Relaxations de la régression logistique : modèles pour l'apprentissage sur une sous-population et la prédiction sur une autre

Farid Beninel*, Christophe Biernacki**

*CREST ENSAI
rue Blaise Pascal, Campus de Ker Lann
35170 Bruz, France
Farid.Beninel@ensai.fr

**Université Lille1, UFR de mathématiques, UMR 6524
59655 Villeneuve d'Ascq, France
Christophe.Biernacki@math.univ-lille1.fr

Résumé. Habituellement en analyse discriminante on a à prédire le groupe d'appartenance à partir des variables de description ou covariables. La règle de prédiction est élaborée en utilisant un échantillon d'apprentissage soumis aux mêmes conditions externes que les individus à prédire. Dans ce travail, on s'intéresse à la prédiction d'individus d'une certaine sous-population utilisant un échantillon d'apprentissage d'une autre sous-population. En assurance-finance, le problème apparaît quand il faut inférer le groupe d'appartenance de *sociétaires-clients* soumis à certaines conditions externes et que la règle est élaborée à partir d'individus soumis à d'autres. On propose différents modèles étendant la discrimination logistique classique. Ces modèles se fondent sur des relations acceptables entre les fonctions scores que l'on associerait à chacune des sous-populations en présence.

1 Introduction

Traditionnellement, l'analyse discriminante procède de la façon suivante (McLachlan 1992) : un échantillon provient d'une population et une partition en plusieurs groupes de cet échantillon est connue. À partir des variables disponibles, une règle de classement est alors établie dans le but de classer tout nouvel élément non étiqueté. Néanmoins, une hypothèse sous jacente à cette procédure est que ces nouveaux individus et les individus constituant l'échantillon d'apprentissage proviennent de la même population. L'analyse discriminante généralisée consiste à étendre le problème de l'analyse discriminante classique lorsque cette hypothèse fondamentale est relaxée.
Dans Biernacki *et al.* (2002) on considère cette extension dans le cas de la discrimination gaussienne multivariée. À partir d'hypothèses simples et raisonnables sur la nature du lien stochastique entre les deux sous-populations d'où proviennent respectivement l'échantillon d'ap-

prentissage et l'échantillon de prédiction, il a été établi que cette liaison était nécessairement affine. Un certain nombre de modèles de contraintes sur les paramètres affines de cette liaison ont ensuite été proposés, permettant ainsi d'obtenir des modèles parcimonieux, généralement faciles à interpréter par le praticien et enfin retrouvant puis généralisant des travaux précurseurs en discrimination généralisée (Van Franecker & Ter Brack 1993).
La méthode proposée a été testée sur des données issues de la biologie. Les sous-populations de prédiction et d'apprentissage correspondaient à deux espèces similaires d'oiseaux de mer différant de par leur localisation géographique (Bretagnolle *et al.* 1998). L'objectif était d'estimer ou de prédire le sexe d'oiseaux de la seconde espèce à partir des seules variables biométriques. La règle de prédiction utilisée est construite à partir des données relatives aux individus de la première espèce renseignés, eux, pour la biométrie et le sexe. Les résultats se sont alors révélés très concluants sur cet exemple.
On comprend aisément que le potentiel d'évolution du concept d'analyse discriminante généralisée soit fort puisqu'il concerne naturellement l'ensemble des méthodes et des types de données considérées en analyse discriminante classique. Dans ce travail, nous focalisons notre attention sur la discrimination ou régression logistique, méthode employée dans de nombreux domaines (dont les assurances) et permettant de traiter des données de différentes natures (continues et/ou catégorielles).
L'idée initiale consiste à utiliser les résultats établis dans le modèle gaussien multivarié et à les transposer au modèle logistique. Afin d'exprimer simplement et avec parcimonie une relation entre les paramètres de la discrimination logistique des deux sous-populations (celle d'apprentissage et celle de prédiction), un certain nombre de modèles de liaison vont être identifiés et explicités. Dans un premier temps, il est donc utile de rappeler succinctement les résultats disponibles dans le cas gaussien.

2 Discrimination gaussienne généralisée

2.1 Problématique

En discrimination généralisée (gaussienne ou non), les données consistent en deux échantillons : un échantillon d'apprentissage, (*i.e.* un échantillon *avec* labels) S provenant d'une sous-population Ω et un échantillon de prédiction, (*i.e.* un échantillon *sans* labels) S^* provenant d'une sous-population Ω^*. La problématique fondamentale repose sur le fait que les sous-populations Ω et Ω^* peuvent être différentes.
Dans le contexte de la discrimination gaussienne multivariée, l'échantillon d'apprentissage S est composé de n couples $(\mathbf{x}_1, z_1)$, ..., $(\mathbf{x}_n, z_n)$ où $\mathbf{x}_i$ est un vecteur de $\mathbb{R}^d$ représentant les caractéristiques numériques décrivant l'individu numéro i ($i = 1, \ldots, n$) et où z_i est le numéro de son groupe d'appartenance. Ainsi, $z_i = k$ avec $k = 1, \ldots, K$ si cet individu appartient au groupe k parmi K groupes possibles. Les n couples $(\mathbf{x}_i, z_i)$ sont supposés être des réalisations *i.i.d.* du couple aléatoire $(\mathbf{X}, Z)$ défini sur Ω de distribution jointe

