Généralisation du lasso aux modèles additifs

Marta Avalos, Yves Grandvalet, Christophe Ambroise

36èmes Journées de Statistique (2004) 83

Généralisation du lasso aux modèles additifs

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 non paramétriques ajustés par des splines cubiques de lissage. Cette méthode généralise le "lasso" proposé par Tibshirani, (1995) pour les modèles linéaires. 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 un algorithme de point fixe, dans lequel une décomposition en valeurs singulières permet de réduire considérablement le nombre de calculs. Notre approche est validée dans une étude expérimentale comparant les performances du lasso à la sélection de variables pas à pas.
Mots clés : régression non paramétrique, sélection de variables, sélection de modèle, pénalisation.

Abstract

We present a new method for function estimation in nonparametric additive models fitted by cubic smoothing splines. The method is a generalization of the "lasso" proposal of Tibshirani (1995), designed for the linear regression context. 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 calculated using a fixed point algorithm, combined with a singular value decomposition that considerably reduces computation. This approach is validated by an experimental study, comparing the lasso performances to those from forward selection.
Key words : nonparametric regression, variable selection, model selection, penalization.

Introduction

La régression par modèle additif non paramétrique (Hastie et Tibshirani, 1990) estime la dépendance entre une variable réponse Y et plusieurs variables explicatives $\mathbf{X} = (X_1, \ldots, X_p)$ d'une façon flexible et interprétable. Elle suppose que l'espérance conditionnelle de la réponse peut s'écrire $\mathbb{E}(Y|x_1, \ldots, x_p) = f_0 + f_1(x_1) + \ldots + f_p(x_p)$. En restreignant les dépendances à une somme de fonctions monovariées, ce modèle evite le problème du "fléau de la dimensionnalité". Cette structure simple permet également de représenter l'effet de chaque variable, ce qui facilite l'interprétation des solutions.

La sélection de modèle consiste à déterminer la structure du modèle la plus adaptée aux données. Elle comporte deux sous–problèmes : la sélection de variables et le choix de la complexité. Des méthodes de sélection de la complexité et de sélection de variables ont été proposées pour les modèles additifs (Hastie et Tibshirani, 1990). Ces dernières reposent sur la sélection d'un sous–ensemble de variables. Dans le contexte de la régression linéaire, le choix de la complexité du modèle peut se limiter à la sélection des variables. Toutefois, les méthodes de rétrécissement permettent souvent d'obtenir des résultats plus stables (Tibshirani, 1995). Le *lasso* (least absolute shrinkage and selection operator), proposé par Tibshirani (1995), consiste à minimiser le coût quadratique 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, effectuant ainsi l'estimation des coefficients et la sélection de variables de façon simultanée.

Grandvalet et Canu (1998) ont adapté le lasso aux modèles additifs ajustés par splines. Celui–ci pénalise les composantes non linéaires des estimateurs des fonctions f_j, mais elle ne pénalise pas les composantes linéaires, elle n'accomplit donc pas de sélection de variables. Nous proposons une nouvelle version du lasso pour les modèles additifs ajustés par des splines cubiques de lissage. Les parties linéaires et non linéaires des fonctions f_j sont pénalisées indépendamment, ce qui permet de distinguer les variables pertinentes des non pertinentes. L'estimation des fonctions se base sur un algorithme de point fixe, où une décomposition en valeurs singulières permet de réduire considérablement le nombre de calculs. Nous montrons expérimentalement les performances et la stabilité de cette méthode. La comparaison avec la sélection de variables pas à pas nous permet de caractériser les conditions d'application de chaque méthode.

