Analyse discriminante sur données binaires lorsque les populations d'apprentissage et de test sont différentes

Julien Jacques*, Christophe Biernacki**

*Laboratoire de Statistiques et Analyse des Données,
Université Pierre Mendès France,
38040 Grenoble Cedex 9, France.
julien.jacques@iut2.upmf-grenoble.fr,
http://www.julien.jacques2.free.fr
**Laboratoire Paul Painlevé UMR CNRS 8524,
Université Lille I,
59655 Villeneuve d'Ascq Cedex, France.
christophe.biernacki@math.univ-lille1.fr

Résumé. L'analyse discriminante généralisée suppose que l'échantillon d'apprentissage et l'échantillon test, qui contient les individus à classer, sont issus d'une même population. Lorsque ces échantillons proviennent de populations pour lesquelles les paramètres des variables descriptives sont différents, l'analyse discriminante généralisée consiste à adapter la règle de classification issue de la population d'apprentissage à la population test, en estimant un lien entre ces deux populations. Ce papier étend les travaux existant dans un cadre gaussien au cas des variables binaires. Afin de relever le principal défi de ce travail, qui consiste à déterminer un lien entre deux populations binaires, nous supposons que les variables binaires sont issues de la discrétisation de variables gaussiennes latentes. Une méthode d'estimation et des tests sur simulations sont présentés, puis des applications dans des contextes biologique et d'assurance illustrent ce travail.

1 Introduction

L'analyse discriminante classique suppose que l'échantillon d'apprentissage et l'échantillon test, qui contient les individus à classer, sont issus d'une même population. Depuis les travaux de Fisher (1936), qui introduit une règle de discrimination linéaire entre deux groupes, de nombreuses évolutions ont été proposées (cf. McLachlan (1992) pour une revue). Toutes ces évolutions concernent la nature de la règle de discrimination : paramètrique, semi-paramètrique ou encore non paramétrique.

Une évolution alternative, introduite par Van Franeker et Ter Brack (1993) puis développée par Biernacki et al. (2002), considère le cas où l'échantillon d'apprentissage et l'échantillon test ne sont pas nécessairement issus d'une même population. Biernacki et al. définissent plusieurs modèles d'*analyse discriminante généralisée* dans un contexte gaussien, et les expérimentent

sur une application biologique dans laquelle les deux populations sont des oiseaux de mer d'une même espèce, mais d'origines géographiques différentes.
Mais dans beaucoup de domaines, comme les assurances ou la médecine, un grand nombre d'applications traite de données binaires. L'objectif de ce papier est d'étendre l'analyse discriminante généralisée, établie dans un contexte gaussien, au cas des données binaires. La différence entre les populations d'apprentissage et de test peut être géographique (comme dans l'application biologique précédemment citée), mais aussi temporelle ou tout autre.
La prochaine section présente les données et le modèle des classes latentes pour les deux populations d'apprentissage et de test. La section 3 fait l'hypothèse que les données binaires sont une discrétisation de variables continues latentes. Cette hypothèse permet d'établir un lien entre les deux populations, qui conduit à proposer huit modèles d'analyse discriminante généralisée pour données binaires. La section 4 traite de l'estimation des paramètres de ces modèles. Dans la section 5, des tests sur simulations, une application dans un contexte biologique puis une application à des données d'assurances illustrent l'efficacité de l'analyse discriminante généralisée vis-à-vis de l'analyse discriminante classique et de la classification automatique. Finalement, la dernière section discute des possibles extensions de ces travaux.

2 Les données et le modèle des classes latentes

Les données consistent en deux échantillons : un étiqueté, S, issu d'une population P, et un non étiqueté, S^*, issu d'une population P^*. Les deux populations P et P^* peuvent être différentes.
L'échantillon d'apprentissage S est composé de n couples $(\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n)$, où $\boldsymbol{x}_i \in \{0,1\}^d$ est le vecteur des variables explicatives pour l'individu i, et $\boldsymbol{z}_i = (z_i^1, \ldots, z_i^K)$ représente l'appartenance de cet individu à l'un des K groupes, avec $z_i^k = 1$ si l'individu i appartient au groupe k et $z_i^k = 0$ sinon.
Ces couples $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ sont supposés être des réalisations indépendantes du couple aléatoire $(\boldsymbol{X}, \boldsymbol{Z})$ de distribution :

$$X^j_{|Z^k=1} \sim \mathcal{B}(\alpha_{kj}) \quad \forall j = 1, \ldots, d \quad \text{et} \quad \boldsymbol{Z} \sim \mathcal{M}(1, p_1, \ldots, p_K). \tag{1}$$

En utilisant l'hypothèse d'indépendance conditionnelle des variables explicatives X^j ($j = 1, \ldots, d$) (Everitt (1984) ; Celeux et Govaert (1991)), la distribution de probabilité de $\boldsymbol{X}$ s'écrit :

$$f(x^1, \ldots, x^d) = \sum_{k=1}^{K} p_k \prod_{j=1}^{d} {\alpha_{kj}}^{x^j} (1 - \alpha_{kj})^{1-x^j}. \tag{2}$$

L'échantillon test S^* est quant à lui composé de n^* individus pour lesquels seules les variables explicatives $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_{n^*}^*$ sont connues (les variables sont les mêmes que pour l'échantillon d'apprentissage). Les appartenances aux classes $\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_{n^*}^*$ sont inconnues. Nous considérons les couples $(\boldsymbol{x}_i^*, \boldsymbol{z}_i^*)$ ($i = 1, \ldots, n^*$) comme des réalisations indépendantes du couple aléatoire $(\boldsymbol{X}^*, \boldsymbol{Z}^*)$ de distribution analogue à $(\boldsymbol{X}, \boldsymbol{Z})$ mais de paramètres différents, notés α_{kj}^* et p_k^*.
L'objectif de la discrimination généralisée est alors d'estimer les n^* étiquettes $\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_n^*$ inconnues en utilisant l'information contenue à la fois dans S et dans S^*. Le défi consiste ainsi à trouver un lien entre les deux populations P et P^*.

Remarque. L'utilisation de la terminologie "test" pour la population P^* et son échantillon S^* est abusive puisque l'échantillon S^* est utilisé pour déterminer la règle de classement. Nous la conservons néanmoins afin de faciliter les comparaisons avec les méthodes classiques de discrimination.

