MCMC algorithms and hierarchical architectures for spatial modeling using Nearest Neighbor Gaussian Processes

Sébastien Coube-Sisqueille

UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR
ÉCOLE DOCTORALE 211

Laboratoire de Mathématiques et de leurs Applications de Pau (LMAP)

MCMC algorithms and hierarchical architectures for spatial modeling using Nearest Neighbor Gaussian Processes

---

Sébastien Coube-Sisqueille

A thesis submitted for the degree of
Doctor of Philosophy in Mathematics

Thesis defense planned for the 19th October 2021 in front of the thesis committee composed by:

Benoit Liquet	Professor	UPPA	Supervisor
Christian Robert	Professor	CEREMADE	Reviewer
Denis Allard	Director of research	INRAE	Reviewer
Kerrie Mengersen	Professor	QUT	Examiner
Denys Pommeret	Professor	I2M	Examiner
Edith Gabriel	Director of research	INRAE	Examiner
Clara Grazian	Senior Lecturer	UNSW	Examiner

Acknowledgments

Apparently, the Athenians had an altar dedicated to the Unknown God. Before I start to give names, I would like to raise an altar to the Unknown Helper: being a terribly distracted person, I certainly forgot people who helped me, and I ask for their forgiveness in advance.

Benoît Liquet and Noëlle Bru, my PhD advisors, deserve my deepest thanks. Noëlle gave me good advice and careful proofreading of the manuscript. The person I worked with the most was Benoît. It is difficult to summarize three years of working relationship, but if I had to pick one word it would be "trust". Benoît encouraged me to investigate the subjects I felt attracted to and left me much freedom to organize myself. Nonetheless, he put me in contact with the right people and he knew how to help me in the critical moments. I was lucky to have him as an advisor.

Benoît also made it possible to finance my contract with a research grant. That is why I think that the E2S (Energy and Environment Solutions) program of the UPPA must be cited here. While we are at it, the administration of the UPPA deserves a word. Bruno Demoisy is the person who took care of conferences, training, and dinners at the restaurant after the conferences and the training. Richard Lacroix helped me with those administrative messes of mine.

I sincerely thank the two reviewers, Xian Robert and Denis Allard, for their fair and constructive comments, and the other members of the jury, Edith Gabriel who presided the defense, Kerrie Mengersen, Denys Pommeret, and Clara Grazian, for their interest and many questions. I also would like to thank the members of my *comité de thèse*, Denys Pommeret again, Nathalie Cail-Milly, and Etienne Prevost, for their good advice.

Sudipto Banerjee must definitely feature here for our fruitful and exciting research collaboration. His insight, knowledge, and kindness, have been truly stimulating for me.

Just a word there for the guys and girls of the lab: Camilo, Fania, Claire, Aurélien, Iosu, Florèn, Bastien, Teo. Special mention for Claire who helped me with the organization of the defense.

I also would like to cite benevolent proofreaders. Sarat Moka, and his direct and salutary comments. Miyer, Cesar, y José, also known as *los tres mosqueteros*, and the rest of the Cristancho-Fajardo cartel.

I would like to mention the cool teachers I met in twenty years and who gave me a Stockholm syndrome so strong I did a PhD. Chronologically, and not exhaustively: Philippe Aurisset, Gilles Canguilhem, Pascal Mano, Richard Galliano, Philippe Mangeot, Pierre Alquier, Nicolas Chopin, and Xian Robert.

Eventually, I have special and loving thoughts for people who are dear to me. Lina, *mi amor*, who endured my perpetual caffeine, stress and lockdown-fueled bad temper while working on her own thesis. My grandmother, thanks to whom I was elegant as I could be for the defense. My brother Raphaël, for being my bro. My father, from whom I inherited my good character, for his unfailing support. My mother, who gave me her *there-must-be-a-better-way-to-*

do-this turn of mind, for her motherly love, the plants, the mushrooms and the aquaponics.

Abstract

Over the past five years, Nearest Neighbor Gaussian Processes (NNGP) arose as a computationally scalable method for spatial statistical models, but remain hampered by problems caused by the behavior of Markov Chain Monte-Carlo (MCMC) algorithms. Several approaches allow to alleviate those issues but they restrict the flexibility of the original model.

This work keeps the "jack of all trades" basic model and tackles its MCMC weak points with several strategies. The robustness and efficiency of high-level parameters estimation is boosted using interweaving strategies. Lower-level operations are parallelized using Chromatic Sampling. Efficient Hamiltonian methods are developed for NNGP models.

In a second time, the versatility of the NNGP model is used in order to tackle nonstationary modeling. An original parametrization and model architecture are proposed in order to ease model interpretation and selection while capturing complex nonstationarity patterns. An innovative MCMC strategy based on Hamiltonian methods and Nested Interweaving is proposed.

Chromatic Sampling; Interweaving; Markov Chain Monte Carlo; Nearest Neighbor Gaussian Process; Nonstationary spatial modeling.

Résumé

Au cours des cinq dernières années, les processus gaussiens des plus proches voisins (NNGP) sont apparus comme une méthode pour adapter les modèles statistiques spatiaux aux données de grande taille, mais ils restent entravés par des problèmes computationels causés par le comportement des algorithmes de Monte-Carlo par chaîne de Markov (MCMC). Plusieurs approches permettent d'atténuer ces problèmes mais elles limitent la flexibilité du modèle original.

Ce travail conserve le modèle de base et son côté "couteau suisse" tout en s'attaquant à ses points faibles MCMC avec plusieurs stratégies. La robustesse et l'efficacité de l'estimation des paramètres de haut niveau sont renforcées par des stratégies d'entrelacement. Les opérations de bas niveau sont parallélisées à l'aide de l'échantillonnage chromatique. Des méthodes Hamiltoniennes efficaces sont développées pour les modèles NNGP.