Le lasso pour les modèles additifs

Considérons le problème de régression classique, où nous disposons d'un ensemble d'apprentissage $\mathcal{L} = \{(\mathbf{x}, \mathbf{y})\}$, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^t$, $\mathbf{y} = (y_1, \ldots, y_n)^t$. Supposons, pour simplifier, que les variables sont centrées. L'estimateur lasso est la solution du problème d'optimisation sous contraintes suivant :

$$\min_{\alpha_1,\ldots,\alpha_p} \left|\left| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j \right|\right|_2^2 \quad \text{sous contrainte} \quad \sum_{j=1}^{p} |\alpha_j| \leq \tau, \tag{1}$$

où τ est le paramètre de rétrécissement qui règle la complexité du modèle. Nous considérons maintenant le cas où les fonctions f_j sont approchées par des splines cubiques de lissage $\widehat{f_j}$. Soient $\mathbf{N}_j$ la matrice de la base naturelle B–spline 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 de B–splines. L'extension du lasso proposée par Grandvalet et Canu (1998) est donnée par

$$\min_{\beta_1,\ldots,\beta_p} \left|\left| \mathbf{y} - \sum_{j=1}^{p} \mathbf{N}_j \beta_j \right|\right|_2^2 \quad \text{sous contrainte} \quad \sum_{j=1}^{p} \sqrt{\beta_j^t \mathbf{\Omega}_j \beta_j} < \tau, \tag{2}$$

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 $\widehat{\beta}_j^t \mathbf{\Omega}_j \widehat{\beta}_j = 0$, la j–ème composante n'est pas éliminée mais linéarisée.

Modification du lasso

Les splines sont une méthode de lissage linéaire : l'estimation de f_j peut s'écrire $\widehat{\mathbf{f}}_j = \mathbf{S}_j \mathbf{y}$, où $\mathbf{S}_j$, matrice de lissage $n \times n$, dépend du paramètre de lissage λ_j et des $\mathbf{x}_j$, mais elle est indépendante de $\mathbf{y}$. La matrice de lissage des splines cubiques a deux valeurs propres égales à 1, correspondant aux fonctions propres constante et linéaire, et $n-2$ valeurs propres dans l'intervalle $]0,1[$, correspondant aux fonctions d'ordre supérieur. Nous la décomposons comme suit : $\mathbf{S}_j = \mathbf{G}_j + \widetilde{\mathbf{S}}_j$. En ajoutant en (2) un terme de pénalisation agissant sur la composante linéaire, il est possible de sélectionner des variables :

$$\min_{\alpha, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_p} \left|\left| \mathbf{y} - \mathbf{x}\alpha - \sum_{j=1}^{p} \mathbf{N}_j \widetilde{\beta}_j \right|\right|_2^2 \quad \text{sous contraintes} \quad \sum_{j=1}^{p} |\alpha_j| \leq \tau_1 \quad \text{et} \quad \sum_{j=1}^{p} \sqrt{\widetilde{\beta}_j^t \mathbf{\Omega}_j \widetilde{\beta}_j} < \tau_2, \tag{3}$$

et $\mathbf{x}_j^t \mathbf{N}_j \widetilde{\beta}_j = 0$, $\mathbf{1}^t \mathbf{N}_j \widetilde{\beta}_j = 0$, où τ_1 et τ_2 sont les paramètres de rétrécissement, $\alpha = (\alpha_1, \ldots, \alpha_p)^t$ et $\widetilde{\beta}_j$ sont les coefficients de la partie linéaire et non linéaire, respectivement.

Algorithme

Le problème d'optimisation (3) est reformulé sous forme lagrangienne, ce qui conduit à un problème de type pénalisation multiple adaptative (Grandvalet et Canu, 1998), qui peut être résolu par un algorithme de point fixe. La quantité de calcul peut y être considérablement réduite grâce à une décomposition en valeurs singulières, évitant des inversions dans la partie itérative de l'algorithme. La matrice de lissage devient $\mathbf{S}_j = \mathbf{U}_j(\mathbf{Z}_j + \lambda_j \mathbf{I})^{-1} \mathbf{U}_j^t$, où $\mathbf{U}_j$ est orthogonale et $\mathbf{Z}_j$ est diagonale définie positive et λ_j est positif. La trace de la matrice de lissage est obtenue par $\mathrm{tr}(\mathbf{S}_j) = \sum_i z_{ij}/(\lambda_j + z_{ij})$. Notons que cette décomposition est effectuée une seule fois pour toutes le valeurs (τ_1, τ_2) testées.

