Université de Strasbourg
M1 - Magistère
Statistique

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

Sujet 18

Soit un échantillon i.i.d. $(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ issu d'une variable parente X dont la loi admet une densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$.

• **Estimateur à noyau multivarié**: Soit H une matrice symétrique définie positive, que l'on paramétrise comme une matrice de variance, à savoir sous la forme:

$$\boldsymbol{H} = \begin{pmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{pmatrix}.$$

Un noyau multivarié est une fonction $K : \mathbb{R}^d \to \mathbb{R}$ intégrable et telle que $\displaystyle\int_{\mathbb{R}^d} K(\boldsymbol{u})d\boldsymbol{u} = 1$.
L'estimateur à noyau de la densité est défini pour $\boldsymbol{x} \in \mathbb{R}^d$ par:

$$f_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(\det \boldsymbol{H})^{-1/2} K\left(\boldsymbol{H}^{-1/2}.\left(\boldsymbol{x} - \boldsymbol{X}_i\right)\right). \tag{1}$$

Un noyau multivarié très utilisé est le noyau gaussien défini pour $\boldsymbol{x} \in \mathbb{R}^d$ par:

$$K(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\,\boldsymbol{x}^t.\boldsymbol{x}\right).$$

C'est son utilisation qui motive la paramétrisation de la matrice $\boldsymbol{H}$ comme une matrice de variance. En utilisant le noyau gaussien, (1) devient:

$$f_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(2\pi)^{-d/2}(\det \boldsymbol{H})^{-1/2} \exp\left(-\frac{1}{2}\,\left(\boldsymbol{x} - \boldsymbol{X}_i\right)^t.\boldsymbol{H}^{-1/2}.\left(\boldsymbol{x} - \boldsymbol{X}_i\right)\right).$$

La fonction `kde` du package `ks` implémente l'estimateur à noyau multivarié avec un noyau gaussien tronqué à un hypercube.

Historiquement[1], la matrice $\boldsymbol{H}$ était choisie dans la classe restreinte $\mathcal{A} = \{h^2 I_d\}$, ce qui fournissait un estimateur plus simple:

$$f_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right).$$

L'intérêt de cette classe restreinte réside dans le fait qu'il y a un seul paramètre de lissage à déterminer. Une autre classe restreinte de paramètres matriciels de lissage est celle des matrices diagonales définies positives[2] $\mathcal{D} = \{\mathrm{diag}(h_1^2, ..., h_d^2), h_1, ..., h_d > 0\}$. La classe la plus générale[3]

[1] Cacoullos, 1966
[2] Epanenchnikov, 1969
[3] Deheuvels, 1977

est la classe $\mathcal{F}$ des matrices symétriques définies positives. Elle permet un lissage différent en fonction des différents directions mais implique une plus grande complexité de calcul et surtout le fait d'avoir à choisir un nombre beaucoup plus grand de paramètres de lissage.
Soient les conditions suivantes:
(A_1): f est de carré intégrable et deux fois différentiable et ses dérivées secondes sont continues bornées et de carré intégrable.
(A_2): le noyau $K : \mathbb{R}^d \to R$ est tel que $\displaystyle\int_{\mathbb{R}^d} K(x)^2 dx < \infty$, K est à symétrie sphérique (ie invariant par rotation centrée en 0) et admet un moment d'ordre 2, donc satisfait $\displaystyle\int z_j K(z) dz = 0$ pour $j = 1, ..., d$, $\displaystyle\int z_j z_k K(z) dz = 0$ pour $j, k = 1, ..., d$, $j \neq k$ et $\displaystyle\int z_j^2 K(z) dz = m_2(K) < \infty$ pour $j = 1, ..., d$.
(A_3): Les matrices $\boldsymbol{H} = \boldsymbol{H}_n$ forment une suite de matrices telle que $h_{j,k} \overset{n\to\infty}{\to} 0$ pour $j, k = 1, ..., d$ et $n(\det \boldsymbol{H})^{1/2} \overset{n\to\infty}{\to} \infty$.
Sous ces conditions, un équivalent asymptotique du MISE est:

$$AMISE(\widehat{f}_{\boldsymbol{H}}) = \frac{\int K(x)^2 dx}{n(\det \boldsymbol{H})^{1/2}} + \frac{(m_2(K))^2}{4} \int_{\mathbb{R}^d} (\mathrm{Tr}(\boldsymbol{H}.\mathrm{Hess} f(x)))^2 dx.$$

