Estimation de densité par noyau bêta bivarié avec structure de corrélation

Sobom M. Somé, Francial G. Libengué & Célestin C. Kokonendji

Université de Franche-Comté
Laboratoire de Mathématiques de Besançon - UMR 6623 CNRS-UFC
16 route de Gray, 25030 Besançon cedex, France.
prenom.nom@univ-fcomte.fr

Résumé. L'objet de cette communication est de présenter le noyau bêta bivarié avec une structure de corrélation, introduite au niveau de la matrice des fenêtres. Ce noyau associé est conçu par une variante de la méthode mode-dispersion et est utilisé pour l'estimation de densités sur $[0,1]^2$. Des propriétés de l'estimateur sont examinées, en particulier les biais de bordure, et ensuite comparées à ceux du cas produit. Une étude par simulation ainsi qu'une application aux données réelles seront présentées, avec une sélection de la matrice des fenêtres optimales effectuée par validation croisée.

Mots-clés. Estimation non-paramétrique, biais de bord, noyau asymétrique.

Abstract. The purpose of this communication is to present the bivariate beta kernel with correlation structure, which is introduced in the bandwith matrix. This associated kernel is built with a mode-dispersion principle and is used for density estimation on $[0,1]^2$. Some properties of the estimator are examined, in particular the boundary bias, and then compared to the product case. A simulation study and an application will be provided with optimal bandwith matrix parameters selected by cross validation.

Keywords. Asymmetric kernel, boundary bias, nonparametric estimation.

1 Introduction

Dans cette communication, nous nous intéressons à l'estimation de densités sur $[0,1]^2$. Les noyaux classiques multivariés (p. ex. normal, Epanechnikov) ne sont pas adaptés pour estimer de telles densités. En effet, ces estimateurs assignent des poids en dehors du support de la densité à estimer, causant ainsi des problèmes de biais de bordure. Pour y remédier, Bouezmarni et Rombouts (2010) ont proposé, dans le même esprit de Chen (1999, 2000), des estimateurs à noyaux produits pour des supports partiellement ou totalement bornés ; cependant, la matrice des fenêtres est diagonale.

Nous proposons ici un noyau bêta bivarié qui possède une structure de corrélation. Pour cela, on choisit une corrélation introduite à partir de deux noyaux bêta univariés et indépendants selon la méthode de Sarmanov (1966) ; voir aussi Lee (1996).

Ce noyau associé est naturellement approprié pour l'estimation non-paramétrique de données (bornées) sur $[0,1]^2$ ayant ou non une structure de corrélation entre les composantes bivariées. C'est pourquoi, dans ce qui suit, nous proposons une définition des noyaux associés multivariés, incluant les noyaux classiques multivariés. Ensuite, nous construisons le noyau bêta bivarié de Sarmanov et présentons certaines propriétés de son estimateur. Des simulations et application illustreront ce travail. À chaque fois, nous observerons les différences avec le noyau bêta produit.

2 Noyau bêta bivarié avec corrélation

Pour tout $d \in \{2,3,\ldots\}$, on notera par $\mathbb{T}_d \left(\subseteq \mathbb{R}^d\right)$ le support de la densité multivariée f à estimer. Soit $x = (x_1,\ldots,x_d)^T \in \mathbb{T}_d$ le point d'estimation et $\mathbf{H}_d = (h_{ij})_{i,j=1,\ldots,d}$ une matrice des fenêtres, symétrique et définie positive. Un *noyau associé multivarié* $K_{x,\mathbf{H}_d}$ est défini comme toute densité de probabilité paramétrée par x et $\mathbf{H}_d$, de support $\mathbb{S}_{x,\mathbf{H}_d}$ et satisfaisant les conditions suivantes :

