Y^{-1}Y)$ ce qui est donc équivalent à projeter $(X - W_{XY}W_Y^{-1}Y)$ sur v_X au sens de C^{-1} . En particulier, la projection W_Z^{-1}-orthogonale de la moyenne empirique centrée $\begin{bmatrix} \mu_j - \bar{\mu} \\ 0 \end{bmatrix}$ sur $\begin{bmatrix} v_X \\ 0 \end{bmatrix}$ est d'abscisse ${}^t v_X C^{-1}(\mu_j - \bar{\mu})$, ce qui donne la même représentation que la projection obtenue par l'AFD de X seul où W_X aurait été remplacée par $C = W_X - W_{XY}W_Y^{-1}W_{YX}$.

4. Mise en œuvre

L'analyse conditionnelle nécessite d'une part le choix (la découverte...) d'un Y convenable, d'autre part l'estimation des divers paramètres en cause. Nous abordons d'abord ce dernier problème en supposant Y choisi, vérifiant les hypothèses (a) et (b) du paragraphe 3. Nous étudions ensuite comment introduire un Y convenable.

4.1. Estimation

L'estimation de W_X, des μ_j et de $\bar{\mu}$ se fait sur l'échantillon d'apprentissage. Par contre, l'estimation de μ_Y, W_Y et W_{XY} se fait sur l'ensemble des données disponibles : période d'apprentissage plus période opérationnelle. Cela est *nécessaire* dans la mesure où la variable Y peut être constante (ou presque) sur l'échantillon d'apprentissage, et même peut être impossible à définir à partir de ce seul échantillon. Cela est *possible* grâce à l'hypothèse (a) comme nous le précisons maintenant.

La moyenne μ_Y étant la même pour l'ensemble des observations, la moyenne empirique de toutes ces données estimera μ_Y le mieux possible. De même pour W_Y qui sera estimé par la matrice des covariances empiriques de Y sur l'ensemble des

observations où cette variable est disponible. En ce qui concerne W_{XY}, on a, pour une observation du groupe j, $W_{XY} = \mathbb{E}\left[(X - \mu_j)\ {}^t(Y - \mu_Y)\right] = \mathbb{E}\left[X\ {}^t(Y - \mu_Y)\right]$, quel que soit j. On peut donc estimer W_{XY} par $\frac{1}{N}\sum_{i=1}^{N} X_i\ {}^t(Y_i - \bar{Y})$ où N est le nombre total de données pour lesquelles Y est disponible et $\bar{Y}$ est la moyenne empirique des Y_i. Donc, ici aussi, grâce à l'hypothèse (a), W_{XY} est estimable au moyen de l'ensemble des données, même sans connaître le groupe de provenance de X.

Remarque 3

Les estimations de W_X, W_{XY} et W_Y étant faites sur des données différentes, on n'est pas assuré que la matrice estimant C soit positive. Si des valeurs propres négatives apparaissent, on peut être tenté de les remplacer par 0; ce n'est pas judicieux car il convient ensuite de calculer C^{-1} (et qu'une inverse généralisée n'est pas pertinente dans la mesure où les plus petites valeurs propres de C sont les plus « importantes » pour la discrimination). Le problème de l'estimation de C reste donc à approfondir pour l'utilisation de la métrique C^{-1}. Nous nous sommes contentés d'utiliser (une estimation de) la métrique W_X^{-1} dans l'exemple présenté au paragraphe 5.

4.2. Choix de la variable complémentaire

Abordons maintenant le problème très ouvert du choix de la variable complémentaire Y. Celui-ci peut être envisagé de deux façons.

(a) On peut d'abord effectuer un choix « externe » en considérant les mesures d'un ou plusieurs indicateurs indépendants du groupe auquel appartient l'unité statistique étudiée et susceptibles d'être bien corrélés avec les effets parasites à maîtriser donc bien représentatifs d'une éventuelle dérive.

(b) On peut aussi chercher à effectuer un choix « interne » en analysant convenablement l'évolution des données (pendant et après apprentissage) afin de découvrir d'éventuelles combinaisons des mesures X qui semblent les plus caractéristiques d'une éventuelle dérive tout en étant indépendantes du groupe d'appartenance. Pour cela, on peut envisager d'utiliser, par exemple, les méthodes proposées dans Caussinus & Ruiz-Gazen (1995) ou des méthodes voisines, d'inspiration similaire. Pour certains cas, celles-ci ne feront apparaître sans doute que la structure en groupe déjà connue mais, dans d'autres cas, elles sont susceptibles de mettre en relief des combinaisons des variables liées à d'autres phénomènes, par exemple une évolution fonction du temps. Il suffira alors de retenir ces dernières combinaisons comme vecteur Y.