Sélection des paramètres de la complexité

Les paramètres de Lagrange (μ, λ) correspondant à (τ_1, τ_2) doivent être optimisés par une procédure de sélection de modèle. La valeur optimale est, par définition, celle qui minimise l'erreur en prédiction. Cette erreur peut être estimée par le critère d'information d'Akaike (AIC), le critère d'information Bayésien (BIC), ou la validation croisée généralisée (GCV) (Hastie et Tibshirani, 1990). Les fonctions AIC, BIC et GCV sont évaluées sur une grille sur μ et λ. Les points minimisant AIC, BIC et GCV, respectivement, sont sélectionnés.

Algorithme :

1. Décomposition en valeurs singulières.
2. Pour chaque (μ, λ), initialiser : $\mu_j = \mu$, $\mathbf{M} = \mu \mathbf{I}_p$, $\lambda_j = \lambda$.
3. Composantes linéaires :
   (a) Estimation des coefficients : $\alpha = \left(\mathbf{x}^t\mathbf{x} + \mathbf{M}\right)^{-1} \mathbf{x}^t\mathbf{y}$, où $\mathbf{M} = \mathrm{diag}(\mu_j)$.
   (b) Reestimation des termes de pénalisation : $\mu_j = \mu \frac{\|\alpha\|_1}{p|\alpha_j|}$.
   (c) Itérer 3.(a) et 3.(b) jusqu'à convergence.
4. Composantes non linéaires :
   (a) Estimation des coefficients :
      i. Backfitting : $\widetilde{\mathbf{r}}_j = \mathbf{y} - \mathbf{G}\mathbf{y} - \sum_{k \neq j} \mathbf{N}_j \widetilde{\beta}_j, \quad \mathbf{N}_j \widetilde{\beta}_j = \widetilde{\mathbf{S}}_j \widetilde{\mathbf{r}}_j, \quad j = 1, \ldots, p.$
      ii. Calcul de $\widetilde{\beta}_j$ à partir des dernières estimations.
   (b) Reestimation des termes de pénalisation : $\lambda_j = \lambda \frac{\sum_{j=1}^p \sqrt{\widetilde{\beta}_j^t \mathbf{\Omega}_j \widetilde{\beta}_j}}{p \sqrt{\widetilde{\beta}_j^t \mathbf{\Omega}_j \widetilde{\beta}_j^t}}$.
   (c) Itérer 4.(a) et 4.(b) jusqu'à convergence.

Les trois méthodes nécessitent une estimation du nombre effectif de paramètres ou degrés de liberté, df. Nous adoptons l'estimation suivante :

$$\mathrm{df}(\mu, \lambda) \approx \sum_{j=1}^p \mathrm{df}_j(\mu, \lambda) = \mathrm{tr}\left[\mathbf{x}\left(\mathbf{x}^t\mathbf{x} + \mathbf{M}(\mu)\right)^{-1}\mathbf{x}^t\right] + \sum_{j=1}^p \mathrm{tr}\left[\widetilde{\mathbf{S}}_j(\lambda)\right]. \qquad (4)$$

Les degrés de liberté liés aux composantes linéaires et non linéaires sont estimés séparément. Le calcul pour les composantes linéaires se base sur la formulation du problème en termes de la pénalisation multiple adaptative. Le nombre de degrés de liberté liés aux composantes non linéaires est approché par la somme des p degrés de liberté sur chaque coordonnée, df_j (Hastie et Tibshirani, 1990) et facilement calculé grâce à la décomposition en valeurs singulières. Notons que les critères AIC, BIC et GCV ne tiennent pas compte de l'effet de la sélection du modèle (Ye, 1998) : les calculs supposent que le modèle sélectionné est connu a priori, ce qui introduit du biais dans l'estimation. L'estimateur (4) présente le même défaut : l'estimation des termes de pénalisation (μ_j, λ_j) n'est pas prise en compte.

