REVUE DE STATISTIQUE APPLIQUÉE

M. FROUARD
A. CARLIER
J. CAUQUIL

Une analyse discriminante sur données longitudinales

Revue de statistique appliquée, tome 47, n° 2 (1999), p. 41-59
<http://www.numdam.org/item?id=RSA_1999__47_2_41_0>

Rev. Statistique Appliquée, 1999, **XLVII** (2), 41-59

UNE ANALYSE DISCRIMINANTE SUR DONNÉES LONGITUDINALES

M. Frouard†, A. Carlier†, J. Cauquil‡

† *Laboratoire de Statistique et Probabilités, UMR C55830, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse Cedex*

‡ *Institut de Recherche Pierre Fabre, 52, rue Léon Blum, 81106 Castres Cedex*

RÉSUMÉ

Dans le cadre des données longitudinales de suivi clinique, nous considérons le problème de la discrimination entre deux populations. Ces deux populations sont définies par rapport à un événement d'intérêt survenant ou non au cours du suivi. Sous des hypothèses sur l'évolution de la distribution des covariables en fonction du temps, et en nous ramenant à un problème de discrimination sur données transversales, nous proposons un estimateur de la fonction linéaire discriminante, et nous montrons la normalité asymptotique du vecteur des coefficients associés. Ces derniers permettront de mettre en évidence les variables influant sur l'événement d'intérêt. Nous illustrons cette méthode sur des données réelles de suivi médical.

Mots-clés : *discrimination, données longitudinales, fonction linéaire discriminante de Fisher, sélection de variables, chaînes de Markov.*

ABSTRACT

In the framework of longitudinal data from clinical trials, we are interested in the discrimination between two populations. These two populations correspond to the occurence or not of an event of interest during the follow-up. Under some assumptions about the change in time of the covariates distribution, we reduce to the discrimination of cross-sectional data at each time, leading to the same discriminant coefficients. Then we give an estimator of these coefficients, and its asymptotic distribution is obtained under the preceding assumptions. These coefficients are used to point out the factors relating the occurence of the event. Finally, we present an application on real clinical data.

Keywords : *discrimination, longitudinal data, Fisher's linear discriminant function, variable selection, Markov chains.*

1. Introduction

Dans le cadre de la pharmaco-vigilance, nous nous intéressons à l'analyse des effets secondaires répertoriés au cours d'études de la tolérance d'un médicament. Ces effets peuvent être liés à des caractéristiques intrinsèques des patients (âge,

antécédents familiaux, ...), ou à une prise concomittante d'autres traitements. Nous cherchons à élaborer des groupes ou profils de patients se distinguant par leur sensibilité à un effet secondaire particulier, afin d'en surveiller ou d'en éviter l'apparition lors d'études futures. Après un essai clinique, les patients sont répartis en deux groupes : ceux qui ont présenté l'effet secondaire au moins une fois au cours de l'essai, et ceux qui ne l'ont pas présenté. Nous supposons que ces groupes peuvent être discriminés à l'aide des covariables relevées au cours du suivi du patient, et nous cherchons à identifier celles qui influent sur l'apparition de l'effet secondaire. Nous utilisons pour cela une méthode dérivée de l'analyse discriminante [2]. Celle-ci permet en effet de construire une combinaison linéaire des covariables, appelée fonction linéaire discriminante, séparant au mieux les deux groupes au sens d'une certaine métrique. Le vecteur λ des coefficients de la fonction discriminante nous permettra d'identifier les variables influentes, et par conséquent les causes possibles d'appartenance à chacun des deux groupes.

Les données de suivi clinique sont longitudinales [5]. La méthode d'analyse discriminante usuelle ne peut donc être utilisée directement dans ce problème. En effet, outre l'existence d'une corrélation entre les mesures d'un même individu, la distribution des covariables dépend du temps, et les hypothèses requises pour l'analyse discriminante ne sont plus vérifiées.

Afin de nous ramener à un problème d'analyse discriminante classique, nous ne conservons qu'une seule observation par patient en procédant de la manière suivante : pour un patient du premier groupe, cette observation correspond à la première apparition de l'effet secondaire; pour un patient du second groupe, elle est choisie aléatoirement de façon à ce que la distribution des temps dans les deux groupes soit identique [6].

Dans la section 2, nous rappelons la formulation générale du problème de discrimination entre deux populations, puis nous considérons une possibilité de mise en œuvre dans le cas de données longitudinales, après avoir introduit quelques notations, et les hypothèses relatives aux covariables.

Dans la section 3, nous proposons un estimateur du vecteur λ des coefficients de la fonction discriminante. Sous certaines hypothèses, et en utilisant les résultats de Das Gupta [4], nous donnons les expressions exactes de la moyenne et de la matrice de variance-covariance de ce vecteur à distance finie, ainsi que sa distribution asymptotique. Celle-ci permet d'établir des tests destinés à la sélection de variables [11].

Dans la section 4, nous présentons l'application de cette méthode à des données réelles de suivi clinique. Nous estimons les coefficients de la fonction discriminante, après transformation des données par des techniques classiques d'analyse factorielle [1].

2. Notations et position du problème

2.1 Principe de la discrimination entre deux groupes

De façon générale [2], l'analyse discriminante concerne l'étude d'observations multidimensionnelles réparties en plusieurs groupes définis *a priori*. On lui attribue les deux objectifs complémentaires suivants :

- l'étude descriptive de la liaison entre l'ensemble des caractéristiques recueillies et la variable d'appartenance aux groupes,
- la construction de règles de classement pour l'affectation de nouvelles observations aux différents groupes, en fonction de leur caractéristiques. Dans ce cas, on se place dans un contexte inférentiel, et on suppose que les deux groupes sont issus de deux populations.

