MAP553 Apprentissage Statistique

PC5 : Estimation adaptative de la régression

1 Le critère C_p de Mallows

Considérons le modèle de régression :

$$y = f + \xi, \quad \text{où} \quad f = \begin{pmatrix} f(X_1) \\ \vdots \\ f(X_n) \end{pmatrix}, \quad y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix},$$

où les X_i sont des points déterministes, les ξ_i sont des variables aléatoires indépendantes de moyenne 0 et de variance σ^2. Soit S une matrice donnée déterministe de taille $n \times n$ et $\hat{f}_{\mathrm{S}} = \mathrm{S}y$ un estimateur linéaire de f. Le risque quadratique de l'estimateur $\hat{f}_{\mathrm{S}}$ est défini par

$$R(f, \hat{f}_{\mathrm{S}}) = \mathbb{E}\|\hat{f}_{\mathrm{S}} - f\|^2 = \|\mathrm{S}f - f\|^2 + \frac{\sigma^2}{n}\mathrm{Tr}(\mathrm{S}^T\mathrm{S}),$$

où $\|f\|^2 = \frac{1}{n}\sum_{i=1}^n f^2(X_i)$.
Considérons une famille d'estimateurs $\{\hat{f}_{\mathrm{S}},\ \mathrm{S} \in \mathcal{S}\}$ où $\mathcal{S}$ est une classe finie de matrices symétriques $n \times n$ (pensez par exemple à différents estimateurs à noyau correspondant à différents choix de fenêtre). Idéalement, on aimerait choisir le "meilleur" estimateur dans cette famille, à savoir celui ayant le plus petit risque. Comme le risque de chacun des estimateurs nous est inconnu (il dépend de f), nous ne pouvons pas directement choisir celui ayant le plus petit risque. Mallows a introduit un estimateur sans biais du risque de $\hat{f}_{\mathrm{S}}$ (à un terme additif près)

$$C_p(\mathrm{S}) = \|\hat{f}_{\mathrm{S}} - y\|^2 + \frac{2\sigma^2}{n}\mathrm{tr}(\mathrm{S}), \qquad \textbf{(}C_p\textbf{ de Mallows)}$$

et il a proposé de sélectionner

$$\tilde{f} = \hat{f}_{\hat{\mathrm{S}}} \ \text{ où } \ \hat{\mathrm{S}} = \operatorname*{argmin}_{\mathrm{S}\in\mathcal{S}} C_p(\mathrm{S}).$$

Lorsque les hypothèses suivantes sur l'ensemble $\mathcal{S}$ sont vérifiées, un résultat de Kneip (1994) permet de contrôler le risque de $\tilde{f}$.

Hypothèses d'ordre

1. $0 \le \mathrm{v}^T\mathrm{S}\mathrm{v} \le \mathrm{v}^T\mathrm{v}, \quad \forall \mathrm{S} \in \mathcal{S}, \quad \forall \mathrm{v} \in \mathbb{R}^n$;
2. Les matrices $\mathrm{S}, \mathrm{S}' \in \mathcal{S}$ commutent : $\mathrm{S}\mathrm{S}' = \mathrm{S}'\mathrm{S}, \quad \forall \mathrm{S}, \mathrm{S}' \in \mathcal{S}$;
3. Pour tout $\mathrm{S}, \mathrm{S}' \in \mathcal{S}$ on a ou bien $\mathrm{S} \ge \mathrm{S}'$ ou bien $\mathrm{S}' \ge \mathrm{S}$.

Sous ces hypothèses nous avons alors *l'inégalité oracle* suivante.

Théorème 1 (Kneip 1994) *Supposons que la classe $\mathcal{S}$ vérifie les hypothèses d'ordre, et supposons que les ξ_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Alors, il existe une constante $d < \infty$ telle que, pour tout f et tout $\varepsilon > 0$, l'estimateur adaptatif $\tilde{f}$ vérifie*

$$\mathbb{E}\|\tilde{f} - f\|^2 \le (1+\varepsilon)\min_{\mathrm{S}\in\mathcal{S}} \mathbb{E}\|\hat{f}_{\mathrm{S}} - f\|^2 + \frac{d}{\varepsilon n}. \tag{1}$$

