

Sujet 7

• **Rappels:** Soit f une fonction de densité à estimer. Soit (X_1, \dots, X_n) un échantillon i.i.d. de variables aléatoires distribuées comme une variable aléatoire X dont la loi admet la densité $f(\cdot)$ par rapport à la mesure de Lebesgue sur \mathbb{R} .

↳ L'estimateur à noyau de la densité (obtenu par convolution avec un noyau) est défini pour $x \in \mathbb{R}$ par:

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

où la fenêtre $h > 0$ est le paramètre de lissage et où $K(\cdot)$ est un noyau positif ie $K : \mathbb{R} \rightarrow \mathbb{R}$ est une fonction intégrable telle que $\int_{\mathbb{R}} K(u)du = 1$ et $K(\cdot) \geq 0$.

• **Implémentation au moyen du logiciel R:**

La fonction `density` du logiciel R détermine l'estimateur de la densité par méthode à noyau. Il est possible de choisir à la fois le noyau et la fenêtre. Voici quelques choix possibles de noyau:

noyau	expression
Epanenchnikov	$\frac{3}{4}(1 - u^2)I(u \leq 1)$
gaussien	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$
triangulaire	$(1 - u)I(u \leq 1)$
rectangulaire	$\frac{1}{2}I(u \leq 1)$
biweight= quartic	$\frac{15}{16}(1 - u^2)^2I(u \leq 1)$
cosinus	$\frac{\pi}{4} \cos(\pi u/2)I(u \leq 1)$

Vous pouvez fournir une valeur de votre choix pour la taille de fenêtre h dans l'argument `bw`. Sinon, des méthodes de choix automatiques de la fenêtre sont implémentées. L'argument `bw="nrd0"` correspond au choix

$$h = 0.9 \frac{\hat{\sigma}}{n^{1/5}}$$

où n est la taille de l'échantillon et où $\hat{\sigma} = \min(S_n, R_n/1.349)$ en notant S_n l'écart-type empirique de l'échantillon et R_n l'étendue interquartile de l'échantillon. L'instruction `bw="nrd"` correspond au choix

$$h = 1.06 \frac{\hat{\sigma}}{n^{1/5}}$$

L'instruction `bw="ucv"` correspond à un choix issu de la méthode de la validation croisée. L'idée est de déterminer un choix "optimal" de la fenêtre h noté h_{opt} de la façon suivante:

$$\begin{aligned} h_{\text{opt}} &= \arg \min_{h>0} \text{ISE}(\widehat{f}_h) \\ &= \arg \min_{h>0} \int (\widehat{f}_h(x) - f(x))^2 dx \\ &= \arg \min_{h>0} \left\{ \int (\widehat{f}_h(x))^2 dx - 2 \int \widehat{f}_h(x) f(x) dx + \int f(x)^2 dx \right\} \\ &= \arg \min_{h>0} \left\{ \int (\widehat{f}_h(x))^2 dx - 2 \int \widehat{f}_h(x) f(x) dx \right\}. \end{aligned}$$

Notons

$$J(h) = \int (\widehat{f}_h(x))^2 dx - 2 \int \widehat{f}_h(x) f(x) dx$$

et remarquons que le premier terme est entièrement connu. Ensuite, remarquons que

$$\int \widehat{f}_h(x) f(x) dx = \mathbb{E} \left[\widehat{f}_h(X) | X_1, \dots, X_n \right]$$

pour une variable aléatoire X de densité f , indépendante de (X_1, \dots, X_n) . On estime alors $\mathbb{E} \left[\widehat{f}_h(X) | X_1, \dots, X_n \right]$ sans biais par la méthode dite du "leave-one-out". Cela fournit le critère suivant à optimiser le critère en h :

$$\widehat{J(h)} = \int (\widehat{f}_h(x))^2 dx - \frac{2}{n} \sum_{i=1}^n \widehat{f}_{-i,h}(X_i)$$

où $\widehat{f}_{-i,h}$ est l'estimateur de f calculé avec la fenêtre h mais en ôtant l'observation i , ce qui donne:

$$\widehat{f}_{-i,h}(x) = \frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n K\left(\frac{x - X_j}{h}\right).$$

Notons que la fonction $\widehat{J(h)}$ est un estimateur sans biais de $\text{MISE}(\widehat{f}_h)$ à une constante indépendante de h près puisque

$$\mathbb{E} \left[\widehat{J(h)} \right] = \text{MISE}(\widehat{f}_h) - \int f(x)^2 dx.$$

L'argument `bw.ucv` de la fonction `density` permet d'implémenter ce choix de fenêtre dans le calcul de l'estimateur à noyau.

Soient les conditions suivantes:

(A₁): f est de carré intégrable et deux fois différentiable et sa dérivée seconde est continue bornée et de carré intégrable.

