Séminaire Mathématique de Béjaïa (LaMOS)
Volume 10, 2011, pp. 55 - 58

12

Approche Bayésienne dans l'estimation de la densité de probabilité par la méthode du noyau

Nabil ZOUGAB et Smail ADJABI

Laboratoire de Modélisation et d'Optimisation des Systèmes (LAMOS)
Université de Béjaïa, Béjaïa 06000, Algérie
Tél. (213) 34 21 51 88

Résumé Le problème fondamental dans l'estimation de la densité de probabilité est le choix du paramètre de lissage. Dans ce travail, nous proposons d'utiliser l'approche Bayésienne pour estimer ce paramètre. Cette approche est une alternative pour les méthodes classiques tel que : les méthodes plug-in et les techniques de validation croisée. Une étude de simulation est conduite pour comparer les performances de l'approche Bayésienne proposée et les méthodes classiques via l'erreur quadratique moyenne intégrée asymptotique.
Mots clés : Approche Bayésienne ; Paramètre de lissage ; Validation croisée ; plug-in ; Simulation.

12.1 Introduction

On dispose d'un échantillon $X_1, X_2, ..., X_n$ de variables aléatoires indépendantes de même loi, de densité de probabilité inconnu f. On définit l'estimateur à noyau de Rosenblatt [4] et Parzen [5] de f par :

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{12.1}$$

$$= \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i), \tag{12.2}$$

ou K est la fonctin noyau satisfaisant $\int_{\mathbb{R}} K(y)dy = 1$, $\int_{\mathbb{R}} yK(y)dy = 0$ and $\int_{\mathbb{R}} y^2K(y)dy = \sigma_K^2 < \infty$ et h est un réel positif appelé paramètre de lissage vérifiant $h \longrightarrow 0$ et $nh \longrightarrow \infty$ quand $n \longrightarrow \infty$. Si le choix du noyau n'est pas un problème dans l'estimation de la densité, il n'en est pas de même pour le choix du paramètre de lissage qui ne dépend que de la taille d'échantillon n. L'objectif de ce travail est double. Premièrement, nous proposons l'approche Bayésienne pour estimer le paramètre de lissage. En suite, nous comparons cette approche proposée aux méthodes classiques, à savoir les méthodes plug-in et validation croisée via une étude de simulation.

12.2 Approche Bayésienne

Dans cette section, nous étudions l'approche Bayésienne pour estimer le paramètre de lissage. L'estimateur Bayésien est obtenu via la loi a posteriori $\pi(h|data)$. En particulier,

nous considérons une séquence de variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes de même loi et de densité de probabilité inconnu f et des réalisations $x_1, x_2, \ldots, x_n$. Alors la fonction vraisemblance est donnée pa :

$$L(x_1, \ldots, x_n; h) = \pi(x_1, \ldots, x_n|h) = \prod_{i=1}^{n} f_h(x_i).$$

La technique de validation croisée consiste à estimer $f(x_i)$ à partir de l'ensemble des points sauf le point x_i, et le résultat est donné par :

$$f_{h,i}(x_i) = \frac{1}{n-1} \sum_{j=1, i\neq j}^{n} K_{x_i,h}(x_j).$$

Alors, la fonction vraisemblance de validation croisée est donnée par :

$$LCV(x_1, \ldots, x_n; h) = \pi(x_1, \ldots, x_n|h) = \prod_{i=1}^{n} f_{h,i}(x_i).$$

Par le théorème de Bayes, la loi a posteriori de h prend la forme suivante :

$$\pi(h|x_1, \ldots, x_n) = \frac{\pi(x_1, \ldots, x_n|h)\pi(h)}{\pi(x_1, \ldots, x_n)} = \frac{\pi(h) \prod_{i=1}^{n} f_{h,i}(x_i)}{\pi(x_1, \ldots, x_n)},$$

ou $\pi(x_1, \ldots, x_n) = \int \pi(x_1, \ldots, x_n|h)\pi(h)dh$. On peut écrire aussi

$$\pi(h|x_1, \ldots, x_n) \propto \pi(x_1, \ldots, x_n|h)\pi(h) = \pi(h) \prod_{i=1}^{n} f_{h,i}(x_i).$$

Parconséquent, la loi a posteriori est

$$\pi(h|x_1, \ldots, x_n) \propto \pi(h) \prod_{i=1}^{n} \frac{1}{n-1} \sum_{j=1, i\neq j}^{n} K_{x_i,h}(x_j).$$

Nous supposons que la loi a priori de h est

$$\pi(h) \propto \frac{1}{1+h^2}, \tag{12.3}$$

Finalement, la loi a posteriori de h est de la forme suivante

$$\pi(h|x_1, \ldots, x_n) \propto \frac{1}{1+h^2} \prod_{i=1}^{n} \frac{1}{n-1} \sum_{j=1, i\neq j}^{n} K_{x_i,h}(x_j). \tag{12.4}$$

La simulation directement de la loi a posteriori (12.4) est très difficile, voir impossible, nous proposons alors d'utiliser les méthodes de Monté Carlo par Chaîne de Markov (MCMC) pour estimer le paramètre de lissage.

