MAP553 APPRENTISAGE STATISTIQUE

Projet individuel

Année 2014
A. Tsybakov

Les rapports sont à envoyer à `alexandre.tsybakov@ensae.fr` au plus tard le 10/12/2014

ADAPTATION PAR LA MÉTHODE DE LEPSKI

Considérons le modèle de régression:

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \ldots, n, \tag{1}$$

où $f : [0,1] \to \mathbf{R}$ est une fonction à estimer, $X_i = i/n$ et les ξ_i sont des variables i.i.d. de loi $\mathcal{N}(0,1)$.

On s'intéresse à l'estimation adaptative de f à partir de l'échantillon $Y_1, \ldots, Y_n$. La *méthode de Lepski* est une méthode d'adaptation basée sur des comparaisons implicites du bias avec la variance. Elle ne fait pas recours à l'estimation sans biais du risque que l'on utilise dans les critères de type C_p ou validation croisée qui ne sont adaptés qu'au risque MISE. A la différence de ces derniers, la méthode de Lepski permet de construire des estimateurs adaptatifs pour un spectre très large des mesures de risque.

Le but de ce projet est d'étudier la méthode de Lepski [2] pour le risque au point fixé (MSE). Cette étude mettra en évidence la différence fondamentale de l'adaptation au point fixé (= adaptation avec le risque MSE) avec celle en norme L_2 (= adaptation avec le risque MISE). Notamment, l'adaptation au point fixé n'est pas possible sans perte logarithmique en vitesse de convergence par rapport à l'estimation non-adaptative. Autrement dit, il y a un inévitable "prix logarithmique à payer" si l'on veut atteindre l'adaptation avec le risque MSE. Ceci est connu sous le nom de *phénomène de Lepski*. La situation est donc très différente de l'estimation avec le risque MISE analysée en détail dans le cours, où il n'y a aucun prix à payer pour l'adaptation.

1 Un estimateur basique de régression et ses propriétés

Soit $m \geq 1$ un entier. Considérons la partition de $[0,1]$ en m intervalles $\Delta_1, \ldots, \Delta_m$ de longeur $h = 1/m$ chacun. L'estimateur *régressogramme* de la fonction f est une fonction constante par morceaux définie par :

$$\hat{f}_h(x) = \frac{1}{m_j} \sum_{i=1}^{n} Y_i I(X_i \in \Delta_j), \text{ pour } x \in \Delta_j, \quad j = 1, \ldots, m,$$

où $m_j = \sum_{i=1}^{n} I(X_i \in \Delta_j)$.

Dans la suite, fixons un point $x_0 \in [0,1]$.

Montrez que la variance de $\hat{f}_h(x_0)$ est de l'ordre de $O(1/nh)$. Montrez que le biais de $\hat{f}_h(x_0)$ est de l'ordre de $O(h^\beta)$ si f appartient à la classe de Hölder $\Sigma(\beta, L)$ pour un $\beta \in]0,1[$ et $L > 0$. La valeur $h^*_\beta = n^{-1/(2\beta+1)}$ est un choix de h réalisant l'équilibre biais–variance (ici et dans la suite, on suppose sans perte de généralité que $1/h$ est un entier).

2 Borne inférieure minimax pour le risque MSE

Montrer que $n^{-2\beta/(2\beta+1)}$ est la vitesse optimale au sens minimax pour le risque MSE sur la classe de Hölder $\Sigma(\beta, L)$ avec $\beta \in]0,1[$ et $L > 0$. Utiliser le résultat de la question précédente et la borne inférieure minimax

$$r^* := \inf_{T_n} \sup_{f \in \Sigma(\beta,L)} \mathbf{E}_f(|T_n(x_0) - f(x_0)|^2) \geq c n^{-2\beta/(2\beta+1)} \tag{2}$$

où $\inf_{T_n}$ désigne le minimum sur tous les estimateurs, $c > 0$ est une constante et $\mathbf{E}_f$ est l'espérance par rapport à la distribution $\mathbf{P}_f$ du vecteur $(Y_1, \ldots, Y_n)$ vérifiant (1).

Prouver (2) en utilisant le schéma suivant.

1. Réduction au problème de "deux hypothèses". Montrer que, pour deux fonctions quelconques f_1 et f_2 dans $\Sigma(\beta, L)$,

   $$r^* \geq \frac{1}{4}(f_1(x_0) - f_2(x_0))^2 \int \min(\pi_{f_1}, \pi_{f_2}),$$

   où π_{f_j} est la densité de $\mathbf{P}_{f_j}$ par rapport à la mesure de Lebesgue.

