APPROCHE BAYESIENNE DANS L'ESTIMATION NON PARAMÉTRIQUE DE LA DENSITÉ DES DONNÉES DE DÉNOMBREMENT PAR NOYAU ASSOCIÉ

Smail Adjabi[a], Nabil Zougab[a], Célestin C. Kokonendji[b]
[a]Laboratoire LAMOS, Université de Béjaia, 06000 Béjaia, Algérie
[b]Université de Franche-Comté, Laboratoire de Mathématiques de Besançon
UMR 6623, CNRS-UFC, 16 route de Gray, 25030 Besançon cedex, France
E-mail : adjabi@hotmail.com, nabilzougab@yahoo.fr, celestin.kokonendji@univ-fcomte.fr

Résumé

L'approche bayesienne pour la sélection de la fenêtre de lissage dans l'estimation de la fonction de masse de probabilité discrète par la méthode du noyau associé est une bonne alternative aux méthodes populaires classiques telles que la méthode plug-in et la technique de validation croisée. Dans ce travail, nous proposons l'approche bayesienne locale pour estimer le paramètre de lissage en considérant ce paramètre comme une quantité aléatoire avec une distribution a priori. En utilisant le critère de l'erreur quadratique intégrée (ISE), l'approche bayesienne est comparée aux méthodes plug-in et validation croisée sur des simulations de données générées par des fonctions discrètes connues et sur des données réelles de comptage. Les résultats montrent la supériorité de l'approche bayesienne sur les méthodes classiques particulièrement pour les échantillons de petite et moyenne taille.

Mots clés : : Fonction de masse de probabilité, Noyau discret associé, Plug-in, Validation croisée, Bayesienne locale, Distribution a priori, Erreur quadratique intégrée, Erreur quadratique moyenne intégrée

Abstract

The Bayesian approach to bandwidth selection in discrete associated kernel estimation of probability mass function (pmf) is a good alternative to the classical popular methods plug-in and the cross validation technique. In this work, we propose a local Bayesian approach to bandwidth selection treating the bandwidth h as a random quantity with a prior distribution. The performance of this proposed approach and the classical methods are compared by using simulations of data generated from known discrete functions using standard ISE (Integrated Square Error) as criterion. The Bayesian method is also applied to a real count data. The simulation results show the superiority of the Bayesian approach over the classical methods, in particular for small and moderate sample size.

Key words : Discrete function ; discrete associated kernel ; Plug-in, Cross validation ; Local bandwidth ; Prior distribution, Integral square error, Mean integral square error

1 Introduction

Pour estimer une fonction de densité de probabilité discrète (dite généralement fonction de masse de probabilité) par l'approche non paramétrique, l'estimateur empirique appelé aussi estimateur à