La discrimination des groupes à partir des caractéristiques disponibles nécessite de définir une notion de distance entre groupes. Comme nous l'avons précisé en introduction, nous nous limitons dans cet article au cas de deux groupes. Soit X une variable aléatoire réelle représentant un caractère continu mesuré sur deux groupes G_1 et G_2. Si X a pour moyennes respectives μ_1 et μ_2 dans les groupes G_1 et G_2, et si son écart-type a pour valeur σ, une mesure de la distance entre G_1 et G_2 est alors donnée par :

$$\frac{|\mu_1 - \mu_2|}{\sigma}.$$

Pour le cas multivarié, notons $X \in \mathbb{R}^p$ le vecteur des p variables explicatives, et Σ la matrice de variance-covariance. On peut en chercher une combinaison linéaire ou fonction linéaire discriminante de Fisher U, qui maximise l'écart entre G_1 et G_2, au sens de la distance de Mahalanobis [11]. Posons $U = \lambda' X$, $\lambda \in \mathbb{R}^p$. Le problème revient à chercher la combinaison linéaire λ des covariables qui différencie au mieux les deux groupes, c'est-à-dire qui réalise le maximum de la quantité :

$$\frac{\lambda'(\mu_1 - \mu_2)}{(\lambda' \Sigma \lambda)^{1/2}}. \tag{1}$$

La résolution des équations normales fournit une solution à une constante multiplicative près. En fixant la constante à 1, cette solution pour le vecteur λ des coefficients discriminants est donnée par $\lambda = \Sigma^{-1}\delta$, où $\delta = \mu_1 - \mu_2$. Comme le fait remarquer McLachlan [11], la fonction discriminante de Fisher est obtenue sans faire d'hypothèses sur la distribution des covariables.

Si les populations dont sont issus G_1 et G_2 sont normalement distribuées et de même dispersion, la règle de décision permettant le classement de nouvelles observations est basée sur cette fonction linéaire discriminante. En effet, considérons un nouvel individu de caractéristiques $X = x$. Le classement de cet individu dans l'un des deux groupes se fait par la comparaison des probabilités *a posteriori* $\Pr(G_j | X = x)$ d'appartenance à chacun des deux groupes :

$$\Pr(G_j | X = x) = \frac{\pi_j f_j(x)}{f_X(x)}, \qquad (j = 1, 2),$$

où

- π_j est la probabilité *a priori* du groupe G_j,
- $f_j(x)$ est la densité du vecteur de caractéristiques X conditionnellement au groupe G_j,

– et $f_X(x)$ est la densité d'un mélange des groupes G_1 et G_2 dans les proportions π_1 et π_2 :

$$f_X(x) = \pi_1 f_1(x) + \pi_2 f_2(x) .$$

La comparaison des probabilités *a posteriori* s'effectue en calculant le logarithme de leur rapport [11] :

$$\log \frac{\Pr(G_1 | X = x)}{\Pr(G_2 | X = x)} = \log \frac{\pi_1}{\pi_2} + \xi(x) ,$$

avec $\xi(x) = \log (f_1(x)/f_2(x))$. Sous l'hypothèse d'un modèle normal homoscédastique pour le vecteur X, on a pour $\xi(x)$:

$$\begin{aligned} \xi(x) &= -\frac{1}{2} (\mu_1 + \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) + \left[\Sigma^{-1} (\mu_1 - \mu_2)\right]' x \\ &= -\lambda_0 + \lambda' x, \end{aligned}$$

où $\lambda_0 = \lambda' (\mu_1 + \mu_2) /2$ est le point équidistant des deux moyennes μ_1 et μ_2. Dans cette dernière expression, $\lambda' x$ correspond à la valeur de la fonction discriminante U pour l'observation x.

Nous disposons ici de données longitudinales provenant de suivis cliniques de patients. Nous renvoyons au livre de Diggle, Liang et Zeger [5] pour une présentation des modèles généralement adoptés dans le traitement de telles données. Notre problème étant d'identifier les variables ayant une influence sur l'apparition de l'effet secondaire, il relève donc de l'inférence sur la liaison entre les caractéristiques et la variable de groupe. Après avoir présenté quelques notations et hypothèses, nous proposons une utilisation de la technique d'analyse discriminante pour résoudre ce problème.

2.2 Notations et hypothèses

Considérons le cas d'un effet secondaire particulier. Au cours d'un essai clinique, les patients sont suivis au cours du temps. On note $t = 0, ..., T$ les instants d'observation, l'instant précédent le début du traitement correspond à $t = 0$. Chaque patient est caractérisé par la série d'observations $\{(X_t, Y_t)_{t=1,...,T}\}$, où :

– la réponse Y_t au temps t est définie par :

$$Y_t = \begin{cases} 1 & \text{si l'événement indésirable est présent au temps } t, \\ 0 & \text{sinon,} \end{cases}$$

– et $X_t \in \mathbb{R}^p$ est le vecteur des covariables à l'instant $t - 1$.

Le vecteur X_t peut contenir :

– des covariables indépendantes du temps (sexe, antécédents,...); pour un même individu, ces variables prennent les mêmes valeurs quel que soit t;

- des covariables dépendant du temps (traitements concomittants, pression artérielle, ...).

Nous faisons l'hypothèse suivante sur la structure de dépendance entre les séries $(X_t)_t$ et $(Y_t)_t$:

[H0] L'occurrence de l'effet secondaire à l'instant t ne dépend que des valeurs de la réponse Y_{t-1} à l'instant $t-1$ et des covariables X_t à l'instant t. En adoptant la méthode proposée par Lauritzen [9] pour modéliser graphiquement la dépendance entre processus, et avec la notation adoptée précédemment pour X_t, nous représentons sur le graphe orienté suivant la structure de dépendance entre les séries $(X_t)_t$ et $(Y_t)_t$:

$$\begin{array}{ccccccc} \ldots & Y_{t-1} & \rightarrow & Y_t & \rightarrow & Y_{t+1} & \ldots \\ & \uparrow & & \uparrow & & \uparrow & \\ \ldots & X_{t-1} & \rightarrow & X_t & \rightarrow & X_{t+1} & \ldots \end{array}$$

Ainsi $(Y_t)_t$ est une chaîne de Markov dont les probabilités de transition dépendent des covariables :

$$\Pr(Y_t = 1 | Y_1, ..., Y_{t-1}, X_1, ..., X_t) = \Pr(Y_t = 1 | Y_{t-1}, X_t).$$

Notons que cette hypothèse n'est pas très restrictive, car une covariable mesurée en t peut représenter un résumé du suivi jusqu'à t.