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... On peut montrer que la fenêtre optimale est de l'ordre de $n^{-2/(d+4)}$. Le AMISE correspondant à ce choix de fenêtre est l'ordre de $n^{-4/(d+4)}$.
La fonction `Hpi` du package `ks` implémente ce choix.

La matrice H peut être choisie en utilisant le principe de de la règle de référence à la loi normale. On remplace f par la densité gaussienne. Cela fournit

$$\boldsymbol{H}_{NS} = \left(\frac{4}{d+2}\right)^{2/(d+4)} \frac{1}{n^{2/(d+4)}} \widehat{\Sigma^2}$$

où $\widehat{\Sigma^2}$ est un estimateur de la matrice de variance.
La fonction `Hns` du package `ks` implémente ce choix.

L'idée de la méthode de la validation croisée est de déterminer un choix "optimal" de la fenêtre $\boldsymbol{H}$ noté $\boldsymbol{H}_{\mathrm{opt}}$ de la façon suivante

$$\begin{aligned}
\boldsymbol{H}_{\mathrm{opt}} &= \arg\min_{h>0} \mathtt{ISE}(\widehat{f}_h) \\
&= \arg\min_{\boldsymbol{H}\in\mathcal{F}} \int (\widehat{f}_{\boldsymbol{H}}(x) - f(x))^2 dx \\
&= \arg\min_{\boldsymbol{H}\in\mathcal{F}} \left\{ \int (\widehat{f}_{\boldsymbol{H}}(x))^2 dx - 2 \int \widehat{f}_{\boldsymbol{H}}(x) f(x) dx + \int f(x)^2 dx \right\} \\
&= \arg\min_{\boldsymbol{H}\in\mathcal{F}} \left\{ \int (\widehat{f}_{\boldsymbol{H}}(x))^2 dx - 2 \int \widehat{f}_{\boldsymbol{H}}(x) f(x) dx \right\}.
\end{aligned}$$

Notons

$$J(\boldsymbol{H}) = \int_{\mathbb{R}^d} (\widehat{f}_{\boldsymbol{H}}(x))^2 dx - 2 \int_{\mathbb{R}^d} \widehat{f}_{\boldsymbol{H}}(x) f(x) dx$$

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

$$\int_{\mathbb{R}^d} \widehat{f}_{\boldsymbol{H}}(x) f(x) dx = \mathbb{E}\left[\widehat{f}_{\boldsymbol{H}}(\boldsymbol{X}) | \boldsymbol{X}_1, ..., \boldsymbol{X}_n\right]$$

pour un vecteur aléatoire $\boldsymbol{X}$ de densité f, indépendant de $(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$. On estime alors sans biais $\mathbb{E}\left[\widehat{f}_{\boldsymbol{H}}(\boldsymbol{X}) | \boldsymbol{X}_1, ..., \boldsymbol{X}_n\right]$ par la méthode dite du "leave-one-out". Cela fournit le critère suivant à optimiser le critère en $\boldsymbol{H}$:

$$\widehat{J(\boldsymbol{H})} = \int_{\mathbb{R}^d} (\widehat{f}_{\boldsymbol{H}}(x))^2 dx - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{-i,\boldsymbol{H}}(\boldsymbol{X}_i)$$

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

$$\widehat{f}_{-i,\boldsymbol{H}}(x) = \frac{1}{(n-1)(\det \boldsymbol{H})^{1/2}} \sum_{j=1, j\neq i}^{n} K\left(\boldsymbol{H}^{-1/2}.\,(\boldsymbol{x} - \boldsymbol{X}_i)\right).$$

La fonction `Hlscv` du package `ks` implémente ce choix.