12.3 Simulation

Nous présentons dans cette section le travail de simulation effectué. Afin d'illustrer les performances de l'approche Bayésienne et les méthodes classiques (plug-in et validation croisée), nous utilisons 3 densités tests. Nous avons choisi des densités présentant différents aspects :

• **D1** Le mélange de deux densités de loi normale : $f_1 \sim \frac{1}{2}\mathcal{N}(1, \frac{16}{49}) + \frac{1}{2}\mathcal{N}(-1, \frac{16}{49})$.

• **D2** Le mélange de trois densités de loi normale : $f_2 \sim \frac{1}{3}\mathcal{N}(-1, 0.5) + \frac{1}{3}\mathcal{N}(0.5, 0.5) + \frac{1}{3}\mathcal{N}(2, 0.5)$.

• **D3** Le mélange de quatre densités de loi normale et de loi gamma : $f_3 \sim \frac{1}{4}\mathcal{N}(6.5, 2) + \frac{3}{8}\mathcal{N}(8, 1) + \frac{1}{8}\mathcal{N}(14, 1.5) + \frac{1}{8}\mathcal{N}(18.5, 1.5) + \frac{1}{8}\mathcal{G}(3, 1)$, ou $\mathcal{G}(y; 3, 1) = \frac{1}{\Gamma(3)}(y - 20)^2 \exp\{-(y - 20)\}$, $(y > 20)$.

Nous utilisons les notations suivantes :

1. n la taille de l'échantillon,
2. N_{sim} nombre de simulations,
3. h_{sj} : le paramètre de lissage obtenu par la méthode de Sheather and Jones (voir Sheather and Jones [3]),
4. h_{scv} : le paramètre de lissage obtenu par la méthode de validation croisée lissée (voir Hall et al. [2]),
5. h_{mcmc} : le paramètre de lissage obtenu par les méthodes MCMC,
6. h^* : le paramètre de lissage théorique ; $h^* = \left[\frac{\int K^2}{\sigma_K^4}\right]^{1/5} \left[\frac{1}{\int f''}\right]^{1/5} n^{-1/5}$,
7. $AMISE(h)$: l'erreur quadratique moyenne intégrée asymptotique ; $AMISE(h) = \frac{h^4}{4}\sigma_K^4 \int f''^2(x)dx + \frac{\int K^2(y)dy}{nh}$,
8. $Eff_{AMISE} = \frac{AMISE(h^*)}{AMISE(h_{opt})}$, $h_{opt} = \{h_{sj}, h_{scv}, h_{mcmc}\}$.

f	N_{sim}	n	$[\overline{h}_{sj}, std]$	$[\overline{h}_{scv}, std]$	$[\overline{h}_{mcmc}, std]$	h^*	$AMISE^*$	Eff_{AMISE}
D1	50	200	(0.258, 0.0008)	(0.315, 0.0006)	(0.261, 0.0010)	0.238	0.0074	**[0.922, 0.655, 0.916]**
		500	(0.216, 0.0002)	(0.268, 0.0003)	(0.210, 0.0008)	0.198	0.0035	**[0.936, 0.775, 0.940]**
		1000	(0.186, 0.0007)	(0.232, 0.0008)	(0.190, 0.0012)	0.172	0.0020	**[0.932, 0.805, 0.935]**
D2	50	200	(0.333, 0.0040)	(0.377, 0.0029)	(0.270, 0.0015)	0.241	0.0073	**[0.732, 0.593, 0.929]**
		500	(0.238, 0.0007)	(0.268, 0.0006)	(0.218, 0.0010)	0.201	0.0035	**[0.916, 0.804, 0.949]**
		1000	(0.199, 0.0004)	(0.216, 0.0005)	(0.180, 0.0005)	0.174	0.0020	**[0.922, 0.861, 0.953]**
D3	50	200	(0.721, 0.0024)	(0.997, 1.509967e-05)	(0.643, 0.0013)	0.527	0.0033	**[0.773, 0.434, 0.871]**
		500	(0.553, 0.0003)	(0.761, 0.0009)	(0.507, 0.0041)	0.440	0.0016	**[0.875, 0.542, 0.922]**
		1000	(0.447, 0.0012)	(0.565, 0.0011)	(0.443, 0.0025)	0.382	0.0009	**[0.912, 0.665, 0.931]**

TABLE 12.1: Comparaison des résultats de simulation.

Les résultats de simulation montrent que l'approche Bayésienne est meilleur que les méthodes classiques au sens de l'erreur quadratique moyenne intégrée asymptotique $AMISE$.

Références

1. B W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1986.
2. P. Hall and J S. Marron and B U. Park, *Smoothed cross validation*, Probability Theory and Related Fields, volume 92, pp. 1-20, 1992
3. S J. Sheather and M C. Jones, *A reliable data-based bandwidth selection method for kernel density estimation*, Journal of the Royal Statistical Society Series B,volume 53, pp. 683-690, 1991.
4. M. Rosenblatt, *Remarks in some nonparametric estimates of a density function*, Annals of Mathematical Statistics, volume 27, pp. 832-837, 1956.
5. E. Parzen,*On estimation of a probability density function and mode*, Annals of Mathematical Statistics, volume 33, pp. 1065-1076, 1962.