(A₂): le noyau $K : \mathbb{R} \rightarrow \mathbb{R}$ est tel que $\int_{\mathbb{R}} K(x)^2 dx < \infty$, K est symétrique et admet un moment d'ordre 2, donc satisfait $\int xK(x)dx = 0$ et $\int x^2K(x)dx < \infty$.

(A₃): Les fenêtres $h = h_n > 0$ forment une suite telle que $h \xrightarrow{n \rightarrow \infty} 0$ et $nh \xrightarrow{n \rightarrow \infty} \infty$. Sous ces conditions, un équivalent asymptotique du MISE est:

$$\text{AMISE}(\widehat{f}_h) = \frac{\int K(x)^2 dx}{nh} + \frac{h^4}{4} \left(\int x^2 K(x) dx \right)^2 \int f''(x)^2 dx.$$

La fenêtre optimale est alors celle qui minimise le AMISE. Cela fournit:

$$h_{\text{opt}} = \left(\frac{\int K(x)^2 dx}{\left(\int x^2 K(x) dx \right)^2 \int f''(x)^2 dx} \right)^{1/5} \frac{1}{n^{1/5}}.$$

Comme f'' est inconnue, il faut l'estimer: c'est le principe du *plug-in* (injection). On estime f'' par un estimateur à noyau avec une fenêtre que l'on qualifie de "pilote" et dont le choix est loin d'être une question triviale. Notons que les problèmes d'estimation de f et de f'' ne sont pas équivalents... Le AMISE correspondant à ce choix de fenêtre est l'ordre de $1/4^{1/5}$.

L'argument `bw.SJ` de la fonction `density` permet d'implémenter ce choix de fenêtre.

• **Critères objectifs:**

Critères ponctuels: pour un estimateur \hat{f} de f , le biais est ponctuellement donné par

$$b_f(\hat{f}(x)) = \mathbb{E}[\hat{f}(x)] - f(x),$$

la variance est ponctuellement donnée par

$$\text{Var}(\hat{f}(x)) = \mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right)^2\right],$$

et l'écart quadratique moyen est ponctuellement donné par

$$R_f(\hat{f}(x)) = \mathbb{E}\left[\left(\hat{f}(x) - f(x)\right)^2\right].$$

Critères globaux: le carré du biais intégré est donné par

$$\int \left(\mathbb{E}[\hat{f}(x)] - f(x)\right)^2 dx,$$

la variance intégrée est donnée par

$$\int \text{Var}(\hat{f}(x)) dx,$$

et l'écart quadratique moyen intégré (MISE = Mean Integrated Squared Error) est donné par

$$\int \mathbb{E}\left[\left(\hat{f}(x) - f(x)\right)^2\right] dx.$$

Exercice 1.

On note $\varphi_{(m,\sigma^2)}$ la densité de la loi gaussienne de paramètres $m \in \mathbb{R}$ et $\sigma^2 > 0$. Considérons les distributions suivantes:

- la loi gaussienne standard $\mathcal{N}(0, 1)$ de densité $\varphi_{(0,1)}$,
- les lois $\Gamma(1, 1/2)$, $\Gamma(2, 1/2)$ et $\Gamma(5, 1)$,
- la loi triangulaire de densité donnée pour $x \in \mathbb{R}$ par

$$g(x) = \max(1 - |x|, 0),$$

- la loi de Student à 2 degrés de liberté, dilatée, translatée: $X = 0.25Y - 0.5 = (Y - 2)/4$ où $Y \sim T(2)$,
- on note $\tilde{\varphi}_{(0,1)}$ la restriction de $\varphi_{(0,1)}$ à $[-1, 2]$ puis l'on considère la loi de densité définie comme suit:

$$f(x) = \frac{\tilde{\varphi}_{(0,1)}(x)}{\int_{-1}^2 \varphi_{(0,1)}(u) du}, \quad x \in \mathbb{R},$$

- la loi de densité donnée par la fonction en escalier suivante:

$$f(x) = \begin{cases} 1/2 & \text{si } 0 \leq x < 1/2, \\ 2 & \text{si } 1/2 \leq x < 2/3, \\ 5/4 & \text{si } 2/3 \leq x \leq 1, \\ 0 & \text{sinon,} \end{cases}$$

- la loi dont la densité est en dents de scie comme ci-dessous:

$$f(x) = g(x + 9) + g(x + 7) + \dots + g(x - 7) + g(x - 9),$$

- le mélange de lois gaussiennes de densité:

$$\frac{1}{2} \varphi_{(0,0.1)}(x) + \frac{1}{2} \varphi_{(1,0.4)}(x).$$

1. Simuler M échantillons de taille n suivant les distributions exposées ci-dessus pour différentes valeurs de n .
2. Déterminer l'estimateur à noyau de la densité en utilisant différents noyaux et différentes tailles de fenêtre implémentés dans le logiciel R.
3. Analyser le comportement de l'estimateur à noyau à l'aide des critères objectifs précédemment présentés.