Prior de référence de Gibbs

Joseph Muré [1,2] & Josselin Garnier [3] & Loic Le-Gratiet [1] & Anne Dutfoy [1]

[1] *Département de Management des Risques Industriels (MRI), EDF R&D, Chatou et Saclay, France / prenom.nom@edf.fr*

[2] *Laboratoire de Probabilités et Modèles Aléatoires (LPMA), Université Paris Diderot (Paris 7), Paris, France / mure@math.univ-paris-diderot.fr*

[3] *Centre de Mathématiques Appliquées (CMAP), École Polytechnique, Palaiseau, France / josselin.garnier@polytechnique.edu*

Résumé. Dans l'esprit des priors de référence, nous proposons une loi *a priori* objective sur les paramètres du noyau de corrélation d'un modèle de krigeage simple. Ce prior étant propre et défini à travers ses densités conditionnelles, la distribution *a posteriori* associée l'est aussi et se prête donc bien à l'échantillonnage de Gibbs, ce qui rend le traitement bayésien viable. Des exemples numériques montrent que le taux de couverture fréquentiste des intervalles de prédiction associés est quasi-optimal. Ce travail peut être étendu au cadre plus général du krigeage universel.

Mots-clés. Processus gaussien, krigeage, prior de Jeffreys, prior de référence de Gibbs, vraisemblance intégrée, taux de couverture fréquentiste, posterior propre.

Abstract. We propose an objective prior distribution on correlation kernel parameters for Simple Kriging models in the spirit of reference priors. Because it is proper and defined through its conditional densities, it and its associated posterior distribution lend themselves well to Gibbs sampling, thus making the full-Bayesian procedure tractable. Numerical examples show it has near-optimal frequentist performance in terms of prediction interval coverage. This work can be extended to the more general Universal Kriging framework.

Keywords. Gaussian process, Kriging, Jeffreys prior, Gibbs reference prior, integrated likelihood, frequentist coverage, posterior propriety.

1 Introduction

Les processus gaussiens sont couramment utilisés pour modéliser une quantité à valeur réelle lorsque cette quantité est observée en un nombre fini de points. Un tel modèle permet de représenter de manière concise l'incertitude quant à la valeur de la quantité aux points où on ne dipose pas d'observation. Dans ce qui suit, on se place dans le cadre de processus gaussiens stationnaires de moyenne nulle, connu dans la littérature géostatistique sous le nom de « krigeage simple » et qui est aussi fréquemment utilisé lors d'expériences numériques ou à des fins d'apprentissage statistique.

Cependant, la loi de probabilité d'un processus gaussien stationnaire ne dépend pas uniquement de sa fonction de moyenne (supposée ici nulle) mais aussi d'un paramètre de variance et d'un noyau de corrélation dépendant lui-même de paramètres.

Nous proposons une loi *a priori* « objective » de type Jeffreys sur ces paramètres, à laquelle nous donnons le nom de « prior de référence de Gibbs ».

C'est parce que l'Estimateur du Maximum de Vraisemblance (EMV) des paramètres d'un modèle de krigeage est peu robuste, la fonction de vraisemblance étant souvent assez plate [1], que la détermination d'un prior sur lesdits paramètres est nécessaire. Certes, l'EMV peut être stabilisé en ajoutant une composante aux éléments diagonaux de la matrice de covariance mais, comme l'a relevé [2], cet ajout équivaut à supposer qu'une partie de la variabilité n'est pas explicable par les entrées du modèle. Alternativement, [1] propose de pénaliser la fonction de vraisemblance, ce qui peut aussi être vu comme l'utilisation du Maximum A Posteriori (MAP) associé à un certain prior. Évidemment, la démarche bayésienne évite le problème de la robustesse de l'estimation des paramètres, puisqu'il suffit d'utiliser la distribution prédictive *a posteriori*.

