Université de Strasbourg
M1 - Magistère
Statistique - études de cas

Ségolen Geffray
geffray@math.unistra.fr
Année 2017/2018

Sujet 9: Etude de l'estimateur de la densité

• **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.

$\mapsto$ Soient $\phi(.)$ et $\psi)$ deux fonctions orthonormales et soit

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j \in \mathbb{N}, k \in \mathbb{Z}.$$

L'estimateur de $f(.)$ par la méthode des séries orthogonales obtenu en utilisant la base d'ondelettes $(\phi, \psi_{j,k} : j \in \mathbb{N}, k \in \mathbb{Z})$ est défini pour $x \in \mathbb{R}$ par:

$$\widehat{f}_{M,n}(x) = \widehat{c}_0\phi(x) + \sum_{j=0}^{M}\sum_{k=0}^{2^j-1} \widehat{w}_{j,k}\psi_{j,k}(x)$$

avec

$$\widehat{c}_0 = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i),$$
$$\widehat{w}_{j,k} = \frac{1}{n}\sum_{i=1}^{n}\psi_{j,k}(X_i),$$

et avec $M \in \mathbb{N}^*$.

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

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

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 $K(.) \geq 0$.

$\mapsto$ Soit $(I_j)_{j=1,...,J}$ est une partition appropriée à la distribution considérée. Supposons que les intervalles I_j sont tous de même longueur notée $|I_j| = h > 0$. L'estimateur de $f(.)$ au moyen d'un histogramme est défini pour $x \in \mathbb{R}$ par:

$$\widehat{f}_{n,h}(x) = \frac{1}{nh}\sum_{j=1}^{J} N_j I(x \in I_j)$$

avec

$$N_j = \sum_{i=1}^{n} I(X_i \in I_j).$$

• **Implémentation au moyen du logiciel** `R`:
Le package `wavethresh` du logiciel `R` permet d'implémenter l'estimateur de la densité par la méthode des séries orthogonales avec une base d'ondelettes, choisir celle implémentée par défaut.
La fonction `hist` du logiciel `R` détermine l'estimateur de la densité par histogramme.
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 mais vous utiliserez le noyau et la fenêtre fournis par défaut par le logiciel.

Critères objectifs:
• Pour un estimateur $\widehat{f}$ de f, le carré du biais intégré est donné par

$$\int (f_J(x) - f(x))^2 \, dx,$$

la variance intégrée est donnée par

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

et l'écart quadratique moyen intégré est donné par

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

Exercice 1.
On se propose de travailler avec les fonctions de densité qui suivent. Notons $\varphi_{(m,\sigma^2)}$ la densité de la loi gaussienne $\mathcal{N}(m, \sigma^2)$ de paramètres $m \in \mathbb{R}$ et $\sigma^2 > 0$.

1. (a) loi uniforme sur $[0, 1]$,
2. (b) loi Beta
3. (c) loi triangulaire de densité donnée pour $x \in \mathbb{R}$ par
$$f(x) = \max(1 - |x|, 0)$$
4. (d) la fonction en escalier suivante:
$$f(x) = \begin{cases} 3/5 & \text{si } 0 \leq x < 1/3, \\ 9/10 & \text{si } 1/3 \leq x3/4, \\ 51/30 & \text{si } 3/4 \leq x \leq 1, \\ 0 & \text{sinon.} \end{cases}$$
5. (e) la fonction linéaire par morceaux suivante:
$$f(x) = \begin{cases} 12x & \text{si } 0 \leq x < 1/4, \\ 6 - 12x & \text{si } 1/4 \leq x < 1/2, \\ 4x - 2 & \text{si } 1/2 \leq x < 3/4, \\ 4 - 4x & \text{si } 3/4 \leq x \leq 1, \\ 0 \quad \text{sinon.} \end{cases}$$

6. (f) notons $\widetilde{\varphi}_{(2,0.8)}$ la restriction de $\varphi_{(2,0.8)}$ à $[0,1]$ puis définissons

$$f(x) = \frac{\widetilde{\varphi}_{(2,0.8)}}{\int_0^1 \varphi_{(2,0.8)}(x)dx}$$

7. la fonction suivante:

$$f(x) = \begin{cases} \dfrac{1}{0.16095}\varphi_{(1,0.7)} & \text{si } 0 \leq x < 1/2, \\ \dfrac{1}{0.16095}\varphi_{(0,0.7)} & \text{si } 1/2 \leq x \leq 1. \end{cases}$$

8. (g) la loi gaussienne $\mathcal{N}(0.5, 0.02)$.

1. Simuler des échantillons de taille n suivant les distributions précédemment exposées.
2. Illustrer de manière empirique à partir de données simulées et analyser le comportement des trois estimateurs (méthode des séries orthogonales, histogramme, convolution) de la densité, en termes de carré du biais intégré, variance intégrée et écart quadratique moyen intégré.

Vous veillerez à faire varier la taille de l'échantillon n simulé.

• **Méthode de la fonction de répartition inverse** (à toutes fins utiles):
On définit la fonction pseudo-inverse de F sur $[0,1]$ par

$$F^{-1}(u) = \inf\{t \in \mathbb{R} : F(t) \geq u\}$$

Proposition 1. *Si U suit une loi uniforme sur $[0,1]$, alors $F^{-1}(U)$ a pour fonction de répartition F.*

Preuve: Commençons par montrer que $F^{-1}(u) \leq t$ ssi $u \leq F(t)$.
Soient $u \in [0,1]$ et $t \in \mathbb{R}$ tels que $u \leq F(t)$. Par définition de la fonction de répartition inverse, on a alors $F^{-1}(u) \leq t$. Réciproquement, si $F^{-1}(u) \leq t$, alors pour tout $y > t$, $F(y) \geq u$ car F est croissante. Et puisque F est continue à droite, $F(t) \geq u$.
En utilisant ce résultat, on en déduit que

$$P(F^{-1}(U) \leq t) = P(U \leq F(t)) = F(t). \qquad \square$$

Ainsi, dans le cas où F^{-1} est explicite, pour générer un échantillon $X_1, \ldots, X_n$ suivant la fonction de répartition F, on génère un échantillon $(U_1, \ldots, U_n)$ de variables uniformément distribuées sur $[0,1]$ et on pose $X_i = F^{-1}(U_i)$.