APPROCHE BAYESIENNE DANS L'ESTIMATION NON PARAMÉTRIQUE DES DENSITÉS HEAVY TAILED PAR NOYAU ASSOCIÉ

Smail ADJABI[a], Yasmina ZIANE[a], Nabil ZOUGAB[a]
[a]Laboratoire LAMOS, Université de Béjaia, 06000 Béjaia, Algérie
E-mail : adjabi@hotmail.com,yasmina-ziane@hotmail.com, nabilzougab@yahoo.fr

Résumé

l'analyse et l'estimation de la densité des données heavy tailed (queue lourde) sont complexes en raison de leurs caractéristiques spécifiques qui sont la décroissance lente vers zéro et présence d'observations rares dans la queue. Dans ce travail, on estime les densités du type heavy tailed par la méthode non paramétrique du noyau associé. Nous utilisons le noyau asymétrique associé Birnbaum-Saunders-power-exponential (BS-PE) pour éviter les effets de bord et réduire le biais de l'estimateur. Le paramètre de lissage qui intervient dans l'estimation de la densité est estimé par l'approche bayesienne adaptative pour les deux fonctions perte : quadratique et entropie. L'évaluation de performance de l'approche bayesienne est réalisée par simulation sur des densités heavy tailed cibles connues et sur des données données réelles : trafic web et les données environnementales. Les résultats obtenus montrent l'avantage de l'approche bayesienne adaptative par rapport a la méthode classique validation croisée non biaisée (UCV) selon le critère de l'erreur quadratique intégrée (ISE).

Mots clés : Données heavy tailed, fonction perte, noyau BS-PE, paramètre de lissage, approche bayesienne, validation croisée.

Abstract

Analysis and estimation of the heavy tailed data density are complex because of their specific characteristics that are slow decrease to zero and presence of rare observations in the tail. In this work, it is estimated the heavy tailed densities type by the non-parametric method of the associated core. We use the associated asymmetric core Birnbaum-Saunders-power-exponential (textrm BS-PE) to avoid edge effects and reduce the bias of the estimator. The smoothing parameter which intervenes in the estimation of the density is estimated by the adaptive Bayesian approach for both loss functions and quadratic entropy. Evaluating performance of the Bayesian approach is achieved by simulation of heavy tailed densities known targets and data on actual data : web traffic and environmental data. The results show the advantage of adaptive Bayesian approach compared to the conventional method unbiased cross validation (UCV) according to the criterion of the integrated square error (ISE)

Key words : Heavy tailed data, loss function, BS-PE kernel Bandwidth, bayesian approach, Cross validation.

1 Introduction

Parzen(1962) a proposé l'estimateur a noyau dans le cas univarié des densités à support non borné. Le choix du noyau dans ce cas est peu important, car il n'influe pas sur la qualité de l'estimateur. Les

noyaux les plus utilisées sont les noyaux symétriques dite aussi classiques comme le noyau gaussien et le noyau Epanechnicov. Cependant, lorsqu'on veut estimer des densités à support non borné, l'estimateur à noyau classique devient non consistant, à cause des effets du bord. Ce problème est dû à l'utilisation des noyaux symétriques qui assignent un poids en dehors du support lorsque le lissage est pris en compte prés du bord. Pour remédier à ce problème, on remplace les noyaux symétriques par des noyaux asymétriques. Cette idée est due à Chen(1999, 2000) où il a proposé les noyaux gamma, gamma modifié et le noyau Bêta pour estimer des densités à support borné. Divers noyaux asymétriques ont été proposé par la suite : noyau inverse gaussien (IG), réciproque inverse gaussien (RIG), noyau log-normal (LN) et Birnbaum-Saunders Généralisé (GBS).

Les performances de l'estimateur de la densité à noyau asymétrique dépend crucialement du paramètre de lissage qui contrôle la qualité du lissage de l'estimateur. Deux catégories de méthodes ont été proposé pour le choix du paramètre de lissage. La première repose sur la minimisation de l'erreur quadratique moyenne intégrée (MISE). L'inconvénient de cette méthode est que le paramètre de lissage optimal dépend de quantités inconnues. La deuxième catégorie est la validation croisée, elle est intéressante en pratique car elle se laisse guider seulement par les observations. Ces deux méthodes ont tendance à fournir des estimateurs sous ou sur lissé lorsque les données sont de petite ou moyenne taille ou encore lorsqu'on veut estimer des fonctions complexes. En plus de ces deux méthodes classiques pour la sélection du paramètre de lissage, il existe l'approche bayesienne qui consiste à considérer le paramètre de lissage h comme une variable aléatoire et lui associer une loi a priori qui sert à compenser le manque d'information.

