Modèles de régression gaussienne pour des distributions en entrée

Nil Venet [1] & François Bachoc [2] & Fabrice Gamboa [2] & Jean-Michel Loubes [2]

[1] *CEA Tech en Occitanie Pyrénées-Méditerranée*
INSA, Bât. 8 135 avenue de Rangueil 31400 Toulouse
nil.venet@cea.fr
[2] *Institut de Mathématiques de Toulouse*
Université Paul Sabatier 118, route de Narbonne F-31062 Toulouse Cedex 9
(francois.bachoc, jean-michel.loubes, fabrice.gamboa)@math.univ-toulouse.fr

Résumé. On s'intéresse ici à la prédiction de processus gaussiens indexés par des distributions de probabilité. Pour cela nous donnons des noyaux définis positifs sur l'espace des distributions, que nous construisons à partir des distances de Monge-Kantorovich. Ces dernières, aussi appelées distances de Wasserstein, ont reçu une attention grandissante, notamment en machine learning, en tant que mesures de dissemblance entre des distributions. Ces noyaux sont les covariances de processus gaussiens dont nous donnons des propriétés de stationnarité et qui généralisent des processus classiques, tel le mouvement brownien fractionnaire. Nous donnons des résultats théoriques et menons des simulations numériques qui étayent les bonnes performances des estimateurs par Krigeage associés à ces processus.

Mots-clés. Processus gaussien, noyau défini positif, Krigeage, distance de Wasserstein, mouvement brownien fractionnaire.

Abstract. In this work we are interested in the forecast of Gaussian processes indexed by probability distributions. For this, we provide a family of positive definite kernels built using Monge-Kantorovich distances. Also known as Wasserstein distances, they have received a growing attention, particularly in machine learning as a discrepancy measure for probability measures. These kernels are the covariances of Gaussian processes with stationarity properties which generalize classical processes, such as the fractional Brownian motion. We give theoretical and numerical results that support the good performances of the Kriging estimator associated to these processes.

Keywords. Gaussian process, positive definite kernel, Kriging, Wasserstein distance, fractional Brownian motion.

1 Introduction

Le Krigeage est une méthode d'estimation pour des données fonctionnelles dont la popularité initiale en statistiques spatiales s'est depuis étendue à de nombreux domaines

(voir [2] et [4]). Elle consiste en la prédiction de la valeur inconnue d'une fonction en un point par une combinaison linéaire de valeurs connues en d'autres points. Les données sont modélisées a priori par un processus aléatoire, souvent gaussien, et l'estimateur de Krigeage est donné par l'espérance du processus conditionné à interpoler les observations. C'est aussi l'estimateur linéaire sans biais de variance minimale du processus choisi.

La loi d'un processus gaussien étant caractérisée par sa moyenne (qu'on supposera toujours nulle ici) et sa fonction de covariance, le choix de cette dernière est cruciale pour assurer les bonnes propriétés du modèle. Typiquement on souhaite que les valeurs du processus à différents points soient d'autant plus corrélées que ceux-ci sont proches. Une démarche naturelle consiste à choisir une distance sur l'espace des entrées et à chercher des fonctions de covariance stationnaires (c'est-à-dire des noyaux définis positifs) sur celui-ci. Pour ces dernières la corrélation entre les valeurs du processus en deux points ne dépend que de la distance entre ces derniers. On préférera utiliser des processus à accroissements stationnaires lorsque les données modélisées présentent des tendances, voire des processus présentant des accroissements stationnaires d'ordre supérieur. S'il existe une littérature importante sur les fonctions de covariances stationnaires sur $\mathbb{R}^d$, $d \geq 1$, notons que la construction de telles fonctions sur des espaces plus généraux est encore un domaine de recherche ouvert.