noyau de Dirac est souvent utilisé par les praticiens en raison de sa simplicité et ses bonnes propriétés asymptotiques. Cependant, cet estimateur n'est pas approprié pour des échantillons de petite ou moyenne taille (voir Senga Kiessé (2008)). Aitchison and Aitken (1976) ont proposé un estimateur à noyau discret pour estimer des fonctions discrètes. Le noyau discret utilisé par les auteurs n'a qu'une seule forme et n'est approprié que pour des données catégorielles et des distributions discrètes finies. Récemment, Kokonendji et al (2007), Senga Kiessé (2008) et Kokonendji and Senga Kiessé (2011) ont introduit la notion de noyau associé ainsi que deux classes de noyaux discrets, à savoir les noyaux discrets standards et les noyaux triangulaires discrets pour estimer des fonctions discrètes à support discret. En pratique, pour utiliser l'estimateur à noyau associé il faut choisir le noyau associé K et le paramètre de lissage h. Pour le noyau, le choix peut être adapté selon le support de la fonction inconnue à estimer. En revanche, le paramètre de lissage est un facteur important et crucial dans l'estimation de la fonction de densité par la méthode du noyau associé. De petites ou de grandes valeurs de h peuvent conduire à une estimation sous ou sur-lissée. Deux catégories de méthodes classiques ont été proposées. La première catégorie repose sur la minimisation de l'erreur quadratique moyenne intégrée (MISE). Cependant, le paramètre de lissage optimal obtenu dépend d'une ou plusieurs quantités inconnues. La deuxième catégorie est de type validation croisée, elle est intéressante en pratique car elle se laisse guider seulement par les observations. Cependant, les techniques de validation croisée peuvent produire plusieurs minimums locaux. Ces deux catégories ont tendance à fournir des estimateurs sous ou sur-lissés lorsque les données sont de petite ou moyenne taille. L'objectif de ce travail est de proposer l'alternative bayesienne pour le choix du paramètre de lissage, en particulier pour des échantillons de taille finie : petite ou moyenne. Le formalisme bayesien est caractérisé par le traitement du paramètre de lissage h comme une variable aléatoire, en lui associant une loi a priori $\pi(.)$. L'estimation bayesienne peut alors être obtenue ainsi via la moyenne de la loi a posteriori. Nous présentons l'approche bayésienne locale, qui consiste à traiter le paramètre de lissage localement (c'est-à-dire en chaque point x) en utilisant le noyau binomial et la distribution bêta comme loi a priori de h. Cette combinaison nous permet d'obtenir explicitement l'estimateur bayesien local. Les performances de cette approche bayesienne sera comparée par simulation et sur des données réelles aux performances des méthodes classiques, en utilisant le critère de l'erreur quadratique intégrée (ISE).

2 Estimateur à noyau associé discret

Définition
Soit $X_1, \ldots, X_n$ un n-échantillon issu d'une variable aléatoire discrète X de fonction de masse de probabilité inconnue f sur le support $\mathbb{T}$ ($\mathbb{T} = \mathbb{N}$ ou $\mathbb{T} = \{0, 1, 2, \ldots, m\}$). L'estimateur à noyau associé discret $\widehat{f}_h(x)$ de $f(x) = Pr(X_i = x)$ est de la forme :

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}) , \quad x \in \mathbb{T} \tag{1}$$

$$= \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i) , \quad x \in \mathbb{T}, \tag{2}$$

où K est la fonction noyau supposée être une densité de masse de probabilité et h est le paramètre de lissage. L'écriture (1) est la plus connue depuis les travaux de Rosenblatt (1956) et Parzen(1962). Quant à la seconde notation (2), elle est due à Chen(1999) dans la cas continu et Kokonendji and Senga kiessé (2011) dans le cas discret, dans le but d'adapter la fenêtre à la cible pour un "type de noyau" donné. $K_{x,h}$ est le noyau associé discret (dépendant de la cible x et du paramètre de lissage

h), supposé être une densité de masse de probabilité de support $\mathbb{S}_x$. Le noyau discret $K_{x,h}$ relié à la variable aléatoire discrète $\mathcal{K}_{x,h}$ est dit noyau associé discret de cible x et de fenêtre h si les conditions suivantes sont satisfaites :

$$x \in \mathbb{S}_x\ , \quad \lim_{h\to 0} \mathbb{E}(\mathcal{K}_{x,h}) = x \quad \text{et} \quad \lim_{h\to 0} \mathbb{V}(\mathcal{K}_{x,h}) \in \mathcal{V}(0)\ , \tag{3}$$

où $\mathcal{V}(0)$ est le voisinage de 0.

3 Choix du noyau

Il existe plusieurs types de noyaux associés discrets : Noyau de Dirac, triangulaire, binomial, Poisson et binomial négatif. Le choix du noyau peut être adapté selon le support de la densité à estimer. Dans ce travail, nous utiliserons le noyau binomial, à cause de sa propriété de sous dispersion. Considérons la distribution binomial $\mathcal{B}(N,p)$, avec $N \in \mathbb{N}^*$ et $p \in (0,1]$, la méthode de transformation Mode-Dispersion (Senga kiessé (2008)) permet la construction d'un noyau dépendant de x et de h : $\mathbb{B}_{x,h}$ associé à la variable aléatoire $\mathcal{B}_{x,h}$ de distribution $\mathcal{B}(x+1,(x+h)/(x+1))$ de support $\mathbb{S}_x = \{0,1,\ldots,x+1\}$ et de probabilité de succès $(x+h)/(x+1)$; ainsi $h \in [0,1]$. Le noyau associé binomial $\mathbb{B}_{x,h}$ est alors de la forme :

$$\mathbb{B}_{x,h}(y) = \frac{(x+1)!}{y!(x+1-y)!}\left(\frac{x+h}{x+1}\right)^{y}\left(\frac{1-h}{x+1}\right)^{x+1-y}\ , \forall y \in \mathbb{S}_x. \tag{4}$$

