Discrimination par modèles additifs parcimonieux

Marta Avalos, Yves Grandvalet, Christophe Ambroise

Discrimination par modèles additifs parcimonieux

Marta Avalos, Yves Grandvalet, Christophe Ambroise

Laboratoire HeuDiaSyC, UMR CNRS 6599,
Université de Technologie de Compiègne
BP 20529 / 60205 Compiègne
{avalos, grandvalet, ambroise}@hds.utc.fr
http://www.hds.utc.fr

Résumé :
Nous proposons une méthode de discrimination non paramétrique conçue pour favoriser l'interprétabilité de la prédiction. D'une part, l'utilisation d'un modèle additif généralisé permet de représenter graphiquement l'effet de chaque variable d'entrée sur la variable de sortie. D'autre part, les paramètres de ce modèle sont estimés par vraisemblance pénalisée, où le terme de régularisation généralise la pénalisation l_1 aux fonctions splines. Cette pénalisation favorise les solutions parcimonieuses sélectionnant une partie de l'ensemble des variables d'entrée, tout en permettant une modélisation flexible de la dépendance sur les variables sélectionnées. Nous étudions l'adaptation de différents critères de sélection analytiques à ces modèles, et nous les évaluons sur deux jeux de données réelles.

Mots-clés : Sélection de variables, réglage de la complexité, régularisation l_1, apprentissage supervisé, interprétabilité, modèles additifs généralisés, splines.

1 Introduction

La précision de généralisation est la mesure habituelle de la performance en apprentissage statistique. Cependant, dans certaines applications, un prédicteur de type "boîte noire" ne sera pas accepté par l'utilisateur final. Par exemple, dans le diagnostic médical, une méthode estimant la probabilité d'un événement (décès, maladie) ne sera généralement pas considérée comme utile par le médecin, alors qu'une méthode qui rend intelligible ce résultat sera plus facilement utilisable. Les modèles interprétables peuvent également être utilisés comme outil d'analyse de données afin d'inspirer des nouvelles idées sur les relations entre les variables et améliorer, ainsi, la compréhension du domaine (Bratko, 1997).

Le modèle de régression logistique est un outil standard en discrimination lorsque la compréhension de l'effet de chaque variable d'entrée sur la variable de sortie est un aspect crucial. Le modèle logistique modélise les probabilités *a posteriori* de deux (ou plusieurs) classes par la transformation, dite logistique, d'une fonction linéaire des variables d'entrée. Cette transformation, connue également sous le nom de *softmax* dans

la littérature connexioniste (Bridle, 1990), assure que les probabilités *a posteriori* appartiennent à l'intervalle $[0, 1]$ et que leur somme vaut 1.

La simplicité du modèle logistique en fait, avec les arbres de décision, une des méthodes de discrimination les plus interprétables. Cependant, l'hypothèse de dépendance linéaire est souvent trop restrictive. Le modèle logistique additif est une généralisation permettant d'identifier et de décrire les effets non linéaires. Il remplace chaque composante linéaire par une fonction plus générale f_j (Hastie & Tibshirani, 1990). La forme non paramétrique des f_j accorde plus de flexibilité au modèle, alors que la structure additive préserve la possibilité de représenter l'effet de chaque variable.

Nous présentons ici un algorithme d'estimation des modèles additifs généralisés qui consiste à pénaliser de manière adaptative les fonctions f_j du modèle. Les bénéfices attendus sont de deux ordres : faciliter la procédure d'estimation en s'affranchissant du besoin de définir des complexités adaptées pour chaque fonction, faciliter l'interprétation du modèle final en favorisant des solutions parcimonieuses, ne sélectionnant qu'une partie des variables d'entrée.

Un rappel du modèle logistique linéaire et du modèle logistique additif est introduit dans la section 2. Nous présentons ensuite notre procédure d'estimation aboutissant à un modèle parcimonieux (section 3). L'algorithme de résolution de la régression logistique additive parcimonieuse est présenté dans la section 4. Des techniques de sélection des paramètres réglant la complexité sont appliquées à ce problème dans la section 5. Deux exemples réels permettent d'illustrer la méthode présentée et de comparer les techniques de sélection de la complexité dans la section 6. Enfin, les conclusions et perspectives sont discutés dans la section 7.

2 Modèles logistique linéaire et logistique additif

2.1 Modèle logistique linéaire

Considérons le vecteur aléatoire $(X, Y) = (X_1, ..., X_p, Y)$, où Y est une variable binaire (codée 0–1). Le modèle de régression logistique s'écrit

$$\text{logit}[P(Y = 1|X = x)] = \log \frac{P(Y = 1|X = x)}{1 - P(Y = 1|X = x)} = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_p x_p, \quad (1)$$

où $x = (x_1, \ldots, x_p)$. Pour un problème de décision où l'objectif est de minimiser le taux d'erreur de classement (coût {0,1}), la frontière de décision est alors définie par l'hyper–plan $\{x|\alpha_0 + \sum_{j=1}^{p} \alpha_j x_j = 0\}$, et la relation inverse donne la probabilité *a posteriori*,

$$P(Y = 1|X = x) = \frac{\exp(\alpha_0 + \sum_{j=1}^{p} \alpha_j x_j)}{1 + \exp(\alpha_0 + \sum_{j=1}^{p} \alpha_j x_j)}, \quad (2)$$

qu'on reconnaît comme la fonction softmax des réseaux de neurones.

Le modèle logistique fait partie de la famille des modèles linéaires généralisés, qui regroupe les modèles qui visent à exprimer l'espérance d'une variable de sortie en fonction d'une combinaison linéaire des variables d'entrée (Fahrmeir & Tutz, 2001). Les outils d'estimation, de diagnostic, et les propriétés théoriques de cette famille de modèles sont bien développés.

La classe des modèles linéaires généralisés est caractérisée par trois composantes. La composante aléatoire identifie la distribution de probabilités de la variable à expliquer (parmi les distributions de la famille exponentielle : gaussienne, gaussienne inverse, Gamma, Poisson, binomiale,...). La composante déterministe du modèle est le prédicteur linéaire : $\nu = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_p x_p$. La troisième composante exprime une relation fonctionnelle entre la composante aléatoire et le prédicteur linéaire : $\nu = g(\mathbb{E}[Y])$, où g est la fonction de lien, monotone et différentiable. Dans le cas concret du modèle logistique, la distribution de probabilité de Y est binomiale (i.e. la distribution d'une variable binaire quelconque) et le lien est la fonction logit.

2.2 Modèle logistique additif

Les modèles additifs généralisés sont une extension des modèles linéaires généralisés, permettant d'identifier et de décrire les effets non linéaires. Cette extension consiste à remplacer chaque composante linéaire du prédicteur linéaire par une fonction plus générale : $\nu = \alpha_0 + f_1(x_1) + \ldots + f_p(x_p)$ (Hastie & Tibshirani, 1990). Dans le cas concret de la logistique additive, le modèle s'écrit :

$$\log \frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} = \alpha_0 + f_1(x_1) + \ldots + f_p(x_p), \tag{3}$$

où les f_j sont des fonctions lisses. Les fonctions f_j étant monovariées, elles sont facilement représentables graphiquement, ce qui permet au modèle additif de garder l'interprétabilité du modèle linéaire. Par ailleurs, en restreignant les dépendances à une somme de fonctions monovariées, ce modèle évite le problème du "fléau de la dimensionnalité", ce qui permet, en particulier, d'avoir des estimateurs peu variables pour des échantillons de taille modérée.

