TD 9 : estimation de densités

Les questions marquées d'un astérisque (*) *sont facultatives.*

Exercice 1. (Distance en variation totale)

1. Rappeler la définition de la distance en variation totale entre deux mesures.

Dans la suite, si f est une fonction à valeurs réelles, on note $f_+(x) = \max(f(x), 0)$ et $f_-(x) = \max(-f(x), 0)$ les parties positive et négative de $f(x)$.

2. Montrer que si P et Q sont deux mesures de probabilité de densités respectives f et g par rapport à une mesure μ, alors
$$\sup_{A \text{ mesurable}} (P(A) - Q(A)) = \int (f-g)_+(x)\mu(\mathrm{d}x)$$
et
$$\sup_{B \text{ mesurable}} (Q(B) - P(B)) = \int (f-g)_-(x)\mu(\mathrm{d}x).$$
Indication : utiliser que $(f-g)(x) = (f-g)_+(x) - (f-g)_-(x)$.

3. Montrer que
$$\int (f-g)_+(x)\mu(\mathrm{d}x) - \int (f-g)_-(x)\mu(\mathrm{d}x) = 0$$
et
$$\int (f-g)_+(x)\mu(\mathrm{d}x) + \int (f-g)_-(x)\mu(\mathrm{d}x) = \|f-g\|_1.$$

4. En déduire $d_{VT}(P,Q) = \int (f-g)_+(x)\mu(\mathrm{d}x) = \frac{1}{2}\|f-g\|_1$.

Exercice 2. (Consistance ponctuelle des estimateurs à noyau) L'objectif de l'exercice est de démontrer le résultat suivant, dont une version a été vue en cours.

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. d'une variable aléatoire réelle X de loi de densité f par rapport à la mesure de Lebesgue. Soit $\hat{f}_{n,K}$ l'estimateur à noyau associé au noyau K et à une taille de fenêtre $h_n > 0$, autrement dit, en notant $K_h(x) = K(x/h)/h$, pour tout $x \in \mathbb{R}$,
$$\hat{f}_{n,K}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x - X_i).$$

Supposons $M = \max(\int_{y\in\mathbb{R}} |y|K(y)\mathrm{d}y, \int_{y\in\mathbb{R}} K(y)^2\mathrm{d}y, \int_{y\in\mathbb{R}} |y|K(y)^2\mathrm{d}y) < +\infty$.

Supposons également f de classe $\mathcal{C}^1$ et $L = \max(\|f\|_\infty, \|f'\|_\infty) < +\infty$. Alors pour tout $x \in \mathbb{R}$, si $h_n \longrightarrow 0$ et $nh_n \longrightarrow +\infty$,
$$\mathbb{E}[(\hat{f}_{n,K}(x) - f(x))^2] \longrightarrow 0.$$

1. Montrer que $\mathbb{E}[\hat{f}_{n,K}(x)] = \int_{y\in\mathbb{R}} K_{h_n}(y)f(x-y)\mathrm{d}y$. On note $(K_{h_n} * f)(x)$ le terme de droite. $K_{h_n} * f$ est appelé le *produit de convolution* de K_{h_n} et f.

2. (Décomposition biais-variance) Montrer que
$$\mathbb{E}[(\hat{f}_{n,K}(x) - f(x))^2] = \mathbb{E}[(\hat{f}_{n,K}(x) - (K_{h_n} * f)(x))^2] + ((K_{h_n} * f)(x) - f(x))^2.$$

3. Montrer que $((K_{h_n} * f)(x) - f(x))^2 \leqslant M^2L^2h_n^2$.
Indication : utiliser que $\int_y K_h(y)dy = 1$ *pour tout* $h > 0$.

4. Justifier la suite d'inégalités suivante :

$$\begin{aligned}
\mathbb{E}[(\hat{f}_{n,K}(x) - (K_{h_n} * f)(x))^2] &= \frac{1}{n}\mathrm{Var}(K_{h_n}(x-X)) \\
&\leqslant \frac{1}{n}\mathbb{E}[K_{h_n}(x-X)^2] \\
&= \frac{1}{nh_n}\int_{y\in\mathbb{R}} K(y)^2 f(x+h_n y)\mathrm{d}y \\
&\leqslant \frac{1}{nh_n}(ML + MLh_n)
\end{aligned}$$

Indications : pour la première égalité, réécrire le terme de variance sous la forme d'une somme de variables i.i.d.. Pour la deuxième, utiliser la formule de König-Huygens (cours / TD 1). Pour la dernière, utiliser que $|f(x+h_n y) - f(x)| \leqslant Lh_n|y|$.

5. En déduire le résultat souhaité.

Exercice 3. On observe les six valeurs suivantes : $0.44, 1.36, 1.01, 0.02, 1.81, 0.48$. On suppose ces observations générées de manière i.i.d. suivant une loi sur $[0,2]$ de densité f. Pour chacun des quatre estimateurs de f suivants :

— L'estimateur par histogrammes de largeur $1/2$,

— L'estimateur à noyau de noyau uniforme (aussi appelé rectangulaire) de largeur de fenêtre 0,5 de la densité f,

— L'estimateur à noyau de noyau $x \mapsto 2\max(1-2|x|, 0)$ (noyau triangulaire de taille de fenêtre $1/2$)

— L'estimateur à noyau gaussien de taille de fenêtre 0,4. Pour rappel, le noyau gaussien de taille de fenêtre σ est la fonction $x \mapsto \exp(-x^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$.

1. Calculer leur valeur en 0,75 et en -0,1.
2. Les tracer (sauf l'estimateur à noyau gaussien).