APPROCHE BAYÉSIENNE LOCALE DANS L'ESTIMATION PAR NOYAUX ASSOCIÉS DISCRETS MULTIVARIÉS

Nawal BELAID[a], Smail ADJABI[a], Célestin C. KOKONENDJI[b], Nabil ZOUGAB[a]
[a]Laboratoire LAMOS, Université de Béjaia, 06000 Béjaia, Algérie
[b]Université de Franche-Comté, LMB UMR 6623 CNRS-UFC, Besançon Cedex, France
E-mail : belaidnawelro@hotmail.fr, adjabi@hotmail.com,
celestin.kokonendji@univ-fcomte.fr, nabilzougab@yahoo.fr

Résumé

Dans ce travail, nous proposons l'approche bayésienne locale pour la sélection de la matrice des fenêtres de lissage dans l'estimation de la fonction de masse de probabilité multivariée par la méthode du noyau associé. Nous traitons la matrice des fenêtres comme une variable aléatoire de loi a priori conjuguée avec le noyau utilisé. La forme explicite de l'estimateur de cette matrice est obtenue par la fonction perte quadratique. Les performances de l'approche proposée sont comparées avec la méthode classique de validation croisée (LSCV) sur des données simulées. Les résultats obtenus montrent que l'approche bayésienne locale est meilleure que LSCV en terme de l'erreur quadratique intégrée.

Mots clé : Loi bêta, noyau binomial, noyau Dirac Discrete Uniform, validation croisée, distribution a priori, erreur quadratique intégrée

Abstract

In this work we propose a Bayesian local approach to select the matrix of bandwidths in discrete multivariate associated kernel estimation of probability mass function. We treat the bandwidths as a random variable with prior conjugate distribution to the kernel used. We obtain the explicit form of this matrix under the quadratic loss function. The performance of the proposed approach are compared with the classical cross validation method (LSCV) on simulated data. The obtained results show that the Bayesian local method performs better than cross-validation in terms of integrated squared error.

Key words : Beta distribution, binomial kernel, cross-validation, Dirac Discrete Uniform kernel, prior distribution, integral square error

1 Introduction

Un des problèmes rencontrés en statistique est celui de l'estimation des fonctions multivariées à support $\mathbb{T}^d \subseteq \mathbb{Z}^d$ avec $(d \geq 2)$. Cette estimation trouve ses applications dans divers domaines tels que l'économie, la médecine, le sport, etc.
Soient $\mathbf{X}_1, \ldots, \mathbf{X}_n$ des vecteurs aléatoires indépendants et identiquement distribués (iid) de fonction de masse de probabilité (fmp) multivariée commune inconnue f à estimer sur $\mathbb{T}^d$; on notera f_n la fmp

associée aux n vecteurs aléatoires iid $\mathbf{X}_1, \ldots, \mathbf{X}_n$. L'estimateur à noyau associé multivarié de $f(\mathbf{x})$ est de la forme :

$$\widehat{f}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{x},H}(\mathbf{X}_i), \qquad \mathbf{x} \in \mathbb{T}^d, \tag{1}$$