L'objectif de ce travail est d'améliorer l'estimateur à noyau associé des densités Heavy tailed en utilisant le noyau Birnbaum Saunders-Puissance Exponentielle (BS-PE) et l'approche bayesienne adaptative pour la sélection du paramètre de lissage. Des études sur des données simulées à partir de densités cibles de type heavy tailed et sur des données réelles sont menées pour comparer l'approche bayesienne et la méthode UCV pour la sélection du paramètre de lissage.

2 Estimateur à noyau associé BS-PE

Soit $X_1, \ldots, X_n$ une séquence de variables aléatoires indépendantes et identiquement distribuées de densité inconnue f. L'estimateur à noyau associé introduit par Chen(1999) est de la forme :

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

où $h > 0$ est le paramètre de lissage, il joue le rôle de paramètre de dispersion autour de la cible x et K est une fonction appelée noyau, elle détermine la forme du voisinage autour du point x. Cependant, lorsqu'il s'agit des densités à support borné ou semi-borné, l'utilisation des noyaux symétriques causent un sérieux problème aux bords (biais aux bords) ce qui engendre des estimateurs non consistants. Pour cette raison, plusieurs auteurs Chen(1999, 2000), Scaillet(2004), Jin et Kawckzak(2003), Marchant et al(2013) et Saulo et al(2013) ont proposé une famille de noyaux asymétriques pour remédier à ce problème. En plus du support des données, le type de données peut fortement influer sur le choix du noyau, comme par exemple les données qui se caractérisent par une queue ou un pôle. D'après Marchant et al(2013) le noyau le plus approprié aux données qui se caractérisent par une queue est le noyau BS-PE de paramètre $\nu = 2$ défini par :

$$K_{x,h}(t) = \frac{\nu}{2^{\frac{1}{2\nu}}\Gamma(\frac{1}{2\nu})\sqrt{4h}}\left(\frac{1}{\sqrt{xt}} + \sqrt{\frac{x}{t^3}}\right)\exp\left(\frac{-1}{2h^{\nu}}\left(\frac{t}{x} + \frac{x}{t} - 2\right)^{\nu}\right), t > 0. \tag{2}$$

L'estimateur de la densité associé au noyau BS-PE est alors de la forme :

$$\widehat{f}_h(x) = \frac{1}{n} \times \frac{\nu}{2^{\frac{1}{2\nu}}\Gamma(\frac{1}{2\nu})\sqrt{4h_i}} \sum_{i=1}^{n} \left(\frac{1}{\sqrt{xX_i}} + \sqrt{\frac{x}{X_i^3}} \right) \exp\left(\frac{-1}{2h_i^{\nu}} \left(\frac{X_i}{x} + \frac{x}{X_i} - 2 \right)^{\nu} \right), x > 0. \quad (3)$$

3 Méthodes de sélection du paramètre de lissage

Validation croisée non biaisée

Le principe de la méthode validation croisée non biaisée consiste à sélectionner la paramètre de lissage h qui minimise l'erreur quadratique intégrée (ISE) donnée par

$$ISE(h) = \int (\widehat{f}_h(x) - f(x))^2 dx = \int \widehat{f}_h^2(x)dx - 2\int \widehat{f}_h(x)f(x)dx + \int f^2(x)dx.$$

Comme la quantité $\int f^2(x)dx$ ne dépend pas de h, la valeur optimale h_{UCV} est

$$h_{UCV} = \arg\min_h UCV(h),$$

où

$$UCV(h) = \int \widehat{f}_h^2(x) - \int \widehat{f}_h(x)f(x)dx. \quad (4)$$

Le terme $\int \widehat{f}_h(x)f(x)dx$ peut être estimé par $\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{h,i}(X_i)$ où $\widehat{f}_{h,i}(X_i)$ est l'estimateur de la densité calculé à partir de l'échantillon privé de l'observation X_i. Le critère à optimiser devient alors

$$UCV(h) = \int \left[\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i) \right]^2 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} K_{X_i,h}(X_j). \quad (5)$$

