Une nouvelle approche non paramétrique pour estimer la fonction de régression spatiale

Camille Ternynck[1] & Sophie Dabo-Niang[1,3] & Anne-Françoise Yao[2]

[1] *Université Lille III - Laboratoire EQUIPPE. Domaine universitaire du Pont de Bois - Maison de la recherche - BP 60149 - 59653 Villeneuve-d'Ascq Cedex - camille.ternynck@univ-lille3.fr / sophie.dabo@univ-lille3.fr*

[2] *Université Blaise Pascal - Laboratoire de Mathématiques UMR 6620 - CNRS Campus des Cézeaux - BP 80026 - 63171 Aubière Cedex - Anne-francoise.Yao@math.univ-bpclermont.fr*

[3] *MODAL team, INRIA Lille-Nord de France*

Résumé. Dans ce travail, nous nous intéressons à l'estimation non paramétrique de la fonction de régression lorsque la variable réponse est scalaire et la variable explicative est multivariée, ceci dans le cadre des données spatialement dépendantes. La particularité de l'estimateur proposé est de combiner deux noyaux permettant de contrôler à la fois la distance entre les observations et celle entre les sites. La convergence presque complète ainsi que la convergence en moyenne d'ordre q (norme L^q) ($q \in \mathbb{N}^*$) de l'estimateur à noyaux sont obtenues en considérant des processus α-mélangeants. Nous présenterons également un prédicteur spatial issu de l'estimation de la régression. Notre exposé sera illustré par des simulations et une application à des données environnementales.

Mots-clés. Estimateur à noyau, Régression spatiale, Champs aléatoires, Variables mélangeantes

Abstract. In this work, we are interested in the nonparametric spatial estimation of the regression function when the response variable is real-valued and the explanatory variable is a vector. The special feature of the proposed estimator is to combine two kernels in order to control both the distance between observations and that between the spatial locations. Almost complete convergence and consistency in L^q norm ($q \in \mathbb{N}^*$) of the kernel estimate are obtained when the sample considered is an α-mixing sequence. We will also present a spatial predictor resulting from the regression estimation. Finally, numerical studies are carried out in order to illustrate the practical behavior of our methodology both for simulated data and for an environmental data set when the data are spatially dependent.

Keywords. Kernel regression estimation, Spatial regression, Random fields, Mixing conditions

1 Introduction

Le principal objectif de la statistique spatiale est de répondre à l'un des besoins de nombreuses disciplines scientifiques, à savoir la prévision spatiale. En effet, dans certaines applications, les données traitées sont enregistrées en des sites spécifiques et il apparaît important de déterminer quels autres sites ont une influence sur le site étudié.

Les premières méthodes de prévision spatiale apparaissent, au début du 20ème siècle, dans le domaine de la géostatistique, sous le nom de "Krigeage" et ont largement été étudiées dans la littérature. Cependant, pour pallier certains problèmes rencontrés par ces méthodes, de nouvelles techniques, issues de la statistique non paramétrique, se développent et sont au coeur de ce travail. En effet, on s'intéressera ici à l'estimation de la fonction de régression spatiale ainsi qu'à la prédiction spatiale par le biais de méthodes non paramétriques. Les premiers résultats dans cette direction sont ceux de Biau et Cadre (2004) et concernent la prédiction à noyau d'un champ aléatoire strictement stationnaire indexé dans $(\mathbb{N}^*)^N$. Par la suite, Dabo-Niang et Yao (2007) ont contribué à cette problématique en s'intéressant à l'estimation de la régression et à la prédiction de champs aléatoires continuement indexés. Dans Menezes et al. (2010), la prédiction non paramétrique à noyau est considérée pour des processus stochastiques spatiaux quand les sites d'observations sont supposés aléatoires. La principale différence entre ces estimateurs est que le dernier est composé d'un noyau sur les sites alors que les deux autres considèrent un noyau sur les valeurs observées du champ spatial.

L'objectif de ce travail est de proposer une nouvelle approche non paramétrique pour l'estimation de la fonction de régression qui sera ensuite utilisée à des fins de prévision spatiale. L'originalité de l'estimateur suggéré est de reposer sur chacun des avantages des estimateurs présentés précédemment. En effet, notre estimateur est basé sur deux noyaux, l'un contrôlant la distance entre les observations et l'autre contrôlant la structure de dépendance spatiale. Cette idée a été introduite dans Dabo-Niang et al. (2013) dans le contexte de l'estimation de la densité et dans Ternynck (2014) dans le cadre de la régression spatiale pour données fonctionnelles.

Nous présenterons notre estimateur et donnerons quelques résultats asymptotiques quand l'échantillon considéré est $\alpha-$mélangeant. Nous proposerons ensuite un prédicteur spatial basé sur cet estimateur. Des simulations et une application montreront le comportement pratique de notre méthode en présence de données spatialement dépendantes.