Ici nous nous intéressons à la prédiction de données fonctionnelles dont les entrées sont des distributions de probabilités. Cette situation apparait par exemple lors de simulations numériques où les entrées ne sont pas déterministes mais de lois connues. Dans la Section 2 nous considérons l'espace des distributions de probabilités muni de la distance de Wasserstein (voir [5]), et nous donnons des fonctions de covariance de deux types : stationnaires et à accroissement stationnaires.

Nous nous penchons dans la Section 3 sur la sélection d'un noyau dans un modèle paramétrique de famille de covariances, dans le cas stationnaire. Nous montrons la consistance et la normalité asymptotique de l'estimation du paramètre de covariance par cette méthode. Puis nous considérons l'estimation par Krigeage sous la covariance estimée et montrons son optimalité asymptotique. Enfin, nous testons dans la Section 4 cette méthode sur un jeu de données simulées, en choisissant une famille paramétriques de covariances stationnaires obtenues précédemment. Nous constatons qu'elle se compare avantageusement à une utilisation de noyaux de covariances qui sont fonctions d'un nombre finis de paramètres associés aux distributions. Nos résultats étayent l'efficacité de la distance de Wasserstein en tant que mesure de dissemblance entre des distributions de probabilités, et délivrent une méthode performante de prédiction de données fonctionnelles dont les entrées sont non déterministes.

Cette communication est entièrement basée sur des travaux à paraitre dans [1], auquel nous renvoyons pour plus de détails, les preuves des résultats et une bibliographie plus importante.

2 Noyaux de covariances sur l'espace de Wasserstein

Considérons l'ensemble $\mathcal{W}_2(\mathbb{R})$ des mesures de probabilité sur la droite réelle qui possèdent un moment d'ordre deux. Pour $\mu, \nu \in \mathcal{W}_2(\mathbb{R})$, notons $\Pi(\mu, \nu)$ l'ensemble des mesures de probabilités sur $\mathbb{R} \times \mathbb{R}$ dont les lois marginales sont μ et ν. Rappelons que la distance de Wasserstein quadratique entre μ et ν est définie par

$$W_2(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int |x - y|^2 \, d\pi(x, y) \right)^{1/2}. \tag{1}$$

Théorème 2.1. *Pour $F : \mathbb{R}^+ \to \mathbb{R}^+$ complètement monotone et $0 < H \leq 1$,*

$$(\mu, \nu) \mapsto F\left(W_2^{2H}(\mu, \nu)\right) \tag{2}$$

est la fonction de covariance d'un processus gaussien stationnaire indexé par $\mathcal{W}_2(\mathbb{R})$.

Rappelons qu'une fonction $F : \mathbb{R}^+ \to \mathbb{R}^+$ complètement monotone est une fonction infiniment dérivable telle que $(-1)^n F^{(n)}$ est à valeurs positives pour tout $n \in \mathbb{N}$. Les fonctions $x^{-\lambda}$, $\ln(1 + 1/x)$ et $e^{-\lambda x}$ sont par exemple complètement monotones pour $\lambda > 0$.

Théorème 2.2. *Pour tout $0 \leq H \leq 1$ et $\sigma \in \mathcal{W}_2(\mathbb{R})$,*

$$K^{H,\sigma}(\mu, \nu) = \frac{1}{2}\left(W_2^{2H}(\sigma, \mu) + W_2^{2H}(\sigma, \nu) - W_2^{2H}(\mu, \nu)\right) \tag{3}$$

est la fonction de covariance d'un processus gaussien à accroissements stationnaires indexé par $\mathcal{W}_2(\mathbb{R})$, non-dégénéré si et seulement si $0 < H < 1$. S'il est centré, ce processus est le champ brownien H-fractionnaire indexé par l'espace de Wasserstein $\mathcal{W}_2(\mathbb{R})$ (voir [3]).