Dans le cas de données transversales, les groupes à discriminer sont associés aux différentes modalités de la réponse. Ici, la série $(Y_t)_t$ des réponses est à valeur dans $\{0,1\}^T$, et peut donc prendre 2^T valeurs différentes. Comme dans le cas transversal, nous aurions pu associer les groupes à discriminer à ces 2^T valeurs. Cependant, l'objectif étant de sélectionner les variables influant sur la première occurrence de l'effet secondaire, nous ne cherchons pas à modéliser la série complète des réponses $(Y_t)_{t=1,...,T}$. Nous considérons les groupes de patients G_1 et G_2 suivants :

- G_1 est l'ensemble des patients ayant présenté l'effet secondaire au moins une fois au cours de l'essai :

$$G_1 = \{\omega \in \Omega \mid \sum_{t=1}^{T} Y_t(\omega) \geqslant 1\},$$

- et G_2 l'ensemble des patients n'ayant jamais présenté l'effet secondaire au cours de l'essai :

$$G_2 = \{\omega \in \Omega \mid \sum_{t=1}^{T} Y_t(\omega) = 0\}.$$

Les groupes G_1 et G_2 sont associés à la survenue ou non d'un effet secondaire au cours du suivi. Si cet effet survient pour la première fois à un instant t, alors il n'est pas pertinent d'utiliser l'information postérieure à cette apparition pour discriminer

les groupes G_1 et G_2. Ainsi, en notant τ le temps d'arrêt qui correspond à cet instant d'apparition, on associe à un individu du groupe G_1 les valeurs des covariables $(X_1, ..., X_\tau)$ mesurées jusqu'à l'apparition de l'effet secondaire. Le temps d'arrêt τ variant d'un individu à l'autre, nous pouvons définir une partition $\{G_{1t}; t = 1, ..., T\}$ de G_1, où G_{1t} est l'ensemble des individus pour lesquels $\tau = t$:

$$G_{1t} = \{\omega \in \Omega \,|\, (Y_1(\omega), ..., Y_{t-1}(\omega), Y_t(\omega)) = (0, ..., 0, 1)\}.$$

Remarquons alors qu'à deux individus ω et ω' tels que $\tau(\omega) \neq \tau(\omega')$, on associe des vecteurs $(X_1, ..., X_\tau)$ de dimensions différentes. De plus, ce temps τ de première occurence ne peut pas être défini pour un individu du groupe G_2. Pour cela, nous procédons, pour chaque instant t, à la discrimination des groupes G_{1t} et G_2, à l'aide des valeurs des covariables $(X_1, ..., X_t)$ mesurées jusqu'à t.

Comme nous l'avons vu dans la section 2.1, la discrimination repose sur des hypothèses relatives à la distribution des covariables conditionnellement aux groupes. Généralement, la discrimination de deux groupes G_1 et G_2 à partir de séries d'observations $X = (X_t)_{t=1,...,T}$ nécessite la spécification des densités conditionnelles $f_j(x)$, $x \in \mathbb{R}^{p \times T}$, $(j = 1, 2)$. McLachlan [11] propose plusieurs références concernant la discrimination entre séries temporelles, traitant par exemple du problème de reconnaissance de la parole sur la base d'enregistrements. Cependant, l'hypothèse H0 formulée précédemment permet ici de ne spécifier que la densité de X_t conditionnellement aux groupes, et non la densité conjointe de $(X_1, ..., X_t)$.

Nous utilisons par la suite les notations $\mathbb{E}(X_t | G_{1t})$ et $\mathbb{E}(X_t | G_2)$ pour désigner respectivement l'espérance de X_t dans le groupe G_{1t} et dans le groupe G_2. Afin de nous ramener aux conditions de l'analyse discriminante (*cf.* 2.1), et d'obtenir des résultats concernant la distribution asymptotique du vecteur λ des coefficients discriminants, nous faisons les hypothèses supplémentaires suivantes :

[H1] Le vecteur aléatoire X_t à valeurs dans $\mathbb{R}^p$ suit une loi normale conditionnellement au groupe. On pose :

$$X_t = \mu_{1t}\, \mathbb{I}_{G_{1t}} + \mu_{2t}\, \mathbb{I}_{G_2} + e_t,$$

avec

$$\mathbb{E}(X_t | G_{1t}) = \mu_{1t},\ \mathbb{E}(X_t | G_2) = \mu_{2t},\ \text{var}(X_t) = \Sigma,\ \text{et}\ e_t \sim \mathcal{N}_p(0, \Sigma).$$

[H2] Les moyennes vérifient de plus le modèle suivant :

$$\mu_{jt} = a_t + b_j, \qquad j = 1, 2,$$

où a_t désigne un paramètre de nuisance, et correspond à un effet temps, et b_j à un effet lié au groupe. Les (X_t) peuvent être considérées comme des courbes de croissance [8], et l'on renvoie à McLachlan [11] pour des références concernant les méthodes de classement adaptées. L'hypothèse de parallélisme des profils moyens de chaque groupe constitue un cas particulier des hypothèses précédentes. Cette hypothèse peut être testée [14].

Notons que pour un groupe G_{1t} donné, l'hypothèse H1 concerne uniquement l'espérance des covariables X_t mesurées à l'instant $t-1$. Pour ce groupe, l'hypothèse H2 permet alors de considérer le cas d'une éventuelle interaction entre les facteurs groupe et temps.

2.3 Recherche de la fonction discriminante

Afin de mettre en évidence les covariables qui ont une influence sur l'apparition de l'effet secondaire, nous procédons pour tout t à la discrimination des groupes G_{1t} et G_2 définis précédemment. Soit λ_t le vecteur des coefficients qui définit la fonction discriminante ainsi obtenue. Nous obtenons alors le résultat suivant :

Proposition 1 : *Sous les hypothèses H0, H1 et H2, le vecteur λ_t des coefficients obtenus par discrimination des groupes G_{1t} et G_2 est indépendant du temps :*

$$\forall\, t = 1, ..., T, \qquad \lambda_t = \Sigma^{-1}(b_1 - b_2) = \lambda\,.$$