Quelle que soit l'utilisation faite de la loi *a priori*, la choisir peut être une gageure si l'on manque d'information. Le bayésianisme objectif, introduit dans ce cadre par [3], permet de répondre à ce problème. Les auteurs de [3] ont calculé le prior de référence dans le présent cadre et établi que le posterior associé est propre. Leur travail a ensuite été étendu par [4], [5] and [6]. Cependant, tous ces travaux font une hypothèse restrictive afin de garantir la propreté du posterior. Elle suppose essentiellement que pour tout noyau de corrélation deux fois dérivable, le nombre de points d'observation ne saurait excéder de plus que 2 la dimension de l'espace. Ainsi, des noyaux de covariance usuels tels que les noyaux de Matérn de paramètre de régularité $\nu > 1$ ne peuvent être utilisés. Cet état de fait, auquel s'ajoute notre volonté de rendre la méthode bayésienne utilisable en pratique nous a conduit à considérer un prior « objectif » autre quoique similaire. Comme il est défini à travers ses densités conditionnelles, favorisant ainsi l'échantillonnage de Gibbs, nous l'avons appelé « prior de référence de Gibbs ».

Nous garantissons théoriquement la propreté du prior de référence de Gibbs portant sur le vecteur des longueurs de corrélation des classes de noyaux de Matérn tensorisée et anisotrope géométrique. De plus, nous fournissons un cadre de démonstration de ce résultat pour d'autres classes de noyaux de corrélation.

Nous échantillonnons les lois *a posteriori* conditionnelles par l'algorithme de Metropolis, ce qui, couplé avec l'algorithme de Gibbs, permet d'échantillonner la loi jointe *a posteriori*. Nous comparons ensuite deux manière d'utiliser cette distribution : la manière bayésienne et l'approche du Maximum A Posteriori.

Cette dernière a l'inconvénient de nécessiter l'estimation d'un paramètre. Cela dit, les comparaisons montrent que l'estimateur MAP est significativement plus robuste que l'EMV vis-à-vis de de la variabilité des réalisations du processus gaussien.

Au-delà de l'inférence des paramètres, le plus important pour nous est notre capacité à rendre compte de l'incertitude quant aux valeurs prises par le processus gaussien aux

points où il n'est pas observé. Des exemples montrent que la méthode baysésienne aboutit à des intervalles de prédiction dont les taux de couverture sont proches de leur niveau théorique, tandis que les intervalles de prédiction issus des méthodes de *plug-in* du MAP et *a fortiori* de l'EMV ont des taux de couverture sensiblement plus faibles.

2 Définition du prior de référence de Gibbs

Introduisons quelques notations. Nous considérons dans un premier temps un noyau de covariance de la forme $\sigma^2 K_\theta$, où $\sigma^2 \in \mathbb{R}_+^*$ est appelé paramètre de variance et K est un noyau de corrélation stationnaire dépendant d'un paramètre $\theta \in \mathbb{R}_+^*$ appelé parfois *longueur de corrélation* : plus θ est grand, plus les valeurs du processus seront corrélées, même en des points distants les uns des autres. Les points où le processus est observé sont notés $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. Leur matrice de corrélation est $\boldsymbol{\Sigma}_\theta$: il s'agit de la matrice dont l'entrée (i, j) est $K_\theta(\boldsymbol{x}_i - \boldsymbol{x}_j)$. L'algorithme du prior de référence de Berger-Bernardo [7] nous conduit à calculer successivement $\pi(\sigma^2|\theta)$, le prior de Jeffreys sur σ^2 sachant θ, puis $\pi(\theta)$, le prior de Jeffreys sur θ relativement à la vraisemblance *intégrée*, c'est-à-dire obtenue en intégrant la vraisemblance contre $\pi(\sigma^2|\theta)d\sigma^2$.