Approche bayesienne adaptative

Soit $x_1, x_2, \ldots, x_n$ les réalisations des variables aléatoires $X_1, X_2, \ldots, X_n$ i.i.d. de densité inconnue f. L'approche que nous allons proposer consiste dans un premier temps à considérer que l'estimateur de la densité à noyau associé adaptatif est de la forme :

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h_i}(x_i), \qquad x \geq 0, \quad (6)$$

où K_{x,h_i} est le noyau associé et h_i le paramètre de lissage adaptatif associé à chaque observations x_i. Nous allons utiliser l'approche bayesienne pour estimer le paramètre de lissage h_i, $i = 1, \ldots, n$ en supposant que c'est une variable aléatoire de loi a priroi $\pi(h_i)$. L'estimateur de Bayes sera alors obtenu à partir de l'estimation de la loi a posteriori de h_i donnée par :

$$\widehat{\pi}(h_i|x_i) = \frac{\widehat{f}_{h_i,i}(x_i, h_i)\pi(h_i)}{\int \widehat{f}_{h_i,i}(x_i, h_i)\pi(h_i)dh_i}, \quad (7)$$

où $\widehat{f}_{h_i,i}(x_i)$ est l'estimateur de la densité à noyau associé adaptative calculé à partir de l'échantillon privé de l'observation x_i.

L'estimateur de Bayes sous la perte quadratique $L(\widehat{h},h)=(\widehat{h}-h)^2$ et sous la perte entropie $L(\tilde{h},h)=\frac{\tilde{h}}{h}-\log(\frac{\tilde{h}}{h})-1$ du paramètre de lissage h_i sont donnés respectivement par

$$\widehat{h}_i=\int h_i\widehat{\pi}(h_i|x_i)dh_i, \tag{8}$$

$$\tilde{h}_i=\left[\int h_i^{-1}\widehat{\pi}(h_i|x_i)dh_i\right]^{-1}. \tag{9}$$

Les expressions des estimateurs (8) et (9) donnent dans certains cas des résultats explicites. Cela est dû à l'utilisation des priors conjugués. Ces résultats ont été proposés par Brewer (2000), Zougab et al (2012) et Ziane et al(2015).
Nous supposons que le paramètre de lissage h_i à une loi a priori conjuguée avec le noyau BS-PE de paramètres α et β donnée par :

$$\pi(h_i)=\frac{\nu}{\Gamma(\alpha)\beta^{\alpha}}\frac{1}{h_i^{\alpha\nu+1}}\exp(-\frac{1}{\beta h_i^{\nu}}),\ h_i>0. \tag{10}$$

En exploitant la conjugalité entre la loi a priori (10) et le noyau BS-PE (2), les estimateurs du paramètre de lissage adaptatif sous la fonction perte quadratique et entropie sont respectivement de la forme :

$$\widehat{h}_i=\frac{1}{\beta^{\frac{1}{\nu}}}\times\frac{\Gamma(\alpha-\frac{1}{2\nu})\sum\limits_{j=1,i\neq j}^{n}\left(\frac{1}{\sqrt{x_ix_j}}+\sqrt{\frac{x_i}{x_j^3}}\right)\left[\frac{(\frac{x_j}{x_i}+\frac{x_i}{x_j}-2)^{\nu}}{2}+1\right]^{-\alpha+\frac{1}{2\nu}}}{\Gamma(\alpha+\frac{1}{2\nu})\sum\limits_{j=1,i\neq j}^{n}\left(\frac{1}{\sqrt{x_ix_j}}+\sqrt{\frac{x_i}{x_j^3}}\right)\left[\frac{(\frac{x_j}{x_i}+\frac{x_i}{x_j}-2)^{\nu}}{2}+1\right]^{-\alpha-\frac{1}{2\nu}}}. \tag{11}$$

$$\tilde{h}_i=\frac{1}{\beta^{\frac{1}{\nu}}}\times\frac{\Gamma(\alpha+\frac{1}{2\nu})\sum\limits_{j=1,i\neq j}^{n}\left(\frac{1}{\sqrt{x_ix_j}}+\sqrt{\frac{x_i}{x_j^3}}\right)\left[\frac{(\frac{x_j}{x_i}+\frac{x_i}{x_j}-2)^{\nu}}{2}+1\right]^{-\alpha-\frac{1}{2\nu}}}{\Gamma(\alpha+\frac{3}{2\nu})\sum\limits_{j=1,i\neq j}^{n}\left(\frac{1}{\sqrt{x_ix_j}}+\sqrt{\frac{x_i}{x_j^3}}\right)\left[\frac{(\frac{x_j}{x_i}+\frac{x_i}{x_j}-2)^{\nu}}{2}+1\right]^{-\alpha-\frac{3}{2\nu}}}. \tag{12}$$