Preuve : Soit t fixé. Afin d'obtenir l'expression de la fonction discriminante de G_{1t} et G_2, calculons pour un individu de caractéristiques $(X_1, ..., X_t) = (x_1, ..., x_t)$ le logarithme du rapport des probabilités *a posteriori*, comme explicité dans la section 2.1. La probabilité *a posteriori* d'appartenir à G_{1t} est égale à :

$$\Pr\left(G_{1t} | \, x_1, ..., x_t\right) = \Pr\left(Y_1 = 0, ..., Y_{t-1} = 0, Y_t = 1 | \, x_1, ..., x_t\right).$$

Sous l'hypothèse H0, on peut décomposer cette probabilité de la façon suivante :

$$\Pr\left(G_{1t} | \, x_1, ..., x_t\right) = \Pr\left(Y_t = 1 | \, Y_{t-1} = 0, ..., Y_1 = 0, x_t\right) \times \prod_{j=2}^{t-1} \Pr\left(Y_j = 0 | \, Y_{j-1} = 0, x_j\right) \Pr\left(Y_1 = 0 | \, x_1\right).$$

D'après la formule de Bayes, le premier terme du membre de droite s'écrit encore :

$$\Pr(Y_t = 1 | \, Y_{t-1} = 0, ..., Y_1 = 0, x_t) = \frac{f_{G_{1t}}(x_t) \Pr(G_{1t})}{\Pr(Y_{t-1} = 0, ..., Y_1 = 0, x_t)}\,.$$

Or, compte-tenu de H0, la probabilité *a priori* du groupe G_{1t} est égale à :

$$\begin{aligned}\Pr(G_{1t}) &= \Pr(Y_t = 1, Y_{t-1} = 0, ..., Y_1 = 0) \\ &= \Pr(Y_t = 1 | \, Y_{t-1} = 0) \times \Pr(Y_{t-1} = 0, ..., Y_1 = 0)\,.\end{aligned}$$

Ainsi, le logarithme du rapport des probabilités *a posteriori* se réduit à :

$$\xi(x_t) = \log \frac{\Pr(G_{1t})}{\Pr(G_2)} + \log \frac{f_{G_{1t}}(x_t)}{f_{G_2}(x_t)} .$$

En utilisant l'hypothèse H1 pour le calcul du rapport des densités conditionnelles par groupe, on obtient la fonction linéaire discriminante :

$$U_t = \lambda_t' x_t \quad \text{avec} \quad \lambda_t = \Sigma^{-1} \delta_t = \Sigma^{-1}(\mu_{1t} - \mu_{2t}).$$

Enfin, à partir de l'expression des moyennes conditionnelles donnée par H2, on en déduit :

$$\delta_t = \mu_{1t} - \mu_{2t} = b_1 - b_2 .$$

D'où le résultat.

La valeur de λ_t ne dépend pas du temps, on la notera par la suite λ.

Remarque : La distance de Mahalanobis Δ^2 entre G_{1t} et G_2 reste constante au cours du temps :

$$\forall\, t = 1, ..., T, \qquad \Delta^2 = \delta_t' \Sigma^{-1} \delta_t = (b_1 - b_2)' \Sigma^{-1} (b_1 - b_2) ,$$

et le point équidistant des moyennes μ_{1t} et μ_{2t} est : $h_t = \lambda'(\mu_{1t} + \mu_{2t})/2$.

En conséquence de la proposition 1, nous proposons dans la section suivante une estimation unique du vecteur λ des coefficients discriminants, sous forme d'une combinaison des estimations obtenues en chaque temps par la discrimination de G_{1t} et G_2.

3. Un estimateur de la fonction discriminante

On dispose des données d'un échantillon de n individus. Soient $n_1 = \text{Card}(G_1)$, le nombre d'individus présentant l'effet secondaire au moins une fois au cours de l'essai, et $n_{1t} = \text{Card}(G_{1t})$ le nombre d'individus l'ayant présenté pour la première fois à l'instant t, avec $\sum_{t=1}^{T} n_{1t} = n_1$. Soit également $n_2 = \text{Card}(G_2) = n - n_1$ le nombre d'individus n'ayant pas présenté l'effet secondaire au cours de l'essai.

La discrimination à t fixé des groupes G_{1t} et G_2 fournit une estimation du vecteur λ des coefficients discriminants, calculée à partir des n_{1t} individus de G_{1t} et des n_2 individus de G_2. Pour estimer λ à un temps $t' \neq t$, on utilisera les $n_{1t'}$ individus supplémentaires du groupe $G_{1t'}$, et à nouveau les n_2 individus de G_2. Comme nous n'avons pas spécifié la structure de covariance des $(X_t)_t$ pour le groupe G_2, les T estimations de λ obtenues par discrimination $(G_{1t})_{t=1,...,T}$ et G_2 peuvent être combinées à la seule condition d'être indépendantes. Chacune des T estimations

de λ doit donc porter sur des individus différents, ce qui nécessite d'utiliser en chaque temps uniquement un sous-groupe G_{2t} du groupe G_2. Ainsi, nous répartissons ces n_2 individus sur les T instants d'observation par un tirage aléatoire sans remise, avec équiprobabilité des individus, de façon à ce que la distribution des individus en fonction du temps soit identique pour les groupes G_1 et G_2 :

$$\forall\, t = 1, ..., T, \qquad \frac{n_{1t}}{n_1} \simeq \frac{n_{2t}}{n_2}\,, \tag{2}$$

où $n_{2t} = \text{Card}(G_{2t})$ et $\sum_t n_{2t} = n_2$. L'ensemble $\{G_{2t}; t = 1, ..., T\}$ forme une partition de G_2. Dans le cas d'un effet secondaire rare, l'effectif de G_2 est très élevé. Par conséquent, le fait de se restreindre dans la pratique à G_{2t} au lieu de G_2 ne diminue pas de façon considérable la précision de l'estimateur de la fontion discriminante.

Nous verrons dans la proposition 2 que le choix de cette répartition permet d'obtenir un estimateur de variance minimum de la différence δ. Nous en déduisons l'estimateur du vecteur λ des coefficients de la fonction discriminante, puis nous en donnons quelques propriétés.