où $K_{\mathbf{x},H}(.)$ est le "noyau associé discret multivarié" dépendant du vecteur cible $\mathbf{x}$ et de la matrice des fenêtres H. Les performances de l'estimateur (1) dépendent de H qui contrôle le degré de lissage et de la forme de l'orientation du noyau. Cette matrice contient $d(d+1)/2$ éléments indépendants à estimer dans $\mathbb{R}^d$, $d \geq 2$. Pour remédier à la complexité du choix de la matrice H, on supposera que cette dernière est une matrice diagonale ; $H = Diag_d(h_j)_{j=1,\ldots,d}$, voir Bouezmarni et Roumbouts (2010). Des méthodes classiques ont été proposées dans la littérature pour le choix de la matrice de lissage telles que plug-in et validation croisée, voir Chacón et Duong (2010, 2011). L'approche alternative aux méthodes classiques est l'approche bayésienne proposée dans le cas continu univarié par Brewer (1998), Zhang et al. (2006) et De Lima et Atuncar (2010) dans le cas continu multivarié.
L'objectif de ce travail est de proposer l'approche bayésienne locale pour le choix de la matrice des fenêtres. Nous ulilisons les noyaux associés discrets multivariés (binomial et Dirac Discrete Uniform) et la distribution bêta comme la loi a priori pour obtenir la forme exacte de la matrice des fenêtres H. Cette méthode a été étudiée par Zougab et al. (2012) dans le cas discret univarié, Gangopadhyay et Cheung (2002) dans le cas continu univarié et par De Lima et Atuncar (2010) dans le cas continu multivarié. Les performances de l'approche bayésienne sont comparées sur des données simulées avec celles de la méthode classique de validation croisée, en terme de l'erreur quadratique intégrée (ISE).

2 Noyau associé discret multivarié

Dans cette section on présente la notion du noyau associé discret multivarié ainsi que le noyau produit. Deux exemples illustratifs seront présentés.

2.1 Définitions

Definition 1 *(Noyau associé discret multivarié) Soient $\boldsymbol{x} \in \mathbb{T}^d \subseteq \mathbb{Z}^d$ et H une matrice des fenêtres avec $\mathbb{T}^d$ le support de la fmp f à estimer. Une fmp $K_{\boldsymbol{x},H}(.)$ de support $\mathbb{S}_{\boldsymbol{x},H} \subseteq \mathbb{Z}^d$ est appelée noyau associé discret multivarié si*

$$\boldsymbol{x} \in \mathbb{S}_{\boldsymbol{x},H} \tag{2}$$

$$\mathbb{E}(\mathcal{Z}_{\boldsymbol{x},H}) = \boldsymbol{x} + \boldsymbol{a}(\boldsymbol{x}, H) \tag{3}$$

$$Cov(\mathcal{Z}_{\boldsymbol{x},H}) = B(\boldsymbol{x}, H), \tag{4}$$

où $\boldsymbol{a}(\boldsymbol{x}, H)$ et $B(\boldsymbol{x}, H)$ tendent vers le vecteur nul et la matrice nulle respectivement quand $H \to 0_d$ (0_d est une matrice carrée nulle d'ordre d) et $\mathcal{Z}_{\boldsymbol{x},H}$ est une vecteur aléatoire discret de loi $K_{\boldsymbol{x},H}$.

Definition 2 *(Noyau associé discret multivarié produit) Soient $\mathbb{T}_1^{[j]}$ le support des marges univariés de f pour $j = 1, \ldots, d$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$ le vecteur cible, et $h_1, h_2, \ldots, h_d$ sont les fenêtres de lissages univariés. $K_{x_j,h_j}^{[j]}$ est le j-ème noyau associé discret univarié de support $\mathbb{S}_{x_j,h_j}$. Le "noyau associé discret multivarié produit" est définit comme suit :*

$$K_{\boldsymbol{x},H}(\cdot) = \prod_{j=1}^{d} K_{x_j,h_j}^{[j]}(\cdot), \;\; \forall x_j \in \mathbb{T}_1^{[j]} \subseteq \mathbb{Z},$$

2.2 Exemples

Exemple 1 *Le noyau catégoriel multivarié "Dirac Discrete Uniform" introduit par Aitchison et Aitken (1976) et Racine et Li (2004) est donné comme suit :*

$$DirDU_{\boldsymbol{x},H,\boldsymbol{c}}(\boldsymbol{y}) = \prod_{j=1}^{d}(1-h_j)^{\boldsymbol{I}_{y_j=x_j}}\left(\frac{h_j}{c_j-1}\right)^{1-\boldsymbol{I}_{y_j=x_j}}, \tag{5}$$

où $\mathbb{S}_{\boldsymbol{x},\boldsymbol{c}} = \times_{j=1}^{d}\{0,1,\ldots,c_j-1\} = \mathbb{T}^d$*, avec* $\times_{j=1}^{d}$ *est le produit cartésien des ensembles* $\{0,1,\ldots,c_j-1\}$*,* $c_j \in \{2,3,\ldots\}\ \forall j = 1,\ldots,d$*,* $H = Diag_d(h_j)$*,* $\boldsymbol{1}_A$ *est la fonction indicatrice de* A*.*