Bien entendu, les deux types de construction de Y ne s'excluent pas, certaines coordonnées de Y pouvant être obtenues de la première façon, les autres étant obtenues de la seconde.

5. Exemple

Dans cet exemple[1] , $p = 18$ mesures sont réalisées sur chaque unité statistique et il y a $g = 3$ groupes (notés 1, 2 et 3 dans les figures). L'étude a porté sur 47 jours et, chaque jour, environ 5 unités de chaque groupe ont été considérées fournissant au total 695 unités.

Ici, le groupe véritable est connu pour les 695 unités, mais, pour tester la méthode, nous supposons que l'échantillon d'apprentissage est seulement constitué des données recueillies en début de période; les données suivantes étant utilisées pour valider les techniques proposées, on les appellera donc *échantillon de validation.* Parmi les nombreux essais effectués, nous avons choisi de présenter essentiellement ci-dessous les résultats pour un apprentissage sur les 5 premiers jours qui nous semblent assez bien représentatifs de l'ensemble des résultats obtenus. L'échantillon d'apprentissage comporte 89 unités statistiques (nous disposions en effet de 6 observations par jour les tout premiers jours et de moins par la suite).

5.1. Analyse Discriminante usuelle

Dans un premier temps, nous considérons une Analyse Discriminante usuelle avec les 89 premières unités statistiques comme échantillon d'apprentissage. La figure 1 montre l'évolution du pourcentage de mal classés au fur et à mesure de la validation, de la 90-ième à la dernière unité. Ce pourcentage augmente au cours du temps pour atteindre 49 % à la fin de la période étudiée. En décomposant la période de validation en 2, on obtient environ 34 % de mal classés dans la première moitié et 64 % dans la seconde, c'est-à-dire guère mieux que par une affectation « au hasard » (qui conduirait à environ deux tiers d'échecs).

La situation peut être illustrée par l'Analyse Factorielle Discriminante. On fournit deux graphiques correspondant au premier plan factoriel de l'AFD :

- la figure 2 représente les observations de l'échantillon d'apprentissage : elles sont parfaitement discriminées selon le groupe;
- la figure 3 représente l'ensemble des 695 observations et met clairement en évidence une grave détérioration des résultats à attendre de l'analyse.

Compte tenu du problème de dérive sous-jacent à ces données, l'utilisation de l'Analyse Discriminante usuelle est particulièrement contre-indiquée sur une longue période.

5.2. Analyse Discriminante Conditionnelle

Sur l'exemple considéré ici, on dispose aussi d'une série de mesures qui, de toute évidence, ne sont pas liées au groupe de l'unité statistique, mais peuvent être

1. Il s'agit de données réelles sur lesquelles, pour des raisons de confidentialité, nous ne donnerons pas davantage de précisions.

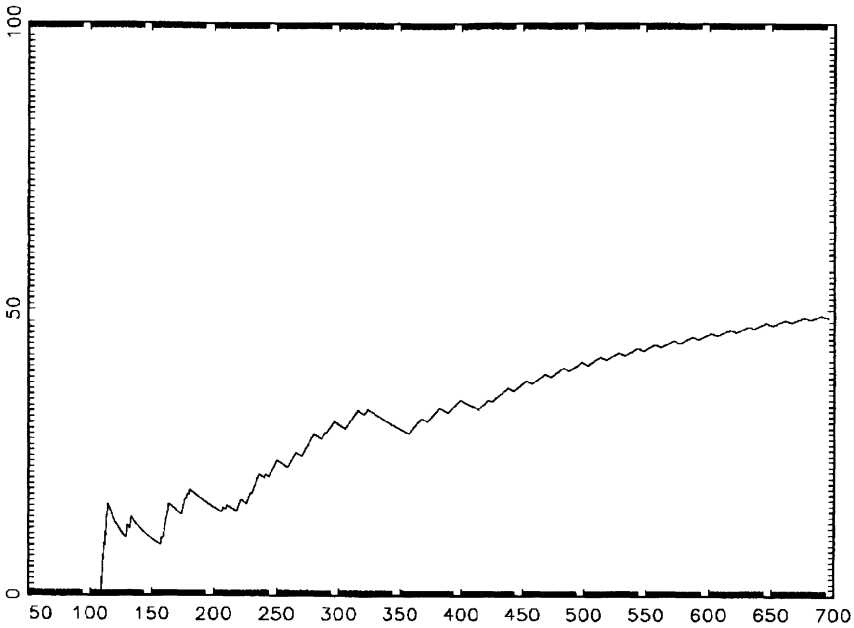

FIGURE 1
Pourcentage de mal classés, au cours de la période de validation, par l'Analyse Discriminante usuelle.

