Régression multiple non-paramétrique par noyaux associés

Sobom M. Somé[1], Tristan Senga Kiéssé[2] & Célestin C. Kokonendji[1]

[1]*Université de Franche-Comté*
Laboratoire de Mathématiques de Besançon - UMR 6623 CNRS-UFC
16 route de Gray, 25030 Besançon cedex, France.
(sobom.some ; celestin.kokonendji)@univ-fcomte.fr
[2]*Université Nantes Angers Le Mans*
Institut de Recherche en Génie Civil et Mécanique GeM UMR 6183 CNRS
Chaire Génie Civil Eco-construction, 44600 Saint-Nazaire.
tristan.sengakiesse@univ-nantes.fr

Résumé. L'objet de ce travail est de proposer une méthode non-paramétrique d'estimation d'une fonction de régression multiple, bien indiqué pour à la fois des variables explicatives continues et des variables explicatives de dénombrement. Le modèle est d'abord présenté, ensuite une définition du noyau associé multivarié le plus général est introduite avec trois cas particuliers. L'estimateur de Nadaraya-Watson utilisant ces noyaux associés est alors présenté à travers une étude par simulation ainsi qu'une application aux données réelles, avec à chaque fois une sélection de la matrice des fenêtres par validation croisée.

Mots-clés. Estimateur de Nadaraya-Watson, matrice des fenêtres, validation croisée.

Abstract. The purpose of this communication is to propose a nonparametric estimator for multiple regression function, appropriated for both continuous and count explanatory variables. The model is first presented, then a definition of the most general multivariate associated kernel is introduced with three particular cases. The Nadaraya-Watson estimator using associated kernels is then provided through simulation studies and real data analysis, in each case with a selection of the bandwidth matrix by cross validation.

Keywords. Bandwidth matrix, cross-validation, Nadaraya-Watson estimator.

1 Introduction

Nous nous intéressons à la régression multiple non-paramétrique sur des données multivariés composée de variables continues (bornées ou non) et de variables de

dénombrement. Par exemple, la relation entre la variable aléatoire réelle à expliquer Y et des prédicteurs $x_1, \ldots, x_d$ est donnée par

$$Y = m(x_1, \ldots, x_d) + \epsilon, \tag{1}$$

où m est une fonction de régression inconnue de $\mathbb{T}_d \subseteq \mathbb{R}^d$ dans $\mathbb{R}$ et ϵ le terme d'erreur de moyenne nulle et de variance finie. Les noyaux multivariés classiques (e.g. gaussiens) ne sont adaptés que pour l'estimation de fonction de régression de données non bornées (i.e. $\mathbb{R}^d$) ; voir Scott (1992). Racine et Li (2004) ont proposé des noyaux multiples composés de noyaux univariés gaussiens pour les variables continues et de noyaux d'Aitchison et Aitken (1976) pour les variables catégorielles ; voir aussi Hayfield et Racine (2007) pour des implémentations et utilisations de ces noyaux multiples sous le logiciel R (2013). Signalons que l'utilisation des noyaux gaussiens (i.e. symétriques) produisent des poids en dehors des variables à support non bornés. Dans le cas univarié continu, Chen (1999, 2000) est l'un des premiers à avoir proposé des noyaux asymétriques (e.g. bêta, gamma) dont les supports coïncident avec celles des densités à estimer ; voir aussi Bertin et Klutchnikoff (2011). Aussi, Libengué (2013) a étudié plusieurs familles de ces noyaux univariés qu'il a appelé noyaux associes univariés ; voir aussi Kokonendji et al. (2007, 2009), Kokonendji et Senga Kiéssé (2011), Zougab et al. (2012, 2013,2014ab), Wansouwé et al. (2014) pour les cas univariés discrets. La version multivarié continu des noyaux associés a été étudié par Kokonendji et Somé (2014).