3.1 Estimation de δ

A t fixé, $\delta = b_1 - b_2$ est estimé par $d_t = \overline{x}_{1t} - \overline{x}_{2t}$, où $\overline{x}_{jt}$ est la moyenne des observations du groupe G_{jt} :

$$\overline{x}_{jt} = \frac{1}{n_{jt}} \sum_{i \in G_{jt}} x_t(i),$$

$x_t(i) \in \mathbb{R}^p$ désignant le vecteur des covariables pour l'individu i au temps t. On obtient ainsi T estimations de δ. On propose alors l'estimateur suivant de δ :

$$d = \sum_{t=1}^{T} \frac{n_{1t}}{n_1}\, d_t. \tag{3}$$

Proposition 2 : *Sous les hypothèses H1 et H2, d est un estimateur sans biais de δ de variance minimum dans la classe des estimateurs sans biais de δ combinaisons linéaires des d_t, distribué selon une loi normale $\mathcal{N}_p(\delta, \Sigma/c_2)$, avec $c_2 = (1/n_1 + 1/n_2)^{-1}$.*

Preuve : Le premier résultat est obtenu en développant d de la façon suivante :

$$d = (b_1 - b_2) + \sum_{t=1}^{T} \left(\frac{1}{n_1} \sum_{i \in G_{1t}} e_t(i) - \frac{n_{1t}}{n_1 n_{2t}} \sum_{i \in G_{2t}} e_t(i) \right).$$

Calculons la variance de d :

$$\begin{aligned}\text{var}(d) &= \text{var}\left(\sum_{t=1}^{T} \frac{n_{1t}}{n_1}\left(\overline{x}_{1t} - \overline{x}_{2t}\right)\right) \\ &= \frac{\Sigma}{n_1} + \text{var}\left(\sum_{t=1}^{T} \frac{n_{1t}}{n_1}\,\overline{x}_{2t}\right).\end{aligned}$$

En supposant que (2) est strictement vérifié, on a alors :

$$\begin{aligned}\text{var}\left(\sum_{t=1}^{T} \frac{n_{1t}}{n_1}\,\overline{x}_{2t}\right) &= \text{var}\left(\sum_{t=1}^{T} \frac{n_{2t}}{n_2}\,\overline{x}_{2t}\right) \\ &= \text{var}\left(\frac{1}{n_2}\sum_{t=1}^{T}\sum_{i\in G_{2t}} x_t(i)\right) \\ &= \frac{\Sigma}{n_2}\,.\end{aligned}$$

D'où :

$$\text{var}(d) = \Sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{\Sigma}{c_2}\,.$$

Remarque : Si on calcule $\overline{x}_{2t}$ en utilisant les n_2 observations au temps t, l'estimateur d reste sans biais, mais il est obtenu par une somme pondérée de variables dépendantes. Ainsi var(d) dépend de la structure de covariance des erreurs $(e_t)_t$ dans le groupe G_2 et son estimation nécessite des hypothèses supplémentaires.

Montrons maintenant que la répartition (2) conduit à un estimateur de variance minimum. Soit une pondération quelconque $(p_t)_{t=1,\ldots,T}$ affectée aux différents temps ($p_t > 0$ et $\sum_{t=1}^{T} p_t = 1$), et un choix quelconque de la répartition des individus du groupe G_2 sur les temps (on note encore n_{2t} l'effectif du groupe G_{2t}). Considérons l'estimateur de d suivant :

$$d = \sum_{t=1}^{T} p_t\, d_t\,.$$

On obtient alors,

$$\begin{aligned}\operatorname{var}(d) &= \sum_{t=1}^{T}(p_t)^2 \operatorname{var}(\overline{x}_{1t}) + \sum_{t=1}^{T}(p_t)^2 \operatorname{var}(\overline{x}_{2t}) \\ &= \left(\frac{1}{n_1}\sum_{t=1}^{T}\frac{n_1}{n_{1t}}(p_t)^2 + \frac{1}{n_2}\sum_{t=1}^{T}\frac{n_2}{n_{2t}}(p_t)^2\right)\Sigma\,.\end{aligned}$$

Dans l'expression ci-dessus, chacune des deux sommes est minorée par 1, et ce minimum est atteint pour $p_t = n_{1t}/n_1 = n_{2t}/n_2$, d'où le résultat.

Enfin, la normalité de d se déduit du fait que d est la somme pondérée des T v.a. normales indépendantes $\overline{x}_{1t} - \overline{x}_{2t}$.

Remarque : En utilisant l'hypothèse de normalité des variables explicatives, on montre que la structure statistique étudiée est de type exponentiel, de statistique exhaustive et totale égale à :

$$S = \left[\sum_{t=1}^{T}\left(\sum_{i\in G_{1t}} x_t(i)\,x_t'(i) + \sum_{i\in G_{2t}} x_t(i)\,x_t'(i)\right), \sum_{t=1}^{T}\sum_{i\in G_{1t}} x_t(i), \sum_{t=1}^{T}\sum_{i\in G_{2t}} x_t(i)\right].$$

L'estimateur d défini par l'équation (3) s'écrivant comme une fonction de cette statistique S, on en déduit d'après le théorème de Lehmann-Scheffe que d est un estimateur optimal de δ dans la classe de tous les estimateurs sans biais.

3.2 Estimation de Σ

La matrice de dispersion intra-classe est estimée de façon usuelle à l'aide de la somme des carrés intra-échantillons. Posons :

$$W = \sum_{j=1}^{2}\sum_{t=1}^{T}\sum_{i\in G_{jt}} (X_t(i) - \overline{x}_{jt})\,(X_t(i) - \overline{x}_{jt})' \in \mathcal{M}_{p\times p}\,,$$

avec $c_1 = n_1 + n_2 - 2T$. Alors W/c_1 est une estimation sans biais de Σ, dont la loi est une distribution de Wishart $W \sim W(\Sigma, c_1, p)$.

Remarque : d et W sont indépendantes.

3.3 Estimation de λ, U *et* Δ^2

En remplaçant dans l'expression de λ chaque quantité par son estimateur, l'estimateur l du vecteur λ des coefficients de la fonction discriminante s'écrit :

$$l = c_1 W^{-1} d\,.$$

Chaque coefficient étant associé à une covariable, nous pouvons ainsi identifier, à partir de l, les covariables intervenant dans la discrimination des groupes G_{1t} et G_{2t}, et qui influencent donc l'apparition de l'effet secondaire. Le principe de la sélection de variables à partir d'une analyse canonique discriminante consiste à éliminer les variables qui ne semblent pas contribuer aux variables canoniques, ou à la fonction discriminante dans le cas de deux groupes. La méthode la plus usuelle [10] passe par l'examen des coefficients discriminants standardisés : un faible coefficient en valeur absolue signifie que la variable associée ne contribue pas à la discrimination. On peut trouver dans [10] une comparaison des méthodes utilisées pour la sélection de variables en analyse discriminante.