• **Critères objectifs**:
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 $\varphi_{(m_1,m_2),(\sigma_1^2,\sigma_2^2,\rho\sigma_1\sigma_2)}$ la densité de la loi gaussienne bivariée de vecteur moyenne égal à $\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$ et de matrice de variance égale à $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. Considérons les distributions suivantes:

- la loi de densité:
$$f(x, y) = 2x\, I(0 < x < 1, 0 < y < 1)$$
- la loi de densité:
$$f(x, y) = \varphi_{(0,0),(1/4,1,0)}(x, y)$$
- la loi de densité:
$$f(x, y) = f_C(x) f_C(y$$
en notant f_C la densité de la loi de Cauchy standard

- la loi de densité:
$$f(x,y) = \psi_{(0,0.2,10)}(x)(x)\psi_{(0.5,0.1,3)}(y)$$
- la loi de densité:
$$f(x,y) = \varphi_{(0,0),(1,1.2,-7/10)}(x,y)$$
- la loi de densité:
$$f(x,y) = \frac{2}{3}\,\varphi_{(0,0),(1,4,1)}(x,y) + \frac{1}{3}\,\varphi_{(0,0),(4/9,4/9,-1/3)}(x,y)$$
- la loi de densité:
$$f(x,y) = \frac{1}{2}\,\varphi_{(-1,1),(4/9,4/9,3/5)}(x,y) + \frac{1}{2}\,\varphi_{(1,-1),(4/9,4/9,3/5)}(x,y)$$
- la loi de densité:
$$f(x,y) = \frac{4}{11}\,\varphi_{(-2,2),(1,1,0)}(x,y) + \frac{3}{11}\,\varphi_{(0,0),(0.8,0.8,-0.72)}(x,y) + \frac{4}{11}\,\varphi_{(2,-2),(1,1,0)}(x,y)$$
- la loi de densité:
$$f(x,y) = \frac{9}{20}\,\varphi_{(-6/5,6/5),(9/25,9/25,3/10)}(x,y) + \frac{9}{20}\,\varphi_{(6/5,-6/5),(9/25,9/25,-3/10)}(x,y) + \frac{1}{10}\,\varphi_{(0,0),(1/4,1/4,1/5)}(x,y)$$
- la loi de densité:
$$\begin{aligned} f(x,y) = {} & \frac{1}{2}\,\varphi_{(0,0),(1,1,0)}(x,y) + \frac{3}{40}\,\varphi_{(0,0),1/16(1,1,-9/10)}(x,y) + \frac{1}{5}\,\varphi_{(1,1),1/4(1,1,-9/10)}(x,y) \\ & + \frac{3}{40}\,\varphi_{(-1,1),1/8(1,1,0)}(x,y) + \frac{3}{40}\,\varphi_{(-1,-1),1/8(1,1,-9/10)}(x,y) + \frac{3}{40}\,\varphi_{(1,-1),1/16(1,1,0)}(x,y) \end{aligned}$$
- la loi de densité:
$$\begin{aligned} f(x,y) = {} & \frac{1}{2}\,\varphi_{(0,0),(1,1,0)}(x,y) + \frac{1}{10}\,\varphi_{(0,0),1/16(1,1,0)}(x,y) + \frac{1}{10}\,\varphi_{(-1,-1),1/16(1,1,0.2)}(x,y) \\ & + \frac{1}{10}\,\varphi_{(-1,1),1/16(1,1,0)}(x,y) + \frac{1}{10}\,\varphi_{(1,-1),1/16(1,1,0.4)}(x,y) + \frac{1}{10}\,\varphi_{(1,1),1/16(1,1,0)}(x,y) \end{aligned}$$

Vous pourrez utiliser les fonctions `contour`, `persp` et `image` du logiciel `R` pour représenter la densité à estimer.

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 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.