Nous proposons, pour l'estimation de la fonction m de régression (1), des noyaux associés multiples composés de noyaux associés discrets univariés (e.g. binomial, triangulaire discret) et continues (e.g. bêta, gamma). Ces noyaux associés sont adaptés à cette situation de mélange d'axes car elles respectent scrupuleusement le support de chaque variable explicative. C'est pourquoi, dans ce qui suit nous proposons d'abord une définition générale des noyaux associés multivariés (discrets ou continus) et présentons trois cas particuliers : associé classique, associé multiple et associé avec structure de corrélation de Sarmanov (1966). Ensuite, à partir de l'estimateur de Nadaraya-Watson multivarié nous définissons l'estimateur à noyaux associés multiples pour les fonctions de régression continues et de dénombrement. Des simulations et applications étudieront l'efficacité de cette méthode.

2 Noyaux associés multivariés

Considérons $\mathbf{X}_1, \cdots, \mathbf{X}_n$ une suite de vecteurs aléatoires indépendants et identiquement distribués (i.i.d.) de densité (discrète et/ou continue) inconnue f sur $\mathbb{T}_d$, un sous ensemble de $\mathbb{R}^d$ ($d \geq 1$). On dira dans la suite du travail que f est une densité par rapport à une mesure $\boldsymbol{\nu} = \nu_1 \otimes \ldots \otimes \nu_d$, où ν_j est une mesure de Lebesgue ou de comptage sur le support univarié correspondant $\mathbb{T}_1^{[j]}$ et donc $\mathbb{T}_d = \otimes_{j=1}^{d} \mathbb{T}_1^{[j]}$, pour $j = 1, \ldots, d$. Un

estimateur à noyaux associés multivarié $\widehat{f_n}$ de f est alors simplement défini par

$$\widehat{f_n}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{x},\mathbf{H}}(\mathbf{X}_i), \quad \forall \mathbf{x} \in \mathbb{T}_d \subseteq \mathbb{R}^d, \tag{2}$$

où $\mathbf{H}$ est une matrice des fenêtres d'ordre $d \times d$ (i.e. symétrique et définie positive) telle que $\mathbf{H} \equiv \mathbf{H}_n \to \mathbf{0}$ quand $n \to +\infty$, et $K_{\mathbf{x},\mathbf{H}}(\cdot)$ est appelé le noyau associé, paramétré par $\mathbf{x}$ et $\mathbf{H}$. Ce noyau $K_{\mathbf{x},\mathbf{H}}(\cdot)$ est une fonction de densité de probabilité (f.d.p.) par rapport à la mesure $\boldsymbol{\nu}$, et est défini précisément comme suit.

Définition 2.1 *Soit $\mathbb{T}_d$ le support de la densité à estimer avec $\mathbb{T}_d \subseteq \mathbb{R}^d$ et $\mathbf{H}$ une matrice des fenêtres. Pour un vecteur cible $\mathbf{x} \in \mathbb{T}_d$, on considère un vecteur aléatoire $\mathcal{Z}_{\mathbf{x},\mathbf{H}}$ de f.d.p. paramétrée $K_{\mathbf{x},\mathbf{H}}(\cdot)$ et de support $\mathbb{S}_{\mathbf{x},\mathbf{H}} \left(\subseteq \mathbb{R}^d\right)$. La fonction $K_{\mathbf{x},\mathbf{H}}(\cdot)$ est appelée "noyau associé multivarié (ou général)" si les conditions suivantes sont satisfaites :*

$$\mathbf{x} \in \mathbb{S}_{\mathbf{x},\mathbf{H}}, \;\; \mathbb{E}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{x} + \mathbf{a}(\mathbf{x},\mathbf{H}) \;\; \textit{et} \;\; \mathrm{Cov}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{B}(\mathbf{x},\mathbf{H}),$$

où $\mathbf{a}(\mathbf{x},\mathbf{H}) = (a_1(\mathbf{x},\mathbf{H}),\ldots,a_d(\mathbf{x},\mathbf{H}))^\top$ et $\mathbf{B}(\mathbf{x},\mathbf{H}) = \left(b_{ij}(\mathbf{x},\mathbf{H})\right)_{i,j=1,\ldots,d}$ tendent respectivement vers le vecteur nul et la matrice nulle quand $\mathbf{H} \to \mathbf{0}$.

