Pénalisation l1 pour les MAG

Marta Avalos, Yves Grandvalet, Christophe Ambroise

37èmes Journées de Statistique (2005) 25.1

Pénalisation l_1 pour les MAG

Marta Avalos & Yves Grandvalet & Christophe Ambroise

Laboratoire HeuDiaSyC UMR CNRS 6599
Université de Technologie de Compiègne
BP 20529 / 60205 Compiègne

Résumé

Nous présentons une nouvelle méthode d'estimation fonctionnelle dans le cadre des modèles additifs généralisés (MAG). Les paramètres de ce modèle sont estimés par vraisemblance pénalisée, où le terme de régularisation généralise la pénalisation l_1 aux fonctions splines. Comme dans le cas linéaire, certains coefficients sont rétrécis, alors que les autres sont annulés exactement, aboutissant ainsi à des modèles parcimonieux, sélectionnant les variables jugées significatives. Les solutions sont calculées par une modification de l'algorithme classique, qui intègre l'optimisation du problème pénalisé. La mise en œuvre de la méthode est illustrée par une application à la prédiction de la concentration plasmatique d'un type d'antirétroviraux chez des patients infectés par le VIH.
Mots clés : modèles additifs généralisés, sélection de variables, réglage de la complexité, modélisation non paramétrique, sélection de modèle, apprentissage supervisé, épidémiologie.

Abstract

We present a new method for function estimation in generalized additive models (GAM). Parameters of this model are estimated via penalized likelihood, where the term of regularization generalizes the l_1–penalization to the splines functions. As in the linear case, it shrinks coefficients, some of them going exactly to zero and hence gives parsimonious models, selecting significant variables. Solutions are computed with a modified version of the standard algorithm, including the optimization of the penalized problem. The implementation of this method is illustrated with an application to plasmatic concentration prediction of a type of antiretroviral, for patients infected by HIV.
Key words : generalized additive models, variable selection, complexity tuning, nonparametric modeling, model selection, supervised learning, epidemiology.

Introduction

La classe des modèles linéaires généralisés (MLG) offre une approche unifiée de nombreux outils d'analyse standards en statistique appliquée. Ils généralisent le modèle linéaire à une classe très large de problèmes concernant la relation entre une variable réponse Y et plusieurs variables explicatives $X_1, \ldots, X_p$, qui inclu, par exemple, le modèle logistique ou des modèles de survie. Comme dans la régression linéaire, les effets des variables explicatives sur la variable réponse sont supposés être linéaires, cependant, la distribution

des réponses ainsi que le lien entre les entrées et la distribution de la sortie peuvent être très généraux. Malgré leur simplicité, les MLG sont inadéquats dans certaines situations.

1) Si le nombre de variables explicatives est élevé (par rapport au nombre d'observations) ou si elles sont très corrélées, alors la variance des estimations sera importante, aboutissant à des prédictions imprécises. Ce problème est habituellement abordé en utilisant des méthodes de sélection pas à pas, qui sont néanmoins, connues pour son instabilité. Une approche différente consiste à imposer une pénalisation sur les grandes fluctuations des paramètres à estimer. La pénalisation au sens de la norme l_1, connue sous le nom de *lasso* (Hastie *et al.*, 2001), consiste à minimiser le coût sous une contrainte sur la norme l_1 des coefficients. Une particularité de cette contrainte est que certains coefficients sont rétrécis, alors que les autres sont annulés exactement. En outre, la forme régulière de la contrainte donne lieu à un modèle moins variable que celui obtenu par sélection pas à pas.