Simulations

Dans les simulations suivantes nous comparons le lasso à la sélection pas à pas. Sur 4 jeux d'essais, nous avons généré 100 échantillons, les résultats sont donnés en termes de la moyenne. Nous comparons l'erreur en prédiction commise par chaque méthode, en l'estimant sur un ensemble de test de taille 6000. Le rapport du nombre de variables correctement estimées pertinentes sur le nombre de variables réellement pertinentes, et le rapport du nombre de variables correctement estimées non pertinentes sur le nombre de variables réellement non pertinentes sont également rapportés, ainsi que la proportion de cas où des variables parfaitement redondantes sont sélectionnées simultanément.

Les variables significatives et les paramètres de lissage de la sélection de variables pas à pas sont sélectionnés par le critère GCV, évalué sur une grille de 5 valeurs pour chaque composante additive (cas linéaire, $\lambda_j = \infty$, compris). Dans le cas du lasso, les

TAB. 1 – Identification des variables pertinentes, des variables non pertinentes et des dépendances entre les variables explicatives. Les valeurs sont des proportions moyennes.

	Pertinence				Non pertinence				Redondance
Méthode	Ex. 1a)	Ex. 1b)	Ex. 1c)	Ex. 2	Ex. 1a)	Ex. 1b)	Ex. 1c)	Ex. 2	Ex. 2
Pas à pas	0.990	0.879	0.645	0.869	0.803	0.817	0.765	0.842	0.007
Lasso (AIC)	0.998	0.941	0.920	0.979	0.156	0.137	0.075	0.283	0.877
Lasso (BIC)	0.998	0.974	0.953	0.999	0.054	0.072	0.035	0.138	0.947
Lasso (GCV)	1.000	0.895	0.868	0.964	0.396	0.335	0.175	0.382	0.760
Lasso*	0.998	0.945	0.911	0.977	0.249	0.190	0.125	0.197	0.830

fonctions AIC, BIC et GCV, sont évaluées sur une grille 5×5, et les performances obtenues sont comparées à la performance optimale (celle obtenue par une méthode de sélection choisissant le modèle d'erreur en prédiction minimale).

Le problème du contrôle de la complexité est particulièrement crucial quand le nombre de degrés de liberté du modèle est de l'ordre de grandeur de l'échantillon. Le nombre de variables réellement explicatives et la concurvité (l'équivalent non paramétrique de la colinéarité) sont des autres sources de difficulté pour l'estimation des fonctions.

Exemple 1. Nous considérons 100 échantillons de taille $n = 50$ constitués de $p = 14$ variables explicatives indépendantes. Les réponses sont générées par $y = \delta_j^1 x_j + \delta_j^2 \sin 2\pi k_j x_j + \varepsilon$, où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, σ^2 tel que $R^2 = 0.95$. Le paramètre k_j contrôle la courbure. Les paramètres $\delta_j = \{0, 1\}$ contrôlent le type de fonction. La variable est non pertinente si $\delta_j^1 = 0$ et $\delta_j^2 = 0$, la fonction est linéaire si $\delta_j^1 = 1$ et $\delta_j^2 = 0$, et non linéaire, si $\delta_j^2 = 1$. Nous considérons 3 cas en fonction de δ_j : a) 4 b) 8 et c) 12 variables pertinentes.

Exemple 2. Nous considérons 100 échantillons de taille $n = 50$ constitués de $p = 16$ variables explicatives générées comme suit : $(X_1, \ldots, X_{12}) \sim \mathcal{N}_{12}(\mathbf{0}, \mathbf{1})$; $X_{13} = X_1 + 1$; $X_{14} = X_2 + X_3$; $X_{15} = X_4^2$; $X_{16} = X_5 X_6$. La réponse est générée par $\mathbf{y} = \mathbf{x}_1 + 2\mathbf{x}_2 + 3\mathbf{x}_3 + \sin \pi\mathbf{x}_4 + \sin \frac{\pi}{2}\mathbf{x}_5 + (\mathbf{x}_6 + \sin \pi\mathbf{x}_6) + (\mathbf{x}_7 + \sin \frac{\pi}{2}\mathbf{x}_7) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, σ^2 tel que $R^2 = 0.95$.