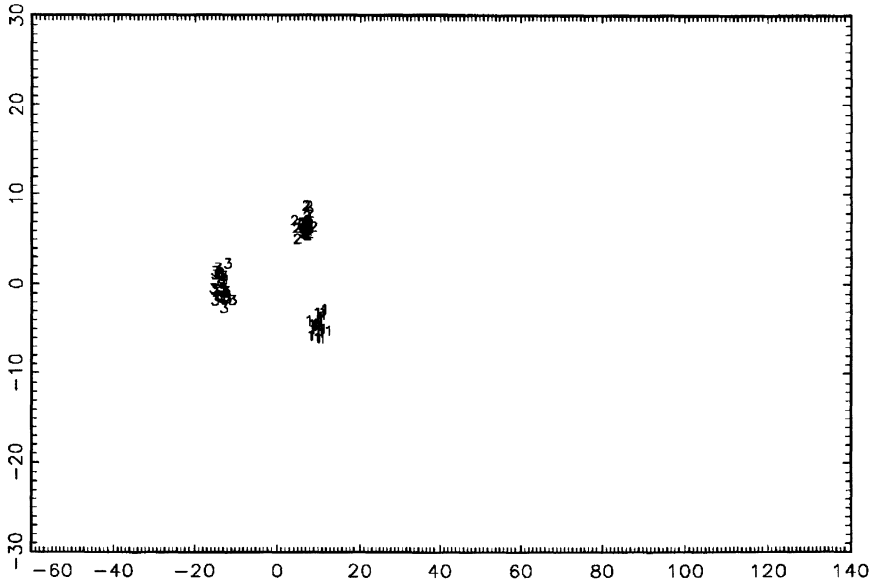

FIGURE 2
Échantillon d'apprentissage projeté sur le premier plan de l'Analyse Factorielle Discriminante usuelle.

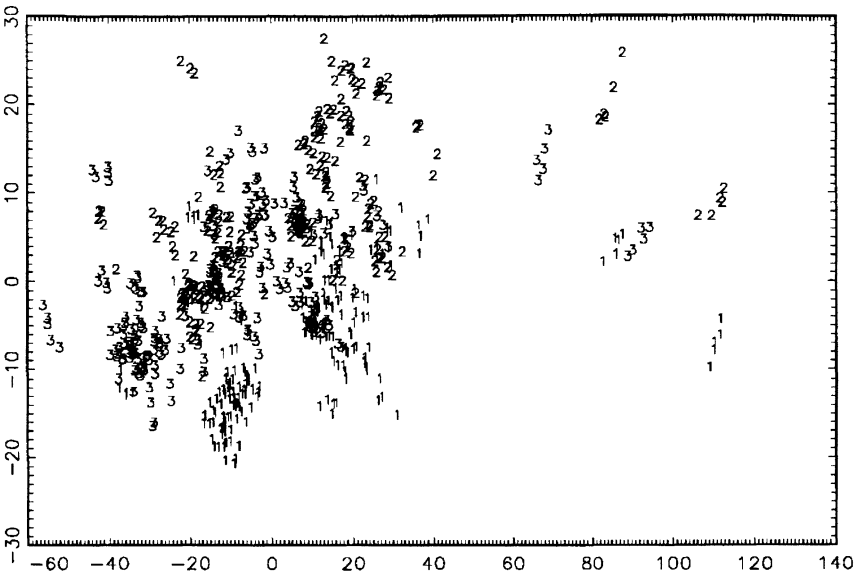

FIGURE 3
Échantillon global projeté sur le premier plan de l'Analyse Factorielle Discriminante usuelle.

liées à une dérive temporelle. On leur fait jouer le rôle du vecteur Y introduit en 3.1 et nous allons voir que cela conduit à une amélioration notable des résultats.