De la Définition 2.1 on déduit trois cas particuliers intéressants de noyaux associés multivariés. Le premier, approprié pour des variables continues de supports non bornés (en particulier $\mathbb{R}^d$), est généralement défini par l'estimateur

$$\widehat{f_n}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i), \;\; \forall \mathbf{x} \in \mathbb{R}^d =: \mathbb{T}_d, \tag{3}$$

avec $K_{\mathbf{H}}(\mathbf{y}) = (1/\det\mathbf{H})K\left(\mathbf{H}^{-1}\mathbf{y}\right)$ ou souvent $K_{\mathbf{H}}(\mathbf{y}) = (1/\det\mathbf{H})^{1/2}K\left(\mathbf{H}^{-1/2}\mathbf{y}\right)$ pour tout $\mathbf{y} \in \mathbb{R}^d$; voir aussi Zougab et al. (2014b). La proposition suivante transforme la fonction noyau continu symétrique K en noyau associé classique.

Proposition 2.2 *Soit $\mathbb{R}^d =: \mathbb{T}_d$ le support de la densité à estimer. Soit K un noyau classique (symétrique) de support $\mathbb{S}_d \subseteq \mathbb{R}^d$, de moyenne $\boldsymbol{\mu} = \mathbf{0}$ et de matrice de variance-covariance $\boldsymbol{\Sigma}$. Pour un vecteur cible $\mathbf{x} \in \mathbb{R}^d$ et une matrice des fenêtres $\mathbf{H}$, le noyau K est tranformé en noyau associé classique :*

$$K_{\mathbf{x},\mathbf{H}}(\cdot) = \frac{1}{\det\mathbf{H}} K\left\{\mathbf{H}^{-1}(\mathbf{x} - \cdot)\right\}$$

sur $\mathbb{S}_{\mathbf{x},\mathbf{H}} = \mathbf{x} - \mathbf{H}\mathbb{S}_d$ avec $\mathbb{E}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{x}$ (i.e. $\mathbf{a}(\mathbf{x},\mathbf{H}) = \mathbf{0}$) et $\mathrm{Cov}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}$, et où $\mathcal{Z}_{\mathbf{x},\mathbf{H}}$ est le vecteur aléatoire de f.d.p. $K_{\mathbf{x},\mathbf{H}}$.

Le second cas particulier de la Définition 2.1, approprié pour des vecteurs aléatoires composées à la fois de variables aléatoires continues (bornées ou non) et de variables aléatoires de dénombrement, est présenté à travers l'estimateur

$$\widehat{f_n}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} K^{[j]}_{x_j,h_{jj}}(X_{ij}), \quad \forall x_j \in \mathbb{T}_1^{[j]} \subseteq \mathbb{R}, \tag{4}$$

où $\mathbb{T}_1^{[j]}$ est le support des marges univariées de f pour $j = 1,\ldots,d$, $\mathbf{x} = (x_1,\ldots,x_d)^\top \in \times_{j=1}^{d}\mathbb{T}_1^{[j]}$, $\mathbf{X}_i = (X_{i1},\ldots,X_{id})^\top$, pour $i = 1,\ldots,n$ et $h_{11},\ldots,h_{dd}$ sont les fenêtres univariés. La fonction $K^{[j]}_{x_j,h_{jj}}$ est le $j$ème noyau associé univarié (discret ou continu) pour les variables X_{ij}, $i = 1,\ldots,n$, de support $\mathbb{S}_{x_j,h_{jj}} \subseteq \mathbb{R}$. La version continu de cet estimateur (4) a été étudié par Bouerzmarni et Rombouts (2010). Le noyau associé multiple dans (4) est un noyau associé à travers la proposition suivante.