On déduit également l'estimateur $\hat{U}_t$ de la fonction discriminante :

$$\forall t = 1, ..., T, \quad \hat{U}_t = l' X_t = c_1 d' W^{-1} X_t \,.$$

Dans la pratique, la fonction discriminante est utilisée pour classer les individus. Considérons un nouvel individu pour lequel on dispose de la valeur x_t des covariables mesurées en $t-1$, et choisissons une valeur seuil c. Si $\hat{U}_t = l' x_t > c$, on affectera par exemple cet individu au groupe G_{1t}, et à G_{2t} sinon. Ainsi, on peut savoir si cet individu risque de présenter l'effet secondaire à l'instant t.

De même, on estime pour tout t la distance de Mahalanobis Δ^2 entre G_{1t} et G_{2t} par :

$$D_p^2 = c_1 d' W^{-1} d \,.$$

3.4 Propriétés de l'estimateur des coefficients discriminants

On utilise dans cette partie les résultats de Das Gupta [4] sur les moments d'une variable aléatoire suivant une loi de Wishart, pour calculer la moyenne et la dispersion exactes de l'estimateur l des coefficients de la fonction discriminante, et pour établir sa distribution asymptotique.

Proposition 3 : *Sous les hypothèses H0, H1 et H2, l'espérance et la variance de l'estimateur l sont les suivantes :*

$$\mathbb{E}(l) = \frac{c_1}{c_1 - p - 1} \Sigma^{-1} (b_1 - b_2) \,, \quad \textit{si } c_1 - p - 1 > 0 \,,$$

$$\mathrm{var}(l) = K \left[\Sigma^{-1} \left(\delta' \Sigma^{-1} \delta \right) + (c_1 - 1) \frac{\Sigma^{-1}}{c_2} + \Sigma^{-1} \delta \delta' \Sigma^{-1} \left(\frac{c_1 - p + 1}{c_1 - p - 1} \right) \right] ,$$

$$\textit{si } c_1 - p - 3 > 0, \textit{ et avec}$$

$$K = \frac{(c_1)^2}{(c_1 - p)(c_1 - p - 1)(c_1 - p - 3)} \,, \qquad c_2 = \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} .$$

Preuve : Rappelons que $l = c_1 W^{-1} d$. En utilisant le fait que les variables d et W sont indépendantes et suivent respectivement une loi normale $\mathcal{N}_p(\delta, \Sigma/c_2)$ et $W(\Sigma, c_1, p)$, on se ramène aux conditions du corollaire 2.1.1 [4] qui donne l'espérance et la variance de leur produit.

Proposition 4 *Sous les hypothèses H0, H1 et H2, la distribution asymptotique de l est donnée par :*

$$\sqrt{c_2}\left(l - \Sigma^{-1}\delta\right) \rightsquigarrow \mathcal{N}_p(0, \Gamma) \quad \textit{quand } c_2 \to \infty,$$

avec

$$\Gamma = \Sigma^{-1}\delta\delta'\Sigma^{-1} + \left(1 + \delta'\Sigma^{-1}\delta\right)\Sigma^{-1}.$$

Preuve : Ce résultat est obtenu en appliquant le corollaire 3.1.1 [4], qui donne la distribution asymptotique du vecteur λ dans le cas de deux groupes, et qui reste valable quand d est défini suivant l'équation (3).

La propriété de normalité asymptotique donnée dans la proposition 4 peut être utilisée pour effectuer des tests sur les coefficients discriminants à partir de grands échantillons. En particulier, on peut montrer dans le cas de deux groupes que tester la nullité de q coefficients revient à tester l'information additionnelle apportée par les q variables correspondantes dans la discrimination des groupes [13].

4. Application à des données réelles

Dans cette section, nous présentons une application de la méthode à des données réelles de suivi clinique, relatives à l'étude de la tolérance d'un anti-dépresseur. Ces données ont été obtenues auprès de l'Institut de Recherche Pierre Fabre de Toulouse.

4.1 Description des données

L'événement indésirable étudié Y est la tachycardie. La réponse d'un individu au temps t du suivi est donnée par :

$$Y_t = \begin{cases} 1 & \text{si le patient présente une tachycardie au temps } t, \\ 0 & \text{sinon.} \end{cases}$$

Afin de mettre en évidence le type des patients susceptibles de présenter une tachycardie, nous cherchons à identifier, parmi les covariables relevées au cours du suivi, celles qui ont une influence sur l'apparition de cet événement indésirable. Les groupes de patients G_1 et G_2 à discriminer sont les suivants :

– G_1 est l'ensemble des patients ayant présenté une tachycardie durant l'essai,

– G_2 est l'ensemble des patients n'ayant pas présenté de tachycardie.

Dans cette application, certaines variables explicatives sont qualitatives, à deux ou plus de deux modalités. L'analyse discriminante ne peut pas être directement réalisée sur les données initiales. On procède alors suivant le principe de la méthode DISQUAL de discrimination sur variables qualitatives [3]. Cette méthode consiste à réaliser, préalablement à la discrimination, l'Analyse Factorielle des Correspondances Multiples (AFCM) sur l'ensemble des variables explicatives qualitatives. L'analyse discriminante est ensuite effectuée, en utilisant pour variables explicatives les facteurs principaux de l'AFCM, et les variables explicatives continues. Certains coefficients de la fonction discriminante seront alors associés aux facteurs de l'AFCM, plutôt qu'aux variables de départ. Chaque facteur de l'AFCM étant une combinaison linéaire des indicatrices des variables qualitatives, on peut en examiner la constitution, afin d'identifier les variables discriminant les groupes.