3 Relation entre les populations d'apprentissage et de test

3.1 Hypothèse gaussienne sous-jacente

Dans un contexte multi-normal, une relation stochastique linéaire entre P et P^* est non seulement justifiée (avec très peu d'hypothèses) mais aussi intuitive (Biernacki et al. (2002)). Dans le cas de données binaires, comme il ne semble pas exister de relation intuitive, une hypothèse supplémentaire est faite : nous supposons que les données binaires sont dues à la discrétisation de variables gaussiennes latentes. Cette hypothèse n'est pas nouvelle en statistique : nous pouvons citer comme exemple les travaux de Thurstone (1927), qui utilise cette hypothèse dans son modèle de jugement comparatif pour choisir entre deux stimulus, ou encore les travaux d'Everitt (1987), qui propose un algorithme de classification pour des données binaires, catégoriques et continues.
Nous supposons donc que les variables explicatives $X^j_{|Z^k=1}$, de distribution de Bernoulli $\mathcal{B}(\alpha_{kj})$, sont issues de la discrétisation de variables continues latentes $Y^j_{|Z^k=1}$ conditionnellement indépendantes et de distribution normale $\mathcal{N}(\mu_{kj}, \sigma^2_{kj})$:

$$X^j_{|Z^k=1} = \begin{cases} 0 & \text{si } \lambda_j Y^j_{|Z^k=1} < \lambda_j s_j \\ 1 & \text{si } \lambda_j Y^j_{|Z^k=1} \geq \lambda_j s_j \end{cases} \qquad \text{pour } j = 1, \ldots, d, \tag{3}$$

où $s_j \in \mathbb{R}$ est le seuil de discrétisation, et où $\lambda_j \in \{-1, 1\}$ est introduit pour ne pas avoir à choisir quelle valeur de X^j, 0 ou 1, correspond à une valeur de Y^j supérieure au seuil s_j. Le rôle de ce nouveau paramètre est d'éviter aux variables binaires d'hériter de l'ordre naturel induit par les variables continues.
Ainsi, nous pouvons déduire la relation suivante entre α_{kj}, et λ_j, μ_{kj} et σ_{kj} :

$$\alpha_{kj} = p(X^j_{|Z^k=1} = 1) = \begin{cases} 1 - \Phi\left(\frac{s_j - \mu_{kj}}{\sigma_{kj}}\right) & \text{si } \lambda_j = 1 \\ \Phi\left(\frac{s_j - \mu_{kj}}{\sigma_{kj}}\right) & \text{si } \lambda_j = -1 \end{cases} \tag{4}$$

où Φ est la fonction de répartition de la loi normale centrée réduite. Il est important de noter que l'hypothèse d'indépendance conditionnelle rend le calcul de α_{kj} très simple. Sans cette hypothèse, il devient très complexe d'évaluer les intégrales multi-dimensionnelles induites par le calcul de α_{kj}, surtout lorsque la dimension d du problème est grande, ce qui est souvent le cas avec des données binaires.
Comme pour la variable $\boldsymbol{X}$, nous supposons aussi que les variables binaires $\boldsymbol{X}^*$ sont dues à la discrétisation de variables continues latentes $\boldsymbol{Y}^*$. Les équations sont les mêmes que (3) et (4), en changeant α_{kj} en α^*_{kj}, μ_{kj} en μ^*_{kj}, σ_{kj} en σ^*_{kj} et s_j en s^*_j. Le paramètre λ^*_j est naturellement supposé égal à λ_j.
Dans un contexte gaussien, Biernacki et al. (2002) définissent une relation entre les variables

continues de la population P et celles de la population P^*, en émettant deux hypothèses relativement naturelles : la transformation entre P et P^* est $\mathcal{C}^1$ et la j-ème composante $Y^{*j}_{|Z^k=1}$ de $\boldsymbol{Y}^*_{|Z^k=1}$ ne dépend que de la j-ème composante $Y^j_{|Z^k=1}$ de $\boldsymbol{Y}_{|Z^k=1}$. Sous ces deux hypothèses, ils montrent que la seule transformation possible est la transformation stochastique linéaire suivante :

$$\boldsymbol{Y}^*_{|Z^{*k}=1} \sim A_k \boldsymbol{Y}_{|Z^k=1} + \boldsymbol{b}_k, \tag{5}$$

où A_k est une matrice diagonale de $\mathbb{R}^{d\times d}$ contenant les éléments a_{kj} ($1 \leq j \leq d$) et $\boldsymbol{b}_k$ est un vecteur de $\mathbb{R}^d$.
En appliquant cette relation à nos variables continues latentes, nous pouvons en déduire la relation suivante entre les paramètres α^*_{kj} et α_{kj} :

$$\alpha^*_{kj} = \Phi\Big(\delta_{kj}\ \Phi^{-1}(\alpha_{kj}) + \lambda_j \gamma_{kj}\Big), \tag{6}$$

où $\delta_{kj} = \text{sgn}(a_{kj})$, sgn$(t)$ désignant le signe de t, et où $\gamma_{kj} = \text{sgn}(a_{kj})\, s_j/\sigma_{kj} + (b_{kj} - s^*_j)/(|a_{kj}|\ \sigma_{kj})$.
Ainsi, conditionnellement au fait que les paramètres α_{kj} sont connus (ils seront estimés en pratique), l'estimation des Kd paramètres continus α^*_{kj} est obtenue par l'estimation des paramètres relatifs au lien entre P et P^* : δ_{kj}, γ_{kj} et λ_j. Le nombre de paramètres continus (γ_{kj}) à estimer est ainsi Kd, ce qui est équivalent à estimer directement α^*_{kj} sans utiliser la population P. Par conséquent, il est nécessaire de réduire ce nombre de paramètres continus à estimer.
Pour cela, nous introduisons un certain nombre de sous-modèles en imposant des contraintes sur la transformation entre les deux populations P et P^*.

3.2 Modèles de contraintes sur la relation

Modèle M_1 **:** σ_{kj} et A_k sont libres et $\boldsymbol{b}_k = 0$ ($k = 1, \ldots, K$). Le modèle est :

$$\alpha^*_{kj} \quad = \quad \Phi\Big(\delta_{kj}\ \Phi^{-1}(\alpha_{kj})\Big) \quad \text{ avec } \quad \delta_{kj} \in \{-1, 1\}.$$

Cette transformation correspond soit à l'identité, soit à une permutation des modalités de $\boldsymbol{X}$.

Modèle M_2 **:** $\sigma_{kj} = \sigma$, $A_k = aI_d$ avec $a > 0$, I_d la matrice identité de $\mathbb{R}^{d\times d}$ et $\boldsymbol{b}_k = \beta \boldsymbol{e}$, avec $\beta \in \mathbb{R}$ et $\boldsymbol{e}$ le vecteur de dimension d composé seulement de 1 (la transformation est indépendante du groupe et de la dimension). Le modèle est :

$$\alpha^*_{kj} = \Phi\Big(\Phi^{-1}(\alpha_{kj}) + \lambda'_j|\gamma|\Big) \quad \text{ avec } \quad \lambda'_j = \lambda_j\ \ \text{sgn}(\gamma) \in \{-1, 1\} \text{ et } |\gamma| \in \mathbb{R}^+.$$

L'hypothèse $a > 0$ est faite pour avoir l'identifiabilité du modèle, et n'induit aucune restriction. La même hypothèse est faite pour les deux modèles suivants.