Ce noyau vérifie bien les conditions (2)-(4), en effet on a

$$\begin{aligned}
\mathbb{E}(\mathcal{Z}_{\mathbf{x},H}) &= \sum_{\mathbf{y}\in\mathbb{T}^d}(y_1,\ldots,y_d)\left[\prod_{j=1}^{d}(1-h_j)^{\mathbf{1}_{y_j=x_j}}\left(\frac{h_j}{c_j-1}\right)^{1-\mathbf{1}_{y_j=x_j}}\right] \\
&= \sum_{y_1=0}^{c_1-1} y_1\left[\prod_{j=1}^{d}(1-h_j)^{\mathbf{1}_{y_j=x_j}}\left(\frac{h_j}{c_j-1}\right)^{1-\mathbf{1}_{y_j=x_j}}\right],\ldots, \\
&\quad \sum_{y_d=0}^{c_d-1} y_d\left[\prod_{j=1}^{d}(1-h_j)^{\mathbf{1}_{y_j=x_j}}\left(\frac{h_j}{c_j-1}\right)^{1-\mathbf{1}_{y_j=x_j}}\right] \\
&= \left(x_1+h_1\left(1-\frac{x_1}{c_1-1}+\frac{h_1c_1}{2}\right),\ldots,x_d+h_d\left(1-\frac{x_d}{c_d-1}+\frac{h_dc_d}{2}\right)\right)^\top \\
&= \mathbf{x}+H\left(1-\frac{x_1}{c_1-1}+\frac{h_1c_1}{2},\ldots,1-\frac{x_d}{c_d-1}+\frac{h_dc_d}{2}\right)^\top \\
&= \mathbf{x}+\mathbf{a}(\mathbf{x},H)\,,
\end{aligned}$$

avec $\mathbf{a}(\mathbf{x},H)\to 0$ quand $H\to 0_d$, et

$$\begin{aligned}
cov(\mathcal{Z}_{\mathbf{x},H}) &= cov\left(\prod_{j=1}^{d}\mathcal{Z}^{[j]}_{x_j,h_j,c_j}\right) \\
&= Diag_d\left(var(\mathcal{Z}^{[j]}_{x_j,h_j,c_j})\right)_j \\
&= H\ Diag_d\left(x_j^2\frac{c_j^2(1-h_j)-c_j}{(c_j-1)^2}-x_j\frac{c_j^2(1-h_j)-c_j}{c_j-1}+\frac{c_j}{2}\left(\frac{2c_j-1}{3}-\frac{h_jc_j}{2}\right)\right)_j \\
&= B(\mathbf{x},H)\,,
\end{aligned}$$

avec $B(\mathbf{x},H)\to 0_d$ lorsque $H\to 0_d$.
L'estimateur de la fmp associé au noyau Dirac Discrete Uniform est alors de la forme :

$$\widehat{f}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}(1-h_j)^{\mathbf{1}_{X_{ij}=x_j}}(\frac{h_j}{c_j-1})^{1-\mathbf{1}_{X_{ij}=x_j}}, \tag{6}$$

Exemple 2 *Le noyau "Binomial" est définit dans le support* $\mathbb{S}_{x,h} = \{0, 1, \ldots, x+1\}$ *avec* $x \in \mathbb{T} = \mathbb{N}$ *et* $h \in]0, 1]$*, comme suit :*

$$B_{x,h}(y) = \frac{(x+1)!}{y!(x+1-y)!} \left(\frac{x+h}{x+1}\right)^{y} \left(\frac{1-h}{x+1}\right)^{x+1-y}, y \in \mathbb{S}_{x,h}. \tag{7}$$