$$x \in \mathbb{S}_{x,\mathbf{H}_d},\ \ \mathbb{E}\left(\mathcal{Z}_{x,\mathbf{H}_d}\right) = x + A(x,\mathbf{H}_d)\ \text{ et }\ \mathrm{Cov}\left(\mathcal{Z}_{x,\mathbf{H}_d)}\right) = \mathbf{B}(x,\mathbf{H}_d), \tag{1}$$

où $\mathcal{Z}_{x,\mathbf{H}_d}$ est un vecteur aléatoire de densité $K_{x,\mathbf{H}_d}$, $A(x,\mathbf{H}_d) \to 0$ et $\mathbf{B}(x,\mathbf{H}_d) \to \mathbf{0}_d$ lorsque $\mathbf{H}_d \to \mathbf{0}_d$. En particulier, les noyaux classiques donnés dans Scott (1992) sont des noyaux associés multivariés. En effet, soit $\mathcal{K}$ un noyau classique sur $\mathbb{S}_d \subseteq \mathbb{R}^d$, centré en $\mu_{\mathcal{K}} = 0$ et de matrice de variance-covariance $\mathbf{\Sigma}_{\mathcal{K}}$; alors, pour tout $x \in \mathbb{T}_d = \mathbb{R}^d$ et $\mathbf{H}_d$ une matrice des fenêtres, $K_{x,\mathbf{H}_d}(\cdot) = (1/\det\mathbf{H}_d)\mathcal{K}\left\{\mathbf{H}_d^{-1}(x-\cdot)\right\}$ est un noyau associé multivarié avec $\mathbb{S}_{x,\mathbf{H}_d} = x - \mathbf{H}_d\mathbb{S}_d$, $A(x,\mathbf{H}_d) = 0$ et $\mathbf{B}(x,\mathbf{H}_d) = \mathbf{H}_d\mathbf{\Sigma}_{\mathcal{K}}\mathbf{H}_d$.

De manière analogue aux cas univariés (voir Kokonendji et Senga Kiéssé (2011) pour les discrets, Kokonendji et Libengué (2013) pour les continus), un noyau associé multivarié $K_{x,\mathbf{H}_d}$ est intrinsèquement lié à la cible $x \in \mathbb{T}_d$ et à la matrice des fenêtres $\mathbf{H}_d$ ayant au plus $d(d+1)/2$ paramètres différents. Par conséquent, on peut construire un noyau associé multivarié à partir de toute densité de probabilité multivariée K_θ qu'on appelle parfois «type de noyau», dépendant de paramètres $\theta \in \Theta \subseteq \mathbb{R}^{d(d+3)/2}$, de support $\mathbb{S}_\theta$, et admettant un moment d'ordre 2. Par exemple, partant d'un type de noyau K_θ, on construit un noyau associé multivarié en résolvant un système $\theta = (x,\mathbf{H}_d)$. Une solution, si elle existe, $\theta(x,\mathbf{H}_d)$ de $\theta = (x,\mathbf{H}_d)$ conduit au noyau associé multivarié, noté $K_{\theta(x,\mathbf{H}_d)}$, de support $\mathbb{S}_{\theta(x,\mathbf{H}_d)}$ et vérifiant les mêmes caractéristiques données en (1) avec les notations $\mathcal{Z}_{\theta(x,\mathbf{H}_d)}$, $A_\theta(x,\mathbf{H}_d)$ et $\mathbf{B}_\theta(x,\mathbf{H}_d)$.

Le noyau bêta bivarié avec corrélation est construit à partir de densité du même nom. En effet, on considère d'abord deux distributions bêta univariées, de densités

$$g_i(t) = \frac{1}{\mathscr{B}(a_i,b_i)} t^{a_i-1}(1-t)^{b_i-1}\mathbb{1}_{[0,1]}(t), \quad i = 1,2, \tag{2}$$

où $a_i > 0$, $b_i > 0$, $\mathscr{B}(a_i, b_i) = \Gamma(a_i + b_i)/\{\Gamma(a_i)\Gamma(b_i)\}$ et $\Gamma(\cdot)$ est la fonction gamma usuelle. On rappelle que g_i a pour moyenne et variance respectivement

$$\mu_i = \frac{a_i}{a_i + b_i} = \mu_i(a_i, b_i) \quad \text{et} \quad \sigma_i^2 = \frac{a_i b_i}{(a_i + b_i)^2(a_i + b_i + 1)} = \sigma_i^2(a_i, b_i). \tag{3}$$

