Centrale **MA2822**
Année 2018-2019 **Statistiques avancées**

TD2 : Satistiques non paramétriques - estimation d'une densité

Exercice 1

Soit $(\phi_k)_{k\geqslant 0}$ une famille de polynômes tels que :

- pour tout $k \geq 0$, ϕ_k est de degré k ;

- pour tout $k, k' \geq 0$, $\int_{-1}^{1} \phi_k(u)\phi_{k'}(u)\mathrm{d}u = \delta_{k,k'}$.

1. Écrire ϕ_0, ϕ_1 et ϕ_2.
2. Montrer que pour tout $\ell \geqslant 0$, $\sum_{m=0}^{\ell} \phi_m(0)\phi_m(u)\mathbb{1}_{[-1,1]}(u)$ est un noyau d'ordre ℓ.

Exercice 2

Soient $(X_1 \dots, X_n)$ des variables aléatoires i.i.d. de loi de densité $p_\star$ par rapport à la mesure de Lebesgue. On rappelle que, pour tout $h > 0$, l'estimateur J_n^h de l'erreur moyenne quadratique intégrée est sans biais :

$$J_n^h = \int_{\mathbb{R}} p_\star^2(x)\mathrm{d}x + \int_{\mathbb{R}} (\widehat{p}_n^h)^2(x)\mathrm{d}x - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right),$$

où

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

Calculer J_n^h dans le cas du noyau gaussien $K : u \mapsto (2\pi)^{-1/2}\mathrm{e}^{-u^2/2}$.

Exercice 3

Pour tout $m \geqslant 1$ on considère $(C_1, \dots, C_m)$ la partition uniforme de $[0, 1)$ définie, pour tout $1 \leqslant i \leqslant m$, par $C_i = [(i-1/m), i/m)$. Soient $(X_1 \dots, X_n)$ des variables aléatoires i.i.d. de loi de densité $p_\star$ par rapport à la mesure de Lebesgue sur $[0, 1)$. On considère l'estimateur par histogramme de la densité $p_\star$:

$$\widehat{p}_n^h : x \mapsto \frac{1}{nh}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{C_j}(X_i)\mathbb{1}_{C_j}(x),$$

où $h = 1/m$.

1. Évaluer l'erreur quadratique moyenne associée à cet estimateur.
2. Évaluer l'erreur quadratique moyenne intégrée associée à cet estimateur.
3. (Bonus) Dans le cas où $p_\star$ est $\mathcal{C}^2$ sur $\mathbb{R}$, donner un développement asymptotique de l'erreur quadratique moyenne intégrée.

Exercice 4

Soient $(X_1 \dots, X_n)$ des variables aléatoires i.i.d. de loi de densité $p_\star$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$. La densité $p_\star$ est estimée par l'estimateur par fenêtre glissante:

$$\widehat{p}_n^{h_n} : x \mapsto \frac{1}{2nh_n}\sum_{i=1}^{n} \mathbb{1}_{(x-h_n, x+h_n)}(X_i).$$

1. Écrire la décomposition biais-variance pour cet estimateur.

2. Montrer que lorsque $h_n \to 0$ et $nh_n \to +\infty$ alors, pour presque tout x, $\widehat{p}_n^{h_n}(x)$ converge dans L^2 (et donc en probabilité) vers $p_\star(x)$.

3. Considérons une suite $(S_n)_{n\geqslant 1}$ de variables aléatoires indépendantes de lois binomiales de paramètres n et p_n avec $np_n \to +\infty$ et $\limsup p_n < 1$. Pour tout $t \in \mathbb{R}$, en remarquant que

$$p_n \frac{t^2}{np_n(1-p_n)} = O\left(\frac{1}{n}\right) \quad \text{et} \quad p_n^2 \frac{t^2}{np_n(1-p_n)} = O\left(\frac{1}{n}\right) ,$$

effectuer un développement limité de $\mathbb{E}[\mathrm{e}^{itZ_n}]$ où

$$Z_n = \frac{S_n - np_n}{\sqrt{np_n(1-p_n)}} .$$

4. Montrer que :

$$Z_n \Rightarrow \mathcal{N}(0,1) .$$

5. Soit x un réel tel que $f(x) > 0$ et tel que f est dérivable sur dans un voisinage $V(x)$, de dérivée bornée sur ce voisinage. Montrer que si $nh_n \to +\infty$ et $nh_n^3 \to 0$ alors :

$$\sqrt{\frac{2nh_n}{\widehat{p}_n^{h_n}(x)}} \left(\widehat{p}_n^{h_n}(x) - p_\star(x)\right) \Rightarrow \mathcal{N}(0,1) .$$