Le noyau binomial $\mathbb{B}_{x,h}$ est un noyau sous dispersé ($\mathbb{V}(\mathcal{B}_{x,h}) = \frac{(x+h)(1-h)}{x+1} < \mathbb{E}(\mathcal{B}_{x,h}) = x+h$), de mode autour de $x+h$ et il satisfait les conditions (3) d'un noyau associé.
L'estimateur de la densité de masse associé au noyau binomial est alors de la forme :

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{B}_{x,h}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x+1)!}{X_i!(x+1-X_i)!}\left(\frac{x+h}{x+1}\right)^{X_i}\left(\frac{1-h}{x+1}\right)^{x+1-X_i}\ , x \in \mathbb{N}.$$

Cet estimateur est meilleur que l'estimateur à noyau de Dirac (estimateur empirique) et l'estimateur à noyau triangulaire pour des échantillons de petite ou moyenne taille (Senga Kiessé (2008)). Il est également meilleur que l'estimateur à noyaux standards (Poisson et binomial négatif) à cause de sa propriété de sous dispersion.

4 Méthodes classiques de sélection du paramètre de lissage

• $AMISE$ estimateur
L'approximation de l'erreur quadratique moyenne intégrée $(MISE)$ notée $AMISE$ (Kokonendji and Senga Kiessé (2011)) s'écrit :

$$AMISE(h) = \frac{1}{n}\sum_{x\in\mathbb{N}} f(x)\mathbb{P}(\mathcal{K}_{x,h}=x) + \sum_{x\in\mathbb{N}}[f\{\mathbb{E}(\mathcal{K}_{x,h})\} - f(x) + \frac{1}{2}\mathbb{V}(\mathcal{K}_{x,h})f^{(2)}(x)]^2\ ,$$

où $f^{(2)}$ est la différence finie d'ordre 2 définie par :

$$f^{(2)} = \begin{cases} \{f(x+2)-2f(x)+f(x-2)\}/4 & \text{si } x \in \mathbb{N}\setminus\{0,1\}, \\ \{f(3)-3f(1)+2f(0)\}/4 & \text{si } x = 1, \\ \{f(2)-2f(1)+f(0)\}/2 & \text{si } x = 0. \end{cases}$$

L'optimal paramètre de lissage est alors :

$$h_{amise} = \arg\min_h AMISE(h).$$

Remarque
On ne peut pas obtenir la forme de h_{amise}, car $AMISE(h)$ dépend de f qui est inconnue. Pour calculer en pratique h_{amise}, on peut remplacer la densité f inconnue dans $AMISE(h)$ par l'estimateur empirique (estimateur associé au noyau de Dirac).
• Validation croisée estimateur
L'optimal paramètre de lissage h_{cv} de h s'obtient par :

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

avec

$$CV(h) = \sum_{x\in\mathbb{N}} \widehat{f}_h^2(x) - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{h,-i}(X_i) = \sum_{x\in\mathbb{N}} \left[\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\right]^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} K_{X_i,h}(X_j), \tag{5}$$

où $\widehat{f}_{h,-i}$ est calculé sans l'observation X_i.