2.3 Estimation

Considérons l'ensemble d'apprentissage $\mathcal{L} = \{(\mathrm{x}_{i1}, \ldots, \mathrm{x}_{ip}, \mathrm{y}_i)\}$, $i = 1, \ldots, n$, réalisations i.i.d. de (X, Y). Notons $\mathbf{X}$ la matrice des observations d'entrée et $\mathbf{y}$ le vecteur des sorties observées. Nous introduisons le problème d'estimation pour le modèle additif par celui du modèle linéaire, plus simple.

2.3.1 Modèle linéaire

L'estimation des paramètres $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)$ est calculée en maximisant la log–vraisemblance du modèle linéaire généralisé. La log–vraisemblance du modèle logistique s'écrit :

$$l(\alpha) = \textstyle\sum_{i=1}^{n} \mathrm{y}_i \log P_i + (1-\mathrm{y}_i)\log(1-P_i), \tag{4}$$

où $P_i = P(Y_i = 1|X = (\mathrm{x}_{i1}, \ldots, \mathrm{x}_{ip}))$. La méthode d'optimisation standard pour résoudre le problème de maximisation de la log–vraisemblance est le *Fisher scoring*, basé

sur l'algorithme de Newton–Raphson (Hastie & Tibshirani, 1990). La connaissance des dérivées de premier et deuxième ordre sont alors nécessaires :

$$\frac{\partial l(\alpha)}{\partial \alpha} = [\mathbf{1}\ \mathbf{X}]^t(\mathbf{y}-\mathbf{P}), \quad \frac{\partial^2 l(\alpha)}{\partial \alpha \partial \alpha^t} = -[\mathbf{1}\ \mathbf{X}]^t\mathbf{W}[\mathbf{1}\ \mathbf{X}], \tag{5}$$

où $\mathbf{P} = (P_1, \ldots, P_n)^t$, $\mathbf{W} = \mathrm{diag}\,[P_1(1-P_1), \ldots, P_n(1-P_n)]$, et $[\mathbf{1}\ \mathbf{X}]$ indique la matrice des données précédée d'une colonne de uns, afin d'incorporer l'estimation de α_0. Le problème d'optimisation étant convexe en α, la maximisation de (4) consiste à résoudre le système de $p+1$ équations $\dfrac{\partial l(\alpha)}{\partial \alpha} = \mathbf{0}$. Ces équations sont non linéaires en α et elles sont résolues itérativement jusqu'à obtention d'un point fixe. Cette mise à jour peut également s'écrire sous une forme légèrement différente :

$$\alpha^{[k+1]} = \alpha^{[k]} - \left(\frac{\partial^2 l(\alpha)}{\partial \alpha \partial \alpha^t}\bigg|_{\alpha^{[k]}}\right)^{-1} \frac{\partial l(\alpha)}{\partial \alpha}\bigg|_{\alpha^{[k]}} = ([\mathbf{1}\ \mathbf{X}]^t\mathbf{W}[\mathbf{1}\ \mathbf{X}])^{-1}[\mathbf{1}\ \mathbf{X}]^t\mathbf{W}\mathbf{z}, \tag{6}$$

où $\mathbf{z} = [\mathbf{1}\ \mathbf{X}]\alpha^{[k]} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{P})$, ($\mathbf{P}$, et donc $\mathbf{W}$ et $\mathbf{z}$, dépendent de α), et k indique l'itération en cours.

Sous cette forme, l'algorithme est appelé moindres carrés pondérés itératifs (IRLS), car à chaque itération, le problème résolu est équivalent à un problème de moindres carrés pondérés :

$$\alpha^{[k+1]} = \arg\min_{\alpha} \left|\left|\mathbf{W}^{1/2}\big(\mathbf{z} - [\mathbf{1}\ \mathbf{X}]\alpha\big)\right|\right|_2^2. \tag{7}$$

Cette analogie explique que $\mathbf{z}$ soit souvent dénommée "réponse de travail".

2.3.2 Modèle additif

Les premières étapes de résolution sont identiques pour les modèles additifs. L'estimation des paramètres α_j est généralisée à celle des fonctions f_j. Ces dernières sont souvent non paramétriques, de type noyau ou splines. Ici, les f_j sont des splines cubiques de lissage, définies comme la solution au problème de régularisation suivant : parmi les fonctions deux fois continûment dérivables, retenons celles minimisant la fonction coût (ici, la log–vraisemblance) (Wahba, 1990) :

$$\min_{f_j \in \mathcal{C}^2} -\sum_{i=1}^{n} \mathrm{y}_i \log P_i + (1-\mathrm{y}_i)\log(1-P_i) + \sum_{j=1}^{p} \lambda_j \int \big(f_j^{''}(t)\big)^2 dt, \tag{8}$$

où

$$P_i = \frac{\exp[\alpha_0 + f_1(\mathrm{x}_{i1}) + \ldots + f_p(\mathrm{x}_{ip})]}{1 + \exp[\alpha_0 + f_1(\mathrm{x}_{i1}) + \ldots + f_p(\mathrm{x}_{ip})]}. \tag{9}$$

Le premier terme de (8) mesure l'ajustement aux données, et le deuxième terme pénalise les solutions de courbure forte. Les paramètres de lissage λ_j déterminent le compromis entre les deux objectifs. Les fonctions obtenues $\widehat{f}_j$ sont des splines cubiques en x_j, avec des nœuds sur les x_{ij}. La contrainte $\sum_i \widehat{f}_{ij} = 0$, $j = 1, \ldots, p$ assure l'unicité de la solution.

L'algorithme IRLS permet de résoudre (8), et le problème à minimiser pour estimer les fonctions f_j devient un problème quadratique pondéré :

$$\min_{f_j \in \mathcal{C}^2} \left|\left| \mathbf{W}^{1/2}\big(\mathbf{z} - \alpha_0 - \sum_{j=1}^{p} f_j\big) \right|\right|_2^2 + \sum_{j=1}^{p} \lambda_j \int \big(f_j^{''}(t)\big)^2 dt, \tag{10}$$

où la réponse de travail $\mathbf{z}$ est maintenant définie par : $\mathbf{z} = \widehat{\mathbf{f}}^{[k]}(\mathbf{x}_1, \ldots, \mathbf{x}_p) + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{P})$, avec $\widehat{\mathbf{f}}^{[k]}(\mathbf{x}_1, \ldots, \mathbf{x}_p) = \widehat{\alpha}_0^{[k]} + \widehat{\mathbf{f}}_1^{[k]}(\mathbf{x}_1) + \ldots + \widehat{\mathbf{f}}_p^{[k]}(\mathbf{x}_p)$.