Modèle M_3 **:** $\sigma_{kj} = \sigma_k$, $A_k = a_k I_d$, avec $a_k > 0$ pour tout $1 \leq k \leq K$, et $\boldsymbol{b}_k = \beta_k \boldsymbol{e}$, avec $\beta_k \in \mathbb{R}$ (la transformation ne dépend que du groupe). Le modèle est :

$$\alpha^*_{kj} = \Phi\Big(\Phi^{-1}(\alpha_{kj}) + \lambda'_{kj}|\gamma_k|\Big) \text{ avec } \lambda'_{kj} = \lambda_j\ \ \text{sgn}(\gamma_k) \in \{-1, 1\} \text{ et } |\gamma_k| \in \mathbb{R}^+.$$

Remarque. L'hypothèse $\sigma_{kj} = \sigma_k$ qui suppose que, conditionnellement au groupe, toutes les variables latentes ont la même variance, n'est qu'un artefact technique permettant de définir le modèle M_3. Même s'il peut sembler difficile que cette hypothèse soit vérifiée en pratique, cela n'a qu'une importance moindre car le modèle M_3 est mis en concurrence avec d'autres modèles.

Modèle M_4 **:** $\sigma_{kj} = \sigma_j$, $A_k = A$, avec $a_{kj} > 0$ pour tout $1 \leq k \leq K$ et $1 \leq j \leq d$, et $\boldsymbol{b}_k = \boldsymbol{\beta}$ avec $\boldsymbol{\beta} \in \mathbb{R}^d$ (la transformation ne dépend que de la dimension). Le modèle est :

$$\alpha^*_{kj} = \Phi\Big(\Phi^{-1}(\alpha_{kj}) + \gamma_j^{'}\Big) \quad \text{avec} \quad \gamma_j^{'} = \lambda_j \gamma_j \in \mathbb{R}.$$

Notons que dans ce modèle M_4, le paramètre γ_j évoluant librement dans $\mathbb{R}$, il n'est plus possible d'identifier séparément les deux paramètres λ_j et γ_j dans le produit $\lambda_j \gamma_j$. Ce dernier est alors considéré comme un nouveau paramètre réel $\gamma_j^{'}$.

Pour chacun des quatre modèles M_i ($i = 1, \ldots, 4$), nous prenons en compte une hypothèse supplémentaire sur les proportions des groupes (p_i, $i = 1, \ldots, K$) : elles sont conservées ou non de P vers P^*. Nous notons M_i le modèle avec proportions inchangées, et pM_i le modèle avec des proportions potentiellement changées. Huit modèles sont ainsi définis.
Notons que le modèle M_2 est toujours inclus dans les modèles M_3 et M_4, et M_1 peut parfois être inclus dans les trois autres modèles.
Finalement, pour choisir automatiquement parmi ces huit modèles de discrimination généralisée, le critère BIC (*Bayesian Information Criterion*, Schwarz (1978)) peut être employé. Il est défini par :

$$\text{BIC} = -2l(\hat{\boldsymbol{\theta}}) + \nu \log(n^*),$$

où $\boldsymbol{\theta} = \{p^*_k, \delta_{kj}, \lambda_j, \gamma_{kj}\}$ pour $1 \leq k \leq K$ et $1 \leq j \leq d$, $l(\hat{\boldsymbol{\theta}})$ est le maximum de la log-vraisemblance correspondant à l'estimation $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$, et ν est le nombre de paramètres continus libres associés au modèle donné (cf. tableau 1). Le modèle qui conduit à la plus petite valeur de BIC est retenu.

	M_1	M_2	M_3	M_4	pM_1	pM_2	pM_3	pM_4
continus	0	1	K	d	$K-1$	K	$2K-1$	$d+K-1$
discrets	Kd	d	Kd	0	Kd	d	Kd	0

TAB. 1 – *Nombre de paramètres continus et discrets à estimer pour les huit modèles de discrimination généralisée.*

Il est maintenant nécessaire d'estimer le paramètre $\boldsymbol{\theta}$. La méthode du maximum de vraisemblance est retenue.

4 Estimation des paramètres

4.1 Les trois étapes de l'estimation

L'analyse discriminante généralisée nécessite trois étapes d'estimation. Nous présentons la situation où les proportions sont inconnues au sein de P^*, le cas contraire étant immédiat.
La première étape consiste à estimer les paramètres p_k et α_{kj} ($1 \leq k \leq K$ et $1 \leq j \leq d$) de la population P à partir de l'échantillon d'apprentissage S. Comme S est étiqueté, les estimateurs du maximum de vraisemblance sont connus et usuels (Everitt (1984) ; Celeux et Govaert (1991)).
La deuxième étape consiste à estimer les paramètres p_k^* et α_{kj}^* ($1 \leq k \leq K$ et $1 \leq j \leq d$) du mélange de Bernoulli à partir de $\boldsymbol{\theta}$ et S^*. Pour estimer les α_{kj}^*, nous devons estimer les paramètres du lien entre P et P^* : une fois δ_{kj}, γ_{kj} et λ_j estimés, un estimateur de α_{kj}^* est déduit par l'équation (6). Cette étape est détaillée ci-dessous.
Enfin, la troisième étape consiste à estimer les appartenances aux groupes des individus de l'échantillon test S^*, et ce par maximum *a posteriori*.

4.2 Estimation des paramètres de lien

Pour la deuxième étape, l'estimation par maximum de vraisemblance peut être réalisée à l'aide de l'algorithme EM (Dempster et al. (1977)) qui est bien adapté au cas des données manquantes (qui sont ici les appartenances aux classes).
La vraisemblance s'écrit :

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n^*} \sum_{k=1}^{K} p_k^* \prod_{j=1}^{d} {\alpha_{kj}^*}^{x_i^{*j}} (1-\alpha_{kj}^*)^{1-x_i^{*j}}.$$

La log-vraisemblance complétée est donnée par :

$$l_c(\boldsymbol{\theta}; z_1^*, \ldots, z_{n^*}^*) = \sum_{i=1}^{n^*} \sum_{k=1}^{K} z_i^{*k} \log\Big(p_k^* \prod_{j=1}^{d} {\alpha_{kj}^*}^{x_i^{*j}} (1-\alpha_{kj}^*)^{(1-x_i^{*j})}\Big).$$

L'étape E. En utilisant la valeur courante $\boldsymbol{\theta}^{(q)}$ du paramètre $\boldsymbol{\theta}$, l'étape E de l'algorithme EM consiste à calculer l'espérance conditionnelle de la log-vraisemblance complétée :

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}) &= E_{\boldsymbol{\theta}^{(q)}}[l_c(\boldsymbol{\theta}; Z_1^*, \ldots, Z_{n^*}^*) | x_1^*, \ldots, x_{n^*}^*] \\ &= \sum_{i=1}^{n^*} \sum_{k=1}^{K} t_{ik}^{(q)} \Big\{ \log(p_k^*) + \log\Big(\prod_{j=1}^{d} {\alpha_{kj}^*}^{x_i^{*j}} (1-\alpha_{kj}^*)^{1-x_i^{*j}} \Big) \Big\} \end{aligned}$$

où

$$t_{ik}^{(q)} = p(Z_i^{*k} = 1 | x_1^*, \ldots, x_{n^*}^*; \boldsymbol{\theta}^{(q)}) = \frac{p_k^{*(q)} \prod_{j=1}^{d} ({\alpha_{kj}^*}^{(q)})^{x_i^{*j}} (1-{\alpha_{kj}^*}^{(q)})^{(1-x_i^{*j})}}{\sum_{\kappa=1}^{K} p_\kappa^{*(q)} \prod_{j=1}^{d} ({\alpha_{\kappa j}^*}^{(q)})^{x_i^{*j}} (1-{\alpha_{\kappa j}^*}^{(q)})^{(1-x_i^{*j})}}$$

est la probabilité conditionnelle que l'individu i appartienne au groupe k.