5 Approche bayesienne

Nous allons proposer l'approche bayesienne locale qui consiste à estimer h en chaque point x pour lequel on veut estimer la densité. Dans cette approche, on supposera que h est une variable aléatoire de loi a priori $\pi(h)$. L'estimateur de Bayes de h sera obtenu à travers la loi a posteriori $\pi(h/observations)$.
Soit $\pi(h)$ la loi à priori de h, la loi a posteriori de h au point d'estimation x est de la forme :

$$\pi(h/x) = \frac{f(x)\pi(h)}{\int f(x)\pi(h)dh}. \tag{6}$$

Cependant, il n'est pas possible d'utiliser l'approche bayesienne directement, car h n'est pas le vrai paramètre du modèle. Pour remédier à ce problème, on introduit la fonction de masse discrète $f_h(x)$ définie par :

$$f_h(x) = \sum_{x\in\mathbb{T}} f(y)K_{x,h}(y) = \mathbb{E}[K_{x,h}(Y)], \tag{7}$$

où $K_{x,h}$ est le noyau discret associé supposé être une densité de masse et Y une variable aléatoire de densité f. Quand $h \to 0$, pour de grandes valeurs de n, $f_h(x)$ est proche de $f(x)$ (Kokonendji and Senga Kiessé, (2011)). La loi a posteriori devient.

$$\pi(h/x) = \frac{f_h(x)\pi(h)}{\int f_h(x)\pi(h)dh}. \tag{8}$$

Puisque f_h est inconnue, on utilise $\widehat{f}_h$ comme un estimateur naturel de f_h. La loi a posteriori est alors de la forme :

$$\widehat{\pi}(h/x, x_1, \ldots, x_n) = \frac{\widehat{f}_h(x)\pi(h)}{\int \widehat{f}_h(x)\pi(h)dh}.$$

Sous la fonction perte quadratique, l'estimateur de Bayes de h est la moyenne de la loi a posteriori donnée par :

$$\widehat{h}_n(x) = \int h\widehat{\pi}(h/x, x_1, \ldots, x_n)dh.$$

Comme $h \in (0, 1]$, la loi a priori choisie pour h est la loi beta de paramètres $\alpha > 0$ et $\beta > 0$ donnée par

$$\pi(h) = \frac{1}{B(\alpha, \beta)} h^{\alpha-1}(1-h)^{\beta-1}, \;\; h \in [0, 1].$$

L'estimateur $\widehat{h}_n(x)$ de h est alors de la forme :

$$\widehat{h}_n(x) = \frac{\sum_{i=1}^{n}\sum_{k=0}^{X_i} \frac{x^k}{(x+1-X_i)!k!(X_i-k)!} B(X_i+\alpha-k+1, x+\beta+1-X_i)}{\sum_{i=1}^{n}\sum_{k=0}^{X_i} \frac{x^k}{(x+1-X_i)!k!(X_i-k)!} B(X_i+\alpha-k, x+\beta+1-X_i)}, \; x \in \mathbb{N}.$$

Propriété
L'estimateur bayésien $\widehat{h}_n(x)$ de h converge presque sûrement vers 0 quand $n \to \infty$ pour $\alpha > 0$ et $\beta = \beta(n) = \beta_n$ avec $\beta_n \to \infty$ quand $n \to \infty$.
Démonstration. On montre que $\frac{\alpha}{\beta_n+\alpha+x+1} \leq \hat{h}_n(x) \leq \frac{x+1+\alpha}{\beta_n}$. En utilisant l'hypothèse $\beta_n \to \infty$ quand $n \to \infty$, $\widehat{h}_n(x) \to 0$ ps.
Remarque
Le taux de convergence obtenu en minimisant le $MISE$ est $1/\sqrt{n}$ (Kokonendji and Senga Kiessé (2011)). En choisissant dans la pratique $0 < \alpha << \beta_n = \sqrt{n}$, l'estimateur local bayesien converge vers 0 avec le même taux de convergence que celui obtenu par la minimisation du $MISE$.