$$\mathbf{X}_{|Z=k} \sim \mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad k = 1, \ldots, K \quad \text{et} \quad Z \sim \mathcal{M}_K(1, \pi_1, \ldots, \pi_K) \tag{1}$$

où $\mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ correspond à la distribution gaussienne d dimensionnelle de moyenne $\boldsymbol{\mu}_k \in \mathbb{R}^d$ et de matrice de variance $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ et $\mathcal{M}_K(1, \pi_1, \ldots, \pi_K)$ correspond à la loi de

l'indice non nul d'un vecteur aléatoire issu d'une loi multinomiale d'ordre 1 et de paramètres $\pi_1, \ldots, \pi_K$. Ainsi, le paramètre π_k représente la proportion du groupe k dans la sous-population Ω et par suite $\sum_{k=1}^{K} \pi_k = 1$.
L'échantillon de prédiction S^* est composé de n^* individus desquels seules les caractéristiques numériques $\mathbf{x}_1^*$,, …, $\mathbf{x}_{n^*}^*$ sont connues (ces caractéristiques sont les mêmes que pour S), les labels correspondant z_1^*, …, $z_{n^*}^*$ étant inconnus. Les n^* couples $(\mathbf{x}_i^*, z_i^*)$ sont supposés être des réalisations i.i.d. du couple aléatoire $(\mathbf{X}^*, Z^*)$ défini sur l'espace Ω^* de distribution jointe

$$\mathbf{X}^*_{|Z^*=k} \sim \mathcal{N}_d(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*), \quad k = 1, \ldots, K \quad \text{et} \quad Z^* \sim \mathcal{M}_K(1, \pi_1^*, \ldots, \pi_K^*). \tag{2}$$

L'objectif étant d'estimer les n^* labels inconnus z_1^*, …, $z_{n^*}^*$ en utilisant l'information provenant des deux échantillons S et S^*, le principal enjeu est alors d'identifier une relation liant les sous-populations Ω et Ω^*.

2.2 Liaison affine entre sous-populations

L'approche proposée pour la discrimination généralisée consiste à établir une application $\boldsymbol{\phi}_k$ de $\mathbb{R}^d$ dans $\mathbb{R}^d$ liant en loi les vecteurs aléatoires du groupe k ($k = 1, \ldots, K$) des sous-populations Ω et Ω^*. Ainsi

$$\mathbf{X}^*_{|Z^*=k} \sim \boldsymbol{\phi}_k(\mathbf{X}_{|Z=k}) = [\phi_{k1}(\mathbf{X}_{|Z=k}), \ldots, \phi_{kd}(\mathbf{X}_{|Z=k})]^{\mathrm{T}}, \tag{3}$$

avec ϕ_{kj} une application de $\mathbb{R}^d$ dans $\mathbb{R}$ ($j = 1, \ldots, d$). Deux hypothèses essentielles sont alors faites sur l'application $\boldsymbol{\phi}_k$. Tout d'abord, il est supposé que la $j$ème composante $\phi_{kj}(\mathbf{X}_{|Z=k})$ de $\boldsymbol{\phi}_k(\mathbf{X}_{|Z=k})$ ne dépend que de la $j$ème composante $X_{j|Z=k}$ de $\mathbf{X}_{|Z=k}$. Cela revient à admettre que ϕ_{kj} est une application de $\mathbb{R}$ dans $\mathbb{R}$ (par simplicité, la notation ϕ_{kj} est conservée), ce qui s'écrit aussi

$$\boldsymbol{\phi}_k(\mathbf{X}_{|Z=k}) = [\phi_{k1}(X_{1|Z=k}), \ldots, \phi_{kd}(X_{d|Z=k}))]^{\mathrm{T}}. \tag{4}$$

En second lieu, cette application ϕ_{kj} est supposée être de classe C^1. Ces deux hypothèses suffisent à établir que la fonction ϕ_{kj} est nécessairement affine (résultat provenant de De Meyer et *al.* 2000), ce qui conduit aux K relations suivantes

$$\mathbf{X}^*_{|Z^*=k} \sim \mathbf{D}_k \mathbf{X}_{|Z=k} + \mathbf{b}_k, \quad k = 1, \ldots, K \tag{5}$$

avec $\mathbf{D}_k$ une matrice diagonale de $\mathbb{R}^{d \times d}$ et $\mathbf{b}_k$ un vecteur de $\mathbb{R}^d$. Comme conséquence immédiate, les paramètres des deux sous-populations Ω et Ω^* sont liés de la façon suivante :

$$\text{pour} \quad k = 1, \ldots, K, \quad \boldsymbol{\mu}_k^* = \mathbf{D}_k \boldsymbol{\mu}_k + \mathbf{b}_k \quad \text{et} \quad \boldsymbol{\Sigma}_k^* = \mathbf{D}_k \boldsymbol{\Sigma}_k \mathbf{D}_k. \tag{6}$$

3 La discrimination logistique classique

3.1 Le modèle logistique

Par commodité le label Z denotera dorénavant une variable binaire dont la valeur 1 correspond au groupe **1** et la valeur 0 correspond au groupe **2**. Les n paires $(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_n, z_n)$ de

l'échantillon d'apprentissage sont des réalisations indépendantes du couple aléatoire $(\mathbf{X}, Z)$ dont la loi jointe est définie par $\mathbf{X}_{|Z=k} \sim f_k(k=0,1)$ et $Z \sim \mathcal{B}(1,\pi)$. Dans le cadre de cette hypothèse sur les données, le modèle logistique porte sur la distribution conditionnelle, *i.e.*

$$\mathbb{P}(Z=1|\mathbf{X}=\mathbf{x}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\beta_0+\boldsymbol{\beta}^T\mathbf{x})}. \tag{7}$$