Le problème fonctionnel (8) a une formulation paramétrique équivalente, basée sur la décomposition des f_j sur une base des fonctions splines. Les bases de splines les plus utilisées sont la base naturelle des polynômes par morceaux (dont l'expression est simple), et la base naturelle B–spline (plus adéquate pour ce qui concerne leur implémentation numérique). Soient $\mathbf{N}_j$ la matrice de la base évaluée en x_{ij}, $\mathbf{\Omega}_j$ la matrice correspondant à la pénalisation de la dérivée seconde, et β_j, les coefficients de $\widehat{f}_j$ sur la base choisie. Le problème (8) s'écrit

$$\min_{\beta_1,\ldots,\beta_p} -\sum_{i=1}^{n} \mathrm{y}_i \log P_i + (1-\mathrm{y}_i)\log(1-P_i) + \sum_{j=1}^{p} \lambda_j \beta_j^t \mathbf{\Omega}_j \beta_j, \tag{11}$$

où

$$P_i = \frac{\exp[\alpha_0 + \mathbf{N}_1\beta_1 + \ldots + \mathbf{N}_p\beta_p]}{1 + \exp[\alpha_0 + \mathbf{N}_1\beta_1 + \ldots + \mathbf{N}_p\beta_p]}. \tag{12}$$

3 Modèle logistique additif parcimonieux

La résolution de (11)–(12) nécessite de définir au préalable p paramètres de lissage λ_j, qui vont régler la complexité des fonctions f_j. Ce pré–requis n'est pas réaliste pour $p \geq 3$, ce qui explique que les modèles additifs soient encore aujourd'hui peu utilisés quand le nombre de variables est important.

Le problème a été abordé par le biais de la pénalisation adaptative (Grandvalet & Canu, 1998), qui incorpore l'estimation des λ_j dans la procédure d'estimation des paramètres :

$$\min_{\beta_1,\ldots,\beta_p} -\sum_{i=1}^{n} \mathrm{y}_i \log P_i + (1-\mathrm{y}_i)\log(1-P_i) + \sum_{j=1}^{p} \lambda_j \beta_j^t \mathbf{\Omega}_j \beta_j, \tag{13}$$

$$\text{sous contraintes} \quad \sum_{j=1}^{p} \frac{1}{\lambda_j} = \frac{p}{\lambda}, \quad \lambda_j > 0, \tag{14}$$

où seul λ doit être défini avant la procédure d'estimation. Le problème (13)–(14), qui peut être motivé par une approche bayésienne hiérarchique, est également fortement lié à la pénalisation l_1 (Grandvalet & Canu, 1998). Les solutions de ce problème ont tendance à être parcimonieuses, en ce sens que, pour certaines variables, $\widehat{\beta}_j^t \mathbf{\Omega}_j \widehat{\beta}_j = 0$. Cependant ce critère ne sélectionne pas de variables, car la composante linéaire de $\widehat{f}_j$ est dans le noyau de $\mathbf{\Omega}_j$. Ainsi, même si $\widehat{\beta}_j^t \mathbf{\Omega}_j \widehat{\beta}_j = 0$, la j–ième variable n'est pas éliminée mais linéarisée.

3.1 Modification pour la sélection de variables

Les modèles additifs ajustés par des splines cubiques de lissage s'inscrivent dans le cadre théorique des espaces hilbertiens des fonctions L_2 (Hastie & Tibshirani, 1990). Cela permet d'appliquer des résultats généraux.

Soit $\mathcal{H}_j$ l'espace de Hilbert des fonctions mesurables centrées de variance finie, et de produit scalaire défini par $\langle f, g\rangle = \mathbb{E}_{X_j}(f(X_j)\cdot g(X_j))$. Considérons la base naturelle des polynômes par morceaux pour les splines cubiques (Hastie *et al.*, 2001) :

$$\mathbf{N}_j(\mathrm{x}) = \{N_{kj}(\mathrm{x})\}_{k=1}^{n} = \{1, \mathrm{x}, d_{1j}(\mathrm{x}) - d_{n-1,j}(\mathrm{x}), \ldots, d_{n-2,j}(\mathrm{x}) - d_{n-1,j}(\mathrm{x})\}$$

$$d_{ij}(\mathrm{x}) = \frac{(\mathrm{x} - \mathrm{x}_{ij})_+^3 - (\mathrm{x} - \mathrm{x}_{nj})_+^3}{\mathrm{x}_{ij} - \mathrm{x}_{nj}}. \tag{15}$$

Chaque sous–espace $\mathcal{H}_j$ admet une décomposition $\mathcal{H}_j^L \oplus \mathcal{H}_j^{NL}$, où L indique le sous–espace des composantes linéaires et NL indique le sous–espace des composantes non linéaires.

L'idée consiste donc à considérer les parties linéaire et non linéaire séparément. En ajoutant en (13) un terme de pénalisation agissant sur la composante linéaire, il est possible de supprimer l'influence de certaines variables sur le modèle. Le problème d'optimisation s'écrit :

$$\min_{\alpha,\widetilde{\beta}_1,\ldots,\widetilde{\beta}_p} -\sum_{i=1}^{n} \mathrm{y}_i \log P_i + (1-\mathrm{y}_i)\log(1-P_i) + \sum_{j=1}^{p} \mu_j \alpha_j^2 + \sum_{j=1}^{p} \lambda_j \widetilde{\beta}_j^t \mathbf{\Omega}_j \widetilde{\beta}_j, \tag{16}$$

sous contraintes

$$\sum_{j=1}^{p} \frac{1}{\mu_j} = \frac{p}{\mu}, \quad \mu_j > 0, \quad \sum_{j=1}^{p} \frac{1}{\lambda_j} = \frac{p}{\lambda}, \quad \lambda_j > 0, \quad \mathbf{x}_j^t \mathbf{N}_j \widetilde{\beta}_j = 0, \quad \mathbf{1}^t \mathbf{N}_j \widetilde{\beta}_j = 0, \tag{17}$$

où μ et λ sont les paramètres qui règlent la complexité du modèle, $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)^t$, $\widetilde{\beta}_j$ sont les coefficients de la partie linéaire et non linéaire, respectivement, et

$$P_i = \frac{\exp[\alpha_0 + \sum_{j=1}^{p} \mathrm{x}_{ij}\alpha_j - \sum_{j=1}^{p} \mathbf{N}_j \widetilde{\beta}_j]}{1 + \exp[\alpha_0 + \sum_{j=1}^{p} \mathrm{x}_{ij}\alpha_j - \sum_{j=1}^{p} \mathbf{N}_j \widetilde{\beta}_j]}. \tag{18}$$

La parcimonie du problème (17)–(16) découle de l'équivalence entre moindres carrés adaptatifs et pénalisation l_1 (Grandvalet & Canu, 1998). Le problème peut également être résolu par l'algorithme IRLS :

$$\min_{\alpha,\widetilde{\beta}_1,\ldots,\widetilde{\beta}_p} \left|\left| \mathbf{W}^{1/2}\big(\mathbf{z} - [\mathbf{1}\ \mathbf{X}]\alpha - \sum_{j=1}^{p} \mathbf{N}_j \widetilde{\beta}_j\big) \right|\right|_2^2 + \sum_{j=1}^{p} \mu_j \alpha_j^2 + \sum_{j=1}^{p} \lambda_j \widetilde{\beta}_j^t \mathbf{\Omega}_j \widetilde{\beta}_j, \tag{19}$$

Algorithme IRLS :