1. Soit $\varphi_1, \ldots, \varphi_M$ un système de M fonctions données et notons $\Phi_j = (\varphi_j(X_1), \ldots, \varphi_j(X_n))^T$. Soit $\mathrm{S}_M \in \mathbb{R}^{n \times n}$ la matrice de projection sur l'espace engendré par $\Phi_1, \ldots, \Phi_M$. Alors,
$$\hat{f}_{\mathrm{S}_M} = \mathrm{S}_M y$$
est l'estimateur des moindres carrés de f. Considérons la famille d'estimateurs $\{\hat{f}_{\mathrm{S}_M}, M = 1, \ldots, n-1\}$. Montrer que la famille $\mathcal{S}$ correspondante vérifie les hypothèses d'ordre et donc l'estimateur adaptatif correspondant $\tilde{f}$ vérifie l'inégalité d'oracle (1) si les ξ_i sont i.i.d. $\mathcal{N}(0, \sigma^2)$.
2. Soit maintenant $\{\varphi_j\}_j$ la base trigonométrique orthonormée sur $[0,1]$ et $X_i = i/n, i = 1, \ldots, n$. Expliciter la trace $\mathrm{tr}(\mathrm{S}_M)$ et la forme du critère C_p dans ce cas.
3. On rappelle que, sous les hypothèses de la question précédente, pour tout $\beta \geq 1, L > 0$, et pour $M_\beta = \lceil n^{1/(2\beta+1)} \rceil$
$$\sup_{f \in W(\beta, L)} \mathbb{E}\|\hat{f}_{\mathrm{S}_{M_\beta}} - f\|^2 \leq \mathrm{C}(\beta, L) n^{-\frac{2\beta}{2\beta+1}},$$
où $W(\beta, L)$ est la classe de Sobolev (cf. Définition 2.10 du cours). Montrer que $\tilde{f}$, l'estimateur adaptatif par la méthode C_p sur l'ensemble $\mathcal{S} = \{\mathrm{S}_M, M = 1, \ldots, n-1\}$, vérifie
$$\sup_{f \in W(\beta, L)} \mathbb{E}\|\tilde{f} - f\|^2 \leq \mathrm{C}'(\beta, L) n^{-\frac{2\beta}{2\beta+1}}, \quad \forall \beta \geq 1, L > 0.$$
Commenter le résultat.

2 Validation croisée pour la régression non-paramétrique

Le critère C_p n'est applicable qu'aux estimateurs linéaires de la régression, i.e., les estimateurs de la forme $\mathrm{S}y$. Considérons maintenant une autre méthode, celle de validation croisée, qui permet d'obtenir un estimateur adaptatif dans une situation plus générale.
On se met dans le même cadre que dans l'Exercice 1, mais on suppose maintenant que les X_i sont aléatoires et
$$\mathbb{E}(\xi_i | X_1, \ldots, X_n) = 0 \quad \text{et} \quad \mathbb{E}(\xi_i \xi_k | X_1, \ldots, X_n) = \sigma^2 \delta_{jk}.$$
Soit $\{\hat{f}_{n,\tau}, \tau \in \mathcal{T}\}$ une famille d'estimateurs basée sur l'échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. et dépendant d'un paramètre $\tau \in \mathcal{T}$. Supposons que $\hat{f}_{n,\tau}$ est entièrement déterminé par la donnée de $(X_1, Y_1), \ldots, (X_n, Y_n)$ et de τ et que les couples (X_i, Y_i) sont i.i.d. Définissons le risque quadratique de l'estimateur $\hat{f}_{n,\tau}$ par
$$r_{n,\tau}(f) = \mathbb{E} \int (\hat{f}_{n,\tau}(x) - f(x))^2 P_X(dx),$$
où P_X désigne la distribution des X_i.
Notre but est de choisir le paramètre τ de façon adaptative, afin qu'il donne une valeur proche du minimum du risque quadratique. Le critère de validation croisée "leave-one-out" est défini par
$$CV(\tau) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_{n,\tau,-i}(X_i))^2,$$
où $\hat{f}_{n,\tau,-i}$ est l'estimateur de la même forme que $\hat{f}_{n,\tau}$ basé sur l'échantillon de taille $n-1$: $(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)$. En supposant que $\hat{f}_{n,\tau,-i}(x)$ a la même distribution que $\hat{f}_{n,\tau,-1}(x)$ pour tout i, x, montrer que
$$\mathbb{E}[CV(\tau)] = r_{n-1,\tau}(f) + \sigma^2.$$
Commenter le résultat.