MAP553 Apprentissage Statistique

PC2 : Estimation d'une densité

Pour une fonction $g \in L^2(\mathbb{R})$, nous noterons $\mathcal{F}[g]$ sa transformée de Fourier. En particulier, si $g \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$,

$$\mathcal{F}[g](\omega) = \int_{-\infty}^{\infty} e^{\mathrm{i}t\omega} g(t)dt, \quad \omega \in \mathbb{R}.$$

Rappels transformée de Fourier

$$\begin{aligned}
\int_{\mathbb{R}} g^2 &= \frac{1}{2\pi} \int_{\mathbb{R}} |\mathcal{F}[g]|^2 \\
\partial^k \mathcal{F}[g](\omega) &= \mathcal{F}[(\mathrm{i}x)^k g(x)](\omega) \quad \text{si} \quad \int_{\mathbb{R}} |x|^k |g(x)|\, dx < +\infty \\
\mathcal{F}[\partial^k g](\omega) &= (-\mathrm{i}\omega)^k \mathcal{F}[g](\omega) \\
\mathcal{F}[g(z - .)](\omega) &= e^{\mathrm{i}z\omega} \mathcal{F}[g](-\omega) \\
\mathcal{F}\left[g(./h)\right](\omega) &= h\, \mathcal{F}[g](h\omega)
\end{aligned}$$

1 Analyse Fourier du risque des estimateurs à noyau

Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de densité de probabilité p par rapport à la mesure de Lebesgue sur $\mathbb{R}$. Considérons l'estimateur à noyau de p défini par

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right).$$

Supposons que $K \in L^2(\mathbb{R})$ et que c'est une fonction paire (un *noyau symétrique*). Soit ϕ la fonction caractéristique associée à la densité p,

$$\phi(\omega) = \int_{-\infty}^{\infty} e^{\mathrm{i}t\omega} p(t)dt, \quad \omega \in \mathbb{R}.$$

Définissons la fonction caractéristique empirique

$$\phi_n(\omega) = \frac{1}{n} \sum_{j=1}^{n} e^{\mathrm{i}X_j\omega}, \quad \omega \in \mathbb{R}.$$

On écrira, pour abréger,

$$\widehat{K}(\omega) = \mathcal{F}[K](\omega).$$

1. Montrer que, pour tout noyau symétrique $K \in L^2(\mathbb{R})$,

$$\mathcal{F}[\hat{p}_n](\omega) = \phi_n(\omega)\widehat{K}(h\omega).$$

2. Montrer que les premier et le deuxième moments de $\phi_n(\omega)$ valent :

$$\mathbb{E}[\phi_n(\omega)] = \phi(\omega),$$
$$\mathbb{E}[\,|\phi_n(\omega)|^2] = \left(1 - \frac{1}{n}\right)|\phi(\omega)|^2 + \frac{1}{n}$$

et, par conséquent,

$$\mathbb{E}[\,|\phi_n(\omega) - \phi(\omega)|^2] = \frac{1}{n}\left(1 - |\phi(\omega)|^2\right).$$

3. Mesurons l'erreur de l'estimateur $\hat{p}_n$ par son *risque quadratique intégré* :

$$\mathrm{MISE} = \mathbb{E}\int(\hat{p}_n(x) - p(x))^2 dx.$$

Montrer que, pour toute densité de probabilité $p \in L^2(\mathbb{R})$ et tout noyau symétrique $K \in L^2(\mathbb{R})$,

$$\mathrm{MISE} \;=\; \frac{1}{2\pi}\left[\int\left|1 - \widehat{K}(h\omega)\right|^2\left|\phi(\omega)\right|^2 d\omega + \frac{1}{n}\int\left|\widehat{K}(h\omega)\right|^2\left(1 - \left|\phi(\omega)\right|^2\right)d\omega\right]$$