6 Application

Application sur des données simulées

Nous comparons les performances de l'estimateur à noyau binomial basé sur l'approche bayesienne locale avec la méthode AMISE et la technique validation croisée pour le choix du paramètre de lissage, en utilisant deux densités cibles discrètes **F1** distribution de Poisson : $\mathcal{P}(8)$ et **F2** mélange de deux distributions de Poisson : $(2/5)\mathcal{P}(0.5) + (3/5)\mathcal{P}(10)$. Le critère de comparaison est une estimation de l'erreur quadratique intégrée (ISE) définie par :

$$\widehat{ISE} = \frac{1}{N_{sim}} \sum_{t=1}^{N_{sim}} \sum_{x \in \mathbb{N}} \left\{ \widehat{f}_h^{[t]}(x) - f(x) \right\}^2, \tag{9}$$

où $\widehat{f}_h^{[t]}$ est l'estimateur à noyau discret obtenu en utilisant l'échantillon t, N_{sim} est nombre de simulations égal à 100 et f la densité cible. Les résultats sont résumés dans le tableau 2.
La performance de l'approche bayesienne dépend du choix des paramètres de la loi à priori α et β. Pour le choix de ces paramètres, nous combinons trois valeurs de α avec trois valeurs de β, le couple (α, β) choisi sera celui qui donnera le plus petit estimateur du ISE donné dans (9). Les résultats sont résumés dans le tableau 1. Dans le tableau 3, on donne le temps d'exécution pour chaque méthode d'estimation du paramètre de lissage ainsi que le rapport du temps d'exécution de l'approche bayesienne sur le temps d'exécution des autres méthodes.

f	α	β	$\widehat{ISE}$	f	α	β	$\widehat{ISE}$
F1	0.5	5	0.00090	**F2**	0.5	5	0.00072
		10	0.00085			10	0.00071
		15	**0.00082**			15	0.00073
F1	1	5	0.00105	**F2**	1	5	0.00089
		10	0.00089			10	0.00072
		15	0.00086			15	**0.00070**
F1	5	5	0.00272	**F2**	5	5	0.00089
		10	0.00166			10	0.00257
		15	0.00129			15	0.00170

Table 1 : Calcul de $\widehat{ISE}$ pour le choix de α et β

On obtient pour **F1** : $(\alpha, \beta) = (0.5, 15)$ et pour **F2** : $(\alpha, \beta) = (1, 15)$.

f	Critère	n	$\widehat{ISE}_{h_{amise}}$	$\widehat{ISE}_{h_{cv}}$	$\widehat{ISE}_{h_{bayes}}$
F1	$\widehat{ISE}$	20	0.06302	0.06305	**0.05536**
		50	0.01547	0.01464	**0.01318**
		100	0.00544	0.00540	**0.00471**
		150	0.00454	0.00440	**0.00409**
		200	0.00214	0.00211	**0.00208**
		500	0.00143	0.00141	**0.00135**
		1000	0.00085	0.00084	**0.00082**
F2	$\widehat{ISE}$	20	0.01586	0.01404	**0.00966**
		50	0.00640	0.00692	**0.00618**
		100	0.00536	0.00519	**0.00493**
		150	0.00488	0.00464	**0.00457**
		200	0.00242	0.00231	**0.00205**
		500	0.00142	0.00138	**0.00121**
		1000	0.00082	0.00075	**0.00070**

Table 2 : Calcul de $\widehat{ISE}$ pour les densités cibles **F1** et **F2**

f	n	t_{amise}	t_{cv}	t_{bayes}	$\frac{t_{amise}}{t_{bayes}}$	$\frac{t_{cv}}{t_{bayes}}$
F1	50	16.06	2.33	**0.24**	31.16	9.70
	100	16.17	4.02	**0.43**	37.60	9.34
	200	16.30	4.14	**0.52**	31.34	7.96
	500	16.33	11.47	**1.28**	12.75	8.96
	1000	16.11	22.87	**2.64**	6.10	8.66
F2	50	16.11	1.17	**0.25**	64.44	4.68
	100	16.14	2.47	**0.49**	32.93	5.04
	200	16.04	4.72	**0.80**	20.05	5.97
	500	16.30	13.15	**2.45**	6.65	5.36
	1000	16.22	30.65	**5.67**	2.86	5.40

Table 3 : Temps d'exécution (en secondes) pour l'estimation du paramètre de lissage par les différentes méthodes

Les résultats obtenus (tableaux 2 et 3) montrent clairement l'avantage que ce soit du point de vue critère ISE ou temps d'exécution de l'approche bayesienne comparativement aux méthodes classiques.