L'étape M. L'étape M de l'algorithme EM consiste à choisir la valeur de $\boldsymbol{\theta}^{(q+1)}$ qui maximise l'espérance conditionnelle $\mathcal{Q}$ calculée à l'étape E :

$$\boldsymbol{\theta}^{(q+1)} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}). \tag{7}$$

Nous décrivons cette étape pour chaque composante de $\boldsymbol{\theta} = \{p_k^*, \delta_{kj}, \lambda_j, \gamma_{kj}\}$.
Pour les proportions, cette maximisation donne l'estimateur suivant

$$p_k^{*(q+1)} = \frac{1}{n^*} \sum_{i=1}^{n^*} t_{ik}^{(q)}.$$

Pour les paramètres continus γ_{kj}, on montre pour chaque modèle que $\mathcal{Q}$ est une fonction de γ_{kj} strictement concave et qui tend vers $-\infty$ lorsqu'une norme du vecteur de paramètre $(\gamma_{11}, \ldots, \gamma_{Kd})$ tend vers ∞ (cf. Jacques (2005)). Nous pouvons donc utiliser un algorithme de type Newton pour trouver l'unique maximum de $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)})$ suivant $\boldsymbol{\theta}$.
Pour les paramètres discrets δ_{kj} et λ_j, si la dimension d et le nombre de groupes K sont relativement petits (par exemple $K = 2$ et $d = 5$), la maximisation est faite en calculant $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)})$ pour toutes les valeurs possibles des paramètres discrets. Si K ou d sont trop grands, le nombre de valeurs possibles de δ_{kj} est trop important (par exemple 2^{20} pour $K = 2$ et $d = 10$), et il est donc impossible de parcourir toutes ces valeurs dans un temps raisonnable. Dans ce cas, nous utilisons un méthode de relaxation, qui consiste à supposer que le paramètre δ_{kj} (respectivement λ_j) n'est pas un paramètre binaire dans $\{-1, 1\}$ mais continu dans $[-1, 1]$, noté $\tilde{\delta}_{kj}$ (Wolsey (1998)). L'optimisation est alors faite sur le paramètre continu (avec un algorithme de type Newton comme pour γ_{kj} puisque $\mathcal{Q}$ est une fonction de $\tilde{\delta}_{kj}$ strictement concave), et la solution $\tilde{\delta}_{kj}^{(q+1)}$ est discrétisée de la façon suivante pour obtenir une solution binaire ${\delta_{kj}}^{(q+1)}$: ${\delta_{kj}}^{(q+1)} = \text{sgn}(\tilde{\delta}_{kj}^{(q+1)})$.

Remarque. En pratique, le bouclage sur les valeurs possibles des paramètres discrets ne se fait pas au sein de l'étape M, mais en dehors de l'algorithme EM : nous estimons les paramètres p_k^* et γ_{kj} à l'aide de l'algorithme EM pour chaque valeur possible des paramètres discrets δ_{kj} et λ_j, puis nous choisissons la solution de vraisemblance maximale.

5 Comparaisons de méthodes sur données simulées et réelles

5.1 Données simulées

Un grand nombre de tests sur simulations a été réalisé pour valider les huit modèles de discrimination généralisée, et nous en présentons un résumé dans cette section. L'objectif de ces tests est de comparer, suivant le taux de mauvais classement (taux d'erreur), l'analyse discriminante classique (*DA*), la classification automatique (aussi appelée *Clustering*, et qui revient à ne travailler que sur la population test en oubliant la population d'apprentissage), à l'analyse discriminante généralisée, pour laquelle nous avons sélectionné deux modèles, le

meilleur modèle suivant le critère BIC (*GDA BIC*) et le meilleur modèle suivant le taux d'erreur (*GDA error*).

Remarques.
- Les taux d'erreur présentés sont les taux d'erreur apparents, puisqu'ils sont évalués à l'aide de l'échantillon S^* qui a également servi à construire la règle de classement. Une estimation non biaisée des taux d'erreur aurait pû être obtenue en utilisant un autre échantillon de P^* indépendant de S^*.
- Pour la classification automatique, l'affectation des individus aux groupes est faite de sorte à minimiser le taux d'erreur obtenu, en utilisant l'étiquettage des échantillons S^* qui est en réalité connu pour les applications présentées ici.

5.1.1 Bon modèle

Dans une première série de tests, les données binaires sont simulées à partir de la discrétisation de données gaussiennes. De plus, les variables gaussiennes utilisées sont indépendantes conditionnellement à l'appartenance à un groupe, et la transformation entre P et P^* est choisie de sorte à être $\mathcal{C}^1$ et de matrice d'homothétie A_k diagonale. Toutes les hypothèses de l'analyse discriminante généralisée sont ainsi respectées.
Les tests sont réalisés pour deux tailles d'échantillon ($n = n^* = 100$ et $n = n^* = 1000$), deux nombres de variables explicatives ($d = 5$ et $d = 10$), et pour huit transformations correspondant aux huit modèles de discrimination généralisée.
Les paramètres des lois gaussiennes et des huit transformations sont donnés en annexe A.
Pour l'algorithme EM, la convergence est fixée à un gain en log-vraisemblance entre deux étapes de l'algorithme inférieur à 10^{-6}, le nombre maximum d'itérations étant fixé à 200. L'initialisation des paramètres est faite de la façon suivante : 0 pour les paramètres γ, $\frac{1}{K}$ pour les proportions. Pour λ_j qui est égal à -1 ou 1, s'il est possible d'énumérer toutes ses valeurs possibles (nous rappelons que la complexité de l'estimation de λ_j peut dépendre du nombre d de variables explicatives - cf. modèle M_2 - mais également du nombre K de groupes - cf. modèle M_3 où $\lambda_{kj}^{'}$ remplace λ_j), nous choisissons la solution de log-vraisemblance maximale. Dans le cas contraire, nous utilisons la méthode de relaxation décrite précédemment et nous initialisons le paramètre λ_j à 0. De même pour le paramètre δ_{kj} dans le modèle M_1.

Les résultats sont présentés dans la figure 1. L'axe des abscisses représente les huit transformations entre P et P^* effectuées, correspondant aux huit modèles pM_1 à pM_4 et M_1 à M_4 (cette convention sera conservée pour les prochaines figures 2 et 3). Ces résultats montrent que l'analyse discriminante généralisée donne des taux d'erreur plus faibles que l'analyse discriminante classique ou que la classification automatique. Ces bons résultats ne sont pas surprenant puisque toutes les hypothèses de l'analyse discriminante généralisée sont respectées.
Nous présentons maintenant quelques autres tests où ces hypothèses sont violées.

