Prévision pac-bayésienne pour le modèle additif sous contrainte de parcimonie

Benjamin Guedj [1], Pierre Alquier [2], Gérard Biau [1] et Éric Moulines [3]

[1] *LSTA, Université Pierre et Marie Curie – Paris VI*
Boîte 158, Tour 15-25, 2ème étage
4 place Jussieu, 75252 Paris Cedex 05, France
benjamin.guedj@upmc.fr
gerard.biau@upmc.fr
[2] *LPMA, Université Paris Diderot – Paris VII*
Boîte 188, 175 rue du Chevaleret
75013 Paris, France
alquier@math.jussieu.fr
[3] *LTCI, Telecom ParisTech*
37/39 rue Dareau, 75014 Paris, France
moulines@telecom-paristech.fr

Résumé. Nous nous intéressons dans ce travail à la construction et à la mise en oeuvre d'une méthode de prévision dans le modèle additif, sous contrainte de parcimonie. L'estimateur de Gibbs est au coeur de notre approche. Cet estimateur est construit à l'aide d'une loi *a priori* favorisant les solutions parcimonieuses. Nous justifions notre démarche en fournissant une inégalité oracle PAC-bayésienne originale et une implémentation faisant appel à des techniques MCMC récentes.

Mots-clés. Modèles additifs, régression, théorie PAC-bayésienne, parcimonie, inégalité oracle, algorithmes MCMC.

Abstract. We adress the issue of prediction in high dimensional additive models, under a sparsity constraint. Our method strongly relies on the Gibbs estimator, which originates from a prior distribution favoring sparse estimates. We provide the reader with a PAC-bayesian oracle inequality and an explicit implementation using recent MCMC techniques.

Keywords. Additive models, regression, PAC-bayesian theory, sparsity, oracle inequality, MCMC algorithms.

1 Introduction

Au cours des dernières années, de nombreuses méthodes d'estimation et de sélection de variables ont été proposées pour résoudre des problèmes de régression en grande dimension sous contrainte de parcimonie. La méthode des moindres carrés avec une pénalité ℓ^1 (LASSO) a de loin été la plus étudiée ; ses propriétés statistiques sont aujourd'hui bien comprises. Plusieurs autres méthodes ont été introduites (Elastic net, Dantzig selector, etc.). Ces estimateurs sont obtenus comme solutions d'un problème convexe (ou linéaire), qui admettent des solutions numériquement efficaces. En particulier, Hastie, Tibshirani et Friedman (2008), Meier, van de Geer et Bühlmann (2009) et Ravikumar, Rafferty, Liu et Wasserman (2009) présentent de telles stratégies pour le modèle additif qui est le cadre de ce travail.

Ces différentes procédures ont des propriétés statistiques intéressantes essentiellement sous des conditions techniques restrictives. Ces conditions, dans un contexte de régression linéaire, impliquent typiquement que tous les sous-ensembles de régresseurs sont approximativement orthogonaux. Ces hypothèses sont relativement naturelles lorsque l'objectif est de déterminer les variables significatives dans un ensemble de régresseurs. Elles le sont beaucoup moins lorsque l'objectif est de construire des algorithmes de prévision.

D'un point de vue théorique, il existe des estimateurs optimaux pour la prévision qui ne requièrent aucune hypothèse sur la matrice de régression. C'est le cas en particulier de l'estimateur des moindres carrés sous la pénalité ℓ^0. Toutefois, le calcul d'un tel estimateur est un problème NP-difficile ; ce problème a suscité un grand nombre de recherches visant à construire des estimateurs ayant de bonnes propriétés théoriques pour la prévision tout en restant numériquement implémentables.

Dans ce travail, nous nous intéressons à l'étude théorique et à la mise en oeuvre d'une stratégie obéissant à ces considérations, pour le modèle additif. Le modèle additif est un avatar important de la statistique contemporaine, voir par exemple Friedman et Stuetzle (1981) et Hastie et Tibshirani (1990) pour une formalisation. Notre stratégie se décline en trois axes.

- L'estimateur de Gibbs, qui dépend du choix d'une distribution *a priori* favorisant les estimateurs parcimonieux (voir Catoni (2004)).
- La théorie PAC-bayésienne, amorcée par Shawe-Taylor et Williamson (1997) et McAllester (1999), et formalisée par Catoni (2007), nous fournit une inégalité oracle.
- L'algorithme Subspace Carlin & Chib, issu de Petralias et Dellaportas (2011), permet d'échantillonner efficacement des lois de grandes dimensions.

