MAP553 Apprentissage Statistique

PC4 : régression non-paramétrique

1 Espace des splines

Soit $0 < X_1 < \cdots < X_n < 1$ et $n \geq \ell$, $n, \ell \in \mathbb{N}$. Notons S_n^ℓ l'espace des splines de degré $2\ell - 1$ avec les nœuds $X_1, \ldots, X_n$.

1. A partir des contraintes régissant la définition de S_n^ℓ, intuiter la dimension de S_n^ℓ.
2. Montrer à partir des conditions (i) et (ii) qu'un spline s s'écrit sous la forme

$$s(x) = \sum_{j=0}^{2\ell-1} c_j x^j + \sum_{i=1}^{n} d_i (x - X_i)_+^{2\ell-1}.$$

3. Déduire de la condition (iii) que

$$c_\ell = \ldots = c_{2\ell-1} = 0 \quad \text{et} \quad \sum_{i=1}^{n} d_i (1 - X_i)^j = 0, \quad j = 0, \ldots, \ell - 1.$$

4. Prouver le résultat intuité à la première question.

2 Estimateur spline

Soit $0 < X_1 < \cdots < X_n < 1$ et $n \geq \ell$, $n, \ell \in \mathbb{N}$. L'estimateur spline est défini comme solution du problème de minimisation :

$$f_n^{sp} = \underset{\int_0^1 (f^{(l)})^2 < \infty}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int_0^1 (f^{(l)})^2 \right].$$

On admettra que ce problème de minimisation admet une unique solution f_n^{sp} et que cette solution appartient à l'espace des splines S_n^ℓ. Considérons une base $\{\varphi_j,\ j = 1, \ldots, n\}$ de S_n^ℓ. Notons $Y = (Y_1, \ldots, Y_n)^T$, $\varphi(x) = (\varphi_1(x), \ldots, \varphi_n(x))^T$ et introduisons la matrice Φ d'éléments $\Phi_{i,j} = \varphi_j(X_i)$, pour $i, j = 1, \ldots, n$.

1. Montrer que $f_n^{sp}(x) = \varphi(x)^T \theta^{sp}$ où θ^{sp} est solution de

$$\theta^{sp} = \arg\min_{\theta \in \mathbb{R}^n} \|Y - \Phi\theta\|^2 + \lambda \theta^T H \theta,$$

avec $H = \int_0^1 [\varphi(x)^{(l)}][\varphi(x)^{(l)}]^T dx$ et $\|.\|$ la norme euclidienne.

2. En déduire que

$$f_n^{sp}(x) = \varphi(x)^T(\Phi^T\Phi + \lambda H)^{-1}\Phi^T Y = \sum_{i=1}^n W_{ni}^{SP}(x)Y_i,$$

où les $W_{ni}^{SP}(x)$ sont des poids que l'on précisera.

3. Supposons que $Y_i = Q(X_i)$, pour $i = 1, \ldots, n$, avec Q un polynôme de degré $\leq \ell - 1$. Montrer que $f_n^{sp}(x) = Q(x)$ pour tout $x \in [0, 1]$.
4. En déduire que les poids $W_{ni}^{SP}(x)$ de l'estimateur spline vérifient :

$$\sum_{i=1}^n W_{ni}^{SP}(x) = 1 \quad \text{et} \quad \sum_{i=1}^n (x - X_i)^j W_{ni}^{SP}(x) = 0, \quad j = 1, \ldots, \ell - 1.$$

3 Lien entre l'estimateur spline et les estimateurs à noyaux

On considère le modèle de régression non-paramétrique

$$Y_i = f(X_i) + \xi_i, \ i = 1, \ldots, n,$$

où f est une fonction de $[0, 1]$ dans $\mathbb{R}$. Les variables aléatoires ξ_i sont supposées indépendantes, avec

$$\mathbf{E}(\xi_i) = 0, \quad \mathbf{E}(\xi_i^2) = \sigma_\xi^2 < \infty,$$

et $X_i = i/n$, pour $i = 1, \ldots, n$.

L'estimateur spline de lissage (cubique) est défini comme solution du problème de minimisation :

$$f_n^{sp} = \arg\min_{f \in W} \left[\frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 + \kappa \int_0^1 (f'')^2 \right] \tag{1}$$

où $\kappa > 0$ est un paramètre de lissage et W est l'ensemble de toutes les fonctions $f : [0, 1] \to \mathbf{R}$ telles que f' est absolument continue, vérifiant la condition de périodicité : $f(0) = f(1), \ f'(0) = f'(1)$. Dans ce cas (1) définit l'estimateur *spline périodique*.

1. Montrer que le problème de minimisation (1) est équivalent au problème de minimisation :

$$\min_{\{b_j\}} \sum_{j=1}^\infty \Big(-2\hat{\theta}_j b_j + b_j^2(\kappa\pi^4 a_j^2 + 1)[1 + O(n^{-1})] \Big), \tag{2}$$

où les b_j sont les coefficients de Fourier de f, le terme $O(n^{-1})$ est uniforme en $\{b_j\}$,

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^n Y_i \varphi_j(X_i),$$

$\{\varphi_j\}_{j=1}^\infty$ est la base trigonométrique, et les a_j sont définis par :

$$a_j = \begin{cases} j^2, & \text{pour } j \text{ pair}, \\ (j-1)^2, & \text{pour } j \text{ impair}. \end{cases}$$

On pourra admettre que les $\{(\varphi_j(X_1), \ldots, \varphi_j(X_n))/\sqrt{n}, \ j = 1, \ldots, n\}$ forment une base orthonormée de $\mathbb{R}^n$.

2. En remplaçant $O(n^{-1})$ par 0 dans (2), trouver la solution de (2) et en déduire l'approximation de l'estimateur spline périodique par l'estimateur par projection avec poids suivant :

$$\tilde{f}_n^{sp}(x) = \sum_{j=1}^{\infty} \lambda_j \, \hat{\theta}_j \, \varphi_j(x)$$

avec

$$\lambda_j = \frac{1}{1 + \kappa \pi^4 a_j^2} \, .$$

3. En remplaçant $O(n^{-1})$ par 0 dans (2), montrer que l'estimateur spline f_n^{sp} est approché (pour κ assez petit) par l'estimateur à noyau de la forme :

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)$$

avec la fenêtre $h = \kappa^{1/4}$ et le noyau de Silverman

$$K(u) = \int_{-\infty}^{\infty} \frac{\cos(2\pi t u)}{1 + (2\pi t)^4} \, dt = \frac{1}{2} \exp\left(-\frac{|u|}{\sqrt{2}}\right) \sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right).$$

Rappel :

$$\mathcal{F}[K](\omega) = \frac{1}{1 + \omega^4}$$