2) Dans le "monde réel", les effets sont généralement non linéaires, alors l'hypothèse de dépendance linéaire est souvent trop restrictive. Les modèles additifs généralisés (MAG) sont une extension des MLG, (Hastie *et al.*, 2001). Ils remplacent chaque composante linéaire par une fonction plus générale f_j, $j = 1, \ldots, p$. 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 proposons une nouvelle méthode qui généralise la pénalisation l_1 aux MAG ajustés par des splines cubiques de lissage. Notre approche maximise la log–vraisemblance sous deux contraintes : une sur la norme l_1 des coefficients des composantes linéaires et une autre sur la norme l_1 des coefficients des composantes non linéaires des splines. Comme dans le cas linéaire, l'estimation fonctionnelle et la sélection de variables sont effectuées de façon simultanée, aboutissant ainsi à des modèles parcimonieux. Cette méthode a été préalablement adaptée à la régression gaussienne (Avalos *et al.*, 2004) et à la régression logistique (Avalos *et al.*, 2005). Ici nous développons ce principe dans le contexte plus général des MAG. Cette nouvelle approche est appliquée à la prédiction de la concentration plasmatique d'un type d'antirétroviraux chez des patients infectés par le VIH, dans une base de données réelles provenant de l'essai Cophar 1 (Brendel *et al.*, 2005).

MLG et MAG

La classe des MLG est caractérisée par trois composantes : la composante aléatoire, la composante déterministe et la fonction de lien. La composante aléatoire identifie la distribution de probabilités de la variable à expliquer, parmi les distributions de la famille exponentielle. La fonction de densité de Y s'écrit ainsi : $h_Y(y;\theta;\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$, où θ est un paramètre canonique, ϕ est un paramètre de dispersion et les fonctions b et c sont spécifiques à chaque distribution. La composante déterministe du modèle est le prédicteur linéaire $\eta = \alpha_0 + X_1\alpha_1 + \ldots, +X_p\alpha_p$. L'espérance μ est une fonction inversible du prédicteur linéaire : $g(\mu) = \eta$, où g, fonction de lien, est monotone et différentiable. La troisième composante exprime ainsi une relation fonctionnelle entre la composante

aléatoire et le prédicteur linéaire, au moyen de la fonction de lien.

Dans les MAG, le lien linéaire est remplacé par une fonction de lien additive : $g(\mu) = \alpha_0 + f_1(X_1) + \ldots + f_p(X_p)$, où α_0 est une constante et les f_j sont des fonctions non paramétriques, telles que $\mathbb{E}(f_j) = 0$, afin d'assurer l'unicité (Hastie *et al.*, 2001).

Estimation dans les MLG. Soient $\{(X_{ij}, Y_i)\}$ $i = 1, \ldots, n$, $j = 1, \ldots, p$ échantillon i.i.d. de $(X_1, \ldots, X_p, Y)$ de taille n et $\{(x_{ij}, y_i)\}$ des observations. Notons $\mathbf{X}$ la matrice des observations des variables d'entrée centrées réduites, et $\mathbf{y}$ le vecteur des sorties observées centrées. Nous introduisons le problème d'estimation pour les MAG par celui des MLG, plus simple. L'estimation des paramètres $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ est calculée en maximisant la log–vraisemblance du modèle linéaire généralisé, qui s'écrit : $l = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} + c$. L'estimateur maximum de vraisemblance est définit par le système d'équations :

$$\frac{\partial l}{\partial \boldsymbol{\alpha}} = \mathbf{X}\mathbf{W}^{-1} g'(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}, \tag{1}$$

où $\mathbf{W} = \mathrm{diag}\,(g'(\mu_i)^2 \mathrm{var}(Y_i))$. Ce système d'équations n'est pas linéaire en $\boldsymbol{\alpha}$, qui intervient dans les μ_i. La méthode d'optimisation standard est l'algorithme des scores de Fisher, qui procède aux itérations $\boldsymbol{\alpha}^{[k+1]} = \boldsymbol{\alpha}^{[k]} - \left(\frac{\partial^2 l}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^t}\right)^{-1} \frac{\partial l}{\partial \boldsymbol{\alpha}}$. Cet algorithme itératif peut être relu de la façon suivante :

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

où $\mathbf{z} = \mathbf{X}\boldsymbol{\alpha}^{[k]} + \frac{\partial \boldsymbol{\eta}^{[k]}}{\partial \boldsymbol{\mu}}(\mathbf{y} - \boldsymbol{\mu}^{[k]})$, et k indique l'itération en cours. Sous cette forme, l'algorithme est appelé moindres carrés pondérés itératifs (MCPI), car à chaque itération, le problème résolu est équivalent à un problème de moindres carrés pondérés.

