Université de Strasbourg
M1 - Magistère
Statistique

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

Sujet 14

• **Rappels**: Soit un échantillon $(X_1, ..., X_n)$ i.i.d. de fonction de répartition F continue, dont admet la densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$, f étant elle-même dérivable, de dérivée continue. On souhaite estimer F, f et f' par la méthode du noyau.
$\mapsto$ 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 $\int_{\mathbb{R}} K(u)du = 1$ 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 $\int_{\mathbb{R}} K(x)^2 dx < \infty$, K est symétrique et admet un moment d'ordre 2, donc satisfait $\int xK(x)dx = 0$ et $\int x^2 K(x)dx < \infty$.
(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 $n^{-4/5}$.
La fonction `kde` du package `ks` implémente ce choix de fenêtre par défaut.

• **Estimateur à noyau de la fonction de répartition**: Soit K un noyau positif et symétrique tel que $\int x^2 K(x)dx < \infty$. Notons $L(x) = \int_{|\infty}^{x} K(u)du$. Supposons que f est différentiable et

que $\int f'(x)^2 dx < \infty$. L'estimateur à noyau de la fonction de répartition est défini pour $x \in \mathbb{R}$ par

$$\widehat{F}_h(x) = \int_{-\infty}^{x} \widehat{f}_h(x)du = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{x} \frac{1}{h}K\left(\frac{u-X_i}{h}\right)du = \frac{1}{n}\sum_{i=1}^{n} L\left(\frac{u-X_i}{h}\right).$$

L'équivalent asymptotique du MISE est

$$AMISE\left(\widehat{F}_h(x)\right) = \frac{\int F(x)(1-F(x))dx}{n} - 2\frac{h}{n}\int xK(x)L(x)dx + \frac{h^4}{4}\left(x^2K(x)dx\right)^2\int f'(x)^2dx.$$

Une fenêtre optimale est choisie en minimisant

$$2\frac{h}{n}\int xK(x)L(x)dx + \frac{h^4}{4}\left(x^2K(x)dx\right)^2\int f'(x)^2dx,$$

le 1er terme du AMISE ne dépendant pas de h. La fonction `kcde` du logiciel `R` permet de calculer l'estimateur à noyau de la fonction de répartition.
Le choix de fenêtre implémenté et utilisé par défaut est la version "plug-in". Son principe est d'injecter un estimateur de la quantité inconnue f' dans l'expression obtenue. On utilise un estimateur à noyau avec une fenêtre "pilote" dont le choix est loin d'être un problème trivial. Notons que le choix d'une fenêtre de l'ordre de $n^{-2/3}$ fournit

$$AMISE\left(\widehat{F}_h(x)\right) = \frac{\int F(x)(1-F(x))dx}{n} + O\left(\frac{1}{n^{4/3}}\right).$$

La fonction `kcde` du logiciel `R` permet de calculer l'estimateur à noyau de la fonction de répartition.
Le choix de fenêtre implémenté et utilisé par défaut est la version "plug-in".

• **Estimateur à noyau de la dérivée de la densité**: L'estimateur à noyau de la dérivée première de f est défini pour $x \in \mathbb{R}$ par

$$\widehat{f'}_h(x) = \frac{1}{nh^2}\sum_{i=1}^{n} K'\left(\frac{x-X_i}{h}\right).$$

Soient les conditions:
(A_1): f' existe, est de carré intégrable, est deux fois différentiable et f''' est continue bornée et de carré intégrable.
(A_2): le noyau $K : \mathbb{R} \to R$ est tel que $\int_{\mathbb{R}} K(x)^2dx < \infty$, K est symétrique et admet un moment d'ordre 2, donc satisfait $\int xK(x)dx = 0$ et $\int x^2K(x)dx < \infty$ et K' existe et est de carré intégrable.
(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:

$$AMISE(\widehat{f'}_h) = \frac{\int K'(x)^2dx}{nh^3} + \frac{h^4}{4}\left(\int x^2K(x)dx\right)^2\int f'''(x)^2dx.$$

La fenêtre optimale est alors celle qui minimise le AMISE. 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... Notons aussi que le choix d'une fenêtre de l'ordre de $n^{-1/7}$ fournit un AMISE l'ordre de $n^{-4/7}$.
La fonction `kdde` du logiciel `R` permet de calculer l'estimateur à noyau de la dérivée de f. Le choix de fenêtre implémenté et utilisé par défaut est la version "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 gaussienne standard $\mathcal{N}(0,1)$ de densité $\varphi_{(0,1)}$,
- 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 de Student à 3 degrés de liberté, dilatée, translatée: $X = 0.25Y - 0.5 = (Y-2)/4$ où $Y \sim T(3)$,
- on note $\widetilde{\varphi}_{(0,1)}$ la restriction de $\varphi_{(0,1)}$ à $[-1,\infty[$ puis l'on considère la loi de densité définie comme suit:

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

- la loi de densité:

  $0.5\varphi_{(0,1)}(x)+0.1\varphi_{(-1,0.1)}(x)+0.1\varphi_{(-0.5,0.1)}(x)+0.1\varphi_{(0,0.1)}(x)+0.1\varphi_{(0.5,0.1)}(x)+0.1\varphi_{(1,0.1)}(x)$.

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 fonction de répartition, de la densité et de la dérivée de la densité en utilisant la fenêtre *plug-in* `hpi` implémentée dans le logiciel `R`, puis 0.1 `hpi` et 10 `hpi`.

3. Analyser le comportement des estimateurs à l'aide des critères objectifs précédemment présentés.