Résultats. Le tableau 1 présente des mesures concernant l'identification correcte de la pertinence/non pertinence des variables. Le lasso optimal (Lasso*) sélectionne plus fréquemment les variables pertinentes et élimine moins fréquemment de variables non pertinentes que la sélection pas à pas. Parmi les techniques du lasso, la GCV est, en général, celle qui a la meilleure performance. Quant à la coïncidence des variables dépendantes, la sélection pas à pas détecte mieux l'information redondante. Parmi les techniques du lasso, la GCV a aussi dans ce cas la meilleure performance.

Les estimations de l'erreur en prédiction sont montrées sur le tableau 2 pour chacune des méthodes étudiées, ainsi que pour la fonction constante. Chaque valeur est la moyenne (écart type) obtenue sur 100 simulations. Parmi les méthodes de sélection des paramètres de la complexité pour le lasso, la GCV obtient toujours les résultats les plus proches des

TAB. 2 – Erreur moyenne de test (écart type), pour les méthodes en compétition.

Méthode	Ex. 1a)	Ex. 1b)	Ex. 1c)	Ex. 2
Constante	2.65 (0.3)	8.60 (0.9)	12.41 (1.4)	15.63 (2.7)
Pas à pas	0.53 (0.7)	2.52 (1.8)	7.49 (3.2)	18.91 (7.0)
Lasso (AIC)	0.94 (0.8)	3.90 (2.2)	7.70 (3.0)	14.00 (4.8)
Lasso (BIC)	0.82 (0.8)	3.56 (2.3)	7.36 (3.1)	13.33 (5.2)
Lasso (GCV)	0.62 (0.5)	2.77 (1.3)	6.21 (1.9)	13.02 (4.2)
Lasso*	0.55 (0.5)	2.44 (1.3)	5.70 (1.8)	12.18 (3.6)

résultats optimaux et elle est la moins variable. La perte occasionnée par la sélection de modèle est relativement faible, sauf pour le modèle le plus simple. La sélection pas à pas est performante lorsque le nombre de variables pertinentes est faible devant nombre total de variables explicatives, quand ces variables sont indépendantes, elle est, néanmoins, plus variable que le lasso GCV. Quand le nombre de variables pertinentes est modéré ou élevé, ou quand il y a de la concurvité, le lasso est mieux.

Conclusions. Les résultats de nos expériences concordent avec ceux déjà obtenus pour la régression linéaire. La concurvité et le nombre de variables réellement pertinentes sont des facteurs déterminants pour le choix de la méthode de sélection de variables. Le lasso conserve bien les variables pertinentes. En revanche, il élimine peu de variables non pertinentes et de variables redondantes. Plausiblement, ces variables restent très pénalisées dans le modèle.

La sélection pas à pas est bien adaptée lorsque le nombre de variables pertinentes est faible et que ces variables sont indépendantes. Le lasso est bien adapté quand le nombre de variables pertinentes est modéré ou élevé, ainsi qu'en présence de concurvité. Parmi les techniques de sélection de la complexité pour le lasso, la GCV est la plus performante (proche de la meilleure performance possible). Une possible explication de ce fait est que la GCV ne nécessite pas d'estimation de la variance de l'erreur, contrairement à AIC et BIC.

Bibliographie

[1] Grandvalet, Y. et Canu, S. (1998) Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In *NIPS 11*, 445–451, MIT Press.

[2] Hastie, T. J. et Tibshirani, R. J. (1990) *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*, Chapman & Hall.

[3] Tibshirani, R. J. (1995) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, 58(1) :267–288.

[4] Ye, J. (1998) On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93 :120–131.