5.1.2 Mauvais modèles

Plusieurs autres séries de tests dans lesquelles les hypothèses de la discrimination généralisée ne sont pas vérifiées ont été réalisées. Nous présentons dans les figures 2 et 3 les résultats d'une de ces séries de tests, dans laquelle l'hypothèse que les variables binaires sont dues à la

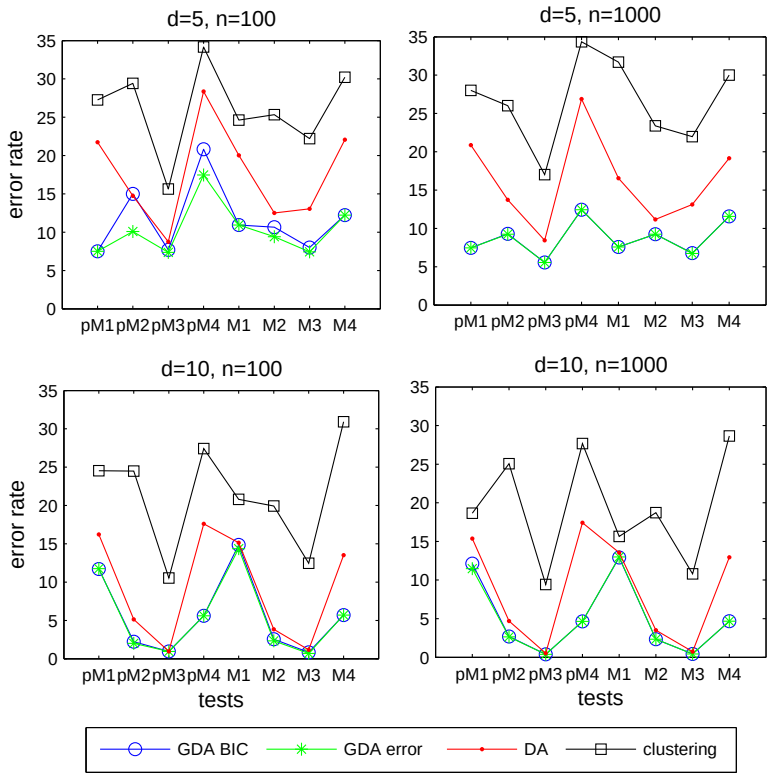

FIG. 1 – *Tests sur données simulées sans bruit. La première ligne correspond à* $d = 5$ *et la seconde à* $d = 10$*. La première colonne correspond à* $n = n^* = 100$ *et la seconde à* $n = n^* = 1000$*.*

discrétisation de variables latentes gaussiennes est perturbée par l'introduction d'un bruit dans les données. Ce bruit consiste en une proportion de données uniformes parmi les données gaussiennes, centrées sur la moyenne des gaussiennes et étendue à plus ou moins deux écart-types. Deux proportions de bruit sont testées : 10% et 30%. Les paramètres des lois gaussiennes sous-jacentes et des transformations sont identiques à ceux des tests du paragraphe précédent.

Pour 10% de bruit (figure 2), l'analyse discriminante généralisée est encore meilleure que les autres méthodes, mais lorsque le bruit est trop important (figure 3), la classification automatique devient la meilleure méthode (suivant le taux de mauvais classement).

Trois autres séries de tests violant les hypothèses de la discrimination généralisée ont également été réalisées. Dans la première série, les données gaussiennes, à partir desquelles sont simulées les données binaires par discrétisation, sont bruitées par l'introduction de données binaires uniformément distribuées. Dans la deuxième série, les données binaires sont simulées

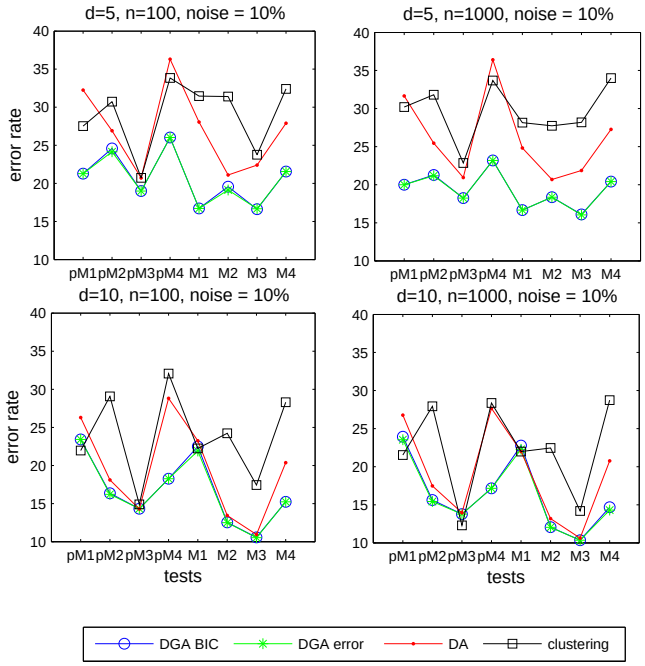

FIG. 2 – *Tests sur données simulées avec* 10% *de bruit. La première colonne correspond à* $n = n^* = 100$ *et la seconde à* $n = n^* = 1000$. *La première ligne correspond à* $d = 5$ *et la seconde à* $d = 10$.

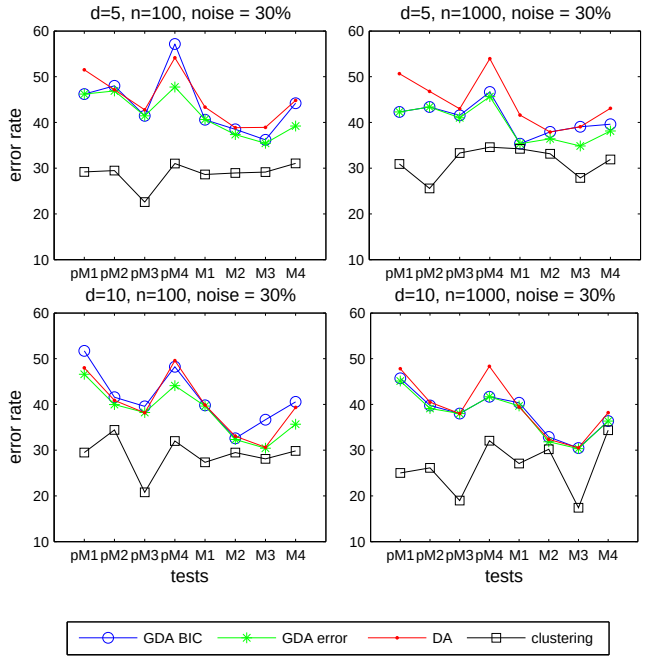

FIG. 3 – *Tests sur données simulées avec* 30% *de bruit. La première colonne correspond à* $n = n^* = 100$ *et la seconde à* $n = n^* = 1000$. *La première ligne correspond à* $d = 5$ *et la seconde à* $d = 10$.

par discrétisation de variables gaussiennes non conditionnellement indépendantes. Finalement, dans une troisième série, la matrice d'homothétie A_k de la transformation entre populations est choisie non diagonale. Dans tous ces tests, les résultats sont similaires à ceux présentés précédement, c'est-à-dire lorsque l'écart au modèle n'est pas trop important, l'analyse discriminante généralisée est meilleure que les autres méthodes, mais lorsque cet écart est trop important, c'est la classification automatique qui doit être préférée.