Notons que $B_{x,h}$ *est la fmp de la loi binomial pour un nombre d'essais* $(x+1)$ *et une probabilité de succès* $(x+h)/(x+1)$*, avec* $\mathbb{E}(\mathcal{Z}_{x,h}) = x + h$ *et* $var(\mathcal{Z}_{x,h}) = (x+h)(1-h)/(x+1)$*. Pour plus de détails, voir Kokonendji et Senga Kiessé (2011).*

L'estimateur de la fmp associé au noyau binomial produit est alors de la forme :

$$\widehat{f}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \frac{(x_j+1)!}{X_{ij}!(x_j+1-X_{ij})!} \left(\frac{x_j+h_{ij}}{x_j+1}\right)^{X_{ij}} \left(\frac{1-h_{ij}}{x_j+1}\right)^{x_j+1-X_{ij}}. \tag{8}$$

3 Approche bayésienne locale pour le choix de la matrice des fenêtres

Nous proposons d'estimer la matrice des fenêtres par l'approche bayésienne locale en utilisant les noyaux binomial produit et Dirac Discrete Uniform donnés dans (5) et (7). Afin d'estimer localement H, nous définissons le modèle $f_n(\mathbf{x}) = \mathbb{E}(K_{\mathbf{x},H}(\mathbf{y}))$ où $\mathbf{y}$ est un vecteur de variables aléatores de loi f, et on considère H comme une matrice de variables aléatoires de loi a priori $\pi(H)$. A partir de la formule de Bayes, la loi a posteriori prend la forme suivante :

$$\pi(H|\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_n) = f_n(\mathbf{x})\pi(H) \left[\int_{\mathcal{M}} f_n(\mathbf{x})\pi(H)dH\right]^{-1},$$

où $\mathcal{M}$ est l'ensemble des matrices diagonales symétriques définies positives. Comme le modèle $f_n(\mathbf{x})$ est inconnu, on le remplace par son estimateur (1). Ainsi une estimation de la loi a posteriori $\pi(H|\mathbf{x})$ est :

$$\widehat{\pi}(H|\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_n) = \widehat{f}_n(\mathbf{x})\pi(H) \left[\int_{\mathcal{M}} \widehat{f}_n(\mathbf{x})\pi(H)dH\right]^{-1}. \tag{9}$$

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

$$\widehat{H}(\mathbf{x}) = \int_{\mathcal{M}} H\widehat{\pi}(H|\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_n)dH.$$

3.1 Les formes explicites de l'estimateur de H

Nous supposons que la loi a priori de $H = Diag_d(h_j)$ est une loi bêta de paramètres α et β donnée par :

$$\pi(h_j) = \frac{1}{\mathbf{B}(\alpha, \beta)} h_j^{\alpha-1}(1-h_j)^{\beta-1}, \;\; j = 1, \ldots, d, \;\; h_j \in (0, 1], \tag{10}$$

où

$$\mathbf{B}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \;\; \alpha, \;\; \beta > 0.$$

En remplaçant (6) et (10) puis (8) et (10) dans (9), les estimateurs de $\widehat{H}(\mathbf{x}) = (\widehat{h}_1(x_1), \ldots, \widehat{h}_d(x_d))$ sous la fonction perte quadratique avec les noyaux Dirac Discrete Uniform et binomial sont donnés respectivement par :