Etant donné un nouvel individu $(\mathbf{x}^*, z^*) \sim (\mathbf{X}, Z)$ soumis aux mêmes conditions que ceux de l'échantillon S et pour lequel seul $\mathbf{x}^*$ est connu, il s'agit de prédire la valeur z^*.
Cette prédicion se base sur la seule information liant variables explicatives et à expliquer que recèle l'échantillon d'apprentissage S. L'affectation au groupe $\{Z=1\}$ ou $\{Z=0\}$ se fait selon la vraisemblance de l'une ou l'autre des deux alternatives, eu égard à la donnée $\mathbf{x}^*$; cette vraisemblance est mesurée par la probabilité $\pi_{\mathbf{x}^*} = \mathbb{P}(Z=1|\mathbf{x}^*)$. On affecte au groupe $\mathbf{1}$ $(z^*=1)$ si $\pi_{\mathbf{x}^*} > \frac{1}{2}$ et au groupe $\mathbf{2}$ $(z^*=0)$ si $\mathbf{x}^* < \frac{1}{2}$; l'indétermination $\pi_{\mathbf{x}^*} = \frac{1}{2}$, *a priori* rare, sera gérée selon le niveau de sévérité que l'on impose quant à admettre au groupe $\mathbf{1}$ *resp.* au groupe $\mathbf{2}$.
L'hypothèse sous jacente au modèle logistique, est que

$$\log\left(\frac{f_1(\mathbf{x}^*)}{f_0(\mathbf{x}^*)}\right) = \tilde{\beta}_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}^*, \tag{8}$$

avec $\tilde{\beta}_0 \in \mathbb{R}$, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_d)^{\mathrm{T}} \in \mathbb{R}^d$. En effet, par la formule de Bayes, on a $\pi_{\mathbf{x}^*} = \pi f_1(\mathbf{x}^*)/[\pi f_1(\mathbf{x}^*) + (1-\pi)f_0(\mathbf{x}^*)]$ et par suite,

$$\log\left(\frac{f_1(\mathbf{x}^*)}{f_0(\mathbf{x}^*)}\right) = \log\left(\frac{(1-\pi)\pi_{\mathbf{x}^*}}{\pi(1-\pi_{\mathbf{x}^*})}\right) = \log\left(\frac{1-\pi}{\pi}\right) + \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}^*. \tag{9}$$

Ainsi, le modèle logistique vaut pour les situations où le rapport des lois marginales f_1 et f_0 peut être considéré à logarithme linéaire, comme cela est le cas par exemple lorsque f_1 et f_0 sont des distributions gaussiennes homoscédatiques. Plus généralement, cela est vérifié pour une large famille de distributions multivariées continues et discrètes (Anderson 1982).

3.2 Estimation des paramètres

Ici, le problème est restreint a l'estimation des paramètres β_0 et $\boldsymbol{\beta}$ du modèle donné en (7), étant donné l'échantillon $S = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_n, z_n)\}$.
La log-vraisemblance, associée aux données, est

$$l(\beta_0, \boldsymbol{\beta}) = \sum_{k=0}^{1} \sum_{i:z_i=k} z_i \log(\pi f_1(\mathbf{x}_i)) + (1-z_i)\log((1-\pi)f_2(\mathbf{x}_i)). \tag{10}$$

On décompose en *resp.* une partie conditionnelle et une partie marginale.

$$l(\beta_0, \boldsymbol{\beta}) = \sum_{k=0}^{1} \sum_{i:z_i=k} \log(\mathbb{P}(Z_i=k|\mathbf{x}_i)) + \sum_{i=1}^{n} \log(\pi f_1(\mathbf{x}_i) + (1-\pi)f_2(\mathbf{x}_i)). \tag{11}$$

En réalité, du fait que les modèles (inconnus) de lois auxquels s'apparentent f_1 et f_2 peuvent être considérés comme indépendantes de $(\beta_0, \boldsymbol{\beta})$, on maximise la seule log-vraiemblance conditionnelle, *i.e.*

$$l_{cond}(\beta_0, \boldsymbol{\beta}) = \sum_{k=0}^{1} \sum_{i:z_i=k} \log(\mathbb{P}(Z_i = k|\mathbf{x}_i)) = \sum_{i=1}^{n} z_i \log(\pi_{\mathbf{x_i}}) + (1-z_i)\ \log(1-\pi_{\mathbf{x_i}}). \quad (12)$$

On trouve dans (Govaert 2003) une justification de cette démarche.
La recherche de $(\beta_0, \boldsymbol{\beta})$ maximisant $l_{cond}(\beta_0, \boldsymbol{\beta})$ consiste en la résolution du système non linéaire (S_1) suivant obtenu par annulation des dérivées partielles :

$$(S_1) : \begin{cases} \frac{\partial l_{cond}(\beta_0, \boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} (z_i - \pi(\mathbf{x}_i, \beta_0, \boldsymbol{\beta})) = 0, \\ \\ \frac{l_{cond}(\beta_0, \boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \mathbf{x}_{i,j} (z_i - \pi(\mathbf{x}_i, \beta_0, \boldsymbol{\beta})) = 0. \qquad j = 1, \ldots, d. \end{cases}$$

Dans le cas général, le système (S_1) admet une unique solution correspondant au maximum. Le cas dégénéré (non unicité) correspond a un modèle surparamétré. Du point de vue computationnel, la résolution du système (S_1) se fait par l'usage de l'algorithme de Newton-Raphson ou ses variantes.