Ce travail s'inscrit dans le prolongement théorique et méthodologique de Alquier et Biau (2011).

2 Notations

Nous considérons une collection $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ de copies *i.i.d.* d'une variable $(\mathbf{X}, Y)$ à valeurs dans $\mathbb{R}^p \times \mathbb{R}$, avec la notation $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ pour tout $i \in \{1, \ldots, n\}$. Nous nous plaçons dans le paradigme $p \gg n$. Le modèle additif s'écrit

$$Y_i = \Psi^\star(\mathbf{X}_i) + W_i = \sum_{j=1}^{p} \Psi_j^\star(X_{ij}) + W_i, \tag{1}$$

où $W = (W_1, \ldots, W_n)$ est une collection de bruits *i.i.d.*, dont nous saurons contrôler les déviations des moments successifs. Rappelons que notre objectif est de construire une stratégie d'estimation de $\Psi^\star$.

Hypothèse 1. Il existe deux constantes positives L et σ^2 telles que pour tout entier $k \geq 2$ et tout $i \in \{1, \ldots, n\}$,

$$\mathbb{E}\left[|W|^k \middle| \mathbf{X}\right] \leq \frac{k!}{2}\sigma^2 L^{k-2}. \tag{2}$$

Il est à noter que cette hypothèse est particulièrement faible : en particulier, elle est satisfaite si W est un bruit gaussien.

La fonction de régression $\Psi^\star$ se décompose additivement sur chacun des régresseurs. La seconde hypothèse consistera à supposer que $\Psi^\star$ est bornée (au sens du supremum).

Hypothèse 2. Il existe une constante $C > \max(1, \sigma)$ telle que $\|\Psi^\star\|_\infty < C$.

Enfin, nous adjoignons à ce modèle une contrainte de parcimonie, c'est-à-dire que nous supposons que la dimension effective $p^\star$ de $\Psi^\star$ est inférieure à p. En d'autres termes, un grand nombre des $\Psi_j^\star$ sont identiquement nulles.

Nous sommes maintenant en mesure de détailler notre procédure d'estimation. Considérons un dictionnaire de fonctions connu, noté $\{\varphi_k\}_{k=1}^K$, avec K un nombre entier connu. Les estimateurs de $\Psi^\star = \sum_{j=1}^p \Psi_j^\star$ que nous envisagerons seront de la forme suivante :

$$\hat{\Psi}_\theta = \sum_{j=1}^{p} \hat{\Psi}_j = \sum_{j=1}^{p} \sum_{k=1}^{m_j} \hat{\theta}_{jk}\varphi_k,$$

où m_j désigne, pour tout $j \in \{1, \ldots, p\}$, l'ordre du développement de $\hat{\Psi}_j$ dans le dictionnaire, et $\hat{\theta}$ est une matrice de taille $p \times K$ de nombres réels. Par conséquent, un modèle sera la donnée d'un sous-ensemble S de $\{1, \ldots, p\}$, désignant les régresseurs actifs, et d'un vecteur m_S de taille $|S|$, des ordres de développement dans le dictionnaire de chacun des estimateurs $\hat{\Psi}_j$. La contrainte de parcimonie nous poussera naturellement à sélectionner des estimateurs avec un ensemble S de petit cardinal.

En lien avec la méthodologie bayésienne classique, nous définissons une loi *a priori* π, chargeant avec une plus grande probabilité les estimateurs parcimonieux. Plus précisément,

$$\pi(\theta) \propto \sum_{S \subset \{1,\dots,p\}} \frac{\alpha^{-|S|}}{\binom{p}{|S|}} \sum_{m_S \in \{1,\dots,K\}^{|S|}} \beta^{-\sum_{\ell \in S} m_\ell} \pi_{(S,m_S)}(\theta), \tag{3}$$

où α, $\beta > 0$ sont des paramètres de pénalisation, et $\pi_{(S,m_S)}$ la distribution uniforme sur la boule de rayon C dans le modèle (S, m_S). Pour un certain paramètre de température $\delta > 0$, nous considérons ensuite la loi *a posteriori* de Gibbs $\hat{\rho}_\delta$ définie comme suit :

$$\frac{\mathrm{d}\hat{\rho}_\delta}{\mathrm{d}\pi}(\theta) = \exp(-\delta R_n(\hat{\Psi}_\theta)), \tag{4}$$

où R_n désigne le risque empirique.