Nous obtenons ainsi

$$\pi(\sigma^2|\theta) \propto \frac{1}{\sigma^2} \; ; \tag{2.1}$$

$$\pi(\theta) \propto \sqrt{\operatorname{Tr}\left[\left(\frac{\partial}{\partial\theta}(\boldsymbol{\Sigma}_\theta)\boldsymbol{\Sigma}_\theta^{-1}\right)^2\right] - \frac{1}{n}\operatorname{Tr}\left[\frac{\partial}{\partial\theta}(\boldsymbol{\Sigma}_\theta)\boldsymbol{\Sigma}_\theta^{-1}\right]^2}. \tag{2.2}$$

Il peut arriver que le noyau de corrélation K dépende d'un paramètre multidimensionnel $\boldsymbol{\theta} = (\theta_1, ..., \theta_r)^\top \in (\mathbb{R}_+^*)^r$. Dans ce cas, la même démarche aboutit à nouveau à $\pi(\sigma^2|\boldsymbol{\theta}) \propto 1/\sigma^2$, mais $\pi(\boldsymbol{\theta})$ est cette fois proportionnel à la racine carrée du déterminant de la matrice dont l'entrée (i, j) est

$$\operatorname{Tr}\left[\frac{\partial}{\partial\theta_i}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial}{\partial\theta_j}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right] - \frac{1}{n}\operatorname{Tr}\left[\frac{\partial}{\partial\theta_i}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]\operatorname{Tr}\left[\frac{\partial}{\partial\theta_j}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]. \tag{2.3}$$

Cette solution fait intervenir le prior de Jeffreys d'un paramètre multidimensionnel, contexte dans lequel il s'avère souvent insatisfaisant [8] : il vaudrait donc mieux séparer les composantes de $\boldsymbol{\theta}$. Pour ce faire, l'algorithme du prior de référence nécessiterait de les ordonner. Dans un contexte où on ne dispose pas d'information *a priori*, une telle opération serait nécessairement arbitraire.

La solution alternative que nous proposons est de calculer le prior de Jeffreys de chaque composante *en supposant les autres composantes connues*. Pour tout entier i compris entre 1 et r, on obtient :

$$\pi(\theta_i \mid \theta_j\ \forall j \neq i) \propto \sqrt{\mathrm{Tr}\left[\left(\frac{\partial}{\partial\theta_i}(\boldsymbol{\Sigma_\theta})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^2\right] - \frac{1}{n}\mathrm{Tr}\left[\frac{\partial}{\partial\theta_i}(\boldsymbol{\Sigma_\theta})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]^2}. \tag{2.4}$$

Le prior de référence de Gibbs est alors défini comme l'unique loi jointe compatible avec ces lois conditionnelles. Pour que cette définition ait un sens, il faut et il suffit que chaque loi conditionnelle soit propre et qu'il existe une loi jointe propre compatible.

Hypothèse 1 *Notons* $\underline{\theta_{r,i}} := \min(\theta_j | 1 \leqslant j \leqslant r, j \neq i)$. *Il existe des nombres réels* $a \geqslant 0$ *et* $b > 1$ *vérifiant l'assertion suivante :*

Pour tout entier i *compris entre 1 et* r, *il existe des nombres réels positifs* T_i, $M_{i,1}$, $M_{i,2}$, $M_{i,3}$, $M_{i,4}$, $M_{i,5}$ *et* $M_{i,6}$ *tels que*

1. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+^*)^r$, $\|\frac{\partial}{\partial\theta_i}\boldsymbol{\Sigma_\theta}\| \leqslant M_{i,1}$.
2. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+^*)^r$, $\|\theta_i^b \frac{\partial}{\partial\theta_i}\boldsymbol{\Sigma_\theta}\| \leqslant M_{i,2}$.
3. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+^*)^r$, $\|\underline{\theta_{r,i}}^{-a} \frac{\partial}{\partial\theta_i}\boldsymbol{\Sigma_\theta}\| \leqslant M_{i,3}$.
4. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+^*)^r$, $\|\underline{\theta_{r,i}}^{-a}\theta_i^b \frac{\partial}{\partial\theta_i}\boldsymbol{\Sigma_\theta}\| \leqslant M_{i,4}$.
5. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+^*)^r$ tel que $\underline{\theta_{r,i}} \leqslant T_i$, $\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\| \leqslant M_{i,5}$.
6. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+^*)^r$ tel que $\underline{\theta_{r,i}} \geqslant T_i$, $\|\underline{\theta_{r,i}}^{-a}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\| \leqslant M_{i,6}$.