1. Fixer μ et λ et initialiser $\alpha_0, \alpha_j, \widetilde{\beta}_j, \mu_j, \lambda_j,\ j = 1, \ldots, p$.
2. Calculer $\mathbf{N}_j, \mathbf{\Omega}_j, \mathbf{f}_j = \mathbf{x}_j\alpha_j + \widetilde{\mathbf{f}}_j$, $\widetilde{\mathbf{f}}_j = \mathbf{N}_j\widetilde{\beta}_j$,
   $f_{ij} = \mathrm{x}_{ij}\alpha_j + \widetilde{f}_j(\mathrm{x}_{ij})$, $j = 1, \ldots, p$, $i = 1, \ldots, n$.
3. Calculer :
   (a) $\mathbf{P} = (P_1, \ldots, P_n)^t$, où $P_i = \dfrac{\exp(\alpha_0 + \sum_{j=1}^p f_{ij})}{1 + \exp(\alpha_0 + \sum_{j=1}^p f_{ij})}$.
   (b) $\mathbf{W} = \mathrm{diag}\,[P_1(1-P_1), \ldots, P_n(1-P_n)]$.
   (c) $\mathbf{z} = \alpha_0 + \sum_{j=1}^p \mathbf{f}_j + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{P})$.
4. Optimiser
   $$\min_{\alpha, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_p} \left|\left| \mathbf{W}^{1/2}\left(\mathbf{z} - [\mathbf{1}\ \mathbf{X}]\alpha - \sum_{j=1}^p \mathbf{N}_j\widetilde{\beta}_j\right)\right|\right|_2^2 + \sum_{j=1}^p \mu_j\alpha_j^2 + \sum_{j=1}^p \lambda_j \left(\widetilde{\beta}_j\right)^t \mathbf{\Omega}_j \left(\widetilde{\beta}_j\right)$$
   sous contraintes $\sum_{j=1}^p \dfrac{1}{\mu_j} = \dfrac{p}{\mu},\ \mu_j > 0,\ \sum_{j=1}^p \dfrac{1}{\lambda_j} = \dfrac{p}{\lambda},\ \lambda_j > 0, , \mathbf{x}_j^t\mathbf{N}_j\widetilde{\beta}_j = 0,\ \mathbf{1}^t\mathbf{N}_j\widetilde{\beta}_j = 0.$
5. Itérer 3. et 4. jusqu'à convergence.

FIG. 1 – Algorithme d'estimation du modèle logistique additif parcimonieux.

sous contraintes

$$\sum_{j=1}^p \frac{1}{\mu_j} = \frac{p}{\mu}, \quad \mu_j > 0, \quad \sum_{j=1}^p \frac{1}{\lambda_j} = \frac{p}{\lambda}, \quad \lambda_j > 0, \quad \mathbf{x}_j^t\mathbf{N}_j\widetilde{\beta}_j = 0, \quad \mathbf{1}^t\mathbf{N}_j\widetilde{\beta}_j = 0. \tag{20}$$

3.2 Relation avec d'autres méthodes de pénalisation l_1

La sélection de variables par pénalisation l_1 est utilisée en régression linéaire généralisée, où elle est connue sous le nom de *lasso* (*least absolute shrinkage and selection operator*) (Tibshirani, 1996), en régression par ondelettes (Chen *et al.*, 1995), et en régression par noyaux (Roth, 2001; Gunn & Kandola, 2002).

La pénalisation au sens de la norme l_1 a été aussi appliqué à la régression additive ajustée des splines cubiques de lissage (Grandvalet & Canu, 1998; Bakin, 1999). Cependant, comme précisé précédemment, les variables sélectionnées ne sont pas éliminées, elles sont linéarisées. En régression, le problème a été résolu en ajoutant un terme de pénalisation agissant sur la composante linéaire (Avalos *et al.*, 2003). Nous développons ici cette idée dans le contexte de la discrimination.

En utilisant le fait que les splines cubiques de lissage constituent un espace de Hilbert à noyau auto–reproduisant (RKHS), Zhang *et al.* (2001) pénalisent les composantes linéaires et non linéaires, en termes des noyaux, séparément. Une approche similaire, dans le cas de splines de lissage plus générales, est due à (Lin & Zhang, 2002), où un seul terme pénalise les composantes linéaires et non linéaires. Dans ces deux derniers travaux, les fonctions f_j sont ajustées par peu de termes mais ceci n'encourage pas la sélection de variables.

Algorithme du point fixe :

1. Composantes linéaires :
   (a) Estimation des coefficients : $\alpha = \left([\mathbf{1}\ \mathbf{X}]^t \mathbf{W} [\mathbf{1}\ \mathbf{X}] + \mathbf{M}\right)^{-1} [\mathbf{1}\ \mathbf{X}]^t \mathbf{W}\mathbf{z}$, où $\mathbf{M} = \mathrm{diag}[\mu_j]$.
   (b) Ré–estimation des termes de pénalisation : $\mu_j = \mu \dfrac{\sum_{k=1}^{p} |\alpha_k|}{p|\alpha_j|}$.
   (c) Itérer 2.(a) et 2.(b) jusqu'à convergence.
2. Composantes non linéaires :
   (a) Matrices de lissage :
   $\mathbf{G}_j = [\mathbf{1}\ \mathbf{x}_j]([\mathbf{1}\ \mathbf{x}_j]^t \mathbf{W} [\mathbf{1}\ \mathbf{x}_j]^t)^{-1} [\mathbf{1}\ \mathbf{x}_j]^t \mathbf{W}$, $\quad \mathbf{G} = (\mathbf{G}_1, \ldots, \mathbf{G}_p)$,
   $\mathbf{S}_j = \mathbf{N}_j (\mathbf{N}_j^t \mathbf{W} \mathbf{N}_j^t + \lambda_j \mathbf{\Omega}_j)^{-1} \mathbf{N}_j^t \mathbf{W}$, $\quad \widetilde{\mathbf{S}}_j = \mathbf{S}_j - \mathbf{G}_j$.
   (b) Estimation des coefficients :
      i. Backfitting : $\widetilde{\mathbf{r}}_j = \mathbf{z} - \mathbf{G}\mathbf{z} - \sum_{k \neq j} \mathbf{N}_j \widetilde{\beta}_j$, $\ \mathbf{N}_j \widetilde{\beta}_j = \widetilde{\mathbf{S}}_j \widetilde{\mathbf{r}}_j$, $\ j = 1, \ldots, p$.
      ii. Calcul de $\widetilde{\beta}_j$ à partir des dernières estimations.
   (c) Ré–estimation des termes de pénalisation : $\lambda_j = \lambda \dfrac{\sum_{j=1}^{p} \sqrt{\widetilde{\beta}_j \mathbf{\Omega}_j \widetilde{\beta}_j}}{p \sqrt{\widetilde{\beta}_j \mathbf{\Omega}_j \widetilde{\beta}_j^t}}$.
   (d) Itérer 3.(b) et 3.(c) jusqu'à convergence.
3. Itérer 1. et 2. jusqu'à convergence.

FIG. 2 – Algorithme d'estimation du problème de minimisation quadratique pondéré sous contraintes.

4 Algorithme

Nous présentons ici notre procédure d'estimation, basée sur les moindres carrés pondérés itératifs (IRLS). La procédure IRLS (figure 1) résout le problème de minimisation quadratique pondéré (19)–(20). L'étape 4, détaillée sur la figure (2), est résolue par l'algorithme décrit par (Avalos *et al.*, 2003). Celui–ci consiste à englober l'algorithme de backfitting dans un algorithme de point fixe. Cependant, le problème de minimisation quadratique de l'algorithme itératif IRLS présente de nouvelles difficultés : les estimations des coefficients linéaires et non linéaires ne sont plus indépendantes car elles interagissent par le biais de la matrice de pondérations $\mathbf{W}$. La quantité de calculs s'en trouve augmentée.