Remarque 1. *C'est l'identité $W_2(\mu, \nu) = \mathbb{E}\left(F_\mu^{-1}(U) - F_\nu^{-1}(U)\right)^2$, (où U est une variable aléatoire uniforme sur $[0, 1]$ et F_μ^{-1}, F_ν^{-1} sont les fonctions quantiles des distributions μ, ν) qui permet l'évaluation par intégration numérique de $W_2(\mu, \nu)$, et donc des noyaux proposés (voir par exemple [5]).*

3 Résultats pour les estimateurs par maximum de vraisemblance et Krigeage

Dans ce paragraphe on suppose qu'on dispose de distributions $\mu_1, \cdots, \mu_n \in \mathcal{W}_2(\mathbb{R})$, pour lesquelles on a observé $y = (Y_{\mu_1}(\omega), \cdots, Y_{\mu_n}(\omega))$, où Y est un processus gaussien centré de covariance K_{θ_0} dans un modèle paramétrique $\{K_\theta, \theta \in \Theta\}$. On s'intéresse dans un premier temps à l'estimateur $\hat{\theta}_{ML}$ de θ_0 par maximum de vraisemblance, tel que $\hat{\theta}_{ML} \in \operatorname{argmin} L_\theta$, avec

$$L_\theta = \frac{1}{n} \ln(\det R_\theta) + \frac{1}{n} y^t R_\theta^{-1} y, \tag{4}$$

où R_θ est la matrice $(K_\theta(\mu_i,\mu_j))_{i,j=1}^n$.

Dans un second temps on s'intéresse à l'estimateur de Y par Krigeage sous la covariance estimée $K_{\hat{\theta}_{ML}}$, défini par

$$\hat{Y}_{\hat{\theta}_{ML}}(\mu) = \mathbb{E}\left(Y_\mu^{\hat{\theta}_{ML}} | (Y_{\mu_1}^{\hat{\theta}_{ML}}, \cdots Y_{\mu_n}^{\hat{\theta}_{ML}}) = (y_1, \cdots, y_n)\right), \tag{5}$$

où $Y^{\hat{\theta}_{ML}}$ est un processus gaussien de covariance $K_{\hat{\theta}_{ML}}$. Les conditions suivantes sont suffisantes pour donner la consistance et la normalité asymptotique de $\hat{\theta}_{ML}$, ainsi que l'optimalité asymptotique de $\hat{Y}_{\hat{\theta}_{ML}}$.

Condition 3.1. *On considère une matrice triangulaire de points d'observations de $\mathcal{W}_2(\mathbb{R})$ $\{\mu_1, ..., \mu_n\} = \{\mu_1^{(n)}, ..., \mu_n^{(n)}\}$ tels que pour tout $n \in \mathbb{N}$ et $1 \leq i \leq n$, μ_i est de support inclus dans $[i, i+K]$, où $K < \infty$ est fixe.*

Condition 3.2. *Le modèle de fonctions de covariance $\{K_\theta, \theta \in \Theta\}$ est tel que*

$$\forall \theta \in \Theta,\ K_\theta(\mu,\nu) = F_\theta\left(W_2(\mu,\nu)\right)\ \text{ et } \quad \sup_{\theta\in\Theta} |F_\theta(t)| \leq \frac{A}{1+|t|^{1+\tau}},$$

avec $A < \infty$ et $\tau > 1$ des constantes.

Condition 3.3. *Nous disposons d'observations $y_i = Y(\mu_i)$, $i = 1, \cdots, n$ du processus aléatoire gaussien Y, centré et de covariance K_{θ_0} pour un $\theta_0 \in \Theta$.*

Condition 3.4. *La suite de matrices $R_\theta = (K_\theta(\mu_i,\mu_j))_{1\leq i,j\leq n}$ est telle que $\lambda_{\inf}(R_\theta) \geq c$ pour une constante $c > 0$, où $\lambda_{\inf}(R_\theta)$ désigne la plus petite valeur propre de R_θ.*