Estimation dans les MAG. L'estimation des paramètres α_j est généralisée à celle des fonctions f_j. Considérons le cas où 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 maximisant la log–vraisemblance (ou, de façon équivalente, minimisant le coût −log–vraisemblance) :

$$\min_{\alpha_0 \in \mathbb{R}, f_j \in \mathcal{C}^2} -l + \sum_{j=1}^{p} \lambda_j \int \left(f_j''(t)\right)^2 dt. \tag{3}$$

Le premier terme de (3) 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}. Les contraintes $\sum_i \widehat{f_j}(x_{ij}) = 0$, $j = 1, \ldots, p$ assurent l'unicité de la solution. Le problème (3) est le lagrangien associé à :

$$\min_{\alpha_0 \in \mathbb{R}, f_j \in \mathcal{C}^2} -l \quad \text{sous contraintes} \quad \int \left(f_j''(t)\right)^2 dt \leq \tau_j, \quad j = 1, \ldots, p, \tag{4}$$

où τ_j est déterminé par λ_j. Le problème fonctionnel (4) a une formulation paramétrique équivalente, basée sur la décomposition des f_j sur une base des fonctions splines. 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 $\boldsymbol{\beta}_j$ le vecteur des coefficients de $\widehat{f_j}$ sur la base choisie, (4) s'écrit :

$$\min_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p} -\sum_{i=1}^{n} l_i \quad \text{sous contraintes} \quad \boldsymbol{\beta}_j^t \mathbf{\Omega}_j \boldsymbol{\beta}_j \leq \tau_j, \quad j = 1, \ldots, p. \tag{5}$$

L'estimation des MAG est réalisée par une modification de l'algorithme MCPI, qui permet l'estimation des fonctions non paramétriques (Hastie *et al.*, 2001).

Pénalisation l_1 pour les MAG

Dans le cadre linéaire, le lasso est la solution du problème d'optimisation suivant :

$$\min_{\alpha_1,\ldots,\alpha_p} -\sum_{i=1}^{n} l_i \quad \text{sous contrainte} \quad \sum_{j=1}^{p} |\alpha_j| \leq \tau, \tag{6}$$

où τ est le paramètre de rétrécissement qui règle la complexité du modèle. La pénalisation l_1 a été préalablement généralisée aux modèles additifs par Grandvalet et Canu (1998) :

$$\min_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p} -\sum_{i=1}^{n} l_i \quad \text{sous contrainte} \quad \sum_{j=1}^{p} \sqrt{\boldsymbol{\beta}_j^t \mathbf{\Omega}_j \boldsymbol{\beta}_j} < \tau, \tag{7}$$

où le terme de pénalisation généralise la pénalisation du lasso $\sum_{j=1}^{p} |\alpha_j| = \sum_{j=1}^{p} \sqrt{\alpha_j^2}$. Cette méthode n'accomplit pas de sélection de variables, car la composante linéaire de $\widehat{f_j}$ est dans le noyau de $\mathbf{\Omega}_j$. Ainsi, même si $\widehat{\boldsymbol{\beta}}_j^t \mathbf{\Omega}_j \widehat{\boldsymbol{\beta}}_j = 0$, la j–ème composante n'est pas éliminée mais linéarisée. Cependant, en décomposant le problème en sa partie linéaire et non linéaire et ajoutant en (7) un terme de pénalisation agissant sur la composante linéaire, il est possible de sélectionner des variables (Avalos *et al.*, 2004). La généralisation de cette idée aux MAG s'écrit :