Nous choisissons pour la discrimination 34 covariables parmi celles qui sont disponibles. Certaines variables sont continues :

- 1 covariable indépendante du temps est mesurée à l'inclusion du patient dans l'essai : le poids (X^1);
- 6 covariables sont mesurées à plusieurs reprises au cours du suivi : la fréquence cardiaque (X^2); des variables qui permettent d'évaluer l'amélioration ou l'aggravation de la dépression sur des échelles spécifiques (X^3 pour l'échelle d'Hamilton, X^4 pour l'échelle MADRS), la diminution de l'anxiété (X^5), la diminution de l'anxiété psychique (X^6), et de l'anxiété somatique (X^7).

D'autres sont qualitatives :

- 15 covariables indépendantes du temps sont mesurées à l'inclusion du patient dans l'essai : le groupe de traitement (placebo, traitement par anti-dépreseur à une dose de 50 mg, 100 mg, ou 200 mg); des covariables binaires indiquant si le patient présentait avant l'essai un certain nombre d'événements indésirables (12 événements indésirables ont été répertoriés parmi lesquels : maux de tête, nausée, insomnie, palpitations...); et deux variables binaires indiquant si le patient était sujet à tachycardie ou bradycardie avant l'essai;
- 12 covariables sont mesurées à plusieurs reprises au cours du suivi : ces variables indiquent l'apparition d'un autre événement indésirable pendant le traitement, ceci afin de mettre en évidence des groupes d'événements apparaissant de façon plus ou moins simultanée.

Ces dernières variables interviendront dans la discrimination par l'intermédiaire des facteurs de l'AFCM.

514 patients ont été suivis pendant 8 semaines, et les données correspondent aux mesures effectuées à $t = 0, 1, 2, 3, 4, 6$ et 8 semaines. Le nombre de patients présents à chaque visite est donné dans le tableau 1.

Ce nombre décroît au cours du temps en raison des sorties d'essai. Celles-ci sont généralement dues à une guérison, ou au contraire à une forte aggravation de la maladie.

TABLEAU 1
Nombre de patients présents chaque semaine

N° semaine	0	1	2	3	4	6	8
Nb patients présents	514	500	431	412	386	339	346

4.2 Résultats

Les individus qui n'ont pas présenté de tachycardie au cours de leur suivi sont considérés comme appartenant au groupe G_2. Par exemple, un patient suivi uniquement pendant les deux premières semaines de l'essai, et n'ayant pas présenté de tachycardie, sera considéré comme appartenant à G_2 jusqu'à la deuxième semaine. Le nombre n_{1t} de premières apparitions de tachycardie à la semaine t, ainsi que le nombre n_2 des patients de G_2 suivis jusqu'au temps t, sont donnés dans le tableau 2 suivant :

TABLEAU 2
Nombre de patients disponibles chaque semaine

N° semaine	1	2	3	4	6	8
Card$(G_{1t}) = n_{1t}$	39	22	9	5	5	2
Card$(G_2) = n_2$	402	337	322	301	259	249

Sur les 500 patients présents la première semaine, 402 ne présenteront pas de tachycardie au cours de leur suivi, 39 présentent déjà une tachycardie à la première semaine; les 59 patients restant présenteront une tachycardie aux semaines suivantes ou sortiront de l'essai.

Les effectifs n_{2t} de patients sélectionnés à la semaine t parmi ceux du groupe G_2, de façon à ce que les distributions des visites sur les temps soient identiques (*cf.* (2)), sont donnés dans le tableau 3 ci-après.

TABLEAU 3
Effectifs des groupes G_{1t} et G_{2t}

N° semaine	1	2	3	4	6	8
Card$(G_{1t}) = n_{1t}$	39	22	9	5	5	2
Card$(G_{2t}) = n_{2t}$	163	92	37	20	20	8

Ces données seront utilisées plus loin pour effectuer les estimations des coefficients discriminants.

Après avoir effectué une AFCM sur l'ensemble des variables qualitatives, nous conservons pour cette application les cinq premiers facteurs $F_1, ..., F_5$ de l'AFCM, qui expliquent 56% de l'inertie totale. Nous disposons donc pour la discrimination de G_1 et G_2 de cinq facteurs et des sept variables $X^1, ..., X^7$. Le tableau 4 présente les coefficients discriminants standardisés, obtenus à l'aide de l'estimateur l (*cf.* 3.3).

TABLEAU 4
Coefficients discriminants standardisés

Variable	λ^j
X^1	0.06
X^2	3.89
X^3	0.25
X^4	- 0.48
X^5	- 0.04
X^6	0.31
X^7	- 0.16
F_1	- 0.28
F_2	- 0.53
F_3	0.35
F_4	- 0.22
F_5	- 0.09

Nous remarquons que le coefficient associé à la variable mesurant la fréquence cardiaque (X^2) est élevé, et traduit ainsi une forte influence de cette variable sur l'apparition d'une tachycardie. Ceci s'explique par le fait qu'une tachycardie correspond à une observation de la fréquence cardiaque supérieure à une valeur seuil de 100 b/mn. Le deuxième facteur de l'AFCM (F_2), ainsi que la variable évaluant l'amélioration de la dépression sur l'échelle MADRS (X^4), semblent également avoir de l'influence dans la discrimination des groupes. Ce dernier coefficient indique que les patients dont l'état de santé s'améliore sont ensuite moins susceptibles de présenter une tachycardie.

Enfin, à partir des coefficients discriminants, nous obtenons pour chaque patient des groupes G_{1t} et G_{2t} la valeur de la fonction discriminante, à l'aide de l'estimateur $\hat{U}_t$ (*cf.* 3.3). La figure 1 représente les histogrammes de cette fonction discriminante en chaque temps et pour chaque groupe. Nous observons sur ces graphiques que les groupes G_{1t} $(Y = 1)$, et G_{2t} $(Y = 0)$, sont bien discriminés : les histogrammes respectifs sont relativement disjoints.