Proposition 2.3 *Soit $\times_{j=1}^{d}\mathbb{T}_1^{[j]} = \mathbb{T}_d$ le support de la densité à estimer f avec $\mathbb{T}_1^{[j]}$ ($\subseteq \mathbb{R}$) les supports des marges univariés (continues ou discrètes) de f. Soit $\mathbf{x} = (x_1,\ldots,x_d)^\top \in \times_{j=1}^{d}\mathbb{T}_1^{[j]}$ et $\mathbf{H} = \mathbf{Diag}(h_{11},\ldots,h_{dd})$ avec $h_{jj} > 0$. Soit $K^{[j]}_{x_j,h_{jj}}$ le noyau associé univarié continu ou discret (voir Définition 2.1 pour $d = 1$) correspondant à la variable aléatoire $\mathcal{Z}^{[j]}_{x_j,h_{jj}}$ de support $\mathbb{S}_{x_j,h_{jj}}(\subseteq \mathbb{R})$, pour tout $j = 1,\ldots,d$. Alors, le noyau associé multiple est aussi un noyau associé :*

$$K_{\mathbf{x},\mathbf{H}}(\cdot) = \prod_{j=1}^{d} K^{[j]}_{x_j,h_{jj}}(\cdot) \tag{5}$$

sur $\mathbb{S}_{\mathbf{x},\mathbf{H}} = \times_{j=1}^{d}\mathbb{S}_{x_j,h_{jj}}$ avec $\mathbb{E}(\mathcal{Z}_{\mathbf{x},\mathbf{H}}) = (x_1 + a_1(x_1,h_{11}),\ldots,x_d + a_d(x_d,h_{dd}))^\top$ et $\mathrm{Cov}(\mathcal{Z}_{\mathbf{x},\mathbf{H}}) = \mathbf{Diag}\left(b_{jj}(x_j,h_{jj})\right)_{j=1,\ldots,d}$. En d'autres termes, les variables aléatoires réelles $\mathcal{Z}^{[j]}_{x_j,h_{jj}}$ sont les composantes indépendantes du vecteur aléatoire $\mathcal{Z}_{\mathbf{x},\mathbf{H}}$.

Le troisième cas particulier de noyaux associés multivariés est construit à partir d'une f.d.p. constituée de produit de f.d.p. univariés et d'une structure de corrélation utilisant la technique de Sarmanov (1966). De tels noyaux associés permettent d'atteindre certains endroits du lissage multidimensionnel. Kokonendji et Somé (2014) ont illustré l'effet de cette technique pour le noyau bêta bivarié. Tout comme Bertin et Klutchnikoff (2011), les propriétés miminax de ce noyau bêta bivarié sont aussi envisageables et plus généralement pour les noyaux associés.

3 Régression multiple par noyaux associés

Considérons une séquence $(\mathbf{X}_1, Y_1),\ldots,(\mathbf{X}_n, Y_n)$ i.i.d. de vecteurs aléatoires sur $\mathbb{T}_d \times \mathbb{R}(\subseteq \mathbb{R}^{d+1})$ de fonction de régression inconnue $m(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ en (1). L'estimateur

de Nadaraya-Watson (Nadaraya, 1964 ; Watson, 1964) $\widehat{m}_n$ de m, en utilisant les noyaux associés multivariés (voir Définition 2.1), est donnée par

$$\widehat{m}_n(\mathbf{x}) = \sum_{i=1}^{n} \frac{Y_i K_{\mathbf{x},\mathbf{H}}(\mathbf{X}_i)}{\sum_{i=1}^{n} K_{\mathbf{x},\mathbf{H}}(\mathbf{X}_i)}, \quad \forall \mathbf{x} \in \mathbb{T}_d \subseteq \mathbb{R}^d, \tag{6}$$

où $\mathbf{H} \equiv \mathbf{H}_n$ est la matrice des fenêtres telle que $\mathbf{H}_n \to \mathbf{0}$ quand $n \to +\infty$. Le noyau associé $K_{\mathbf{x},\mathbf{H}}$ est choisi de manière approprié pour tenir compte de la structure (discrète et continue) des données. Contrairement à l'estimateur à densité $\widehat{f_n}$, l'estimateur $\widehat{m}_n$ de fonction de régression (6) n'a pas besoin d'être normalisé. Pour des supports multivariés composés de supports univariés discrets et continus, on donne l'estimateur $\widehat{m}_n$ en fonction des noyaux associés multiples (5) comme suit :

$$\widehat{m}_n(\mathbf{x}) = \sum_{i=1}^{n} \frac{Y_i \prod_{j=1}^{d} K^{[j]}_{x_j,h_{jj}}(X_{ij})}{\sum_{i=1}^{n} \prod_{j=1}^{d} K^{[j]}_{x_j,h_{jj}}(X_{ij})}, \quad \forall \mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{T}_d := \times_{j=1}^{d} \mathbb{T}_1^{[j]}. \tag{7}$$