L'algorithme de point fixe (figure 2) résout le problème d'estimation des paramètres de pénalisation (étapes 1(b) et 2(c)) (Grandvalet & Canu, 1998). Un algorithme efficace pour trouver les solutions du lasso a été proposé par (Osborne *et al.*, 2000). Celui–ci pourrait être appliqué à l'estimation des composantes linéaires (étapes 1(a)–1(c)).

L'algorithme de backfitting (étape 2(b)), est le plus utilisé pour l'ajustement des modèles additifs (Hastie & Tibshirani, 1990). Il consiste à estimer les fonctions f_j de manière itérative sur les résidus partiels $\mathrm{z}_i - \sum_{k \neq j} f_k(\mathrm{x}_{ik})$. Comme les splines sont une méthode de lissage linéaire, l'estimation de f_j peut s'écrire $\widehat{\mathbf{f}}_j = \mathbf{S}_j \mathbf{z}$, où $\mathbf{S}_j$, matrice de lissage, est indépendante de $\mathbf{z}$. Les contraintes d'orthogonalité (20) sont respectées en décomposant la matrice de lissage des splines cubiques comme suit : $\mathbf{S}_j = \mathbf{G}_j + \widetilde{\mathbf{S}}_j$, où $\mathbf{G}_j$ est la matrice de projection sur l'espace des fonctions propres constante et linéaires et $\widetilde{\mathbf{S}}_j$ est la matrice de rétrécissement, correspondant à l'espace des fonctions propres d'ordre supérieur.

5 Réglage de la complexité

La sélection d'un modèle de complexité adaptée est une étape clé pour les modèles additifs, comme pour les autres modèles d'apprentissage statistique. Elle consiste à choisir, parmi une famille de modèles, celui qui minimise une estimation de l'erreur de généralisation. Cette étape est difficile à mettre en œuvre pour les modèles additifs. En effet, comme ces modèles sont composés d'autant de fonctions que de variables, l'espace de recherche de la complexité est de dimension p. Notre modèle ne présente quant à lui que deux paramètres de réglage de la complexité, ce qui facilite considérablement sa mise en œuvre dès que le nombre de variables est supérieur à deux. Cette simplification est la conséquence des contraintes sur μ_j et λ_j (20), qui permettent à chacune d'explorer un espace de dimension $p-1$ dans la procédure d'estimation des paramètres. Il ne reste plus qu'à fixer les valeurs de μ et λ par estimation de l'erreur de généralisation.

5.1 Estimation du nombre effectif de paramètres

Dans les statistiques paramétriques, la complexité d'un modèle ajusté par moindres carrés est mesurée par la nombre de paramètres. Cette mesure n'est plus satisfaisante quand le critère d'ajustement est modifié. Par exemple, il ne tient pas compte des contraintes sur les paramètres quand ces derniers sont pénalisés.

La notion de nombre effectif de paramètres (ou nombre de degrés de liberté) généralise la mesure de complexité à l'ensemble des prédicteurs linéaires. Elle est moins générale que la dimension de Vapnik–Chervonenkis (Vapnik, 1995), en revanche, elle est facilement calculable. Les prédicteurs linéaires s'écrivent sous la forme $\widehat{f(\mathrm{x})} = \sum_i^n K(\mathrm{x}, \mathrm{x}_i)\mathrm{y}_i$. Ils sont donc linéaires vis à vis du vecteur d'observations, et donc, de forme relativement générale, incluant par exemple, les régresseurs à noyaux et les discriminateurs des plus proches voisins.

Le nombre effectif de paramètres s'exprime simplement comme $df = \sum_i K(\mathrm{x}_i, \mathrm{x}_i)$, (Hastie & Tibshirani, 1990). Cette somme correspond exactement au nombre de paramètres pour les modèles ajustés par moindres carrés. Chacun des éléments $K(\mathrm{x}_i, \mathrm{x}_i)$ mesure la contribution de y_i dans le calcul de $\widehat{f}(\mathrm{x}_i)$. Pour un apprentissage par coeur $K(\mathrm{x}_i, \mathrm{x}_i) = 1$, et le nombre effectif de paramètres est égal à la taille de l'échantillon.

5.1.1 Nombre effectif de paramètres associé aux composantes linéaires

La régression par moindres carrés pénalisés s'écrit $\widehat{\mathbf{f}} = \mathbf{H}_\mu \mathbf{y} = \mathbf{X}\widehat{\alpha} = \mathbf{X}(\mathbf{X}^t\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^t\mathbf{y}$, ce qui permet de calculer simplement le nombre degrés de liberté $df = \mathrm{tr}\,[\mathbf{H}_\mu]$. Une reformalisation des solutions du lasso ($|\widehat{\alpha}_j| = \widehat{\alpha}_j^2/|\widehat{\alpha}_j|$) permet une formulation de type pénalisation quadratique et l'estimation de df suivante (Tibshirani, 1996) :

$$df(\mu) = \mathrm{tr}\left[\mathbf{X}\left(\mathbf{X}^t\mathbf{X} + \mu\mathbf{A}^-\right)^{-1}\mathbf{X}^t\right], \tag{21}$$

où $\mathbf{A} = \mathrm{diag}(|\widehat{\alpha}_j|)$, et $\mathbf{A}^-$ indique la pseudo–inverse. Dans cette estimation, la pénalisation sur les variables jugées non pertinentes n'est pas prise en compte. Une modifica-

tion est donc proposée (Fu, 1998) :

$$df(\mu) = \mathrm{tr}\left[\mathbf{X}\left(\mathbf{X}^t\mathbf{X} + \mu\mathbf{A}^-\right)^{-1}\mathbf{X}^t\right] - p_0, \tag{22}$$

où $\mathbf{A} = \mathrm{diag}(2|\widehat{\alpha}_j|)$, et p_0 est le nombre de coefficients estimés nuls. Cependant, quand $\mathbf{X}$ n'est pas orthogonale, le nombre de degrés de liberté lié aux coefficients nuls ne coïncide pas avec le nombre de coefficients annulés. Notre définition ne considère que les colonnes de $\mathbf{X}$ et $\mathbf{A}$ pour lesquelles les coefficients $\widehat{\alpha}_j$ sont non nuls ($\bar{\mathbf{X}}$ et $\bar{\mathbf{A}}$, respectivement) :

$$df(\mu) = \mathrm{tr}\left[\bar{\mathbf{X}}\left(\bar{\mathbf{X}}^t\bar{\mathbf{X}} + \mu\bar{\mathbf{A}}^{-1}\right)^{-1}\bar{\mathbf{X}}^t\right]. \tag{23}$$

Cette définition induit une prédiction plus conservatrice. En effet, on peut montrer que l'expression (23) est plus grande que celle de Fu (22). Elle est également plus avantageuse numériquement, car la dimension des matrices en (22) est plus élevée.

En abordant le problème sous la forme de la pénalisation adaptative, et tenant compte des pondérations appliquées dans l'algorithme IRLS, nous proposons l'estimation suivante :

$$df(\mu) = \mathrm{tr}\left[\bar{\mathbf{X}}\left(\bar{\mathbf{X}}^t\mathbf{W}\bar{\mathbf{X}} + \bar{\mathbf{M}}\right)^{-1}\bar{\mathbf{X}}^t\mathbf{W}\right], \tag{24}$$

où $\bar{\mathbf{M}}$ comprend les colonnes de $\mathbf{M} = \mathrm{diag}(\mu_j)$ pour lesquelles les coefficients $\widehat{\alpha}_j$ sont non nuls. Les définitions (21)–(23) supposent la linéarité du modèle, ce qui implique que les matrices $\mathbf{A}$ de (21) et (22) ou la matrice $\bar{\mathbf{A}}$ de (23) sont supposées ne pas être affectées par les observations y_i. Cette hypothèse simplificatrice n'étant pas respectée, les définitions (21)–(23) ne proposent qu'une borne inférieure du nombre effectif de paramètres. Notre définition étant la plus conservatrice, elle est la plus proche de la réalité.