Condition 3.5. $\forall \alpha > 0, \liminf_{n\to\infty} \inf_{\|\theta-\theta_0\|\geq\alpha} \frac{1}{n}\sum_{i,j=1}^n \left[K_\theta(\mu_i,\mu_j) - K_{\theta_0}(\mu_i,\mu_j)\right]^2 > 0.$

Condition 3.6. *$\forall t \geq 0$, $F_\theta(t)$ est $\mathcal{C}^1$ en θ et vérifie* $\sup_{\theta\in\Theta} \max_{i=1,\cdots,p} \left|\frac{\partial}{\partial\theta_i} F_\theta(t)\right| \leq \frac{A}{1+t^{1+\tau}}$, *où A, τ sont définis dans la Condition 3.2.*

Condition 3.7. *Pour tout $t \geq 0$, $F_\theta(t)$ est $\mathcal{C}^3$ en θ et $\forall q \in \{2,3\}$, $\forall i_1 \cdots i_q \in \{1, \cdots p\}$,*

$$\sup_{\theta\in\Theta} \max_{i=1,\cdots,p} \left|\frac{\partial}{\partial\theta_{i_1}} \cdots \frac{\partial}{\partial\theta_{i_q}} F_\theta(t)\right| \leq \frac{A}{1+|t|^{1+\tau}}.$$

Condition 3.8. $\forall (\lambda_1 \cdots, \lambda_p) \neq (0, \cdots, 0)$,

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i,j=1}^n \left(\sum_{k=1}^n \lambda_k \frac{\partial}{\partial_{\theta_k}} K_{\theta_0}\left(\mu_i,\mu_j\right)\right)^2 > 0.$$

Théorème 3.9. *Sous les conditions 3.1 à 3.5 l'estimateur par maximum de vraisemblance est consistant, c'est-à dire :*

$$\hat{\theta}_{ML} \xrightarrow[n\to\infty]{\mathbb{P}} \theta_0.$$

Théorème 3.10. *Soit M_{ML} la matrice de taille $p \times p$ définie par*

$$(M_{ML})_{i,j} = \frac{1}{2n} Tr\left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_j} \right),$$

où R_θ est définie dans (4). Sous les conditions 3.1 à 3.8, l'estimateur par maximum de vraisemblance est asympotiquement normal. Plus précisément :

$$\sqrt{n} M_{ML}^{1/2} \left(\hat{\theta}_{ML} - \theta_0 \right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I_n).$$

De plus $\quad 0 < \liminf_{n\to\infty} \lambda_{min}(M_{ML}) \leq \limsup_{n\to\infty} \lambda_{max}(M_{ML}) < +\infty.$

Théorème 3.11. *Sous les conditions 3.1 à 3.8, l'estimateur par Krigeage sous $\hat{\theta}_{ML}$ est asymptotiquement optimal :*

$$\forall \mu \in \mathcal{W}_2(\mathbb{R}), \ \left| \hat{Y}_{\hat{\theta}_{ML}}(\mu) - \hat{Y}_{\theta_0}(\mu) \right| = o_{\mathbb{P}}(1).$$

4 Simulations numériques

Dans ce paragraphe nous évaluons la performance des méthodes statistiques proposées au paragraphe précédent sur des données simulées. Notons $m_k(\nu)$ le k-ième moment d'une distribution de probabilité ν et considérons la fonction $F : \mathcal{W}_2(\mathbb{R}) \to \mathbb{R}$ telle que

$$F(\mu) = \frac{m_1(\nu)}{0.05 + \sqrt{m_2(\nu) - m_1(\nu)^2}}, \tag{6}$$

qu'on va chercher à interpoler. Nous simulons aléatoirement et de manière indépendante des distributions $\nu_1, \cdots, \nu_{100}$ qui sont des gaussiennes de moyennes et de variances tirées uniformément, perturbées aléatoirement afin d'exhiber des irrégularités (voir [1] pour plus de détails). Nous considérons les $(\nu_i, F(\nu_i))_{i=1}^{100}$ comme des données d'apprentissage et choisissons par maximum de vraisemblance les paramètres $\hat{\sigma}^2, \hat{\ell}, \hat{H}$ pour le modèle gaussien paramétrique