Certaines propriétés asymptotiques de ces estimateurs (6) et (7) seront présentées en utilisant les dérivées et les différences finies. À travers l'erreur moyenne quadratique et le coefficient de détermination, des études numériques (par simulation et sur données réelles) dans le cas bivarié montrent l'efficacité de cette méthode. Le choix de la matrice des fenêtres optimale se fera à chaque fois par validation croisée avant une amélioration par des approches bayésiennes. Ce choix implique naturellement des calculs fastidieux ; notamment, dans le cas général (6) avec structure de corrélation où l'on a à sélectionner une matrice symétrique à $d(d+1)/2$ paramètres.

Bibliographie

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

[2] Bertin, K. et Klutchnikoff, N. (2011), Minimax properties of beta kernel estimators, *Journal of Statistical Planning and Inference*, **141**, 2287–2297.

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

[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] Hayfield, T. et Racine, J. S. (2007), Nonparametric Econometrics : The np Package, *Journal of Statistical Software*, **27**, 1–32.

[7] Kokonendji, C. C. et Senga Kiéssé, T. (2011), Discrete associated kernels method and extensions, *Statistical Methodology*, **8**, 497–516.
[8] Kokonendji, C. C., Senga Kiéssé, T. et Demétrio, C. G. B. (2009), Appropriate kernel regression on a count explanatory variable and applications, *Advances and Applications in Statistics*, **12**, 99–125.
[9] Kokonendji, C. C., Senga Kiéssé, T. et Zocchi, S. S. (2007), Discrete triangular distributions and non-parametric estimation for probability mass function, *Journal of Nonparametric Statistics*, **19**, 241–254.
[10] Kokonendji, C. C. et Somé, S. M. (2014), On multivariate associated kernels for smoothing some density functions, *Preprint du Laboratoire de Mathématiques de Besançon no. 2014/02*, soumis pour publication.
[11] Libengué, F. G. (2013), *Méthode Non-Paramétrique par Noyaux Associés Mixtes et Applications*. Unpublished Ph.D. Thesis ; Université Franche-Comté, Besançon, France & Université de Ouagadougou, Burkina Faso, Juin 2013.
[12] Nadaraya, E. A. (1964), On estimating regression, *Theory of Probability and its Applications*, **9**, 141–142.
[13] R Development Core Team (2013), R : A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, ISBN 3-900051-07-0, `http://cran.r-project.org/`.
[14] Racine, J. et Li, Q. (2004), Nonparametric estimation of regression functions with both categorical and continuous data, *Journal of Econometrics*, **119**, 99–130.
[15] Sarmanov, O. V. (1966), Generalized normal correlation and two-dimensionnal Frechet classes, *Doklady (Soviet Mathematics)*, **168**, 596–599.
[16] Scott, W. D. (1992), *Multivariate Density Estimation*, John Wiley & Sons, New York.
[17] Wansouwé, W. E., Kokonendji, C. C. et Kolyang, D. T. (2014), Disake : an R package for discrete associated kernel estimator,
`http://cran.r-project.org/web/packages/Disake`.
[18] Watson, G. S. (1964), Smooth regression analysis, *Sankhya Series A*, **26**, 359–372.
[19] 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.
[20] 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.
[21] Zougab, N., Adjabi, S. et Kokonendji, C. C. (2014a), Bayesian approach in nonparametric count regression with binomial kernel, *Communications in Statistics - Simulation and Computation*, **43**, 1052–1063.
[22] Zougab, N., Adjabi, S. et Kokonendji, C. C. (2014b), Bayesian estimation of adaptive bandwidth matrices in multivariate kernel density estimation, *Computational Statistics and Data Analysis*, **75**, 28–38.