Théorème 1 *Si le paramètre* $\boldsymbol{\theta}$ *est à valeurs dans* $(\mathbb{R}_+^*)^r$, *où* r *est un entier supérieur ou égal à 2, et si l'hypothèse 1 est vérifiée, alors il existe une unique loi de probabilité jointe* $\pi(\boldsymbol{\theta})$ *compatible avec les lois conditionnelles exprimées en (2.4).*

Discutons rapidement les majorations intervenant dans l'hypothèse 1.

- La majoration 1 garantit que $\|\frac{\partial}{\partial\theta_i}\boldsymbol{\Sigma_\theta}\|$ reste borné quand θ_i est au voisinage de 0.
- La majoration 2 fait en sorte que $\|\frac{\partial}{\partial\theta_i}\boldsymbol{\Sigma_\theta}\|$ décroisse « rapidement » quand θ_i tend vers l'infini. Plus b est grand, plus les queues des lois *a priori* conditionnelles (2.4) sont fines. Si b est inférieur ou égal à 1, nous ne pouvons garantir que les lois *a priori* conditionnelles sont propres.
- La majoration 5 ne pose pas problème : à supposer que l'une des composantes « connues » de $\boldsymbol{\theta}$ est suffisamment petite, elle nous permet de contrôler $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$ et l'empêche d'affecter le comportement de $\frac{\partial}{\partial\theta_i}\boldsymbol{\Sigma_\theta}$, ce qui est important pour assurer la propreté des lois conditionnelles (2.4).
- La majoration 6 est la plus difficile à démontrer. Si toutes les composantes de $\boldsymbol{\theta}$ tendent simultanément vers l'infini, $\boldsymbol{\Sigma_\theta}$ tend vers la matrice dont toutes les entrées sont égales à 1 et la norme de son inverse $\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\|$ tend vers l'infini. La majoration 6 permet de contrôler son rythme de croissance. L'idée est la suivante : relativement à la densité conditionnelle $\pi(\theta_i \mid \theta_j\ \forall j \neq i)$, $\underline{\theta_{r,i}}$ est une constante multiplicative, ce qui veut dire que nous pouvons diviser la densité conditionnelle par $\underline{\theta_{r,i}}^a$ pour

nous faciliter la tâche sans changer la loi conditionnelle. Cependant, cette opération augmente la valeur de la densité quand $\underline{\theta_{r,i}}$ est proche de 0, ce qui nécessite de « revérifier » les majorations 1 and 2 dans un tel cas, aboutissant respectivement aux majorations 3 et 4. En bref, a doit être suffisamment grand pour assurer la majoration 6 tout en étant suffisamment petit pour ne pas contrevenir aux majorations 3 et 4.

3 Le cas particulier des noyaux de Matérn

Proposition 2 *Si le noyau de corrélation est de Matérn tensorisé ou anisotrope géométrique, alors l'hypothèse 1 est vérifiée dès lors que l'hypothèse 2 l'est.*

L'hypothèse 2 n'est pas très restrictive, mais elle interdit notamment les ensembles de points d'observation qui peuvent s'écrire sous forme de produit cartésien.

Hypothèse 2 *Les points d'observation étant notés $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, pour tous entiers i et j tels que $1 \leqslant i < j \leqslant n$, aucune composante du vecteur $\boldsymbol{x}_i - \boldsymbol{x}_j$ n'est nulle.*

Le tableau suivant donne les taux de couverture moyens d'intervalles de prédiction en des points non observés dans quatre cas de figure : 1) les paramètres σ^2 et $\boldsymbol{\theta}$ sont connus ; 2) l'EMV 3) le MAP associé à ces deux paramètres est utilisé (*plug-in*) ; 4) la démarche bayésienne est utilisée et aucune estimation des paramètres n'est requise.