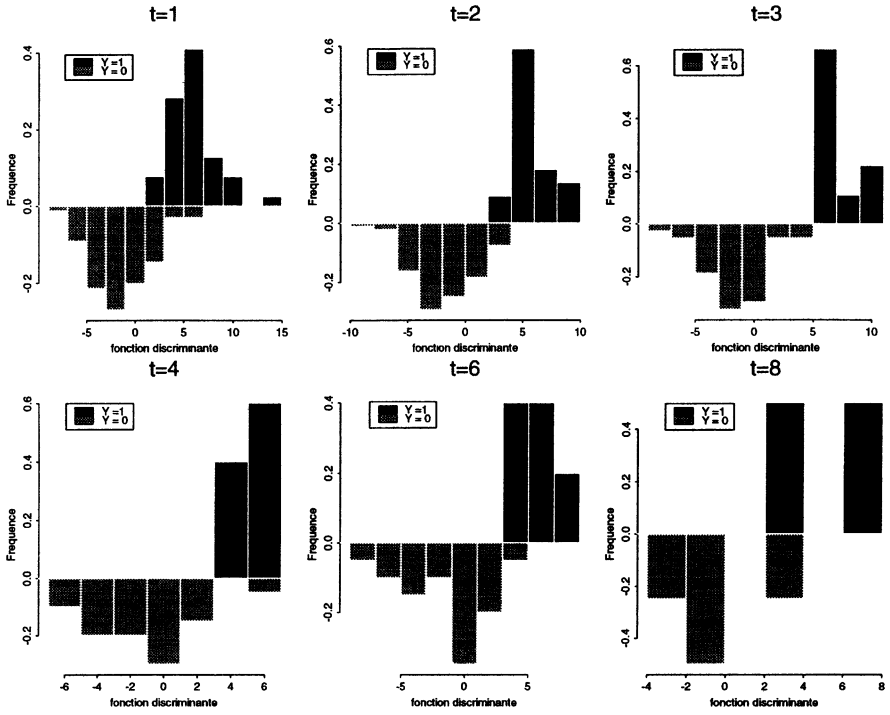

FIGURE 1
Histogrammes de la fonction discriminante

4.3 Discussion

L'application de la méthode proposée dans cet article, à des données réelles de suivi clinique, montre une façon de prendre en compte l'effet temporel des covariables dans une analyse discriminante. Son intérêt réside dans le fait qu'elle repose essentiellement sur l'hypothèse H2. Cette hypothèse signifie d'une part que l'on peut dissocier l'effet temps et l'effet groupe, à l'instant précédent l'apparition de l'événement; et d'autre part, que la moyenne des covariables conditionnellement aux groupes est fonction du temps. La figure 2 montre l'évolution de la distribution de la variable X^5 pendant les trois premières semaines, dans chacun des groupes.

Les coefficients de la fonction discriminante permettent d'identifier les variables qui ont une influence sur l'apparition de l'événement d'intérêt, indépendamment de l'instant auquel elle se produit.

Remarquons également que cette méthode est généralisable à plus de deux groupes, ce cas se présentant par exemple quand on cherche à expliquer la gravité d'un événement.

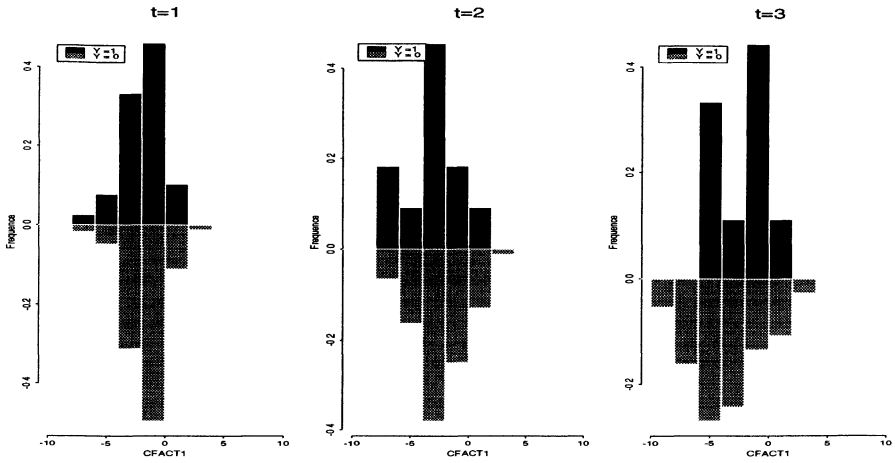

FIGURE 2
Distribution de la baisse de l'anxiété (X^5) en fonction du temps

Bien que nous ayons représenté les histogrammes de la fonction discriminante, nous nous sommes contentés ici de décrire la liaison entre un ensemble de caractéristiques et la variable d'appartenance aux groupes. Il pourrait être intéressant de procéder à l'estimation du taux d'erreur de classement associé à cette fonction discriminante, afin de la valider pour de nouveaux patients.

Références

[1] BECKER R.A., CHAMBERS J.M., WILKS A.R. (1988). *The new S language*. Belmont, California : Wadsworth and Brooks.

[2] CELEUX G. (1990). *Analyse discriminante sur variables continues*. INRIA, collection didactique.

[3] CELEUX G., NAKACHE J.P. (1994). *Analyse discriminante sur variables qualitatives*. Polytechnica.

[4] DAS GUPTA S. (1968). Some aspects of discriminant function coefficients. *Sankhya A*, **30**, 387-400.

[5] DIGGLE, LIANG, ZEGER (1994). *Analysis of longitudinal data*. Oxford University Press.

[6] FROUARD M. (1997) *Discrimination sur base de données, application à la pharmacovigilance. Un outil méthodologique*. Rapport technique de l'Institut de Recherche Pierre Fabre.

[7] HAND D.J. (1981). *Discrimination and classification*. New York : Wiley.

[8] KSHIRSAGAR A.M., SMITH W.B. (1995). *Growth curves*. Statistics : textbooks and monographs, vol. **145**.

[9] LAURITZEN S.L. (1996). *Graphical models*. New York : Oxford University Press.

[10] MCKAY R.J., CAMPBELL N.A. (1982). Variable selection techniques in discriminant analysis I. Description. *BR. J. Math. Statist. Psychol.*, **35**, 1-29.

[11] MCLACHLAN G.J. (1992). *Discriminant analysis and statistical pattern recognition*. New York : Wiley.

[12] Panel on discriminant analysis, classification and clustering (1989). *Statistical Science*, **4**, 34-69.

[13] RAO C.R. (1973). *Linear statistical inference and its application*. Second edition. New York : Wiley.

[14] TOMASSONE R., DANZART M., DAUDIN J.J., MASSON J.P. (1988). *Discrimination et classement*. Masson.