2. Montrer que

   $$\int \min(\pi_{f_1}, \pi_{f_2}) \geq 1 - \frac{1}{2}\sqrt{\chi^2(\pi_{f_1}, \pi_{f_2})}$$

   où

   $$\chi^2(\pi_{f_1}, \pi_{f_2}) = \int \frac{(\pi_{f_1} - \pi_{f_2})^2}{\pi_{f_2}}$$

   est la divergence du χ^2 entre π_{f_1} et π_{f_2}.

3. Conclure en choisissant deux fonctions appopriées f_1 et f_2 dans $\Sigma(\beta, L)$, de la même manière que dans le Chapitre 2 de [4], et en évaluant la divergence du χ^2 entre les densités π_{f_1} et π_{f_2}.

3 Adaptation par la méthode de Lepski. Cas de deux régularités

Supposons maintenant que l'on ne connait pas la régularité β, mais on sait seulement que $\beta \in B$ où B est un sous-ensemble de $]0,1[$. Considérons le cas le plus simple quand B ne contient que deux éléments: $B = \{\beta_1, \beta_2\}$, $0 < \beta_1 < \beta_2 < 1$.

1. Peut-on construire un estimateur ayant la vitesse $n^{-2\beta/(2\beta+1)}$ à la fois pour $\beta = \beta_1$ et $\beta = \beta_2$ (il s'agit toujours de la vitesse pour le risque MSE sur la classe de Hölder $\Sigma(\beta, L)$)? S'inspirant du Chapitre 15 de [1] (Corollaire 15.3), montrer que la réponse à cette question est négative. Autrement dit, il n'existe pas d'estimateur adaptatif optimal pour l'estimation dans un point fixé.

2. Néanmoins, il existe un estimateur adaptatif avec perte logarithmique de vitesse. Cet estimateur est défini par la méthode de Lepski:

$$\hat{f}^*(x_0) = \begin{cases} \hat{f}_{h_1}(x_0) & \text{si } |\hat{f}_{h_1}(x_0) - \hat{f}_{h_2}(x_0)| > c_0 h_1^{\beta_1}, \\ \hat{f}_{h_2}(x_0) & \text{sinon,} \end{cases}$$

où $h_1 = ((\log n)/n)^{1/(2\beta_1+1)}$, $h_2 = n^{-1/(2\beta_2+1)}$ et $c_0 > 0$ est une constante appropriée. Montrer que l'estimateur de Lepski $\hat{f}^*$ vérifie

$$\max_{\beta=\beta_1,\beta_2} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f(|\hat{f}^*(x_0) - f(x_0)|^2/\psi_n(\beta)) \le C,$$

pour une constante $C < \infty$, où

$$\psi_n(\beta) = \begin{cases} ((\log n)/n)^{2\beta_1/(2\beta_1+1)} & \text{si } \beta = \beta_1, \\ n^{-2\beta_2/(2\beta_2+1)} & \text{si } \beta = \beta_2. \end{cases}$$

4 Métode de Lepski : cas général

Soit maintenant $B = [\beta_{\min}, \beta_{\max}]$ avec $0 < \beta_{\min} < \beta_{\max} < 1$. Considérons un ensemble discret

$$\mathcal{B} = \{\beta_{\min} = \beta_1 < \beta_2 < \ldots < \beta_N = \beta_{\max}\}$$

tel que $\beta_j - \beta_{j-1} \asymp 1/(\log n)$. Notons $h_\beta = ((\log n)/n)^{1/(2\beta+1)}$.

L'estimateur de Lepski est alors défini par:

$$\hat{f}^*(x_0) = \hat{f}_{h_{\hat{\beta}}}(x_0),$$

où

$$\hat{\beta} := \max\Big\{\beta \in \mathcal{B} \ : \ |\hat{f}_{h_\beta}(x_0) - \hat{f}_{h_{\beta'}}(x_0)| \le c_0 h_{\beta'}^{\beta'}, \ \forall \beta' < \beta, \ \text{t.q. } \beta' \in \mathcal{B}\Big\}.$$

Montrer que

$$\sup_{\beta\in B} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f(|\hat{f}^*(x_0) - f(x_0)|^2((\log n)/n)^{-2\beta/(2\beta+1)}) \le C$$

pour une constante $C < \infty$.

Nota: Les réponses à toutes les questions doivent comporter les démonstrations complètes et détaillées.

Bibliographie

1. Korostelev, A., Korosteleva, O. (2011) *Mathematical Statistics.* Graduate Studies in Mathematics, vol. 119. Amercian Mathematical Society. Providence, RA.

2. Lepskii, O.V. (1990) On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and its Applications*, vol. 35, pp. 454-466.

3. Lepski, O., Mammen, E., Spokoiny, V.(1997) Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Annals of Statistics*, vol. 25, pp. 929–947.

4. Tsybakov, A.B. (2009) *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer, New York e.a.