$$\min_{\boldsymbol{\alpha},\widetilde{\boldsymbol{\beta}}_1,\ldots,\widetilde{\boldsymbol{\beta}}_p} -\sum_{i=1}^{n} l_i \quad \text{sous contraintes} \quad \sum_{j=1}^{p} |\alpha_j| \leq \tau_{\mathrm{L}}, \quad \sum_{j=1}^{p} \sqrt{\widetilde{\boldsymbol{\beta}}_j^t \mathbf{\Omega}_j \widetilde{\boldsymbol{\beta}}_j} < \tau_{\mathrm{NL}} \tag{8}$$

et $\mathbf{x}_j^t \mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j = 0$, $\mathbf{1}^t \mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j = 0$, afin d'assurer l'orthogonalité des espaces linéaires et non linéaires. Les paramètres τ_{L} et τ_{NL} sont les paramètres de rétrécissement, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^t$ et $\widetilde{\boldsymbol{\beta}}_j$ sont les coefficients de la partie linéaire et non linéaire, respectivement.

L'estimation de ces modèles est réalisée par une modification de l'algorithme MCPI pour les MAG, qui permet l'optimisation du problème sous contraintes par un algorithme de point fixe (Avalos *et al.*, 2004).

Sélection des paramètres de la complexité. Alors que dans les MAG classiques on a besoin de p paramètres de réglage de la complexité (λ_j dans (3) ou τ_j dans (4)), dans

notre approche les complexités individuelles sont distribuées automatiquement, seuls les paramètres réglant la complexité globale (τ_{L} et τ_{NL} dans (8)) doivent être optimisés par une procédure de sélection de modèle. L'erreur en prédiction peut être estimée par des critères analytiques, cependant l'efficacité de ces méthodes n'est pas identique à l'ensemble des MAG. En effet, des critères d'information ont montré de bons résultats dans le cas gaussien (Avalos *et al.*, 2004), qu'on ne retrouve pas dans le cas binomial (Avalos *et al.*, 2005). La validation croisée, qui fait moins d'approximations, peut être alors utilisée pour le choix des paramètres de la complexité.

Application à des données réelles

Les caractéristiques pharmacocinétiques (absorption, distribution et élimination) de certains antirétroviraux, tels que les inhibiteurs de protéase, sont très variables. La relation concentration/effet (thérapeutique ou toxique) est alors meilleur que la relation dose/effet. L'essai Cophar 1 vise à établir une fourchette de concentrations plasmatiques dans laquelle le traitement est efficace et sa toxicité réduite (Brendel *et al.*, 2005). Nous appliquons notre approche aux données de l'essai Cophar 1 concernant la molécule indinavir. La variable réponse, de type gaussien, est la log–concentration plasmatique d'indinavir. Les données sont constituées de 14 variables d'entrée, centrées et réduites : 1. sexe, 2. âge, 3. poids, 4. indice de masse corporelle (IMC), 5. surface corporelle, 6. nombre de CD4, 7. stade de la maladie, 8. durée des traitements (T), 9. durée du nouveau traitement (NT), 10. et 11. nombre de molécules d'un certain type inclues dans la multithérapie (M1 et M2), 12. dose quotidienne (posologie), 13. dose par prise, et 14. présence ou absence de traitement par ritonavir, qui ralentit l'élimination. Il y a 42 sujets, dont 1 avec certaines valeurs manquantes, exclu. La figure 1 montre des effets importants de l'âge, le poids, la posologie, l'IMC, la durée du nouveau traitement, et M1 (une sur–estimation due aux faibles effectifs pour une des valeurs est néanmoins probable). Le stade et la surface corporelle ont des effets très faibles. Les variables durée des traitements, M2, dose, CD4, sexe et ritonavir ont été éliminées. Le poids et la posologie ont des effets linéaires négatifs sur la concentration plasmatique. Elle diminue avec le poids, ce qui est raisonnable, et avec la posologie, ce qui est expliqué par le fait que les doses les plus faibles correspondent aux traitements avec ritonavir, et donc à une élimination plus lente. L'âge, l'IMC et la durée NT semblent avoir un effet quadratique sur la concentration. Par exemple, les concentrations les plus élevées se trouvent dans la fenêtre de valeurs de l'IMC considérées normales, et les plus faibles se trouvent dans les régions de surpoids et de souspoids.

