Université de Strasbourg
M1 - Magistère
Statistique

Ségolen Geffray
geffray@math.unistra.fr
Année 2018/2019

Sujet 9

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

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

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

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

• **Implémentation au moyen du logiciel** `R`:
La fonction `kde` du package `ks` détermine l'estimateur de la densité avec un noyau gaussien tronqué. Soient les conditions suivantes:
(A_1): f est de carré intégrable et deux fois différentiable et sa dérivée seconde est continue bornée et de carré intégrable.
(A_2): le noyau $K : \mathbb{R} \to R$ est tel que $\displaystyle\int_{\mathbb{R}} K(x)^2 dx < \infty$, K est symétrique et admet un moment d'ordre 2, donc satisfait $\displaystyle\int xK(x)dx = 0$ et $\displaystyle\int x^2 K(x)dx < \infty$.
(A_3): Les fenêtres $h = h_n > 0$ forment une suite telle que $h \overset{n\to\infty}{\to} 0$ et $nh \overset{n\to\infty}{\to} \infty$. Sous ces conditions, un équivalent asymptotique du MISE est:

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

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 $1/^{4/5}$.
La fonction `kde` du package `ks` implémente ce choix de fenêtre par défaut.

• **Estimateur à noyau beta**: Supposons que la densité f est à support compact, disons dans $[0, 1]$ sans perte de généralité. Dans ce cas, l'estimateur à noyau est fortement biaisé au bord

du support. Pour remédier à ces problèmes de bord, Chen (1999) a introduit le noyau beta. Rappelons que l'estimateur à noyau s'écrit usuellement

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)$$

où l'on note

$$K_{x,h}(X_i) = \frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

pour une fenêtre $h > 0$ et pour un noyau K. Le noyau beta s'écrit, pour $0 \leq x, y, \leq 1$,

$$K_{x,h}(y) = f_{B(\alpha_{1,h}(x),\alpha_{2,h}(x)}(y)$$

en notant $f_{B(a,b))}$ pour $a, b > 0$ la densité de la loi Beta de 1ère espèce et avec

$$\begin{array}{llll} \alpha_{1,h}(x) = \rho_h(x) & \alpha_{2,h}(x) = (1-x)/h & x \in [0, 2h[\\ \alpha_{1,h}(x) = x/h & \alpha_{2,h}(x) = (1-x)/h & x \in [2h, 1-2] \\ \alpha_{1,h}(x) = x/h & \alpha_{2,h}(x) = \rho_h(1-x) & x \in]1-2h, 1] \end{array}$$

avec

$$\rho_h(x) = 2h^2 + 5/2 - (4h^4 + 6h^2 + 9/4 - u^2 - u/h)^{1/2}.$$

Lorsque f admet une dérivée seconde continue, Chen (1999) a établi que la fenêtre qui fournit un AMISE optimal est en $O\left(\frac{1}{n^{2/5}}\right)$ et que le AMISE correspondant est alors en $O\left(\frac{1}{n^{4/5}}\right)$.
La fonction `kde.boundary` du logiciel `R` permet de calculer l'estimateur à noyau beta. Pour effectuer les calculs avec le noyau présenté ci-dessus, choisir l'argument `boundary.kernel="beta"`. Cette fonction utilise une fenêtre calculée d'après la méthode du *plug-in*.

• **Critères objectifs**:
Critères ponctuels: pour un estimateur $\widehat{f}$ de f, le biais est ponctuellement donné par

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

la variance est ponctuellement donnée par

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

et l'écart quadratique moyen est ponctuellement donné par

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

Critères globaux: le carré du biais intégré est donné par

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

la variance intégrée est donnée par

$$\int \mathrm{Var}\left(\widehat{f}(x)\right) dx,$$

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

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

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 uniforme sur $[0, 1]$,
- les lois $B(1/2, 1/2)$, $B(2, 2)$ et $B(2, 5)$ où $B(a, b)$ désigne la loi Beta de 1$^{\text{ère}}$ espèce de paramètre (a, b) pour $a, b > 0$,
- la loi triangulaire de densité donnée pour $x \in \mathbb{R}$ par

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

- on note $\widetilde{\varphi}_{(0,1)}$ la restriction de $\varphi_{(0,0.5)}$ à $[0, 1]$ puis l'on considère la loi de densité définie comme suit:

$$f(x) = \frac{\widetilde{\varphi}_{(0,0.5)}(x)}{\int_{-1}^{\infty} \varphi_{(0,0.5)}(u)du}, \quad x \in \mathbb{R},$$

- on note $\widetilde{\varphi}_{(0,1)}$ la restriction de $\varphi_{(0.2,0.5)}$ à $[0, 1]$ puis l'on considère la loi de densité définie comme suit:

$$f(x) = \frac{\widetilde{\varphi}_{(0.2,0.5)}(x)}{\int_{-1}^{2} \varphi_{(0.2,0.5)}(u)du}, \quad x \in \mathbb{R},$$

- 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 la fonction linéaire par morceaux suivante:

$$f(x) = \begin{cases} 10x & \text{si } 0 \leq x < 1/4, \\ 4 - 6x & \text{si } 1/4 \leq x < 1/2, \\ 2 - 2x & \text{si } 1/2 \leq x \leq 1, \\ 0 \quad \text{sinon,} \end{cases}$$

- la loi dont la densité est donnée par la fonction définie par morceaux comme suit:

$$f(x) = \begin{cases} 64x^2 & \text{si } 0 \leq x < 1/4, \\ 6 - 12x & \text{si } 1/4 \leq x < 1/2, \\ 4x - 2 & \text{si } 1/2 \leq x < 3/4, \\ 1/2 & \text{si } 3/4 \leq x \leq 1, \\ 0 \quad \text{sinon,} \end{cases}$$

- la loi de densité:

$$f(x) = \begin{cases} \dfrac{1}{c}\,\varphi_{(1,0.4)}(x) & \text{si } 0 \leq x < 1/2, \\ \dfrac{1}{c}\,\varphi_{(0,0.4)}(x) & \text{si } 1/2 \leq x \leq 1, \end{cases}$$

où $c = \displaystyle\int_0^{1/2} \varphi_{(1,0.4)}(u)du + \int_{1/2}^{1} \varphi_{(1,0.4)}(u)du.$

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 avec le noyau gaussien (tronqué) et le noyau beta en utilisant la fenêtre *plug-in* implémentée dans le logiciel `R`.

3. Analyser le comportement des deux estimateurs à noyau ainsi obtenus à l'aide des critères objectifs précédemment présentés.