5.1.2 Nombre effectif de paramètres associé aux composantes non linéaires

Par analogie au modèle linéaire, le nombre de degrés de liberté est défini, pour les méthodes de lissage linéaires unidimensionnelles, comme la trace de la matrice de lissage (Hastie & Tibshirani, 1990). Dans le cas multidimensionnel additif, cela correspond à la trace de la matrice $\mathbf{R}$ qui génère la prédiction, $\widehat{\mathbf{f}} = \mathbf{R}\mathbf{y}$, où $\widehat{\mathbf{f}} = \widehat{\mathbf{f}}_1 + \ldots + \widehat{\mathbf{f}}_p$. Cette matrice correspond à la dernière itération de l'algorithme backfitting et le calcul de sa trace peut s'avérer difficile.

La somme des traces des matrices de lissage individuelles ne correspond pas exactement à la trace de la matrice $\mathbf{R}$, mais elle en est une borne supérieure (Buja *et al.*, 1989; Hastie & Tibshirani, 1990). Afin de ne pas tenir compte des valeurs propres correspondantes aux fonctions constante et linéaires, l'estimation est effectuée sur les matrices de rétrécissement $\widetilde{\mathbf{S}}_j$ (définies dans la figure 2). Le nombre effectif de paramètres associé aux composantes non linéaires est ainsi estimé par :

$$df(\lambda) = \textstyle\sum_j^p \mathrm{tr}\left[\widetilde{\mathbf{S}}_j(\lambda_j)\right] = \sum_j^p \mathrm{tr}\left[\mathbf{S}_j(\lambda_j) - \mathbf{G}_j\right]. \tag{25}$$

Comme dans le cas précédant, le coût de l'estimation des termes de pénalisation individuels n'est pas intégré dans l'estimation de df. Le nombre de degrés de liberté $df(\mu, \lambda)$ est égal à $df(\mu) + df(\lambda)$.

5.2 Critères de sélection de modèle

Plusieurs critères analytiques, basés sur le nombre effectif de paramètres, peuvent être utilisés pour la sélection de modèle. Nous avons utilisé ici le critère d'information d'Akaike (AIC), basé sur l'espérance de l'information de Kullback–Leibler, le critère d'information d'Akaike corrigé (AICc), qui corrige le biais du premier dans les problèmes de petite taille et le critère d'information Bayésien (BIC), dont l'argumentation est de type bayésienne :

$$\begin{aligned}
\mathrm{AIC}(\mu,\lambda) &= -2l(\mu,\lambda) + 2df(\mu,\lambda),\\
\mathrm{AICc}(\mu,\lambda) &= -2l(\mu,\lambda) + \frac{2ndf(\mu,\lambda)}{n - df(\mu,\lambda) - 1},\\
\mathrm{BIC}(\mu,\lambda) &= -2l(\mu,\lambda) + \log(n)df(\mu,\lambda).
\end{aligned} \tag{26}$$

La validation croisée peut également être utilisée pour le choix du paramètre de lissage. Cette méthode de rééchantillonnage demande cependant de nombreux calculs. La validation croisée généralisée (GCV) est une approximation analytique de la validation croisée (Wahba, 1990). Cette méthode, initialement développée pour la sélection du paramètre de lissage des splines cubiques pour la fonction coût quadratique, a été adaptée a des nombreuses méthodes telles que les machines à vecteurs de support (Lin *et al.*, 2000) ou les modèles additifs généralisés (Gu, 1992). Pour ces derniers, elle est calculée de manière itérative sur les réponses de travail de l'algorithme IRLS :

$$\mathrm{GCV}(\mu,\lambda) = \frac{\left|\left|\mathbf{W}^{1/2}\left(\mathbf{z} - \alpha_0 - \sum_{j=1}^{p}\widehat{\mathbf{f}}_j(\mathbf{x}_j)\right)\right|\right|_2^2}{n(1 - df(\mu,\lambda)/n)^2}. \tag{27}$$

6 Application à des données réelles

6.1 Difformité vertébrale post–opératoire

Dans un premier temps, nous reprenons l'étude de cas *kyphosis* utilisée par (Hastie & Tibshirani, 1990; Tibshirani, 1996) pour illustrer les différences entre la régression logistique par modèle additif standard, et la régression logistique pénalisée. La variable réponse indique la présence ou l'absence de difformité vertébrale post–opératoire (*kyphosis*) chez des enfants. Il y a 83 exemples, dont 18 étiquetés *kyphosis*. Les variables d'entrée sont l'âge des enfants (X_1), le nombre de vertèbres touchées par l'opération (X_2) et la position de la première vertèbre concernée (X_3). Les données sont centrées réduites, afin de rendre les pénalisations comparables. Les critères de sélection sont évalués sur une grille 6×6 de valeurs de (μ, λ) à échelle logarithmique.

La figure (3) montre les coefficients linéaires en fonction du paramètre de pénalisation linéaire, à gauche, et la norme des coefficients non linéaires en fonction du paramètre de pénalisation non linéaire, à droite. Les lignes verticales pointillées indiquent la valeur sélectionnée par les critères AIC, AICc, GCV et BIC. Pour la partie linéaire, BIC (ligne pointillée peu dense) a choisi un modèle plus simple que les autres critères (ligne

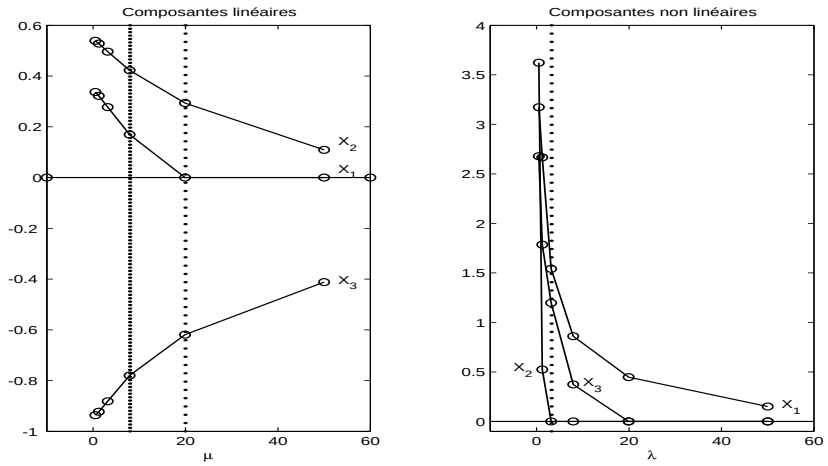

FIG. 3 – Coefficients des composantes linéaires, α_j, et norme des coefficients des composantes non linéaires, $(\widetilde{\beta}_j^t \mathbf{\Omega}_j \widetilde{\beta}_j)^{1/2}$, en fonction des paramètres de la complexité correspondants. Le graphique de gauche correspond à $\lambda = 1.2$, et celui de droite à $\mu = 4.2$, mais l'allure des courbes est similaire pour tous les μ et λ. Les lignes verticales indiquent la sélection effectuée par les critères.

pointillée dense). Pour la partie non linéaire, les quatre méthodes ont effectué le même choix (ligne pointillée).