Conclusions

Les MAG permettent de modéliser de façon flexible et interprétable une classe très large de problèmes concernant la relation entre une variable réponse et plusieurs variables explicatives. Nous avons proposé une nouvelle technique d'estimation qui, d'une part, permet l'application de ces modèles quand le nombre de variables explicatives est élevé (par

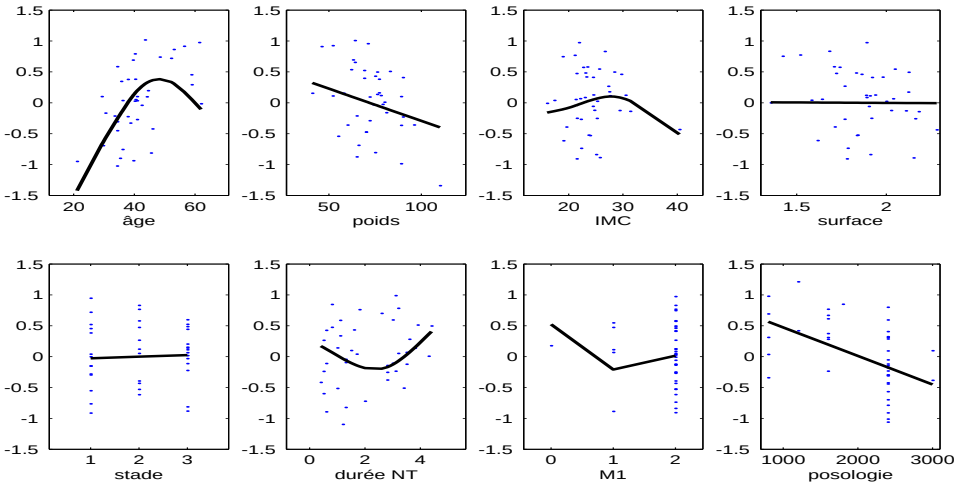

FIG. 1 – Estimations des fonctions additives pour la valeur de (μ, λ) sélectionnée par une validation croisée à 10 blocs (lignes continues) et résidus partiels (points).

rapport aux observations) et, d'autre part, favorise l'élimination des variables les moins pertinentes. La complexité de chaque variable est distribuée de façon automatique, seule la complexité globale, contrôlée par deux paramètres, doit être réglée par une méthode de sélection de modèle. Son application à la prédiction de la concentration plasmatique permet de révéler des effets non linéaires et de sélectionner les variables pertinentes.

Remerciements

Les auteurs remercient le groupe Cophar 1–ANRS 102, dont les données illustrent les méthodes développées, et les membres de l'unité INSERM U738 (et tout particulièrement France Mentré et Xavière Panhard), par leur aide dans l'interprétation des résultats.

Bibliographie

[1] Avalos, A., Grandvalet, Y. et Ambroise, C. (2004) Généralisation du lasso aux modèles additifs. In *XXXVIèmes Journées de Statistique*, Montpellier.
[2] Avalos, A., Grandvalet, Y. et Ambroise, C. (2005) Discrimination par modèles additifs parcimonieux. *Revue d'Intelligence Artificielle* . A paraître en septembre 2005.
[3] Brendel, K., *et al.* (2005) Population pharmacokinetic analysis of indinavir in HIV–patients treated with stable antiretroviral therapy. A paraître.
[4] Grandvalet, Y. et Canu, S. (1998) Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In *NIPS 11*, 445–451, MIT Press.
[5] Hastie, T. J., Tibshirani, R. J. et Friedman, J. (2001) *The Elements of Statistical Learning*, Springer Series in Statistics, Springer.