L'Analyse Discriminante Conditionnelle est conduite comme en situation réelle : à chaque nouvelle observation, les estimations sont remises à jour avant affectation, de sorte que la procédure n'est toujours basée que sur les données disponibles.

La figure 4 montre l'évolution du pourcentage de mal classés au cours de la validation. Comparons les figures 1 et 4. En tout début de période, les pourcentages sont calculés sur un très petit nombre d'unités et n'ont guère de signification. Jusqu'aux environs de la 250-ième unité, l'analyse usuelle est meilleure que l'analyse conditionnelle car le phénomène de dérive n'est pas encore trop sensible. Au-delà, lorsque le phénomène de dérive devient sensible, l'analyse conditionnelle s'avère meilleure que l'analyse usuelle et de plus en plus efficace, pour atteindre un pourcentage global de mal classés de l'ordre de 26 % seulement (contre 49 % pour l'analyse usuelle). En décomposant la période de validation en 2, comme pour l'analyse usuelle, on obtient environ 31 % de mal classés dans la première moitié (du même ordre qu'avec l'analyse usuelle) mais 21 % dans la seconde, ce qui correspond à l'efficacité de l'analyse usuelle au début de la période de validation.

Les figures 5 et 6 représentent le premier plan de l'Analyse Factorielle Discriminante Conditionnelle (voir fin du paragraphe 3.2) obtenu à l'issu de la validation, respectivement pour l'échantillon d'apprentissage et pour l'ensemble de l'échantillon. Ces figures montrent que, contrairement à l'analyse usuelle, l'analyse conditionnelle oublie partiellement l'apprentissage initial (comparer les figures 2 et

5) pour intégrer le phénomène de dérive (comparer les figures 3 et 6). Cela correspond à un *apprentissage poursuivi tout au long de l'utilisation de la méthode.*

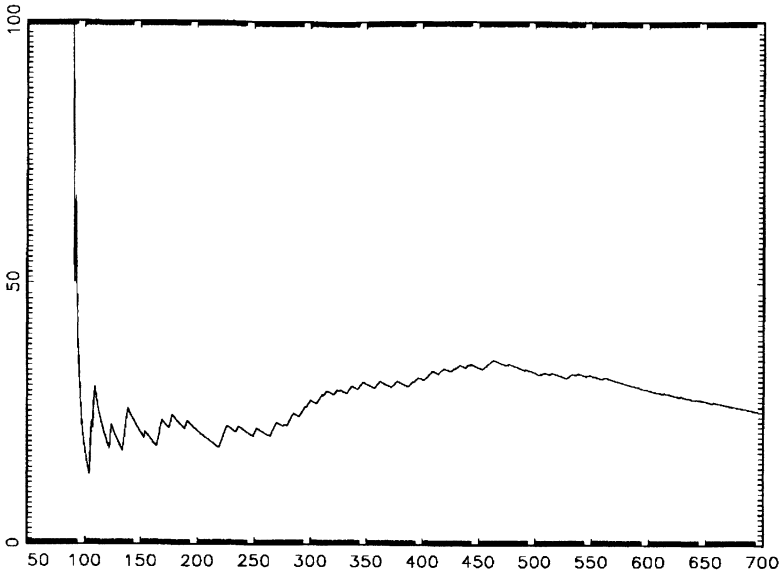

FIGURE 4
Pourcentage de mal classés, au cours de la période de validation, par l'Analyse Discriminante Conditionnelle.

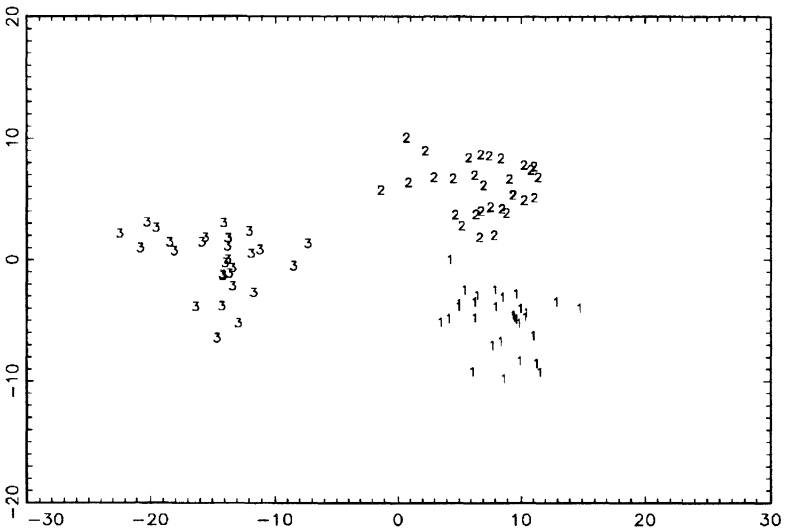

FIGURE 5
Échantillon d'apprentissage projeté sur le premier plan de l'Analyse Factorielle Discriminante Conditionnelle à la dernière étape de la validation.