Tous ces tests nous permettent de conclure que l'analyse discriminante généralisée donne de meilleurs classements que l'analyse discriminante classique ou que la classification automatique lorsque les populations d'apprentissage et de test ne sont pas identiques, et qu'elle est relativement robuste aux hypothèses suivantes : les données binaires sont issues de la discrétisation de variables latentes gaussiennes, et la transformation linéaire entre P et P^* s'effectue avec une matrice d'homothétie A_k diagonale.

Nous présentons maintenant une application sur un jeu de données réelles dans un contexte biologique.

5.2 Données réelles biologiques

Les premières motivations pour lesquelles l'analyse discriminante généralisée a été développée sont des applications biologiques (Biernacki et al. (2002) ; Van Franeker et Ter Brack (1993)), dans lesquelles l'objectif est de prédire le sexe d'oiseaux à partir de variables biométriques. De très bon résultats ont été obtenus sous une hypothèse multi-normale.
L'espèce d'oiseaux considérée est l'espèce *Calanectris diomedea* (Thibault et al. (1997)). Cette espèce peut être divisée en trois sous-espèces, parmi lesquelles les *borealis*, qui vivent dans des îles de l'Atlantique (Açores, Canaries, etc.), et les *diomedea*, qui vivent dans des îles de la Méditerranée (Baléares, Corse, etc.). Les oiseaux de la sous-espèce *borealis* sont généralement plus grands que ceux de la sous-espèce *diomedea*. C'est pourquoi l'analyse discriminante classique n'est pas adaptée pour prédire le sexe des oiseaux *diomedea* à partir d'un échantillon d'apprentissage issu de la population des *borealis*.
La figure 4 illustre les différences entre deux variables morphologiques (tailles des ailes et de la queue), pour les deux espèces *diomedea* et *borealis*.

Un échantillon de *borealis* ($n = 206$, 45% de femelles) est constitué à partir d'information fournies par plusieurs musées nationaux. Cinq variables morphologiques sont mesurées : profondeur et longueur du bec, longueur du tarse (os de la patte), longueur des ailes, longueur de la queue. De même, un échantillon de *diomedea* ($n = 38$, 58% de femelles) est mesuré sur le même ensemble de variables morphologiques. Dans cet exemple deux groupes sont présents, les mâles et les femelles, et tous les oiseaux sont de sexe connus (par dissection).
Pour procurer une application à notre travail, nous discrétisons les variables biométriques en variables binaires (petites ou grandes ailes, ...).
Nous choisissons pour population d'apprentissage l'espèce *borealis* et pour population test l'espèce *diomedea*. Les trois méthodes de classification, discrimination généralisée, discrimination classique et classification automatique sont testées. Les résultats sont présentés dans le tableau 2.
Si nous comparons les résultats obtenus en terme de taux d'erreur, la discrimination générali-

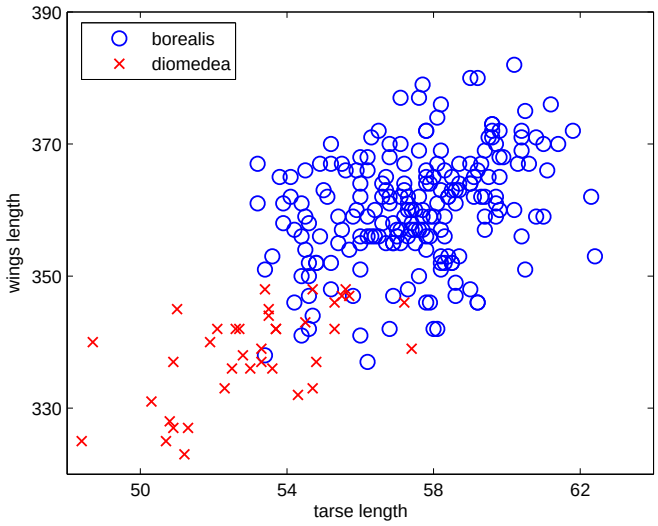

FIG. 4 – *Longueur des ailes et de la queue pour les deux espèces diomedea et borealis.*

	pM_1	pM_2	pM_3	pM_4	M_1	M_2	M_3	M_4	DA	Clustering
Taux	57.9	26.3	23.7	21.0	57.9	23.7	**15.8**	18.4	42.1	23.7
Nombre	22	10	9	8	22	9	**6**	7	16	9
BIC	269.7	222.5	**220.5**	237.0	267.3	221.6	221.5	233.6	648.4	-

TAB. 2 – *Taux (en $\%$) et nombre de mauvais classements, et valeurs du critère BIC pour la population test diomedea avec la population d'apprentissage borealis.*

sée est la meilleure méthode avec un taux d'erreur de 15.8% pour le modèle M_3. Cette erreur est plus faible que celles obtenues par discrimination classique (42.11%) et par classification automatique (23.7%). Si l'on utilise le critère BIC pour choisir un modèle, quatre sont mis en valeur (pM_3, M_3, M_2 et pM_2) parmi lesquels celui qui conduit au taux d'erreur le plus faible (M_3).

Remarque. Le meilleur des taux de mauvais classements (15.8%) est le même que celui obtenu par Biernacki et al. (2002) en travaillant sur les variables continues. Bien que la discrétisation de variables continues engendre indéniablement une perte d'information, cela n'a pas d'effet ici sur le classement des oiseaux de l'espèce *diomedea*.

Cette application illustre l'intérêt de la discrimination généralisée vis-à-vis des procédures classiques de discrimination linéaire ou de classification automatique. En effet, en adaptant la règle de classement issue de l'échantillon d'apprentissage à l'échantillon test en fonction des différences entre les deux populations d'apprentissage et de test, la discrimination généralisée

permet de classer les individus de la population test de façon plus efficace qu'en appliquant directement la règle de classement issue de l'échantillon d'apprentissage, ou encore en oubliant l'échantillon d'apprentissage et en effectuant une classification automatique directement sur l'échantillon test.
Remarquons aussi que l'hypothèse gaussienne conditionnellement au sexe était relativement acceptable sur les variables biométriques sous-jacentes. Néanmoins, il y avait une forte corrélation entre ces différents caractères, violant alors les hypothèses d'indépendance conditionnelle de nos modèles.

5.3 Données réelles d'assurances

La discrimination généralisée peut être une alternative intéressante aux méthodes classiques pour traiter des problèmes de discrimination en assurance, lorsque les populations d'apprentissage et de test différent par exemple suivant des zones géographiques (extension d'un marché parisien à une région de province), suivant des périodes temporelles ou encore suivant les risques assurés (automobile, vie...).