2 Les noyaux d'ordre ℓ

Soit $\ell \geq 1$ un entier. On dit que $K : \mathbb{R} \to \mathbb{R}$ est un *noyau d'ordre* ℓ si les fonctions $u \mapsto u^j K(u), j = 0, 1, \ldots, \ell$, sont intégrables et vérifient

$$\int K(u)du = 1, \qquad \int u^j K(u)du = 0, \quad j = 1, \ldots, \ell.$$

1. Soit $\beta > 0$ et $K \in L^2(\mathbb{R})$ un noyau symétrique vérifiant $\widehat{K} \in L^\infty(\mathbb{R})$. Montrer que la condition

$$\sup_{\omega \in \mathbb{R}\setminus\{0\}} \frac{\left|1 - \widehat{K}(\omega)\right|}{|\omega|^\beta} \leq A \tag{1}$$

est équivalente à

$$\exists\ \omega_0, A_0 < \infty : \qquad \sup_{0 < |\omega| \leq \omega_0} \frac{\left|1 - \widehat{K}(\omega)\right|}{|\omega|^\beta} \leq A_0.$$

2. Montrer que, pour un entier β la condition (1) est vérifiée si K est un noyau d'ordre $\beta - 1$ et

$$\int |u|^\beta \left|K(u)\right| du < \infty.$$

3. Pour quelles valeurs de β les noyaux suivants vérifient-ils la condition (1) :

$$\begin{aligned} K(u) &= \tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{u^2}{2}) \ \ \text{(le noyau gaussien)}, \\ K(u) &= \frac{\sin u}{\pi u} \ \ \text{(le noyau sinc)}, \ \ [\,\text{rappel : } \widehat{K}(\omega) = \mathbf{1}_{[-1,1]}(\omega)\,] \\ K(u) &= \tfrac{3}{4}(1-u^2)I(|u|\le 1) \ \ \text{(le noyau parabolique ou d'Epanechnikov)}. \end{aligned}$$

Même question pour les noyaux définis par leurs transformées de Fourier ($\alpha > 0$) :

$$\begin{aligned} \widehat{K}(\omega) &= \frac{1}{1+|\omega|^\alpha} \qquad \text{(le noyau "spline")}, \\ \widehat{K}(\omega) &= (1-|\omega|^\alpha)_+ \ \ \text{(le noyau de Pinsker)}. \end{aligned}$$

4. Chercher l'ordre maximal du noyau de Silverman

$$K(u) = \tfrac{1}{2}\ \exp(-|u|/\sqrt{2})\sin(|u|/\sqrt{2}+\pi/4).$$

Indication : on admettra que la transformée de Fourier de K est de la forme

$$\widehat{K}(\omega) = \frac{1}{1+\omega^4}\ .$$

3 Vitesses de convergence sur les classes de Sobolev

Supposons que la densité p appartient à une *classe de Sobolev* définie par :

$$\mathcal{P}_{\mathcal{S}}(\beta, L) = \left\{ p \,\Big|\, p \ge 0, \int p(x)dx = 1 \text{ and } \int |\omega|^{2\beta}\big|\phi(\omega)\big|^2 d\omega \le 2\pi L^2 \right\},$$

où $\beta > 0$ et $L > 0$ sont des constantes fixées et $\phi = \mathcal{F}[p]$ désigne la fonction caractéristique associée à p.

1. Soit $\beta \in \mathbb{N}$. Commenter sur le lien entre la condition $\int |\omega|^{2\beta}\big|\phi(\omega)\big|^2 d\omega \le 2\pi L^2$ dans la définition de la classe de Sobolev et la condition sur la dérivée de p :

$$\int \left(p^{(\beta)}(u)\right)^2 du \le L^2.$$

La classe de Sobolev est ainsi une classe de densités suffisamment régulières.

2. Supposons que, pour un $\beta > 0$, la condition (1) est vérifiée et que $K \in L^2(\mathbb{R})$ est un noyau symétrique. Montrer que, si $p \in \mathcal{P}_{\mathcal{S}}(\beta, L)$, alors

$$\mathbb{E}\int (\hat{p}_n(x) - p(x))^2\, dx \le C_1 h^{2\beta} + \frac{C_2}{nh}$$

avec les constantes $C_1 > 0$ et $C_2 > 0$ que l'on précisera.

3. Proposer un choix optimal de la fenêtre h et montrer qu'avec ce choix,

$$\sup_{p\in\mathcal{P}_{\mathcal{S}}(\beta,L)} \mathbb{E}\int (\hat{p}_n(x) - p(x))^2\, dx \le C n^{-\frac{2\beta}{2\beta+1}}$$

où $C > 0$ est une constante ne dépendant que de L, A, β et du noyau K.