Choix des paramètres de la loi a prioi
Choix de α : la variance de la loi a priori est $\mathrm{Var}(h_i)=\{\Gamma(\alpha-2/\nu)\Gamma(\alpha)-\Gamma^2(\alpha-1/\nu)\}/\{\beta^{2/\nu}\Gamma^2(\alpha)\}$. Elle est positive pour $\nu>0$ fixé si $\{\Gamma(\alpha-2/\nu)\Gamma(\alpha)-\Gamma^2(\alpha-1/\nu)\}>0$, c'est-à-dire si $\alpha>2/\nu$.
Choix de β : a partir de l'espérance de la loi a priori $\mathbb{E}(h_i)=\{\Gamma(\alpha-1/\nu)\}/\{\beta^{1/\nu}\Gamma(\alpha)\}$, on peut remarquer qu'une grande valeur de β doit dépendre de la taille de l'échantillon pour assurer la convergence de l'estimateur. Le taux de convergence de l'erreur quadratique moyenne intégrée (MISE) optimal de h donné dans Marchant et al(2013) est $1/n^{2/5}$. En choisissant $\beta=\beta_n=n^{2\nu/5}$, on obtient le même taux de convergence que celui du MISE optimal.

4 Application numérique

Etude de simultion

On compare les performances de l'approche bayesienne adaptative avec la méthode classique globale UCV pour la sélection du paramètre de lissage h dans le contexte de l'estimation de la densité

de probabilité par la méthode du noyau associé BS-PE sur des données heavy tailed à support non négatif en utilisant deux densités cibles **D1** : loi lognormal(1,1) et **D2** : loi de Burr(1,3,1). On simule des échantillons de taille $n = 10, 25, 50, 100, 200, 500$ et 1000 avec un nombre de répétitions $N_{sim} = 100$. L'évaluation de performances sera basée sur le critère ISE. Les paramètres de la loi a priori sont $\alpha = 2.5$ et $\beta = n^{2/5}$.
Les résultats obtenus du ISE moyen ($\overline{ISE}$) sont donnés dans le tableau 1. Le tableau 2 présente les résultats du temps d'exécution de chaque méthode (bayesienne adaptative et la méthode UCV) pour le modèle **D1** et pour une seule simulation.

f	n	$\overline{ISE}_{UCV}$	$\overline{ISE}_{Bayes-quadratique}$	$\overline{ISE}_{Bayes-entropie}$
D1	10	0.03150	**0.02644**	0.02783
	25	0.01339	**0.01037**	0.01099
	50	0.01145	**0.00750**	0.00780
	100	0.00780	**0.00545**	0.00575
	200	0.00593	**0.00440**	0.00464
	500	0.00184	**0.00150**	0.00158
	1000	0.00142	**0.00104**	0.00110
D2	10	0.09747	**0.05542**	0.05560
	25	0.05567	**0.03817**	0.03928
	50	0.02584	**0.01802**	0.01830
	100	0.02275	**0.01181**	0.01178
	200	0.00862	0.00741	**0.00740**
	500	0.00540	0.00442	**0.00440**
	1000	0.00431	**0.00239**	0.00242

TABLE 1 – Résultats de simulation de ISE($\overline{ISE}$) basés sur 100 répétitions pour **D1** et **D2**.

n	t_{UCV}	$t_{quadratic}$	$t_{entropy}$
10	0.09800	**0.00100**	0.00299
25	0.13899	0.00999	**0.00800**
50	0.55900	0.01399	**0.00999**
100	0.76200	0.03899	**0.02300**
200	0.91100	0.43400	**0.10900**
500	3.24400	0.46200	**0.43100**
1000	7.55100	1.92300	**1.46200**