La figure (4) montre l'effet de chaque variable sur la fonction logit, estimé par 4 modèles : logistique additive avec 3 degrés de liberté pour chaque composante, M1 ; logistique linéaire pénalisée (lasso), M2 ; et logistique additive pénalisée, pour les paramètres de la complexité sélectionnés par les différents critères, M3 (AIC, AICc et GCV) et M4 (BIC).

Etant donnée la difficulté de sélectionner les paramètres de lissage quand plusieurs variables sont considérées dans le modèle additif, celui–ci est souvent appliqué comme un outil d'analyse exploratoire. Ainsi, le modèle M1 (Hastie & Tibshirani, 1990) suggère des termes quadratiques, lesquels sont intégrés dans le modèle paramétrique M2 (Tibshirani, 1996). La répartition automatique de la complexité est alors possible, aboutissant à un modèle linéaire en X_2 et X_3 et quadratique en X_1. La méthode que nous proposons permet de distribuer la complexité de chaque variable de façon automatique, sans l'intermédiaire d'une approximation paramétrique.

Les courbes M3 et M4 sont similaires à la courbe M2 pour les variables X_1 et X_2, et plus complexes pour la troisième variable. En observant l'estimation obtenue par le modèle additif M1, on s'attend à que le modèle paramétrique conserve le terme quadratique. En fait, cet exemple montre que le lasso appliqué sur le modèle paramétrique peut produire des résultats contre–intuitifs. Ici, dans la base $\{1, x, x^2\}$ une composante en $\alpha(x+1)^2$ est jugée plus "complexe" qu'une composante en αx^2. En effet, la première se développant en $\alpha(x^2+2x+1)$, elle est pénalisée par 3α, alors que la première

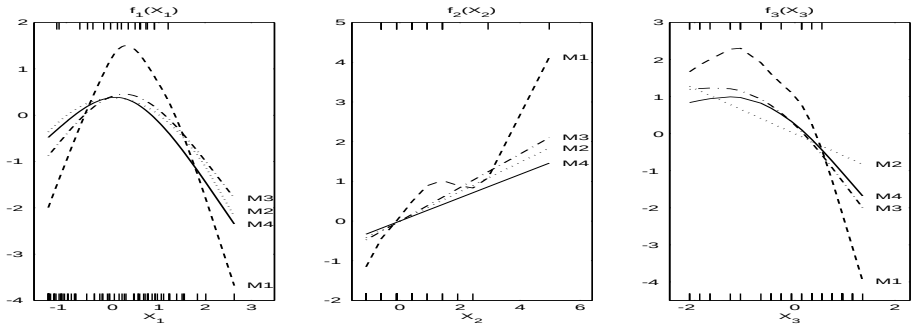

FIG. 4 – Composantes additives ajustées par : le modèle logistique additif (M1, ligne discontinue) ; le modèle logistique lasso (M2, ligne pointillée) ; le modèle logistique parcimonieux sélectionné par AIC, AICc et GCV (M3, ligne point-tirets) ; le modèle logistique parcimonieux sélectionné par BIC (M4, ligne continue). Les bâtons en haut et en bas des graphiques indiquent les observations en présence et absence de kyphosis, respectivement.

est pénalisée par α. Dans la base $\{1, x, (x+1)^2\}$ le phénomène inverse serait observé. Notre algorithme, pour lequel la pénalisation de la partie non linéaire n'est affectée que par la courbure de la fonction, est insensible à ce problème de représentation.

La représentation graphique permet de déduire que : 1) le risque de kyphosis (le odds ratio, plus précisément) augmente jusqu'à l'âge moyen (7 ans), et ensuite il décroît ; 2) le risque augmente quand le nombre de vertèbres touchées augmente (l'augmentation du risque est de 4.5% par vertèbre, selon M3, et de 3%, selon M4) ; et 3) le risque, élevé, stagne jusqu'à une certaine position (vertèbre numéro 10), et décroît ensuite rapidement.

6.2 Risque cardio–vasculaire

Le projet INDANA (Individual Data Analysis of Antihypertensive Intervention Trials) s'inscrit dans le cadre de la prédiction individualisée du risque cardio–vasculaire chez des patients présentant une hypertension artérielle, en vue d'aider la décision des médecins praticiens dans le domaine de la prévention cardio–vasculaire (Gueyffier *et al.*, 1995). La base de données INDANA réunit les données individuelles de 10 essais thérapeutiques (contrôlés randomisés) conduits pour évaluer l'efficacité des traitements anti–hypertenseurs. Cette base de données a été mise en forme et est maintenue dans l'Unité de Pharmacologie Clinique de L'Université de Lyon 1 (chef de projet : F. Gueyffier).

Le modèle logistique additif a été préalablement utilisé sur un des essais de la base INDANA (Shep). La sensibilité (proportion des décédés bien classés) et la spécificité (proportion des non décédés bien classés) sur une validation croisée à 10 blocs stratifiée ont mesuré la performance de la méthode. Les valeurs de sensibilité et de spécificité obtenues par la régression additive sont de 66.36% et 66.37%, respectivement.

TAB. 1 – Valeurs de (μ, λ) choisies par les techniques de sélection de la complexité, ainsi que leur erreur, sensibilité et spécificité sur l'ensemble de test.

	AIC	AICc	BIC	GCV	CV	CV_v
(μ, λ)	$(10, 10)$	$(10, 10)$	$(100, 100)$	$(10, 0.01)$	$(0.01, 0.1)$	$(1, 10)$
Erreur	0.71	0.71	0.78	0.95	0.63	0.69
Sensibilité	66.7%	66.7%	86.7%	40.0%	66.7%	60.0%
Spécificité	61.3%	61.3%	28.2%	66.1%	69.4%	71.0%

Sur ces données, le modèle logistique additif a donné les meilleurs résultats par rapport aux autres méthodes testées, à savoir *balanced–bagging*, *C4.5*, *Fôret–Floue–T–norme*, *Fôret–Floue–stricte*, *Framingham*, *GloBoost*, *Pocock*[1].

La régression logistique additive parcimonieuse est appliquée ici au groupe de contrôle d'un des essais (coope). Les données extraites sont constituées de 10 variables d'entrée : âge, pression systolique, pression diastolique, cholestérol, uricémie, indice pondéral, taille, sexe, tabagisme et facteur de risque (antécédent d'angor, d'infarctus myocardiaque, d'accident cardio–vasculaire, ou hypertrophie ventriculaire). Les trois dernières variables sont binaires et elles sont modélisées et pénalisées linéairement. La variable de sortie est le décès cardio–vasculaire. Il y a 413 exemples, dont 43 décès.

Deux tiers de la base coope (274 exemples, dont 28 décès) constitue l'ensemble d'apprentissage, et un tiers (139 exemples, dont 15 décès) constitue l'ensemble de test. Le critère de comparaison est la proportion d'exemples mal classés. Cette erreur est calculée en appliquant le coût $\{1, 10\}$, afin d'équilibrer les données. La sensibilité et spécificité de chaque méthode sont également calculées.