Les données dont nous disposons concernent le nombre de sinistres matériels des cinq dernières années (depuis 2005), parmi des clients ayant un unique contrat d'assurance automobile et étant assurés depuis au moins cinq années. Ces données proviennent d'un grand groupe d'assurance français.
Les clients sont décrits par les cinq variables explicatives binaires suivantes : département de résidence (Ile de France ou province), situation familiale (célibataire ou non), situation professionnelle (profession supérieure ou autre), sexe et type de paiement (annuel ou non).
La population d'apprentissage P est constituée de clients « infidéles »et la population test P^* est contituée de clients « fidèles ». La fidélité est définie à partir de l'année d'adhésion et de l'âge du client de la façon suivante :

$$\text{taux de fidélité} = \frac{2005 - \text{année d'adhésion}}{\text{âge} - 17}.$$

Les clients considérés comme infidèles ont un taux de fidélité inférieur à 0.5, et les clients considérés comme fidèles ont un taux supérieur à 0.5.
L'échantillon d'apprentissage S est constitué de 112755 individus, et l'échantillon test S^* est constitué de 144277 individus.

L'objectif de notre étude est de discriminer les clients sans sinistre matériel (appelé groupe 1) de ceux ayant au moins un sinistre (appelé groupe 2). Pour cela, nous considérons les coûts de mauvais classement suivant :

- $C(1,2)$: coût de classer sans sinistre un client qui aura au moins un sinistre. Ce coût peut correspondre à un coût financier moyen de sinistre (estimé à partir du passé).
- $C(2,1)$: coût de classer avec sinistre un client qui n'aura pas de sinistre. Ce coût peut correspondre à un coût financier moyen de perdre le client (passage à la concurrence) en essayant de lui appliquer une pénalité financière de risque de sinistre. On peut l'estimer à partir du passé et le mettre à jour en comparant avec les prix de la concurrence.

En fait, seul le ratio $C(1,2)/C(2,1)$ est suffisant pour établir les règles de classement. On appelera ce ratio « coût moyen d'un ou plusieurs sinistres en unité $C(2,1)$ ».

La figure 5 donne en fonction de ce ratio le risque moyen (en unité $C(2,1)$) associé aux régles de décision des huit modèles de discrimination généralisée, de la discrimination classique (D.A.) et de la classification automatique (Clust.).

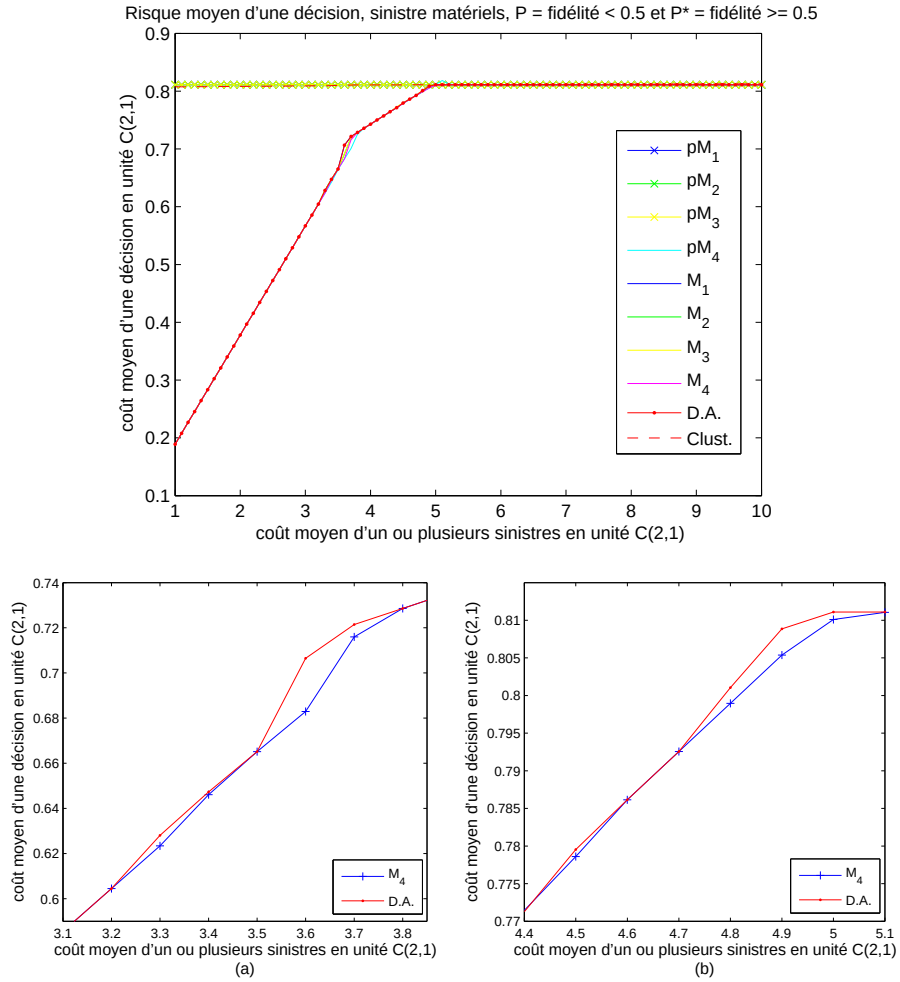

FIG. 5 – *Risque moyen d'une décision en fonction du coût moyen d'un ou plusieurs sinistres.*

On peut constater sur la figure 5 une légère mais intéressante diminution du risque moyen pour quelques valeurs du coût moyen (entre 3 et 4, figure 5(a), et autour de 5, figure 5(b)) essentiellement grâce au modèle de discrimination généralisée M_4, qui est de plus le modèles selectionné par le critère BIC (tableau 3). En dehors de ces plages, le modèle M_4 donne les mêmes performances que la discrimination classique.

	pM_1	pM_2	pM_3	pM_4	M_1	M_2	M_3	M_4	DA
BIC	744235	742935	742944	740295	747345	744629	744641	**740284**	747345

TAB. 3 – *Valeurs du critère BIC pour la discrimination généralisée et la discrimination classique.*

6 Conclusion

L'analyse discriminante généralisée étend l'analyse discriminante classique en permettant aux échantillons d'apprentissage et de test d'être issus de populations différentes mais liées. Notre contribution consiste à adapter les travaux précurseurs réalisés dans un contexte gaussien au cas des données binaires. Deux applications en biologie et en assurance illustrent ce travail. En utilisant les modèles de discrimination généralisée définis dans ce papier, nous obtenons des classements meilleurs que ceux obtenus par analyse discriminante classique ou par classification automatique.

Les perspectives méthodologiques de ces travaux sont nombreuses.
Tout d'abord, nous avons défini le lien entre les deux populations en utilisant la fonction de répartition de la loi normale centrée réduite. Bien qu'il put paraître initialement difficile de trouver un lien entre populations binaires, un lien simple a été obtenu. Il n'aurait pas été facile de l'imaginer, mais il est aisément compréhensible *a posteriori*. Cela pourrait être intéressant d'essayer d'autres types de fonction de répartition (les raisons théoriques devront êtres développées et des tests devront être effectués).
Grâce à cette contribution, l'analyse discriminante généralisée est maintenant développée pour des données continues et des données binaires. Pour permettre de traiter un maximum de cas pratiques, il est important d'étudier le cas de variables catégoriques (à plus de deux modalités), puis ensuite le cas de variables mélangées (variables binaires, continues et catégoriques dans un même problème).
Finalement, il sera aussi très intéressant d'étendre les autres méthodes de discrimination classique comme la discrimination non-paramètrique.