2 Estimation à noyaux de la fonction de régression spatiale

On considère un processus spatial $(Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}) \in \mathbb{R}^d \times \mathbb{R}, \mathbf{i} \in \mathbb{Z}^N)$, défini sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, tel que les variables $Z_{\mathbf{i}}$ sont de même distribution que la variable $Z = (X, Y)$ de densité inconnue $f_{X,Y}$ sur $\mathbb{R}^{d+1}$. La fonction de densité de X sur $\mathbb{R}^d$ est $f(\cdot)$.

Par souci de simplicité, on supposera que la variable Y est bornée. Nous nous intéressons au modèle de régression $r(x) = \mathbb{E}(Y|X = x)$, où $r(\cdot)$ est une fonction inconnue, à valeurs réelles, définie par $r(x) = \varphi(x)/f(x)$ où $\varphi(x) = \int y f_{XY}(x, y)dy,\ x \in \mathbb{R}^d$. Le processus est observé sur le domaine rectangulaire $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, \ldots, i_N),\ 1 \leq i_k \leq n_k,\ k = 1, \ldots, N\}$. On note $\mathbf{n} = (n_1, \ldots, n_N)$ et on définit $\widehat{\mathbf{n}} := n_1 \times \ldots \times n_N$ la taille de l'échantillon. Dans la suite, on écrira $\mathbf{n} \to \infty$ si $\min_{k=1,\ldots,N} n_k \to \infty$. Sans perte de généralité, on supposera que $n_1 = n_2 = \ldots = n_N = n$.

En considérant les sites normalisés, l'estimateur à noyaux de la fonction de régression $r(\cdot)$, pour chaque $x \in \mathbb{R}^d$ fixé, localisé en un site $\mathbf{j}$, est défini par

$$r_{\mathbf{n}}(x_{\mathbf{j}}) = \begin{cases} \dfrac{\varphi_{\mathbf{n}}(x_{\mathbf{j}})}{f_{\mathbf{n}}(x_{\mathbf{j}})} & \text{si} \quad f_{\mathbf{n}}(x_{\mathbf{j}}) \neq 0; \\ \dfrac{1}{\widehat{\mathbf{n}}} \displaystyle\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} & \text{sinon,} \end{cases}$$

où

$$\begin{aligned} \varphi_{\mathbf{n}}(x_{\mathbf{j}}) &= \frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|), \\ f_{\mathbf{n}}(x_{\mathbf{j}}) &= \frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|), \end{aligned}$$

avec $a_{\mathbf{n},\mathbf{j}} = \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|)$ et $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|) = K_2\left(\rho_{\mathbf{n}}^{-1} \left\|\frac{\mathbf{j}-\mathbf{i}}{\mathbf{n}}\right\|\right)$ (où $\frac{\mathbf{i}}{\mathbf{n}} = (\frac{i_1}{n}, \frac{i_2}{n}, \ldots, \frac{i_N}{n})$). De plus, K_1 et K_2 sont des noyaux définis respectivement sur $\mathbb{R}^d$ et $\mathbb{R}$ à support $[0, 1]$, $b_{\mathbf{n}}$ et $\rho_{\mathbf{n}}$ sont deux fenêtres qui tendent vers 0 quand $\mathbf{n} \to \infty$. Pour chaque site $\mathbf{j}$, l'estimateur $f_{\mathbf{n}}(x_{\mathbf{j}})$ est une fonction du nombre $k_{\mathbf{n}} = k_{\mathbf{n},\mathbf{j}}$ de voisins de $\mathbf{j}$ situés à une distance au plus égale à $d_{\mathbf{n}} > 0$ telle que $d_{\mathbf{n}} \to \infty$ quand $\mathbf{n} \to \infty$.

Conditions et résultats de convergence

Nous supposons que le processus $(Z_{\mathbf{i}})$ satisfait une certaine condition de α-mélange. Nous considérons également les hypothèses suivantes :

A1 : *Les fonctions de densité $f_{X,Y}$ et f sont continues sur $\mathbb{R}^{d+1}$ et $\mathbb{R}^d$ respectivement.*

A2 : *Les fonctions de densité et de régression sont lipschitziennes.*

A3 : *Les fonctions $K_1(\cdot)$ et $K_2(\cdot)$ sont des densités lipschitziennes à supports compacts sur $\mathbb{R}^d$ et $\mathbb{R}$ respectivement et $\int \|t\| K_1(t)dt < \infty$.*

A4 : *Il existe des constantes C_1 et C_2 avec $0 < C_1 < C_2 < \infty$ telles que*

$$C_1 \mathbf{1}_{[0,1]}(\|t\|) \quad \leq \quad K_i(t) \leq C_2 \mathbf{1}_{[0,1]}(\|t\|) \qquad\qquad i = 1, 2.$$

A5 : *La densité jointe* $f_{X_\mathbf{i}X_\mathbf{j}}$ *de* $(X_\mathbf{i}, X_\mathbf{j})$ *existe, est bornée et*

$$\forall u, v \in \mathbb{R}^d, \exists C > 0, \qquad |f_{X_\mathbf{i}X_\mathbf{j}}(u,v) - f_{X_\mathbf{i}}(u) f_{X_\mathbf{j}}(v)| < C.$$

A6/A7 : Nous faisons également des hypothèses sur les coefficients de mélange.