4 La discrimination logistique étendue à un mélange de deux sous-populations

4.1 Les données

On dispose des deux échantillons $S = \{(\mathbf{x}_i, z_i) : \ i = 1, \ldots, n\}$ et $S^* = \{(\mathbf{x}_i^*, z_i^*) : \ i = 1, \ldots, n^*\}$; les tailles respectives sont n et n^*. Les paires $(\mathbf{x}_i, z_i)$ sont des réalisations indépendantes du couple aléatoire $(\mathbf{X}, Z)$ restreint à la sous-population Ω ; les paires $(\mathbf{x}_i^*, z_i^*)$ sont, elles, des réalisations indépendantes du couple $(\mathbf{X}, Z)$ restreint a une autre sous-population Ω^*.

4.2 La problèmatique

Il s'agit de mettre en évidence une fonction d'affectation aux groupes pour les individus de Ω^* en utilisant les échantillons d'apprentissage S et S^* ; ce deuxième échantillon est supposé de *petite taille*.
L'utilisation (comme données complémentaires) de l'échantillon $S \subset \Omega$ pour prédire dans Ω^* est pour pallier à l'insuffisance en nombre des observations constituant S^* et se fonde sur un lien supposé entre les restrictions (aux deux sous-populations) du vecteur de covariables, *i.e.* $\mathbf{X}_{|\boldsymbol{\Omega}}$ et $\mathbf{X}_{|\boldsymbol{\Omega^*}}$.
Le lien entre restrictions du vecteur de covariables implique un lien entre fonctions scores. Ainsi, l'utilisation d'un lien acceptable entre les fonctions scores des deux sous-populations permet de se servir de l'information que recèlent les échantillons S et S^* pour affecter, aux

groupes, des individus de Ω^*. Soient Ψ ($\Omega \to \{0,1\}$) et $\Psi^*(\Omega^* \to \{0,1\})$ les deux fonctions score. On fait l'hypothèse que les deux fonctions font intervenir « les mêmes covariables », *i.e.*

$$\Psi(\mathbf{x}) = \log\left(\frac{\pi_{\mathbf{x}}}{1-\pi_{\mathbf{x}}}\right) = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \quad \text{et} \quad \Psi^*(\mathbf{x}^*) = \log\left(\frac{\pi_{\mathbf{x}^*}}{1-\pi_{\mathbf{x}^*}}\right) = \beta_0^* + \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{x}^*.$$

4.3 L'héritage gaussien

Dans le cas où les populations Ω et Ω^* sont gaussiennes homoscédatiques conditionnellement aux groupes et notant les matrices de variances communes $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ et $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_1^* = \boldsymbol{\Sigma}_2^*$, on obtient aisément le lien suivant entre les paramètres logistiques et les paramètres gaussiens des deux sous-populations :

$$\begin{aligned}
\Omega \quad &: \quad \beta_0 = \frac{1}{2}(\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1) \text{ et } \boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\
\Omega^* \quad &: \quad \beta_0^* = \frac{1}{2}(\boldsymbol{\mu}_2^{*\mathrm{T}}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}_2^* - \boldsymbol{\mu}_1^{*\mathrm{T}}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}_1^*) \text{ et } \boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{*-1}(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*).
\end{aligned}$$

En fonction des hypothèses sur les paramètres de liaison entre les deux gaussiennes, il est possible d'exhiber des relations entre β_0^* et β_0 d'une part, et $\boldsymbol{\beta}^*$ et $\boldsymbol{\beta}$ d'autre part, liaisons où l'hypothèse gaussienne n'apparaît plus explicitement. C'est ce genre de modèles que nous allons maintenant décrire ci-dessous.

4.4 Les modèles de liaison entre fonctions scores

Les modèles de liaisons sont donnés par $c \in \mathbb{R}$ et Λ une matrice diagonale d'ordre d tels que $\beta_0^* = \beta_0 + c$ et $\boldsymbol{\beta}^* = \Lambda\boldsymbol{\beta}$. Comme évoqué ci-dessus, les modèles de liaisons retenus sont ceux faisant intervenir les mêmes variables dans chacun des deux modèles et ces modèles ont été inspirés par le cas gaussien homoscédatique décrit précédement.