De plus, g_i est unimodale pour $a_i > 1$ et $b_i > 1$, de mode et dispersion respectivement

$$M_i(a_i, b_i) = \frac{a_i - 1}{a_i + b_i - 2} \quad \text{et} \quad D_i = \frac{1}{a_i + b_i - 2} = D_i(a_i, b_i). \tag{4}$$

Concernant la notion de dispersion, on peut se rapporter à Jørgensen (1997), Jørgensen et Kokonendji (2011, 2013). La densité g_θ de bêta bivarié avec corrélation est donnée par

$$g_\theta(v) = g_1(v_1)g_2(v_2)\left[1 + \left\{\frac{v_1 - \mu_1(a_1, b_1)}{\sigma_1(a_1, b_1)}\right\}\left\{\frac{v_2 - \mu_2(a_2, b_2)}{\sigma_2(a_2, b_2)}\right\}\rho\right]\mathbb{1}_{[0,1]^2}(v), \tag{5}$$

avec $v = (v_1, v_2)^T$ et $\theta := \theta(a_1, b_1, a_2, b_2, \rho) \in \Theta \subseteq \mathbb{R}^5$. Le paramètre de corrélation ρ est celui introduit par Sarmanov (1966) et appartient à l'intervalle

$$\left[-\left(\max\left\{\frac{v_1 - \mu_1(a_1, b_1)}{\sigma_1(a_1, b_1)}\right\}\left\{\frac{v_2 - \mu_2(a_2, b_2)}{\sigma_2(a_2, b_2)}\right\}\right)^{-1}, \left|\left(\min\left\{\frac{v_1 - \mu_1(a_1, b_1)}{\sigma_1(a_1, b_1)}\right\}\left\{\frac{v_2 - \mu_2(a_2, b_2)}{\sigma_2(a_2, b_2)}\right\}\right)^{-1}\right|\right];$$

voir aussi Lee (1996). Le vecteur moyen et la matrice de covariance de g_θ sont respectivement

$$\mu = (\mu_1, \mu_2)^T \quad \text{et} \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}.$$

On a l'unimodalité de la bivariée g_θ de (5) pour $a_i > 1$ et $b_i > 1$, qui sont les mêmes conditions qu'en univariées g_i de (2). Cependant, le mode $M(a_1, a_2, b_1, b_2, \rho) := M_\rho$ de g_θ n'admet pas de forme explicite ; néanmoins, nous avons vérifié numériquement qu'il est légèrement décalé par rapport au vecteur mode en (4) des marges univariées $M_0 = (M_1(a_1, b_1), M_2(a_2, b_2))^T$. Pour la matrice de dispersion $\mathbf{D}_\rho$ de (5) et avec les D_i données en (4), on considère

$$\mathbf{D}_\rho = \begin{pmatrix} D_1 & (D_1D_2)^{1/2}\,\rho \\ (D_1D_2)^{1/2}\,\rho & D_2 \end{pmatrix}.$$

Ensuite, on construit le noyau bêta bivarié avec corrélation ou dit de Sarmanov BS_θ par une variante de la méthode mode-dispersion, établie dans le cas unidimensionnel par Kokonendji et Libengué (2011, 2013) et qui généralise les cas particuliers de Chen (1999, 2000). Au lieu de M_ρ, on considère M_0 avec $\rho = 0$ pour la construction du noyau associé issue de (5) ; le contrecoup de ce choix est compensé dans la réduction