Sous les conditions énoncées précédemment, nous obtenons les résultats de convergence suivants :

Théorème 1 $r_\mathbf{n}(\cdot)$ *converge presque complètement vers* $r(\cdot)$ *et*

$$|r_\mathbf{n}(x_\mathbf{j}) - r(x_\mathbf{j})| = O\left(b_\mathbf{n} + \sqrt{\frac{\log \widehat{\boldsymbol{n}}}{\widehat{\boldsymbol{n}}\, b_{\boldsymbol{n}}^d\, \rho_\mathbf{n}^N}}\right) \quad p.c.$$

Théorème 2 $r_\mathbf{n}(\cdot)$ *converge en moyenne d'ordre* q *vers* $r(\cdot)$ *et*

$$\|r_\mathbf{n}(x_\mathbf{j}) - r(x_\mathbf{j})\|_q \;=\; O\left(b_\mathbf{n} + \sqrt{\frac{1}{\widehat{\mathbf{n}} b_\mathbf{n}^d \rho_\mathbf{n}^N}}\right),\ q > 1.$$

3 Prédiction spatiale

Soit $(Z_\mathbf{i})_\mathbf{i}$ un processus spatial que nous voulons prédire en des sites non observés. Nous supposons que la valeur $Z_\mathbf{i}$ du champ dépend seulement des valeurs prises par ce champ dans un voisinage de $\mathbf{i}$ ne contenant pas $\mathbf{i}$. On utilisera les mêmes notations que précédemment, avec $X_\mathbf{i} = \widetilde{Z}_\mathbf{i}$ et $Y_\mathbf{i} = Z_\mathbf{i}$. Le vecteur $\widetilde{Z}_\mathbf{i}$, dont les d composantes sont les $\{Z_\mathbf{i},\, \mathbf{i} \in \mathcal{V}_\mathbf{i}\}$ concatenées et ordonnées selon un ordre arbitraire, est construit à partir d'un voisinage $\mathcal{V}_\mathbf{i}$ de la forme $\mathbf{i} + \mathcal{V}$, où $\mathcal{V}$ est un ensemble borné fixé qui ne contient pas $\mathbf{0} = (0, 0, \ldots, 0)$. Nous supposons que le champ spatial est observé sur $\mathcal{O}_\mathbf{n}$ contenu dans $\mathcal{I}_\mathbf{n}$.
En se basant sur l'estimateur de la fonction de régression, la prédiction $(Z_\mathbf{i})_{\mathbf{i}\in(\mathbb{Z})^N}$ en un site $\mathbf{i_0}$ s'écrit

$$\widehat{Z}_{\mathbf{i_0}} = r_\mathbf{n}(\widetilde{Z}_{\mathbf{i_0}}) = \frac{\displaystyle\sum_{\substack{\mathbf{i}\in\mathcal{O}_\mathbf{n} \\ \mathcal{V}_\mathbf{i}\subset\mathcal{O}_\mathbf{n}}} Z_\mathbf{i} K_1\left(\frac{\widetilde{Z}_{\mathbf{i_0}} - \widetilde{Z}_\mathbf{i}}{b_\mathbf{n}}\right) K_{2,\rho_\mathbf{n}}(\|\mathbf{i_0} - \mathbf{i}\|)}{\displaystyle\sum_{\substack{\mathbf{i}\in\mathcal{O}_\mathbf{n} \\ \mathcal{V}_\mathbf{i}\subset\mathcal{O}_\mathbf{n}}} K_1\left(\frac{\widetilde{Z}_{\mathbf{i_0}} - \widetilde{Z}_\mathbf{i}}{b_\mathbf{n}}\right) K_{2,\rho_\mathbf{n}}(\|\mathbf{i_0} - \mathbf{i}\|)}.$$

Ce prédicteur spatial est une adaptation de celui proposé dans les travaux de Biau et Cadre (2004).

Nous avons comparé les performances de notre estimateur à celles d'autres estimateurs rencontrés dans la littérature : notre méthode s'avère plus efficace lorsque les données sont spatialement dépendantes.

Références

[1] Biau, G. and Cadre, B. (2004). Nonparametric spatial prediction. *Statistical Inference for Stochastic Processes*, 7(3) :327–349.

[2] Dabo-Niang, S., Hamdad, L., Ternynck, C., and Yao, A.-F. (2013). A kernel spatial density estimation allowing for the analysis of spatial clustering : application to Monsoon Asia Drought Atlas data. Submitted, in revision.

[3] Dabo-Niang, S. and Yao, A.-F. (2007). Kernel regression estimation for continuous spatial processes. *Mathematical Methods of Statistics*, 16(4) :298–317.

[4] Menezes, R., García-Soidán, P., and Ferreira, C. (2010). Nonparametric spatial prediction under stochastic sampling design. *Journal of Nonparametric Statistics*, 22(3) :363–377.

[5] Ternynck, C. (2014). Spatial regression for functional data with spatial dependency. *Journal de la Société Française de Statistique*, accepté, à paraître.