- $(\mathbf{M.1})$: $\beta_0^* = \beta_0$ et $\boldsymbol{\beta}^* = \boldsymbol{\beta}$; la discrimination logistique est identique pour les populations Ω et Ω^*.
- $(\mathbf{M.2})$: $\beta_0^* = \beta_0$ et $\boldsymbol{\beta}^* = \lambda\boldsymbol{\beta}$; les discriminations logistiques des populations Ω et Ω^* diffèrent uniquement au travers du paramètre scalaire λ.
- $(\mathbf{M.3})$: β_0^* libre et $\boldsymbol{\beta}^* = \boldsymbol{\beta}$; les discriminations logistiques des populations Ω et Ω^* diffèrent uniquement au travers du paramètre scalaire β_0^*.
- $(\mathbf{M.4})$: β_0^* libre et $\boldsymbol{\beta}^* = \lambda\boldsymbol{\beta}$; les discriminations logistiques des populations Ω et Ω^* diffèrent cette fois au travers du couple de paramètres scalaires β_0^* et λ.
- $(\mathbf{M.5})$: $\beta_0^* = \beta_0$ et $\boldsymbol{\beta}^*$ libre ; les discriminations logistiques des populations Ω et Ω^* diffèrent seulement au travers du paramètre vectoriel $\boldsymbol{\beta}^*$.
- $(\mathbf{M.6})$: β_0^* libre et $\boldsymbol{\beta}^*$ libre ; il y a donc autant de paramètres libres $(d+1)$ dans la régression logistique associée à Ω et à celle associée à Ω^*. Dans ce cas limite, il n'existe plus de lien stochastique entre les discriminations logistiques des deux populations.

4.5 Estimation des paramètres

L'estimation des paramètres β_0 et $\boldsymbol{\beta}$ associés à la régression logistique de la population Ω se fait tout d'abord de façon standard et ne nécessite alors aucune explication spécifique.

Dans un second temps, et conditionnellement à la connaissance des paramètres de Ω, les paramètres de transition entre les fonctions scores de Ω et de Ω^* doivent être estimés. En notant $\boldsymbol{\theta}$ ce ou ces paramètres, l'estimation peut se réaliser par maximisation de la log-vraisemblance conditionnelle qui est donnée par :

$$l_{cond}(\boldsymbol{\theta}) = \sum_{i=1}^{n^*} z_i^* \log(\pi(\mathbf{x}_i^*, \boldsymbol{\theta})) + (1 - z_i^*) \log(1 - \pi(\mathbf{x}_i^*, \boldsymbol{\theta})). \tag{13}$$

On peut interpréter les $z_1^*, \ldots, z_{n^*}^*$ comme des réalisations des variables de Bernoulli indépendantes $Z_1^*, \ldots, Z_{n^*}^*$ telles que $Z_i^* \sim \mathcal{B}(\pi_{\mathbf{x}_1^*})$, ceci conditionnellement aux réalisations $x_1^*, \ldots, x_{n^*}^*$.
On donne, pour l'ensemble des modèles, le système d'équations non linéaires (en les composantes de $\boldsymbol{\theta}$) correspondant et une condition nécessaire et suffisante d'unicité de la solution.

Le modèle $(\mathbf{M.1})$ Il n'y a aucun paramètre à estimer car on utilise simplement les estimations associées au score de Ω^*.

Le modèle $(\mathbf{M.2})$ Le paramètre $\boldsymbol{\theta} = \lambda \in \mathbb{R}$ est estimé par la solution de l'équation non linéaire à l'inconnue λ

$$\frac{\partial l_{cond}(\lambda)}{\partial \lambda} = \sum_{i=1}^{n^*} \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* (z_i^* - \pi(\mathbf{x}_i^*, \lambda)) = 0. \tag{14}$$

L'expression de la dérivée seconde est donnée par

$$\frac{\partial^2 l_{cond}(\lambda)}{\partial \lambda^2} = -\sum_{i=1}^{n^*} (\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^*)^2 \pi(\mathbf{x}_i^*, \lambda)(1 - \pi(\mathbf{x}_i^*), \lambda). \tag{15}$$

On établit aisément que $\frac{\partial^2 l_{cond}(\lambda)}{\partial \lambda^2} \leq 0$, *i.e.* que $l_{cond}(\lambda)$ est concave. Considérons la condition (16) ci après :

$$\exists i \in S^* : \quad \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* \neq 0. \tag{16}$$

Sous cette condition $l_{cond}(\lambda)$ est strictement concave, d'où l'unicité de solution de l'equation (14), *i.e.* unicité de l'estimation du seul paramètre intervenant dans le modèle. La non vérification de la condition (16) entraine que la fonction $l_{cond}(\lambda)$ est indépendante des observations $\mathbf{x}_i^*$ et de λ d'où la non unicité. En conséquence toutes les estimations (i.e., choix) de λ génèreront des modèles qui affecteront les observations $\mathbf{x}_i^*$ au même groupe d'appartenance.
La condition (16) est peu restrictive ; elle est vérifiée notamment lorsque les observations $\{\mathbf{x}_i^* : \quad i = 1, \ldots, n^*\}$ engendrent un espace de dimension supérieure ou égale au nombre de covariables d.

Le modèle $(\mathbf{M.3})$ On estime $\boldsymbol{\theta} = \beta_0^* \in \mathbb{R}$ solution de l'équation non linéaire à l'inconnue β_0^*

$$\frac{\partial l_{cond}(\beta_0^*)}{\partial \beta_0^*} = \sum_{i=1}^{n^*} (z_i^* - \pi(\mathbf{x}_i^*, \beta_0^*)) = 0. \tag{17}$$

Ce modèle, s'apparente à ce qui se pratique habituellement, *i.e.* ayant obtenu l'estimation d'un modèle logistique, on essaie de décaler le seuil d'affectation à la réponse positive $\{Z = 1\}$, ce qui revient à substituer à l'intercept β_0 la valeur β_0^*.
L'expression de la dérivée seconde est donnée par

$$\frac{\partial^2 l_{cond}(\beta_0^*)}{\partial {\beta_0^*}^2} = -\sum_{i=1}^{n^*} \pi(\mathbf{x}_i^*, \beta_0^*)(1 - \pi(\mathbf{x}_i^*, \beta_0^*)). \tag{18}$$

On vérifie donc, sans peine, que $l_{cond}(\beta_0^*)$ est strictement concave et que par conséquent l'estimation $\widehat{\beta}_0^*$ est unique.