$$\begin{aligned}\widehat{h}_j(x_j)_{DirDU} &= \sum_{i=1}^{n}\left(\frac{\mathbf{B}(1-(\mathbf{1}_{X_{ij}=x_j})+\alpha+1,(\mathbf{1}_{X_{ij}=x_j})+\beta)}{c_j-1^{1-(\mathbf{1}_{X_{ij}=x_j})}}\prod_{\substack{k=1\\k\neq j}}^{d}\frac{\mathbf{B}(1-(\mathbf{1}_{X_{ik}=x_k})+\alpha,(\mathbf{1}_{X_{ik}=x_k})+\beta)}{(c_k-1)^{1-(\mathbf{1}_{X_{ik}=x_k})}}\right)\\ &\quad\times\left(\sum_{i=1}^{n}\prod_{s=1}^{d}\frac{\mathbf{B}(1-(\mathbf{1}_{X_{is}=x_s})+\alpha,(\mathbf{1}_{X_{is}=x_s})+\beta)}{(c_s-1)^{1-(\mathbf{1}_{X_{is}=x_s})}}\right)^{-1},\end{aligned}$$

et

$$\begin{aligned}\widehat{h}_j(x_j)_B &= \sum_{i=1}^{n}\sum_{k_1=0}^{X_{ij}}\frac{(x_j+1)!x_j^{k_1}\mathbf{B}(X_{ij}-k_1+\alpha+1,x_j-X_{ij}+\beta+1)}{(x_j+1-X_{ij})!k_1!(X_{ij}-k_1)!(x_j+1)^{x_j+1}}\\ &\quad\times\prod_{\substack{m=1\\m\neq j}}^{d}\sum_{k=0}^{X_{im}}\frac{(x_m+1)!x_m^{k}\mathbf{B}(X_{im}-k+\alpha,x_m-X_{im}+\beta+1)}{(x_m+1-X_{im})!k!(X_{im}-k)!(x_m+1)^{x_m+1}}\\ &\quad\times\left(\sum_{i=1}^{n}\prod_{s=1}^{d}\sum_{k=0}^{X_{is}}\frac{(x_s+1)!x_s^{k}\mathbf{B}(X_{is}-k+\alpha,x_s-X_{is}+\beta+1)}{(x_s+1-X_{is})!k!(X_{is}-k)!(x_s+1)^{x_s+1}}\right)^{-1}.\end{aligned}$$

4 Application numérique

Dans cette section, on présente une étude de simulation pour comparer les performances de l'approche bayésienne locale avec la méthode classique globale de validation croisée (LSCV), qui consiste à minimiser la fonction $LSCV(H)$ donnée par :

$$LSCV(H) = \sum_{x\in\mathbb{T}_d}\left[\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}K^{[j]}_{x_j,h_j}(X_{ij})\right]^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{\ell\neq i}\prod_{j=1}^{d}K^{[j]}_{X_{ij},h_j}(X_{\ell j}).$$

Par conséquent, on a :

$$H_{opt} = \arg\min_{H\in\mathcal{M}} LSCV(H).$$

On considère deux densités cibles discrètes multivariées, leurs expressions sont données comme suit : **F1** est le produit de deux distributions de Poisson de moyenne identique $\mu = 4 : f(x_1,x_2) = \frac{e^{-4}4^{x_1}}{x_1!}\times\frac{e^{-4}4^{x_2}}{x_2!}$, $(x_1,x_2)\in\mathbb{N}^2$; **F2** est le produit d'un mélange de deux distributions de Poisson de moyennes $\mu_1 = 5$ et $\mu_2 = 12 : f(x_1,x_2) = \prod_{k=1}^{2}\left(\frac{2}{4}\frac{e^{-5}5^{x_k}}{x_k!}+\frac{2}{4}\frac{e^{-12}12^{x_k}}{x_k!}\right)$, $(x_1,x_2)\in\mathbb{N}^2$. Le critère de comparaison est une estimation de l'ISE définie par :

$$ISE = \sum_{\mathbf{x}\in\mathbb{N}^2}[\widehat{f}_n(\mathbf{x}) - f(\mathbf{x})]^2.$$

Les résultats obtenus pour un nombre de simulations égale à 100 sont résumés dans la Table 1. Dans la Table 2 on donne le temps d'exécution pour chaque méthode d'estimation de la matrice des fenêtres.
Les résultats donnés dans les Table 1 et Table 2 indiquent clairement l'avantage de l'approche bayésienne locale par rapport à la méthode classique LSCV en terme de ISE et en temps d'exécution dans le cas du noyau Dirac Discrete Uniform.