Nous testons les méthodes AIC, AICc, BIC, GCV, ainsi que validation croisée stratifiée sur 10 sous–ensembles. Pour cette dernière, nous rapportons les résultats obtenus selon deux critères : l'erreur de classement (avec les coûts $\{1, 10\}$), CV, d'une part, et la vraisemblance, CV_v, d'autre part.

Les méthodes analytiques et de rééchantillonnage sont évaluées sur une grille 5×5 de valeurs de (μ, λ), régulièrement espacées sur une échelle logarithmique. Les résultats sont présentés dans le tableau 1.

Les deux versions de la validation croisée réalisent la meilleure performance. Celle de CV est proche de la performance de test optimale obtenue, 0.62 pour les valeurs $(0.01, 10)$ de (μ, λ). La CV basée sur l'erreur de classification sous–estime la pénalisation sur les parties non linéaires, tandis que la CV basée sur la vraisemblance sur–estime la pénalisation sur les parties linéaires. L'estimation en termes de probabilités ou de frontière de décision n'aboutit pas nécessairement à la même solution (Friedman, 1997).

Les méthodes analytiques basées sur la vraisemblance, AIC, AICc et BIC, induisent des erreurs élevées, mais pour les deux premières, elles ne sont pas éloignées de celle de CV_v. Elles sous–estiment la complexité du modèle, contrairement à ce qu'on s'atten-

[1] Ces résultats sont disponibles sur
`http ://www.grappa.univ-lille3.fr/~torre/Recherche/Indana`

drait (section 5.1.1). La GCV est la méthode analytique qui a l'erreur est la plus grande. Les exemples "décès" sont particulièrement mal classés par cette méthode, ce conduit à une erreur de classement (avec les coûts $\{1, 10\}$) très élevée.

7 Conclusions

Le modèle logistique additif est une méthode de discrimination flexible et interprétable. Néanmoins, la difficulté de l'ajustement de la complexité des fonctions monovariées limite son utilisation à des situations où le nombre de variables est réduit. Nous avons proposé une nouvelle technique d'estimation qui permet l'application de ce modèle quand le nombre de variables est important, en distribuant la complexité de chaque variable de façon automatique. La sélection de la complexité globale se présente comme un problème à résoudre étant donné que les méthodes de rééchantillonnage requièrent un temps de calcul important et les méthodes analytiques ne sont pas performantes.

En perspectives, nous notons que Friedman *et al.* (2000) donnent une interprétation du *boosting* basée sur la régression additive logistique. Il faut noter que, dans leur interprétation, le modèle est additif sur le domaine des classifieurs de base, et non pas sur celui des caractéristiques. La procédure d'estimation est également différente, puisque les modèles ne sont pas ré-estimés par *backfitting*. Malgré ces différences, deux pistes évoquées dans (Friedman *et al.*, 2000) méritent d'être investiguées dans le cadre des modèles additifs tels que ceux présentés ici.

Bien que très différentes par leurs objectifs, ces voies d'exploration se traduisent par une modification mineure de l'algorithme IRLS, consistant à éditer la matrice des pondérations $\mathbf{W}$. La première proposition (section 4 de (Friedman *et al.*, 2000)) vise à augmenter la marge des points correctement classés en présence de données non–séparables. Le moyen envisagé consiste à borner les W_{ii}. La seconde suggestion (section 9 de (Friedman *et al.*, 2000)) consiste simplement à accélérer la procédure d'estimation, en ignorant les points pour lesquels les pondérations W_{ii} sont petites. Ces deux modifications sont extrêmement facile à mettre en œuvre, mais il reste à analyser leurs impacts quand elles sont accomplies à l'intérieur de la boucle de *backfitting*.

Références

AVALOS M., GRANDVALET Y. & AMBROISE C. (2003). Regularization methods for additive models. In M. R. BERTHOLD, H. J. LENZ, E. BRADLEY, R. KRUSE & C. BORGELT, Eds., *5th International Symposium on Intelligent Data Analysis.*, p. 509–520 : Springer. LNCS.

BAKIN S. (1999). *Adaptive Regression and Model Selection in Data Mining Problems*. PhD thesis, School of Mathematical Sciences, The Australian National University, Canberra.

BRATKO I. (1997). Machine learning : between accuracy and interpretability. In G. DELLA RICCIA, Ed., *Learning, networks and statistics. ISSEK'96 workshop*, CISM Courses Lect.382, p. 163–177 : Springer.

BRIDLE J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, relationships to statistical pattern recognition. In F. FOGELMAN SOULIÉ & J. HÉRAULT, Eds., *Neuro-computing : Algorithms, Architectures and Applications*, p. 227–236 : Springer.

BUJA A., HASTIE T. J. & TIBSHIRANI R. J. (1989). Linear smoothers and additive models. *The Annals of Statistics*, **17**, 453–510.

CHEN S., DONOHO D. & SAUNDERS M. (1995). *Atomic decomposition by basis pursuit*. Rapport interne 479, Department of Statistics, Stanford University.

FAHRMEIR L. & TUTZ G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd edition*. Springer Series in Statistics. New York : Springer.

FRIEDMAN J., HASTIE T. & TIBSHIRANI R. (2000). Additive logistic regression : a statistical view of boosting. *The Annals of Statistics*, **28**(2), 337–407.

FRIEDMAN J. H. (1997). On bias, variance, 0/1 loss, and the curse of dimensionality. *Data Mining and Knowledge Discovery*, **1**(1), 55–77.

FU W. J. (1998). Penalized regression : the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**(3), 397–416.

GRANDVALET Y. & CANU S. (1998). Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In M. KEARNS, S. SOLLA & D. COHN, Eds., *Advances in Neural Information Processing Systems 11*, p. 445–451 : MIT Press.

GU C. (1992). Cross–validating non–Gaussian data. *Journal of Computational and Graphical Statistics*, **1**, 169–179.

GUEYFFIER F., BOUTITIE F., BOISSEL J. P., COOPE J., CUTLER J., EKBOM T., FAGARD R., FRIEDMAN L., PERRY H. M. & POCOCK S. (1995). INDANA : a meta–analysis on individual patient data in hypertension. protocol and preliminary results. *Therapie*, **50**(4), 353–362.

GUNN S. R. & KANDOLA J. S. (2002). Structural modeling with sparse kernels. *Machine Learning*, **10**, 581–591.

HASTIE T. J. & TIBSHIRANI R. J. (1990). *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.

HASTIE T. J., TIBSHIRANI R. J. & FRIEDMAN J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Series in Statistics. New York : Springer.

LIN Y., WAHBA G., ZHANG H. & YOONKYUNG L. (2000). *Statistical Properties and Adaptive Tuning of Support Vector Machines*. Rapport interne 1022, University of Winconsin.

LIN Y. & ZHANG H. H. (2002). *Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models*. Rapport interne 1072r, University of Winconsin – Madison and North Carolina State University.

OSBORNE M. R., PRESNELL B. & TURLACH B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, **9**(2), 319–337.

ROTH V. (2001). Sparse kernel regressors. In G. DORFNER, H. BISCHOF & K. HORNIK, Eds., *Artificial Neural Networks–ICANN 2001*, p. 339–346 : Springer, LNCS 2130.

TIBSHIRANI R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, **58**(1), 267–288.

VAPNIK V. (1995). *The Nature of Statistical Learning Theory*. Springer Series in Statistics. New York : Springer.

WAHBA G. (1990). *Spline Models for Observational Data*. Number 59 in Regional Conference Series in Applied Mathematics. Philadelphia, PA. : SIAM.