Modèles de Processus Gaussiens des plus proches voisins non stationnaires : architecture hiérarchique et méthodes MCMC.

Sébastien Coube-Sisqueille * Sudipto Banerjee † Benoît Liquet ‡

Résumé

La modélisation spatiale non stationnaire est une approche intéressante et prometteuse, mais elle souffre de plusieurs problèmes : son coût computationel, la complexité et le manque de lisibilité de modèles hiérarchiques à plusieurs étages, et la difficulté de sélectionner un modèle. Nous répondons à ces trois problèmes en introduisant un modèle non stationnaire utilisant les processus gaussiens des plus proches voisins (NNGP, pour *Nearest Neighbor Gaussian Process*).
Les NNGP, précis et économiques, sont un bon départ pour répondre au problème du temps de calcul. Nous étudions le comportement des NNGP utilisant une fonction de covariance non stationnaire analytiquement et empiriquement.
Nous introduisons une architecture de modèle lisible afin de faciliter la compréhension des résultats et la sélection de modèles. En particulier, nous créons une famille de modèles cohérente qui rassemble les processus spatiaux avec portée stationnaire, les processus non stationnaires avec des paramètres de portée circulaires, et ceux avec des paramètres de portée elliptiques.
Nous tirons parti de notre architecture hiérarchique et de l'utilisation des NNGP en proposant un algorithme de Langevin ajusté par un pas de Metropolis. Nous améliorons cet algorithme en utilisant la méthode de l'entremêlement de paramétrisations.
Nous implémentons nos méthodes en R et les testons avec des jeux de données synthétiques pour trouver des règles empiriques concernant le choix des hyperparamètres et la sélection de modèle. Nous les utilisons pour analyser un jeu de donnés de pollution au plomb aux États-Unis d'Amérique.
Mots-clefs : Processus gaussien des Plus Proches Voisins – Modèle spatial non stationnaire – MCMC

Développement

Nous observons une variable d'intérêt sur une collection de sites spatiaux. Nous partons d'une modélisation spatiale stationnaire et considérons trois extensions où différents paramètres peuvent varier dans l'espace : la variance marginale du processus latent , les paramètres de portée (potentiellement elliptiques) du processus latent, et, quand les observations sont gaussiennes, la variance du bruit Gaussien. Nous utilisons des fonctions de covariance non stationnaires classiques, définies par Paciorek [2003]

Nous utilisons les processus Gaussiens des plus proches voisins (NNGP) [Datta et al., 2016] afin d'approximer les densités spatiales non stationnaires. La densité NNGP est définie en utilisant une densité conditionnelle récurrente "élaguée" :

$$\tilde{f}(w(s_i)|w(s_1,\ldots,s_{i-1}),\theta) = f(w(s_i)|w(pa(s_i)),\theta), \tag{1}$$

$pa(s_i)$ étant les parents du site s_i dans un Graphe Dirigé Acyclique (DAG) dont les sommets sont identifiés avec les observations, $\tilde{f}(\cdot)$ la densité NNGP, et $f(\cdot)$ étant la densité gaussienne non approximée avec des paramètres de covariance θ. Les parents sont souvent choisis comme les plus proches voisins du point, donnant le nom de la méthode.

*Université de Pau et des Pays de l'Adour, ¸
†University of California, Los Angeles ¸
‡Université de Pau et des Pays de l'Adour, ¸

Un package R [Guinness, 2018] comprend des fonctions de covariance non-stationnaires, mais à notre connaissance il n'y a pas eu à ce jour d'étude théorique ou empirique des propriétés des NNGP quand on utilise une fonction de covariance non stationnaire.
Un premier aspect de notre travail a été de préciser des propriétés algébriques des NNGP utilisant une fonction de covariance non stationnaire. Nous avons démontré plusieurs formules de factorisation qui permettent une implémentation efficace, en particulier avec des langages haut niveau tels que R.
Nous avons également testé différentes heuristiques de construction des NNGP, et retrouvons que les heuristiques qui donnent les meilleurs résultats pour des modèles stationnaires [Guinness, 2018] sont également celles qui se comportent le mieux dans le cas non stationnaire.