Le modèle $(\mathbf{M.4})$ On estime $\boldsymbol{\theta} = (\beta_0^*, \lambda) \in \mathbb{R}^2$ par maximisation de l_{cond}. Cela revient à résoudre le système non linéaire

$$(S.4) : \left\{ \begin{array}{l} \frac{\partial l_{cond}(\beta_0^*,\lambda)}{\partial \beta_0^*} = \sum_{i=1}^{n^*} (z_i^* - \pi(\mathbf{x}_i^*, \beta_0^*, \lambda)) = 0, \\ \frac{\partial l_{cond}(\beta_0^*,\lambda)}{\partial \lambda} = \sum_{i=1}^{n^*} \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* (z_i^* - \pi(\mathbf{x}_i^*, \beta_0^*, \lambda)) = 0. \end{array} \right.$$

Soit $\mathbf{H}$ le hessien associé à la log-vraisemblance, en posant $H_{11} = \frac{\partial^2 l(\beta_0^*,\lambda)}{\partial {\beta_0}_*^2}$, $H_{22} = \frac{\partial^2 l(\beta_0^*,\lambda)}{\partial \lambda^2}$ et $H_{12} = \frac{\partial^2 l(\beta_0^*,\lambda)}{\partial \beta_0^* \partial \lambda}$, on établit que

$$H_{11} = -\sum_{i=1}^{n^*} \pi(\mathbf{x}_i^*, \beta_0^*, \lambda)(1 - \pi(\mathbf{x}_i^*, \beta_0^*, \lambda)), \tag{19}$$

$$H_{22} = -\sum_{i=1}^{n^*} (\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^*)^2 \pi(\mathbf{x}_i^*, \beta_0^*, \lambda)(1 - \pi(\mathbf{x}_i^*, \beta_0^*, \lambda)), \tag{20}$$

$$H_{12} = -\sum_{i=1} \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* \pi(\mathbf{x}_i^*, \beta_0^*, \lambda)(1 - \pi(\mathbf{x}_i^*, \beta_0^*, \lambda)). \tag{21}$$

On montre que le hessien $\mathbf{H}$ est semi défini négatif d'où la concavité de $l_{cond}(\beta_0^*, \lambda)$ et donc l'existence de solution du système (S. 4).
Considérons la condition (22) ci après :

$$\exists i, i' \in \{1, \ldots, n^*\} : \quad \boldsymbol{\beta}^{\mathrm{T}} (\mathbf{x}_{i'}^* - \mathbf{x}_i^*) \neq 0. \tag{22}$$

Sous cette condition la concavité est stricte et par conséquent la solution du système (S. 4) est unique. Là aussi, la non vérification de la condition d'unicité du maximum entraine, pour tout choix de (β_0^*, λ), l'affectation de toutes les observations à un même groupe d'appartenance.

Le modèle $(\mathbf{M.5})$ On estime $\boldsymbol{\theta} = \boldsymbol{\beta}^*$ en résolvant le système

$$(S.5) : \quad \frac{\partial l_{cond}(\Lambda)}{\partial \beta_j^*} = \sum_{i=1}^{n^*} \mathbf{x}_{i,j}^* \, (z_i^* - \pi(\mathbf{x}_i^*, \boldsymbol{\beta}^*)) = 0, \quad j = 1, \ldots, d.$$

Le hessien H est défini par son terme générique

$$H_{j,l} = -\sum_{i=1}^{n^*} \mathbf{x}_{i,j}^* \, \mathbf{x}_{i,l}^* \, \pi(\mathbf{x}_i^*, \boldsymbol{\beta}^*) \, (1 - \pi(\mathbf{x}_i^*, \boldsymbol{\beta}^*)) \quad j, l \in \{1, \ldots, d\} \tag{23}$$

La matrice H est semi définie négative (*i.e.* $l_{cond}(\boldsymbol{\theta})$ est concave) quelles que soient les données $\mathbf{x}_1^*, \ldots, \mathbf{x}_{n^*}^*$; ce qui garantit l'existence de l'estimation.
Convenons que $\dim$ désigne la dimension et $\mathcal{E}$ l'éspace engendré et considérons la condition (24) ci après :

$$\dim(\mathcal{E}(\mathbf{x}_1^*, \ldots, \mathbf{x}_{n^*}^*)) = d. \tag{24}$$

Sous cette condition, on a unicité de solution du système (S.5).

Le modèle $(\mathbf{M.6})$ Il s'agit simplement d'une estimation standard des paramètres logistiques $\boldsymbol{\theta} = (\beta_0^*, \boldsymbol{\beta}^*)$ à partir de S^*.