La moyenne est faite pour des jeux de 30 points d'observation répartis aléatoirement dans le cube unité $[0, 1]^3$ et des réalisations de processus gaussiens de moyenne nulle et de noyau de covariance de Matérn anisotrope géométrique de paramètre de variance $\sigma^2 = 1$, de paramètre de régularité $\nu = 5/2$ et dont le vecteur des longueurs de corrélation est donné par le tableau (colonne de gauche).

Long. correl.	Vrai	MLE	MAP	Bayésien
0.4 – 0.8 – 0.2	0.95	0.88	0.91	0.95
0.5 – 0.5 – 0.5	0.95	0.89	0.90	0.94
0.7 – 1.3 – 0.4	0.95	0.90	0.92	0.95
0.8 – 0.3 – 0.6	0.95	0.89	0.91	0.95
0.8 – 1.0 – 0.9	0.95	0.90	0.92	0.94

TABLE 1 – Moyenne relativement à 500 ensembles de 30 points d'observation aléatoirement choisis selon une loi uniforme sur le cube unité et réalisations de processus gaussiens (avec paramètre de variance 1 et paramètre de régularité 5/2) du taux de couverture des intervalles de prédiction de niveau 95% sur l'ensemble du cube. « Vrai » signifie que la prédiction est fondée sur la connaissance des vraies valeurs des paramètres.

Le tableau ci-dessous donne les longueurs moyennes des intervalles de prédiction dans la même situation que précédemment.

Long. correl.	Vrai	MLE	MAP	Bayésien
0.4 – 0.8 – 0.2	2.23	2.05 (-8)	2.13 (-4)	2.59 (+16)
0.5 – 0.5 – 0.5	1.69	1.55 (-8)	1.58 (-6)	1.84 (+9)
0.7 – 1.3 – 0.4	1.09	1.02 (-6)	1.07 (-2)	1.21 (+11)
0.8 – 0.3 – 0.6	1.63	1.51 (-7)	1.56 (-4)	1.82 (+12)
0.8 – 1.0 – 0.9	0.71	0.66 (-7)	0.69 (-3)	0.76 (+8)

TABLE 2 – Moyenne obtenue sur 500 jeux de 30 points d'observation aléatoirement choisis selon une loi uniforme sur le cube unité et réalisations d'un processus gaussien (de paramètre de variance 1 et de paramètre de régularité 5/2) de la grandeur suivante : la moyenne sur le cube des longueurs des intervalles de prédiction de niveau 95%. Les nombres entre parenthèses représentent en pourcents l'augmentation lorsqu'on utilise les approches *plug-in* EMV ou MAP ou l'approche bayésienne en comparaison de ce qui est obtenu lorsque les « vraies » valeurs des paramètres sont connues.

Bibliographie

[1] R. Li et A. Sudjianto. Analysis of computer experiments using penalized likelihood in Gaussian Kriging models. *Technometrics*, 47(2) : 111–120, 2005.

[2] I. Andrianakis et P. G. Challenor. The effect of the nugget on Gaussian process emulators of computer models. *Computational Statistics & Data Analysis*, 56(12) : 4215–4228, 2012.

[3] J. O. Berger, Victor De Oliveira, et Bruno Sansò. Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456) : 1361–1374, 2001.

[4] Rui Paulo. Default priors for Gaussian processes. *Annals of Statistics*, 33(2) : 556–582, 2005.

[5] C. Ren, D. Sun et S. K. Sahu. Objective Bayesian analysis of spatial models with separable correlation functions. *Canadian Journal of Statistics*, 41(3) : 488–507, 2013.

[6] M. Gu. *Robust Uncertainty Quantification and Scalable Computation for Computer Models with Massive Output.* PhD thesis, Duke University, 2016.

[7] J. O. Berger et J. M. Bernardo. On the Development of Reference Priors. *Bayesian statistics*, 4(4) : 35–60, 1992.

[8] C. P. Robert, N. Chopin et J. Rousseau. Harold Jeffreys's Theory of Probability Revisited. *Statistical Science*, 24(2) : 141–172, 2009.