Application sur des données réels

Les données décrivent le nombre de buts marqués par chaque joueur durant une saison de football (championnat et coupe d'Algérie).
Pour comparer les différentes méthodes d'estimation du paramètre de lissage, on utilise l'empirique erreur quadratique intégrée $ISE^0 = \sum_{x\in\mathbb{N}}(\widehat{f}_h(x) - f_0(x))^2$, où f_0 est l'estimateur empirique de la densité.

Méthode	AMISE estimateur	CV technique	Approche bayesienne
Critère ISE$^0(\times 10^4)$	9.51	10.41	**5.17**
Temps d'exécution t_h	16.11	3.01	**0.43**

Table 4. ISE et temps d'execution pour les différentes méthodes d'estimation du paramètre de lissage

On constate (tableau 4) que l'estimateur bayesien possède le plus petit ISE^0 et le plus petit temps d'exécution comparativement aux autres méthodes (AMISE et Validation croisée). La figure 1 montre que les données sont multi-modal, mélange de trois distributions de Poisson.

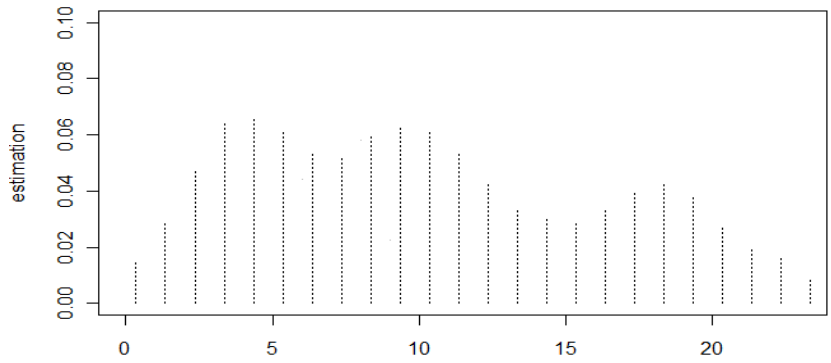

FIGURE 1 – Estimation de la densité des données : nombre de buts par la méthode du noyau associé utilisant le noyau binomial et la loi a priori beta de paramètres $(\alpha, \beta) = (0.5, 15)$

Bibliographie

[1] Abdous, B. and Kokonendji, C. C. (2009), *Consistency and asymptotic normality for discrete associated-kernel estimator*, African Diaspora Journal of Mathematics, 8 :2, 63-70.
[2] Aitchison, J. and Aitken, C. G. G. (1976), *Multivariate binary discrimination by the kernel method*, Biometrika, 63, 413-420.
[3] Chen, S. X. (1999), *Beta kernels estimators for density functions*, Comput. Statist. Data Anal. 31, 131-145.
[4] Kokonendji, C. C. Senga Kiessé, T. and Zocchi, S. S. (2007), *Discrete triangular distributions and non-parametric estimation for probability mass function*, Journal of Nonparametric Statistics, 19 :6, 241-254.
[5] Kokonendji, C.C. and Zocchi, S. S. (2010), *Extensions of discrete triangular distributions and boundary bias in kernel estimation for discrete functions*, Statistics and Probability Letters, 80, 1655-1662.
[6] Kokonendji, C. C. and Senga Kiessé, T. (2011), *Discrete associated kernels method and extensions*, Statistical Methodology, 8, 497-516.
[7] Parzen, E. (1962), *On estimation of a probability density function and mode*, Annals of Mathematical Statistics, 33, 1065-1076.
[8] Rosenblatt, M. (1956), *Remarks on some nonparametric estimates of a density function*, Annals of Mathematical Statistics, 27, 832-837
[9] Senga Kiessé, T. (2008) *Approche non-paramétrique par noyaux associés discrets des données de dénombrement* Thèse de Doctorat, Université de Pau, France.
[10] Zougab, N. Adjabi, S. and Kokonendji, C.C. (2012), *Binomial kernel and Bayes local bandwidth in discrete functions estimation. Journal of Nonparametric Statistics*, 24 :3, 783-795.