4.6 L'algorithmique

Rappelons tout d'abord que l'estimation des paramètres logistiques associés à Ω sont réalisés de façon classique. Du point de vue computationnel, l'estimation du paramètre $\boldsymbol{\theta}$ liant Ω à Ω^* peut s'obtenir pour l'ensemble des modèles étudiés par l'usage de l'algorithme de Newton-Raphson ou ses variantes. De façon équivalente, on peut procéder à une régression logistique sous contrainte, la contrainte dépendant du modèle considéré. Cette variante peut être intéressante pour l'utilisateur disposant d'un logiciel de régression logistique. Nous décrivons maintenant cette seconde option.
L'estimation se fait avec les estimations $\hat{\beta}_0$ et $\hat{\boldsymbol{\beta}}$ ainsi que le *deuxième échantillon d'apprentissage* $S^* = \{(\mathbf{x}_i^*, z_i^*) : \quad i = 1, \ldots, n^*\}$. La procédure logistique est mise en œuvre avec des options variables selon le modèle étudié. Considérons la matrice diagonale construite à partir de $\widehat{\boldsymbol{\beta}}$, *i.e.* $\text{diag}(\widehat{\boldsymbol{\beta}}) = \text{diag}(\widehat{\beta}_1, \ldots, \widehat{\beta}_d)$ et les observations transformées $\widetilde{\mathbf{x}}_i^* = \text{diag}(\widehat{\boldsymbol{\beta}})\mathbf{x}_i^*$ $i = 1, \ldots, n^*$. Notons aussi la variable somme (des covariables) $\mathbf{Y} = \mathbf{1}\widetilde{\mathbf{X}}$ et $\mathbf{y}_i^* = \mathbf{1}^{\mathrm{T}}\tilde{\mathbf{x}}_i^*$ $(i = 1, \ldots, n^*)$ ses réalisations sur le deuxième échantillon d'apprentissage. La notation $\mathbf{1}$ représente un vecteur colonne unitaire de $\mathbb{R}^d$.

Modèle $(\mathbf{M.2})$ Il s'agit de l'estimation du paramètre $\lambda \in \mathbb{R}$ intervenant dans le modèle

$$\text{logit}(\mathbb{P}(Z^* = 1|y^*)) = \widehat{\beta}_0 + \lambda y^*, \tag{25}$$

à partir des données $\{(y_i^*, z_i^*) : \quad i = 1, \ldots, n^*\}$. La procédure logistique appliquée ici, contraint l'intercept d'être égal à $\widehat{\beta}_0$.

Modèle $(\mathbf{M.3})$ On estime $\beta_0^* \in \mathbb{R}$ intervenant dans le modèle

$$\text{logit}(\mathbb{P}(Z^* = 1|y^*)) = \beta_0^* + y^*. \tag{26}$$

La procédure logistique appliquée ici contraint le coefficient associé à «la covariable » y^* d'être égal à l'unité.

Modèle $(\mathbf{M.4})$ On estime $(\beta_0^*, \lambda) \in \mathbb{R}^2$ intervenant dans le modèle

$$\text{logit}(\mathbb{P}(Z^* = 1|y^*)) = \beta_0^* + \lambda y^*. \tag{27}$$

Il s'agit là de mettre en oeuvre la procédure logistique sans poser de contraintes sur les paramètres du modèle.

Modèle $(\mathbf{M.5})$ On estime $\boldsymbol{\beta}^* \in \mathbb{R}^d$ intervenant dans le modèle

$$\text{logit}(\mathbb{P}(Z^* = 1|\widetilde{\mathbf{x}}^*)) = \widehat{\beta}_0 + \boldsymbol{\beta}^{*\mathrm{T}}\widetilde{\mathbf{x}}^*, \tag{28}$$

à partir de l'échantillon transformé $\{(\widetilde{\mathbf{x}}^*, z_i^*) : \quad i = 1, \ldots, n^*\}$. La procédure de régression logistique contraint l'intercept d'être égal à l'estimation (ou intercept) $\hat{\beta}_0$.

4.7 Applications et perspectives d'applications

La mise en oeuvre des modèles d'extension de la régression logistique étudiés, dans les domaines de l'assurance, la banque et le marketing constitue une perspective à court terme. On s'intéresse, notamment à l'actualisation de fonctions scores.
Ainsi le début d'application, présentée ici, se basera sur les données biologiques traités dans (Biernacki et *al.* 2002) par l'analyse discriminante généralisée. Le choix de ces données permet la comparaison des résultats obtenus à ceux obtenus et connus par ailleurs.

Les données considérées consistent en trois échantillons d'oiseaux de mer, provenant de trois sous espèces de l'espèce *Calanectris Diomedea* :

- Le premier est constitué d'oiseaux *Borealis* composé de 93 femelles $(SEX = 2)$ et 113 mâles $(SEX = 1)$;
- le second est constitué d'oiseaux *Diomedea* composé de 22 femelles et 16 mâles ;
- le troisième est constitué d'oiseaux de l'espèce *Edwards* composé de 44 mâles et 48 femelles.

Les trois sous espèces se distinguent par leur répartition géographique (Thibault et *al.* (1997))
Le problème est de prédire le sexe (Variable SEX) à partir de cinq variables consistant en des mesures biométriques :

- BECH et BECL : deux mesures relatives au bec ;
- TARSE : longueur du tarse ;
- AILE : envergure des ailes ;
- QUEUE : longueur de la queue.