du biais de l'estimateur ci-dessous. Ainsi, il convient de placer la cible x sur M_0 et la matrice des fenêtres $\mathbf{H}_2$ sur $\mathbf{D}_\rho$. En résolvant alors le système d'équations à cinq inconnues $(M_0, \mathbf{D}_\rho) = (x, \mathbf{H}_2)$, on obtient la reparamétrisation $\theta = \theta(a_1, b_1, a_2, b_2, \rho)$ de g_θ en

$$\theta(x, \mathbf{H}_2) = \left(\frac{x_1}{h_{11}} + 1, \frac{1 - x_1}{h_{11}} + 1, \frac{x_2}{h_{22}} + 1, \frac{1 - x_2}{h_{22}} + 1, \frac{h_{12}}{(h_{11}h_{22})^{1/2}} \right), \quad \forall x, \mathbf{H}_2. \tag{6}$$

Les expressions $\mu_i(a_i, b_i)$ et $\sigma_i^2(a_i, b_i)$ de (3) deviennent

$$\widetilde{\mu}_i = \frac{x_i + h_{ii}}{1 + 2h_{ii}} = \widetilde{\mu}_i(x_i, h_{ii}) \quad \text{et} \quad \widetilde{\sigma}_i^2 = \frac{(x_i + h_{ii})(1 + h_{ii} - x_i)}{(1 + 2h_{ii})^2(1 + 3h_{ii})} h_{ii} = \widetilde{\sigma}_i^2(x_i, h_{ii}).$$

Enfin, le noyau bêta bivarié de Sarmanov est donné par $BS_{\theta(x,\mathbf{H}_2)}(\cdot) = g_{\theta(x,\mathbf{H}_2)}(\cdot)$ avec $\theta(x, \mathbf{H}_2)$ défini en (6). De (1) et des propriétés de g_θ en (5), on déduit les caractéristiques du noyau associé $BS_{\theta(x,\mathbf{H}_2)}$ de la manière suivante :

$$\mathbb{S}_{\theta(x,\mathbf{H}_2)} = [0,1]^2, \quad A_\theta(x, \mathbf{H}_2) = (A_1, A_2)^T \text{ avec } A_i = \frac{(1 - 2x_i)h_{ii}}{(1 + 2h_{ii})} = A_i(x_i, h_{ii}),$$

$$\mathbf{B}_\theta(x, \mathbf{H}_2) = (B_{ij})_{i,j=1,2} \text{ avec } B_{ii} = \widetilde{\sigma}_i^2(x_i, h_{ii}) \text{ et } B_{12} = \frac{h_{12}}{(h_{11}h_{22})^{1/2}} \widetilde{\sigma}_1(x_1, h_{11}) \widetilde{\sigma}_2(x_2, h_{22}). \tag{7}$$

3 Estimateur à noyau bêta bivarié de Sarmanov

Supposons $X_1, \ldots, X_n$ une suite de vecteurs aléatoires i.i.d, de densité f inconnue sur $[0,1]^2$. On rappelle que les composantes X_{ij} des vecteurs $X_i = (X_{i1}, X_{i2})^T$ peuvent être indépendantes ou non. L'estimateur à noyau bêta bivarié de Sarmanov $\widehat{f_n}$ de f est défini par :

$$\widehat{f_n}(x) = \frac{1}{n} \sum_{i=1}^{n} BS_{\theta(x,\mathbf{H}_2)}(X_i), \quad \forall x \in [0,1]^2. \tag{8}$$