$$\left\{ K_{\sigma^2,\ell,H}(\nu_1, \nu_2) = \sigma^2 \exp\left(-\frac{W_2(\nu_1, \nu_2)^{2H}}{\ell} \right), H \in [0,1], \sigma \in C, l \in C' \right\}, \tag{7}$$

dont les éléments sont des fonctions de covariance d'après le Théorème 2.2. Notons que ce modèle est le pendant d'un modèle classiquement employé en statistiques spatiales, où la distance de Wasserstein est remplacée par la distance euclidienne. Nous estimons maintenant $F(\nu)$ par Krigeage du processus Gaussien de covariance $K_{\hat{\sigma}^2,\hat{\ell},\hat{H}}$.

Nous générons ensuite un ensemble de distributions de test $(\nu_{t,i})_{i=1}^{500}$ de la même manière que les ν_i et nous observons les deux critères de qualité "root mean square error" et "confiance interval ratio"

$$RMSE^2 = \frac{1}{500}\sum_{i=1}^{500}\Big(F(\nu_{t,i}) - \hat{F}(\nu_{t,i})\Big)^2 \text{ et } CIR_\alpha = \frac{1}{500}\sum_{i=1}^{500}\mathbf{1}_{\left\{\left|F(\nu_{t,i})-\hat{F}(\nu_{t,i})\right|\leq q_\alpha\hat{\sigma}(\nu_{t,i})\right\}},$$

où q_α est le $\left(\frac{1}{2}+\frac{\alpha}{2}\right)$-quantile de la gaussienne standard. Nous comparons dans la Table 1 les valeurs de ces critères pour notre méthode ("distribution") et deux méthodes classiques ("Legendre" et "PCA") où les noyaux utilisés opèrent sur un nombre finis de paramètres associés aux distributions. La méthode "Legendre" projette les fonctions de densité sur une base de polynômes de Legendre, tandis que la méthode "PCA" utilise une approximation par composantes principales (voir [1]). Comme on le constate notre méthode présente des valeurs de $RMSE$ très inférieures à celles des méthodes classiques, et un $CIR_{0.9}$ proche de 0.9.

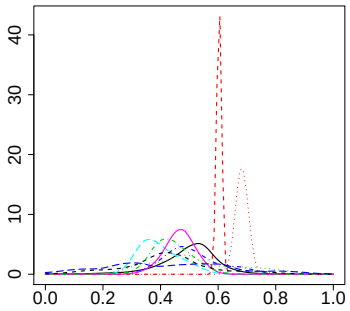

FIGURE 1 – Densités de probabilités de 10 des distributions d'apprentissage

modèle	RMSE	$CIR_{0.9}$
"distribution"	0.094	0.92
"Legendre" ordre 5	0.49	0.92
"Legendre" ordre 10	0.34	0.89
"Legendre" ordre 15	0.29	0.91
"PCA" ordre 5	0.63	0.82
"PCA" ordre 10	0.52	0.87
"PCA" ordre 15	0.47	0.93

TABLE 1 – Comparaison du modèle proposé avec d'autres modèles

Bibliographie

[1] François Bachoc, Fabrice Gamboa, Jean-Michel Loubes, and Nil Venet. Gaussian process regression model for distribution inputs. *arXiv preprint arXiv :1701.09055*, 2017.

[2] Noel A. C. Cressie. *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1991. A Wiley-Interscience Publication.

[3] J. Istas. Manifold indexed fractional fields. *ESAIM Probab. Stat.*, 16 :222–276, 2012.

[4] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2006.

[5] Cédric Villani. *Optimal transport : old and new*, volume 338. Springer Science & Business Media, 2009.