Dans un premier temps, on s'intéresse à la prédiction du sexe d'oiseaux *Diomedea* en considérant l'échantillon des *Borealis*. L'echantillon S^* est donc constitué d'une partie d'oiseaux *Diomedea* (l'autre partie étant réservée aux tests) et l'échantillon S est constitué de l'ensemble des oiseaux *Borealis*.
Le tableau 1 récapitule les résultats des simulations consistant pour chaque taille n, à tirer au hasard 100 sous-échantillons *Diomedea* et à estimer les six modèles à partir de chacun de ces échantillons, combiné avec l'échantillon *Borealis*. Pour ces simulations, on restreint l'évaluation de la qualité d'un modèle à la seule utilisation du pourcentage de bien classés, calculé sur les échantillons tests.

Dans un deuxième temps, le même type de simulations est réalisé pour la prédiction du sexe d'individus de la sous espèce *Edwards* à partir d'oiseaux *Borealis* (échantillon S) et d'oiseaux *Edwards* (échantillon S^*). Les résultats sont donnés par le tableau 2.

Les résultats donnés dans les deux tableaux montrent la supériorité manifeste des modèles faisant intervenir les deux échantillons en présence.
Par ailleurs, les résultats obtenus ici et dans le cadre de la «régression logistique généralisée»confirment ceux obtenus sur les mêmes données dans le cadre de la discrimination gaussiènne généralisée (Biernacki et *al.* (2002)).

Modèle	M_1	M_2	M_3	M_4	M_5	M_6
$n = 10$	58.27	82.71	87.96	80.45	65.60	53.94
$n = 20$	54.59	86.20	85.82	84.28	71.61	52.67
$n = 30$	58.75	90.37	87.91	90.37	77.53	57.91

TAB. 1 – *Tableau donnant le pourcentage moyen d'individus Diomedea bien classés, pour les valeurs* $10, 20, 30$ *de la taille* n *de* S^* *et comme taille d'échantillon test* $(38 - n)$.

Modèle	M_1	M_2	M_3	M_4	M_5	M_6
$n = 10$	51.76	85.27	90.01	83.73	70.09	51.35
$n = 20$	51.73	86.85	86.93	86.75	78.57	49.70
$n = 30$	51.85	89.06	90.44	88.82	83.33	53.76
$n = 50$	52.01	88.91	89.84	89.01	85.07	52.47

TAB. 2 – *Tableau donnant le pourcentage moyen d'individus Edwards bien classés, pour les valeurs* $10, 20, 30, 50$ *de la taille* n *de* S^* *et comme taille d'échantillon test* $(92 - n)$.

5 Conclusion

Les premières expériences numériques montrent l'utilité de modèles étendant la régression logistique et permettant de considérer l'information de présence d'un mélange de populations. Des expériences mumériques plus ambitieuses sont envisagées, notamment l'étude de l'évolution des critères de choix de modèles (BIC, AIC, ...) en fonction des modèles et des tailles des échantillons S^* et S.
Ici, l'estimation des paramètres de transition $\boldsymbol{\theta}$ (*i.e.* (β_0^*, λ) ou $(\beta_0^*, \boldsymbol{\beta}^*)$, selon le modèle) se fait conditionnellement aux paramètres associés à la sous-population Ω (*i.e.* β_0 et β) ; on envisage comme perspective, l'estimation jointe des paramètres de Ω et ceux de transition.

Remerciements : Les auteurs remercient Madame Valérie Molina (attachée de l'INSEE auprés de l'ENSAI, Bruz) pour ses remarques et suggestions concernant l'optimisation du temps de simulation.

6 Références

Anderson, J.A. (1982). Logistic discrimination. In *Handbook of Statistics* (Vol. 2), P.R. Krishnaiah and L. Kanal (Eds.). Amsterdam : North-Holland, pp. 169–191.

Biernacki, C., Beninel, F. & Bretagnolle, V. (2002). *A Generalized Discriminante Rule when Training Population and Test Population Differ on their Descriptive Parameters*, Biometrics, **58**, 2, 387–397.

Bretagnolle, V., Genevois, F. & Mougeot, F. (1998). *Intra and Intersexual Function in the Call of a non Passerine Bird*, Behaviour, 135 : 1161-502.

De Meyer, B., Roynette, B., Vallois, P. & Yor, M. (2000). On independent times and positions for Brownian motion. Technical Report 1, Les prépublications de l'Institut Élie Cartan, Institut Élie Cartan, Vandœuvre lès Nancy, France.

Govaert, G. (2003). *Analyse des données*. Lavoisier serie "traitement du signal et de l'image", Paris, pp.362.

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

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

Van Franecker, J.A. & Ter Brack, C.J.F. (1993). A Generalized Discriminant for Sexing Fulmarine Petrels from External Measurements, *Auk*, 110 : 492-502.

Summary

Usually in discriminant analysis we are faced with the prediction of labels of individuals from a population given their descriptive parameters and using a unique learning sample. Individuals to predict and individuals of the learning sample are submitted to the same external conditions. The problem here, is to predict labels of individuals from a subpopulation using a learning sample from another one. In insurance and finance, this problem occurs in the prediction of the risk groups using characteristic parameters. Individuals to predict and the learning ones come respectively from two subpopulations of *member-cum-consumers* corresponding to geographical areas ... In this work we extend the idea used in generalized discriminant analysis (consisting of the use of the two learning samples to predict) to logistic discrimination, less restrictive concerning the type of descriptive variables. We propose different models of generalized logistic discrimination based on acceptable relations between the score function on the first population and the score function of the second one.