Une première propriété de (8) est

$$\mathbb{E}\left\{\widehat{f_n}(x)\right\} = \mathbb{E}\left\{f\left(\mathcal{Z}_{\theta(x,\mathbf{H}_2)}\right)\right\}, \tag{9}$$

où $\mathcal{Z}_{\theta(x,\mathbf{H}_2)}$ est un vecteur aléatoire de densité $BS_{\theta(x,\mathbf{H}_2)} = g_{\theta(x,\mathbf{H}_2)}$. Supposons que f soit de classe $\mathscr{C}^2([0,1]^2)$. De (7), (9) et des developpements de Taylor de f successivement autour de $\mathbb{E}\left(\mathcal{Z}_{\theta(x,\mathbf{H}_2)}\right)$ et de x, le biais ponctuel de l'estimateur $\widehat{f_n}$ de f est donné par

$$\begin{aligned} \text{Biais}\{\widehat{f_n}(x)\} &= \mathbb{E}\left\{f\left(\mathcal{Z}_{\theta(x,\mathbf{H}_2)}\right)\right\} - f(x) \\ &= A_1 f_1(x) + A_2 f_2(x) + \frac{1}{2}\left\{(A_1^2 + B_{11}) f_{11}(x) + 2(A_1 A_2 + B_{12}) f_{12}(x) \right. \\ &\quad \left. + (A_2^2 + B_{22}) f_{22}(x)\right\} + o(h_{11}^2 + 2h_{12}^2 + h_{22}^2), \end{aligned} \tag{10}$$

où $f_i = \partial f/\partial x_i$, $f_{ii} = \partial^2 f/\partial x_i^2$ et $f_{12} = \partial^2 f/\partial x_1 \partial x_2$. L'annulation du terme du gradient de f dans (10) permet de réduire le biais induit ; voir, p. ex., Chen (1999, 2000), Kokonendji et Libengué (2011, 2013), Bouezmarni et Rombouts (2010). Si de plus f est bornée sur $[0,1]^2$, alors la variance ponctuelle devient

$$\mathrm{Var}\{\widehat{f_n}(x)\} = \frac{1}{n}\|BS_{\theta(x,\mathbf{H}_2)}\|_2^2 f(x) + \mathrm{o}\left(\frac{1}{n(h_{11}h_{22}-h_{12}^2)^{r_2}}\right), \tag{11}$$

où $r_2 = r_2\left(BS_{\theta(x,\mathbf{H}_2)}\right) > 0$ est le plus grand réel tel que $\|BS_{\theta(x,\mathbf{H}_2)}\|_2^2 \leq c_2(x)(h_{11}h_{22}-h_{12}^2)_n^{-r_2}$ et $0 \leq c_2(x) \leq +\infty$. Pour déterminer la valeur de r_2, nous utiliserons une variante bivarié de la fonction R de Brown et Chen (1998).

On mesure la similarité entre l'estimateur $\widehat{f_n}$ et la vraie densité f par *l'erreur moyenne quadratique intégrée asymptotique* (en anglais «Asymptotic Mean Integrated Squared Error (AMISE)»). Pour l'estimateur $\widehat{f_n}$ donné en (8), on a

$$\begin{aligned}\mathrm{AMISE}(\widehat{f_n}) \;=\; \int_{[0,1]^2} \Big(\Big[& A_1 f_1(x) + A_2 f_2(x) + \frac{1}{2}\left\{(A_1^2 + B_{11}) f_{11}(x) + 2(A_1 A_2 + B_{12}) f_{12}(x) \right. \\ & \left. + (A_2^2 + B_{22}) f_{22}(x)\right\}\Big]^2 + \frac{1}{n}\|BS_{\theta(x,\mathbf{H}_2)}\|_2^2 f(x)\Big)\, dx. \end{aligned} \tag{12}$$

Le choix de la matrice des fenêtres optimale se fait par validation croisée. Dans le cas multivarié, le critère de validation croisée par les moindres carrées est une simple extension du cas univarié ; voir Chacón et Duong (2011), Zougab et al. (2012, 2013) pour des alternatives. À partir de (6), (10), (11) et (12), et en prenant $h_{12} = 0$, on obtient les formules équivalentes du cas bivarié produit, ou sans structure de corrélation, de Bouezmarni et Rombouts (2010). Nous montrerons numériquement l'effet de la structure corrélation dans cette étude.