Remerciements

Nous remercions les rapporteurs ayant examiné cet article ; leurs remarques nous ont permis à travers leurs remarques de progresser dans nos recherches.
Nous remercions également Farid Beninel pour nous avoir fourni les données d'assurances.

Références

Biernacki, C., F. Beninel, et V. Bretagnolle (2002). A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics 58,2*, 387–397.

Celeux, G. et G. Govaert (1991). Clustering criteria for discrete data and latent class models. *Journal of classification 8*, 157–176.

Dempster, A., N. Laird, et D. Rubin (1977). Maximum likelihood from incomplete data (with discussion). *Journal of the Royal Statistical Society Series B 39*, 1–38.

Everitt, B. (1984). *An introduction to latent variables models*. London: Chapman & Hall.

Everitt, B. (1987). A finite mixture model for the clustering of mixed-mode data. *Statistics and Probability Letters 6*, 305–309.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*, 179–188.

Jacques, J. (2005). *Contributions à l'analyse de sensibilité et à l'analyse discriminante généralisée*. Thèse de doctorat, Université Joseph Fourier de Grenoble.

McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New-York: Wiley.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Thibault, J.-C., V. Bretagnolle, et C. Rabouam (1997). Cory's shearwater calonectris diomedia. *Birds of Western Paleartic Update 1*, 75–98.

Thurstone, L. (1927). A law of comparative judgement. *Amer. J. Psychol. 38*, 368–389.

Van Franeker, J. et C. Ter Brack (1993). A generalized discriminant for sexing fulmarine petrels from external measurements. *The Auk 110(3)*, 492–502.

Wolsey, L. (1998). *Integer Programming*. Wiley.

A Paramètres des simulations

Les valeurs des paramètres pour tous les tests sur simulations ont été choisies arbitrairement.

A.1 Dimension 5

Les données binaires sont simulées à partir d'une discrétisation d'un mélange de deux gaussiennes de centres et de matrice de variance :

$$\begin{aligned} \boldsymbol{\mu}_1 &= (-2, -1, -1.5, 1.4, -1.2)' \\ \boldsymbol{\mu}_2 &= (1, 1.5, 2, 2.1, 0.9)' \\ \Sigma_1 &= \Sigma_2 = 3I_5 \end{aligned}$$

où I_5 est la matrice identité de dimension 5.
Les proportions de ce mélange dans la population P sont $p_1 = \frac{1}{2}$ et $p_2 = \frac{1}{2}$. Dans la population P^* elles sont soit les mêmes (proportions inchangées), soit $p_1 = \frac{3}{10}$ et $p_2 = \frac{7}{10}$ (proportions différentes).
Les paramètres utilisés pour les transformations entre P et P^* sont :

- paramètres de la transformation 1 (elle respecte les hypothèse du modèle M_1) :
  $A_1 = \text{diag}(0.5, 2, 3, -2, 3)$, $A_2 = \text{diag}(2, -2, 2, 3, -3)$ et $\boldsymbol{b}_1 = \boldsymbol{b}_2 = (0, 0, 0, 0, 0)'$,
- paramètres de la transformation 2 (elle respecte les hypothèse du modèle M_2) :
  $A_1 = A_2 = 2I_5$ et $\boldsymbol{b}_1 = \boldsymbol{b}_2 = -\mathbf{e}_5$ où $\mathbf{e}_5$ est le vecteur unité de dimension 5.
- paramètres de la transformation 3 (elle respecte les hypothèse du modèle M_3) :
  $A_1 = 3I_5$, $A_2 = \frac{1}{2}I_5$, $\boldsymbol{b}_1 = 2\,\mathbf{e}_5$ et $\boldsymbol{b}_2 = 0.6\,\mathbf{e}_5$,
- paramètres de la transformation 4 (elle respecte les hypothèse du modèle M_4) :
  $A_1 = \text{diag}(0.5, 2, 3, 2, 4)$, $A_2 = \text{diag}(2, -2, 2, 3, -3)$ et $\boldsymbol{b}_1 = \boldsymbol{b}_2 = (-1, -2, -2, -3, -2)'$.

A.2 Dimension 10

Les données binaires sont simulées à partir d'une discrétisation d'un mélange de deux gaussiennes de centres et de matrice de variance :

$$\begin{aligned} \boldsymbol{\mu}_1 &= (-2, -1, -1.5, 1.4, -1.2, -1.8, 1.6, 0.5, -1.4, -1.8)', \\ \boldsymbol{\mu}_2 &= (1, 1.5, 2, 2.1, 0.9, 1.6, -2.3, 0.8, 0.5, 2)', \\ \Sigma_1 &= \Sigma_2 = 3I_{10}. \end{aligned}$$

Les proportions du mélange sont les mêmes qu'en dimension 5.
Les paramètres utilisés pour les transformations entre P et P^* sont :

- paramètres de la transformation 1 :
  $A_1 = \text{diag}(0.5, 2, 3, -2, 3, -0.5, 3, 2, -2, 0.5)$, $A_2 = \text{diag}(2, -2, 2, 3, -3, -2, 3, -0.5, 3, 2)$,
  et $\boldsymbol{b}_1 = \boldsymbol{b}_2 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$,
- paramètres de la transformation 2 :
  $A_1 = A_2 = 2I_{10}$ et $\boldsymbol{b}_1 = \boldsymbol{b}_2 = -\mathbf{e}_{10}$,
- paramètres de la transformation 3 :
  $A_1 = 2I_{10}$, $A_2 = 0.5\,I_{10}$, $\boldsymbol{b}_1 = -0.5\,\mathbf{e}_{10}$ et $\boldsymbol{b}_2 = 0.5\,\mathbf{e}_{10}$,
- paramètres de la transformation 4 :
  $A_1 = A_2 = \text{diag}(0.5, 2, 3, 2, 4, 0.5, 3, 2, 4, 2)$ et $\boldsymbol{b}_1 = \boldsymbol{b}_2 = (-1, 1, 2, 1, 2, -2, 3, 2, 1, -4)'$.

Analyse discriminante généralisée sur données binaires

Summary

Standard discriminant analysis supposes that both the training labelled sample and the test sample which has to be classed are issued from the same population. When these samples are issued from populations for which descriptive parameters are different, generalized discriminant analysis allows us to adapt the classification rule issued from the training population to the test population, by estimating a link between this two populations. This paper extends existing work available in a multi-normal context to the case of binary data. To raise the major challenge of this work which is to define a link between the two binary populations, we suppose that binary data are issued from the discretization of latent Gaussian data. Estimation method and tests on simulated data are presented, and two applications in biological and insurance context illustrate this work.