Notre estimateur $\hat{\Psi}^{\mathrm{Gibbs}}$ est alors une réalisation d'une variable tirée suivant la loi $\hat{\rho}_\delta$.

3 Inégalité oracle PAC-bayésienne et implémentation

Nous présentons ici notre principal résultat théorique. Nous désignons par R le risque quadratique classique.

Théorème. Soit $w = 8C\max(L, C)$ et $\delta = \frac{n}{2w}$. Pour tout réel $\varepsilon \in]0, 1[$, avec probabilité au moins $1 - 2\varepsilon$,

$$R(\hat{\Psi}^{\mathrm{Gibbs}}) - R(\Psi^\star) \leq \Xi \inf_{\substack{S \subset \{1,\dots,p\} \\ m \in \{1,\dots,K\}^{|S|}}} \inf_\theta \left\{ R(\hat{\Psi}_\theta) - R(\Psi^\star) + \frac{|S|}{n} + \frac{\log(n)}{n} \sum_{j \in S} m_j + \frac{1}{n} \log\left(\frac{1}{\varepsilon}\right) \right\},$$

où Ξ est une constante positive ne dépendant que de σ, C, L, α et β.

Ce théorème admet l'interprétation suivante : le risque théorique de notre estimateur est borné par le meilleur risque atteignable sur l'ensemble des modèles, plus un terme $\frac{|S|}{n} + \frac{\log(n)}{n} \sum_{j \in S} m_j$ dépendant de la taille du modèle.

Il s'agit enfin d'échantillonner selon la loi $\hat{\rho}_\delta$. Cette loi, de dimension très grande, représente une gageure pour les algorithmes MCMC classiques, pour lesquels la vitesse de convergence chute drastiquement lorsque la dimension de la loi à échantillonner augmente. Nous proposons donc une version de l'algorithme Subspace Carlin & Chib, dont l'idée essentielle consiste à échantillonner, tout au long de la chaîne, dans des voisinages correctement choisis du modèle courant.

L'algorithme (codé à l'aide du logiciel R) ainsi que ses performances sur données simulées et réelles seront présentés oralement.

Bibliographie

[1] Alquier, P. and Biau, G. (2011), *Sparse Single-Index Model*, preprint. `http://www.lsta.upmc.fr/BIAU/ab.pdf`
[2] Catoni, O., *Statistical Learning Theory and Stochastic Optimization*, Lectures on Probability Theory and Statistics, École d'été de Probabilités de Saint-Flour XXXI, Springer, 2004.
[3] Catoni, O., *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of IMS Lecture Notes Monograph Series. Institute of Mathematical Statistics, 2007.
[4] Friedman, J. H. and Stuetzle, W. (1981), *Projection Pursuit Regression*, Journal of the American Statistical Association, 76:817–823.
[5] Hastie, T. J. and Tibshirani, R. J., *Generalized Additive Models*, Chapman & Hall/CRC, 1990.
[6] Hastie, T. J., Tibshirani, R. J., and Friedman, J. H., *The Elements of Statistical Learning, 2nd Edition*, Springer-Verlag, 2008.
[7] McAllester, D. (1999), *Some PAC-Bayesian Theorems*, COLT'98 Proceedings, Machine Learning 37(3): 355-363.
[8] Meier, L., van de Geer, S. and Bühlmann, P. (2009), *High-dimensional Additive Modeling*, The Annals of Statistics, 37: 3779-3821.
[9] Petralias, A. and Dellaportas, P. (2011), *An MCMC model search algorithm for regression problems*, preprint. `http://stat-athens.aueb.gr/~apet/model_choice_17_website.pdf`
[10] R Development Core Team (2011), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
[11] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009), *Sparse Additive Models*, Journal of the Royal Statistical Society, Series B, 71(5):1009-1030.
[12] Shawe-Taylor, J. and Williamson, R. C. (1997), *A PAC Analysis of a Bayes Estimator*, COLT'97 Proceedings.