Dans le but d'imposer une cohérence spatiale ou spatio-temporelle à un champ de paramètres d'une covariance non stationnaire, nous utilisons un processus log-gaussien (log-GP) comme Heinonen et al. [2016]. Le champ latent $\theta(\mathcal{S})$ est analysé comme :

$$log(\theta(s)) = w_\theta(s) + X_\theta(s)\beta_\theta^T \quad \forall\ s \in \mathcal{S} \quad and \quad w_\theta(\mathcal{S}) \sim \mathcal{N}(0, \theta_\theta). \tag{2}$$

Le processus gaussien $w_\theta(\cdot)$ permet de modéliser des variations ayant une cohérence spatiale. Les coefficients de régression linéaires β_θ paramétrisent des effets fixes (entre autres un intercept), et θ_θ est un jeu de paramètres de covariance qui paramétrisent le prior log-GP.
La parmétrisation logarithmique est facile à interpréter, ce qui rend l'utilisation du modèle plus intuitive. Premièrement, les paramètres de covariance telles qu'une portée ou une variance sont des nombres positifs alors qu'un processus gaussien peut prendre toutes les valeurs. Prendre les logarithmes de ces paramètres garantit la validité du prior. Dans le prolongement de cet argument, les paramètres de covariance sont des tailles : une variance est la taille d'une distribution, une portée est la largeur d'un kernel.
Cela ne règle pas le problème des covariances dont les paramètres de portée elliptique sont des matrices définies positives. Nous introduisons un prior original utilisant le logarithme matriciel. Rappelons que le logarithme d'une matrice définie positive est obtenu en passant les valeurs propres au logarithme, et qu'il est bijectif entre les matrices définies positives et les matrices symétriques. Un processus Gaussien multi-varié est alors utilisé comme prior pour les coordonnées des log-matrices dans l'espace vectoriel des matrices symétriques, donnant une extension cohérente au prior log-Gaussien.

Une grande partie du travail de cet article a été de trouver des stratégies MCMC adaptées à de grands jeux de données spatiales. Nous utilisons un algorithme de Langevin ajusté par un pas de Metropolis inspiré du Monte-Carlo hybride de Heinonen et al. [2016]. Dans le cas des variances du bruit et du processus latent, ce pas est simple à implémenter grâce aux formules de factorisation que nous avons obtenues pour les NNGP non stationnaires.
Pour les paramètres de portée, la méthode a besoin du gradient du facteur de Cholesky de la matrice de précision du processus latent par rapport aux paramètres de portée. Ce facteur de Cholesky est défini ligne par ligne avec (1). Bien que la formule du gradient soit fastidieuse, elle peut être implémentée relativement efficacement et permet donc d'utiliser l'algorithme de Langevin pour échantillonner les paramètres de portée. Nous remarquons aussi que le gradient que nous avons obtenu pourrait servir à d'autres méthodes, telles que l'approche de maximum de vraisemblance développée par le package R Guinness [2018].
Nous utilisons les méthodes d'entremêlement de paramétrisation développées par Yu and Meng [2011] et appliquées aux NNGP par Coube and Liquet [2020] afin d'améliorer le comportement des chaînes MCMC.

Références

Sébastien Coube and Benoît Liquet. Improving performances of mcmc for nearest neighbor gaussian process models with full data augmentation. *arXiv preprint arXiv :2010.00896*, 2020.

Abhirup Datta, Sudipto Banerjee, Andrew O. Finley, and Alan E. Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514) :800–812, 2016.

Guinness. Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics*, 60(4) :415–429, 2018. doi : 10.1080/00401706.2018.1437476. URL `https://doi.org/10.1080/00401706.2018.1437476`.

Katzfuss Guinness. *GpGp : Fast Gaussian Process Computation Using Vecchia's Approximation*, 2018. URL `https://CRAN.R-project.org/package=GpGp`.

Markus Heinonen, Henrik Mannerström, Juho Rousu, Samuel Kaski, and Harri Lähdesmäki. Non-stationary gaussian process regression with hamiltonian monte carlo. In *Artificial Intelligence and Statistics*, pages 732–740, 2016.

Christopher Joseph Paciorek. *Nonstationary Gaussian processes for regression and spatial modelling*. PhD thesis, Citeseer, 2003.

Yaming Yu and Xiao-Li Meng. To center or not to center : That is not the question—an ancillarity–sufficiency interweaving strategy (asis) for boosting mcmc efficiency. *Journal of Computational and Graphical Statistics*, 20(3) :531–570, 2011.