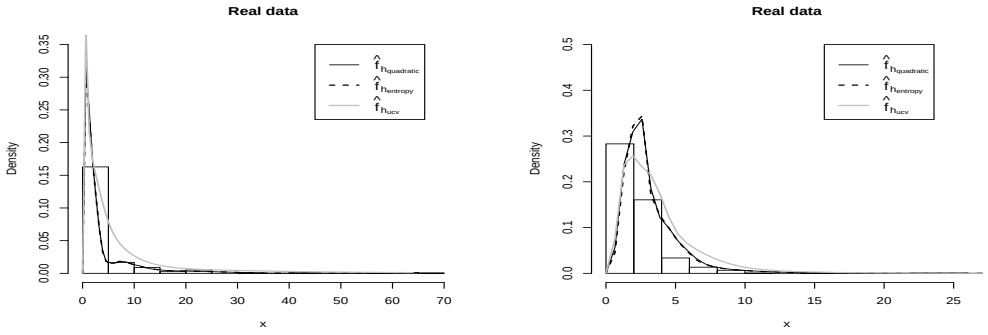
TABLE 2 – Temps d'exécution (en secondes) pour une simulation pour **D1**.

Les résultats donnés dans les tableaux 1 et 2 indiquent clairement l'avantage de l'approche bayesienne adaptative par rapport à la méthode classique UCV en terme de ISE moyen ($\overline{ISE}$) et en temps d'exécution.

Application sur les données réelles

On estime la densité des données réelles par la méthode du noyau associé BS-PE avec les deux méthodes de sélection du paramètre de lissage bayesienne adaptative et la méthode UCV.

- **Le trafic web :** Ces données représentent la taille de différents fichiers web (pdf, html, image, vidéo,...) mesurées en kilo octet, recueillies à partir du serveur de la coupe du monde France 1998 du mois de juin pour un nombre de requêtes $n = 312$.
- **Les données environnementales (SO2) :** Les données concernent l'étude des concentrations (SO2) (en $ppb = ppm \times 1000$) observées chaque heure à une station de surveillance située à Santiago (Chili) du mois de Mars 2002.



(a) (b)

FIGURE 1 – (a) Estimateur de la densité du trafic web, (b) Estimateur de la densité des données environnementales (SO2).

A partir de la figure 1, on remarque que les deux méthodes sont capables de reproduire l'unimodalité des deux types de données. Cependant, nous observons que le meilleur lissage est obtenu par l'approche bayesienne et ceci pour les deux types de données.

Bibliographie

[1] Brewer, M. J. (2000). *A bayesian model for local smoothing in kernel density estimation*, Statistics and Computing, 10, 299-309.
[2] Chen, S. X. (1999), *Beta kernels estimators for density functions*, Computational Statistics and Data Analysis, 31, 131-145.
[3] Chen, S. X. (2000), *Gamma kernel estimators for density functions*, Annals of the Institute of Statistical Mathematics, 52, 471-480.
[4] Jin, X. and Kawczak, J. (2003), *Birnbaum-saunders and lognormal kernel estimators for modelling durations in high frequency financial data*, Annals of Economics and Finance, 4, 103-124.
[5] Marchant, C. Bertin, K. Leiva, V. and Saulo, H. (2013). *Generalized birnbaumï-saunders kernel density estimators and an analysis of financial data*, Computational Statistics and Data Analysis, 63, 1-15.
[6] Parzen, E.(1962). *On estimation of a probability density function and mode*, Ann. Math. Statist., 33 : 1065-1076, 1962.
[7] Saulo, H. Leiva, V. Ziegelmann, F.A. and Marchant, C. (2013). *A nonparametric method for estimating asymmetric densities based on skewed birnbaum-saunders distributions applied to environmental data*, Stochastic Environmental Reseach and Risk Assessment, 27, 1479-1491.
[8] Scaillet, O. (2004), *Density estimation using inverse and reciprocal inverse gaussian kernels*, Journal of Nonparametric Statistics, 16, 217-226.
[9] Ziane, Y. Adjabi, S. and Zougab, N (2015). *Adaptive bayesian bandwidth selection in asymmetric kernel density estimation for nonnegative heavy-tailed data*, Journal of Applied Statistics, 42(8), 1645-1658.
[10] Zougab, N. Adjabi, S. and Kokonendji, C C. (2012). *Adaptive smoothing in associated kernel discrete functions estimation using bayesian approach*, Journal of Statistical Computation and Simulation, 83, 2219-2231.