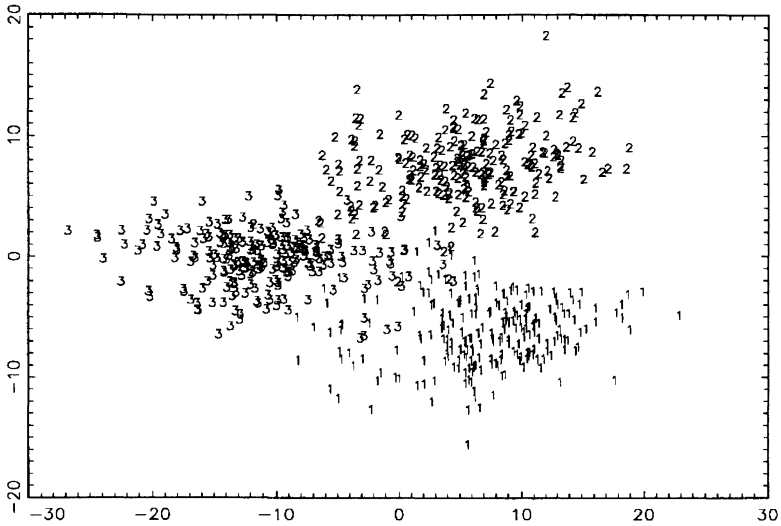

FIGURE 6
Échantillon global projeté sur le premier plan de l'Analyse Factorielle Discriminante Conditionnelle à la dernière étape de la validation.

Remarque 4

Comme on l'a dit, d'autres exemples ont été analysés en modifiant la durée d'apprentissage. Les résultats de l'Analyse Discriminante usuelle sont assez variables, parfois meilleurs que ceux présentés, mais pas meilleurs que ceux de l'Analyse Discriminante Conditionnelle. Par contre, les résultats de cette dernière restent stables.

6. Perspectives

On peut retenir que l'Analyse Discriminante Conditionnelle, telle qu'elle est proposée, est une technique relativement facile à mettre en œuvre et peu coûteuse en temps calcul. Les résultats sont encourageants sur l'exemple traité, mais il reste quelques points à approfondir, par exemple l'estimation de C^{-1}.

D'autre part, dans le cas où des mesures Y « naturelles » ne seraient pas disponibles, on peut envisager de calculer des variables Y en interne, c'est-à-dire à partir de X. Nous avons suggéré plus haut d'utiliser une Analyse en Composantes Principales avec une métrique particulière comme dans Caussinus & Ruiz-Gazen (1995). En fait, les résultats ainsi obtenus sur l'exemple ci-dessus sont encourageants, même s'ils s'avèrent moins performants que ceux obtenus à partir des variables externes. Là encore, un travail complémentaire serait utile.

Notons pour terminer que les principes développés ici sont aussi utilisables en matière de régression, c'est-à-dire lorsque la variable à prédire n'est plus qualitative mais quantitative.

Bibliographie

CAUSSINUS, H. & RUIZ-GAZEN, A. (1995). Metrics for finding typical structures by means of Principal Component Analysis. Data Science and Its Application, Harcourt Brace Japan, p.177–192.

FERRÉ, L. (1995). Improvement of some multidimensional estimates by reduction of dimensionality. Journal of Multivariate Analysis, 54, p.147–162.

GOUPY, J.-L. (1989). Erreur de dérive et choix de l'ordre des essais d'un plan d'expériences factoriel. Rev. Statistique Appliquée, 1989, XXXVII (1), 5–22.

MCLACHLAN, G.J. (1992). Discriminant analysis and statistical pattern recognition. Wiley, New York.

SAPORTA, G. (1990). Probabilités, analyse des données et statistique. Technip, Paris.