MAP553 Apprentissage Statistique

PC3 : Estimation d'une densité

1 Estimateur par projection d'une densité de probabilité

Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de densité de probabilité $p \in L_2([0,1], dx)$. Pour $N \in \mathbb{N}$ fixé, l'estimateur par projection de p (Tchentsov, 1962) est défini par

$$\widehat{p}_n(x) = \sum_{j=1}^{N} \hat{c}_j \varphi_j(x), \quad \text{avec} \quad \hat{c}_j = \frac{1}{n} \sum_{i=1}^{n} \varphi_j(X_i),$$

$\{\varphi_j\}_{j=1}^{\infty}$ étant une base orthonormée de $L_2([0,1], dx)$.

1. Montrer que $\hat{c}_j$ sont des estimateurs sans biais des coefficients de Fourier

$$c_j = \langle p, \varphi_j \rangle = \int_0^1 p(x)\varphi_j(x)dx$$

et donner la variance de $\hat{c}_j$.

2. Exprimer le risque quadratique intégrée (MISE) de l'estimateur $\widehat{p}_n$ en fonction de $p(\cdot)$ et de la base $\{\varphi_j\}_{j=1}^{\infty}$. On le notera MISE($N$).

3. Supposons maintenant que $\{\varphi_j\}_{j=1}^{\infty}$ est la base orthonormée trigonométrique sur $[0,1]$ définie par

$$\varphi_1(x) = 1, \ \varphi_j(x) = \sqrt{2}\cos(2\pi k x) \text{ si } j = 2k, \ \varphi_j(x) = \sqrt{2}\sin(2\pi k x) \text{ si } j = 2k+1$$

pour $k = 1, 2, \ldots$.

   (a) Montrer que le MISE de $\widehat{p}_n$ est borné par

$$\frac{N}{n} + \rho_N\,, \quad \text{avec} \quad \rho_N = \sum_{j=N+1}^{\infty} c_j^2.$$

   (b) Pour $\beta > 0, L > 0$, on définit la classe de Sobolev périodique :

$$W^{per}(\beta, L) = \Big\{ f \in L_2([0,1], dx) \text{ telle que } \sum_{j=1}^{\infty} a_j^2 \langle f, \varphi_j \rangle^2 \leq L^2/\pi^{2\beta} \Big\},$$

avec $a_j = j^\beta$ pour j pair et $a_j = (j-1)^\beta$ pour j impair. Déduire du (a) que, uniformement sur la classe de densités de probabilité appartenant à $W^{per}(\beta, L)$, $\beta > 0, L > 0$, la valeur optimale de la MISE est $O(n^{-\frac{2\beta}{2\beta+1}})$, pour un choix de l'ordre $N = N_n$ de l'estimateur que l'on précisera.

4. On s'intéresse maintenant au choix adaptatif de N. Utilisons le principe d'estimation sans biais du risque. Montrer que la variable aléatoire

$$\hat{J}(N) = \frac{1}{n-1} \sum_{j=1}^{N} \left[\frac{2}{n} \sum_{i=1}^{n} \varphi_j^2(X_i) - (n+1)\hat{c}_j^2 \right]$$

vérifie

$$\mathbb{E}_p(\hat{J}(N)) = \mathrm{MISE}(N) - \int p^2.$$

En déduire une méthode de choix adaptatif de N.

2 Estimation multidimensionnelle : le fléau de la dimension

L'estimateur de Parzen – Rosenblatt admet une version multidimensionnelle. Supposons que l'on dispose de n vecteurs aléatoires $\mathbf{x}_1, \ldots, \mathbf{x}_n$ tels que tous les $\mathbf{x}_i$ sont i.i.d. et de densité p sur $\mathbb{R}^d$. Soit $\mathbf{x} = (x_1, \ldots, x_d)$ un point fixé de $\mathbb{R}^d$. L'estimateur à noyau de la valeur $p(\mathbf{x})$ est défini par

$$\hat{p}_n(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{X_{ij} - x_j}{h} \right)$$

où K est un noyau intégrable sur $\mathbb{R}$ tel que $\int K = 1$, $h > 0$ est une fenêtre et X_{ij} désigne la $j$ème coordonnée du vecteur aléatoire $\mathbf{x}_i$. On supposera dans la suite que le noyau K est à support compact, que la fonction $|K|$ est uniformément bornée et que p vérifie la condition de Hölder :

$$|p(\mathbf{x}) - p(\mathbf{x}')| \leq L \sum_{j=1}^{d} |x_j - x'_j|^\beta, \quad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d,$$

où $0 < \beta < 1$ et L sont deux constantes.

1. Évaluer la variance de $\hat{p}_n(\mathbf{x})$. On pourra admettre qu'il existe une constante $A > 0$ telle que toute densité de probabilité (β, L)-Hölder est uniformément bornée par A (voir la preuve du Théorème 2.1 du cours).
2. Évaluer le biais de $\hat{p}_n(\mathbf{x})$.
3. En déduire une majoration pour le risque quadratique MSE au point $\mathbf{x}$. Chercher ensuite la valeur h_n^* fournissant le minimum par rapport à h du majorant ainsi obtenu et donner la vitesse de convergence optimale. Noter que la vitesse décroît de façon exponentielle en d (le *fléau de la dimension*).