TABLE 1 – Valeurs de ISE ($ISE_{LSCV}, ISE_{Bayes-local}$) pour $d = 2$.

f	n	$DirDU$	$binomial$
F1	20	(4.17e-02,**9.04e-05**)	(5.24e-03,**5.43e-07**)
	100	(1.67e-02,**1.65e-05**)	(4.29e-04,**1.02e-07**)
	200	(8.7e-03,**6.68e-06**)	(2.10e-04,**7.98e-08**)
	500	(2.27e-03,**4.89e-08**)	(1.67e-04,**7.33e-10**)
F2	20	(4.91e-02,**3.84e-05**)	(3.92e-03,**2.32e-06**)
	100	(3.72e-03,**9.51e-08**)	(6.03e-04,**2.51e-08**)
	200	(1.22e-03,**5.22e-08**)	(3.68e-04,**1.24e-09**)
	500	(1.01e-03,**1.94e-09**)	(1.48e-05,**4.02e-10**)

TABLE 2 – Comparaison du temps d'exécution (en secondes) pour une simulation pour F1 (t_{LSCV},$t_{Bayes-local}$).

f	n	$DirDU$	$binomial$
F1	20	(7.453,**5.567**)	(**8.678**,10.213)
	100	(13.521,**10.679**)	(**15.120**,18.542)
	500	(45.567,**34.546**)	(**48.897**,53.895)

Bibliographie

[1] Aitchison, J. Aitken, C.G.G. (1976). *Multivariate binary discrimination by the kernel method*, Biometrika, 63, 413–420.

[2] Bouezmarni, T. Roumbouts, J.V.K. (2010). *Nonparametric density estimation for multivariate bounded data*, Journal of Statistical Planning and Inference, 140, 139–152.

[3] Brewer, M.J. (1998). *A modelling approach for bandwidth selection in kernel density estimation*, In : Proceedings of COMPSTAT Physica Verlag, Heidelberg, 203–208.

[4] Chacón, J.E. Duong, T. (2011). *Unconstrained pilot selectors for smoothed cross-validation*, Australian and New Zealand Journal of Statistics, 53, 331–351.

[5] De Lima, M.S. Atuncar, G.S. (2010). *A Bayesian method to estimate the optimal bandwidth for multivariate kernel estimator*, Journal of Nonparametric Statistics, 23, 137–148.

[6] Duong, T. (2004). *Bandwidth Selectors for Multivariate Kernel Density Estimation*, Ph.D. Thesis Manuscript to University of Western Australia, Perth, Australia.

[7] Gangopadhyay, A.K. Cheung, K.N. (2002). *Bayesian approach to the choice of smoothing parameter in kernel density estimation*, Journal of Nonparametric Statistics, 14, 655–664.

[8] Kokonendji, C.C. Senga Kiessé, T. (2011). *Discrete associated kernels method and extensions*, Statistical Methodology, 8, 497–516.

[9] Racine, J.S. Li, Q. (2004). *Nonparametric estimation of regression functions with both categorical and continuous data*, Journal of Econometrics. 119, 99–130.

[10] Zhang, X. King, M.L. Hyndman, R.J. (2006). *A Bayesian approach to bandwidth selection for multivariate kernel density estimation*, Computational Statistics and Data Analysis, 50, 3009–3031.

[11] Zougab, N. Adjabi, S. Kokonendji, C.C., (2012). *Binomial kernel and Bayes local bandwidth in discrete functions estimation*, Journal of Nonparametric Statistics, 24, 783–795.