4 Conclusion

Dans cette communication, nous avons considéré le noyau bêta bivarié avec corrélation de Sarmanov (1966). Cette structure de corrélation peut être remplacée par d'autres et utilisée pour d'autres familles de noyaux associés continus ou discrets. La structure de corrélation introduite permet d'atteindre certains endroits de l'espace étudié que la situation de non corrélation ne l'autorise pas. Notons que ces noyaux associés avec structure de corrélation sont mieux adaptés aux données bivariées $X_i = (X_{i1}, X_{i2})^T$ ayant une corrélation entre les différentes composantes X_{i1} et X_{i2}. Cependant, le choix des fenêtres par validation croisée utilisée dans ce travail mériterait d'être améliorée ou remplacée, à cause de la faible vitesse de convergence de l'algorithme. Nous signalons qu'en dimension deux, pour construire un noyau associé non-classique avec corrélation, nous avons besoin d'une loi bivariée à cinq paramètres. Plus généralement, pour

un support $\mathbb{T}_d \left(\subseteq \mathbb{R}^d\right)$ de f à estimer par noyaux associés non-classiques, on aura besoin d'au moins $d(d+3)/2$ paramètres pour une construction appropriée, p. ex. en utilisant la méthode mode-dispersion.

Bibliographie

[1] Bouezmarni, T. et Rombouts, J. V. K. (2010), Nonparametric density estimation for multivariate bounded data, *Journal of Statistical Planning and Inference* **140**, 139–152.
[2] Brown, B. M. et Chen, S. X. (1998), Beta Bernstein smoothing for regression curves with compact support, *Scandinavian Journal of Statistics* **26**, 47–59.
[3] Chacón, J. E. et Duong, T. (2011), Unconstrained pilot selectors for smoothed cross validation, *Australian & New Zealand Journal of Statistics* **53**, 331–351.
[4] Chen, S. X. (1999), A beta kernel estimation for density functions, *Computationnal Statistics and Data Analysis* **31**, 131–145.
[5] Chen, S. X. (2000), Probability density function estimation using gamma kernels, *Annals of the Institute of Statistical Mathematics* **52**, 471–480.
[6] Jørgensen, B. (1997), *The Theory of dispersion models*, Chapman & Hall, London.
[7] Jørgensen, B. et Kokonendji, C. C. (2011), Dispersion models for geometric sums, *Brazilian Journal of Probability & Statistics* **25**, 263–293.
[8] Jørgensen, B. et Kokonendji, C. C. (2013), Discrete dispersion and their Tweedie asymptotics, en préparation.
[9] Kokonendji, C. C. et Libengué, F. G. (2011), Méthode des noyaux associés continus et estimation de densité, *43èmes Journées de Statistique de la SFDS*, Tunis, 6 pages.
[10] Kokonendji, C. C. et Libengué, F. G. (2013), Non-classical associated kernels for non-standard density estimators, Soumis pour publication, 32 pages.
[11] Kokonendji, C. C. et Senga Kiéssé, T. (2011), Discrete associated kernels method and extensions, *Statistical Methodology* **8**, 497–516.
[12] Lee, M-L. T. (1996), Properties and applications of the Sarmanov family of bivariate distributions, *Communications in Statistics-Theory and Methods* **25**, 1207–1222.
[13] Sarmanov, O. V. (1966), Generalized normal correlation and two-dimensionnal Frechet classes, *Doklady (Soviet Mathematics)* **168**, 596–599.
[14] Scott, W. D. (1992), *Multivariate Density Estimation*, John Wiley & Sons, New York.
[15] Zougab, N., Adjabi, S. et Kokonendji, C. C. (2012), Binomial kernel and Bayes local bandwith in discrete functions estimation, *Journal of Nonparametrics Statistics* **24**, 783–795.
[16] Zougab, N., Adjabi, S. et Kokonendji, C.C. (2013), A Bayesian approach to bandwidth selection in univariate associate kernel estimation, *Journal of Statistical Theory and Practice* **7**, 8–23.