Université de Strasbourg — Ségolen Geffray
M1 - Magistère — geffray@math.unistra.fr
Statistique - études de cas — Année 2017/2018

Sujet 5: Etude de l'estimateur de la densité, utilisation des polynômes de Laguerre

• **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.

$\mapsto$ Soit $(\varphi_j(.))_{j\in\mathbb{N}}$ une base orthonormée de l'espace des fonctions de carrés intégrables. L'estimateur de $f(.)$ par la méthode des séries orthogonales est défini pour $x \in \mathbb{R}$ par:

$$\widehat{f}_{J,n}(x) = \sum_{j=0}^{J} \widehat{\theta}_{j,n}\varphi_j(x)$$

avec J à choisir parmi les entiers naturels et avec

$$\widehat{\theta}_{j,n} = \frac{1}{n}\sum_{i=1}^{n} \varphi_j(X_i).$$

$\mapsto$ 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 $\int_{\mathbb{R}} K(u)du = 1$ et $K(.) \geq 0$.

$\mapsto$ Soit $(I_j)_{j=1,\dots,J}$ 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 `orthopolynom` du logiciel `R` permet d'implémenter les polynômes de Laguerre.
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. Voici des choix possibles de noyau:

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

Concernant le choix de la fenêtre, vous pouvez faire librement varier h. Des méthodes de choix automatiques de la fenêtre sont également implémentées. L'argument `bw="nrd0"` correspond au choix

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

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

$$h = 1.06\frac{\widehat{\sigma}}{n^{-1/5}}.$$

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.
Ici, la base orthonormée de $L_2(\mathbb{R})$ l'espace des fonctions de carré intégrable sur $\mathbb{R}$ sera celle des polynômes de Laguerre. 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 $\sigma^2 > 0$.

(a) loi uniforme sur $[0,1]$,

(b) loi triangulaire de densité donnée pour $x \in \mathbb{R}$ par

$$f(x) = \max(1 - \lvert x-1\rvert, 0)$$

(c) la fonction en escalier suivante:

$$f(x) = \begin{cases} 3/5 & \text{si } 0 \leq x < 1/3, \\ 9/10 & \text{si } 1/3 \leq x < 3/4, \\ 51/30 & \text{si } 3/4 \leq x \leq 1, \\ 0 & \text{sinon.} \end{cases}$$

(d) la fonction linéaire par morceaux suivante:

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

(e) la loi gaussienne $\mathcal{N}(10, 1)$.

(f) loi Gamma

(g) loi de Cauchy translatée et tronquée: notons g la loi de Cauchy translatée de x_0 de densité donnée pour x dans $\mathbb{R}$ par

$$g(x) = \frac{1}{\pi(1 + (x - x_0)^2)}.$$

Notons $\widetilde{g}$ la restriction de cette loi de Cauchy à $\mathbb{R}^+$. Vous pouvez alors travailler avec la densité définie par

$$f = \frac{\widetilde{g}}{\int_0^\infty g(u)du}.$$

(h) mélange de deux gaussiennes:

$$f(x) = 0.7\varphi_{(10,1)}(x) + 0.3\varphi_{(15,2)}(x)$$

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 évaluer 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 $X_i = F^{-1}(U_i)$.

• **Méthode pour simuler un échantillon i.i.d. issu d'une loi de mélange:**
Pour générer un échantillon $(X_1, \ldots, X_n)$ i.i.d. de variables aléatoires suivant la loi de mélange de densité donnée pour $x \in \mathbb{R}$ par

$$f(x) = \sum_{k=1}^{K} p_k \varphi_{(m_k, \sigma_k^2)}(x)$$

avec, pour $k = 1, ..., K$, $m_k \in \mathbb{R}$, $\sigma_k^2 > 0$, $0 < p_k < 1$ et $\sum_{k=1}^{K} p_k = 1$, on peut utiliser la procédure qui suit.

Indépendamment pour $i = 1, ..., n$, on tire $\mathbf{Z}_i = \begin{pmatrix} Z_i^{(1)} \\ \vdots \\ Z_i^{(K)} \end{pmatrix}$ de sorte que $\sum_{k=1}^{K} Z_i^{(k)} = 1$ et $Z_i^{(k)}(\Omega) = \{0,1\}$ avec $\mathbb{P}(Z_i^{(k)} = 1) = p_k$ et $\mathbb{P}(Z_i^{(k)} = 0) = 1 - p_k$ pour $i = 1, ..., n$ et $k = 1, .., K$.
Cela revient à tirer, indépendamment pour $i = 1, ..., n$, une variable discrète J_i à valeurs dans $\{1, ..., K\}$ avec $\mathbb{P}(J_i = k) = p_k$ pour $k = 1, ..., K$. Conditionnellement à $\{J_i = j\}$, on tire X_i selon une loi $\mathcal{N}(m_j, \sigma_j^2)$ de densité notée $\varphi_{m_j, \sigma_j^2}$.
On obtient ainsi une réalisation de l'échantillon $(X_1, \ldots, X_n)$ dont la loi marginale est la loi de mélange souhaitée.

NB: dans le cas particulier où $K = 2$, indépendamment pour $i = 1, .., n$,

- on tire J_i selon une loi $\mathcal{B}(p_1)$
- si $J_i = 1$, on tire X_i selon une loi $\mathcal{N}(m_1, \sigma_1^2)$,
  si $J_i = 0$, on tire X_i selon une loi $\mathcal{N}(m_2, \sigma_2^2)$.