Dans un deuxième temps, la polyvalence du modèle NNGP est utilisée pour aborder la modélisation non stationnaire. Une paramétrisation et une architecture de modèle originales sont proposées afin de faciliter l'interprétation et la sélection des modèles tout en capturant des structures de non-stationnarité complexes. Une stratégie MCMC innovante basée sur les méthodes hamiltoniennes et l'entrelacement imbriqué est proposée.

Échantillonnage Chromatique; Entrelacement; Modélisation Spatiale Non Stationnaire; Monte-Carlo par chaîne de Markov; Processus Gaussiens des Plus Proches Voisins.

Contents

Chapter 1

Introduction

This introduction starts with a wide shot of the thesis, its hypotheses, and its contributions, presented in section 1.1. After that, the methods that are used in the dissertation are presented in detail in order to explain the approach of the thesis. Section 1.2 introduces spatial modeling. Section 1.3 concerns Vecchia's approximations and Nearest Neighbor Gaussian Processes (NNGP). The matter of MCMC implementation of spatial models is exposed in section 1.4, with an emphasis on NNGP models. The introduction ends with a detailed structure of the developments of the thesis, in section 1.5.

1.1 Overview of the decisions and contributions made in the thesis

The purpose of this work is to find ways to improve the computational behavior of NNGP spatial models and to use the resulting upgrades to propose new applications for NNGPs. The next paragraphs outline the choices and the method of the thesis; in order to get "the big picture", many points about spatial modeling and Vecchia's approximation are overlooked, but they should respectively be clarified in sections 1.2 and 1.3.

Context of the thesis. This thesis builds on the current momentum of Vecchia's approximations and more specifically their NNGP variant in the domain of spatial statistics. Even though Vecchia's approximations were introduced more than thirty years ago (Vecchia, 1988) and studied ever since (Stein et al., 2004), there has been a bloom of publications on the subject in the past five years (Datta et al., 2016; Katzfuss and Guinness, 2017; Guinness, 2018; Finley et al., 2019; Taylor-Rodriguez et al., 2019; Katzfuss et al., 2020; Peruzzi et al., 2020; Katzfuss et al., 2020; Zilber and Katzfuss, 2021). Those developments include software, such as Guinness and Katzfuss (2018), Katzfuss et al. (2020) and Finley et al. (2017).

Vecchia's approximations are a family of sequential approximations of Gaussian densities. They are mostly used in spatial statistics (Datta et al., 2016; Guinness, 2018; Katzfuss et al., 2020), even though they also have applications in other cases where Gaussian Processes are relevant, such as the exploration of a parameter space in computer experiments (Katzfuss et al., 2020). Assume that a Gaussian process (GP) $x(\cdot)$ indexed on a domain $\mathcal{D}$, typically $\mathbb{R}^2$ or the sphere, is observed at a finite collection of locations $\mathcal{S} \subset \mathcal{D}$ with cardinal n. Assign an order to $\mathcal{S}$, and write the ordered set as $(s_1, \ldots, s_n)$. Vecchia's approximation approaches the joint Gaussian density of the process $f(\cdot)$ using a pruned conditional recursive factorization:

$$\begin{aligned} f(x(\mathcal{S})) &= f(x(s_1))\Pi_{i=2}^{n} f\left(x(s_i) \mid x(s_1, \ldots, s_{i-1})\right) \\ &\approx f(x(s_1))\Pi_{i=2}^{n} f\left(x(s_i) \mid x(pa(s_i))\right), \end{aligned} \tag{1.1}$$

where $pa(s_i) \subseteq (s_1, \ldots, s_{i-1})$. The idea is that if $pa(s_i)$ is much smaller than $(s_1, \ldots, s_{i-1})$ but chosen carefully, the conditional density will be cheap to compute but still accurate. A good joint approximation should then arise from the aggregation of conditional approximations. Given the fact that the locations are ordered, $pa(\cdot)$ denotes the parents of a node in a Directed Acyclic Graph (DAG, a graph where the connections are oriented and where there is no cycle) that is used to define Vecchia's approximation. In general, the parents of s_i are chosen as its nearest spatial neighbors among all its predecessors $(s_1, \ldots, s_{i-1})$, even if more subtle tactics exist (Guinness, 2018). A careful reader may object that defining an approximate density through recursive conditional factorization does not require the density to be Gaussian (Lauritzen, 1996); however, I have always found explicitly written that Vecchia's approximation targets Gaussian densities, except in Stein (2012), who does not specify the exact nature of the process. In Gaussian settings, the accuracy of Vecchia's approximations depends on the ordering and parent-picking heuristics but is highly satisfactory (Guinness, 2018), and the method performs well in benchmarks against other state-of-the-art methods (Heaton et al., 2019).

In space-time models, Gaussian data observed at a spatial site $s \in \mathcal{S}$ is often analyzed as:

$$z(s) = X(s)\beta^T + w(s) + \epsilon(s),$$

where $z(\cdot)$ is the observed variable, $X(\cdot)$ are covariates, β is a vector of regression coefficients, $\epsilon(\cdot)$ is a Gaussian white noise, and $w(\cdot)$ is a Gaussian Process latent field that intends to capture a spatially coherent error. The problem is that multivariate Gaussian density is not affordable in high dimension, and is approached using Vecchia's approximation.

NNGPs are a special case of Vecchia's approximation applied to the density of the latent field $w(\cdot)$, that is in 1.1 "$x(\cdot)$" is replaced by "$w(\cdot)$". The density that is factorized and approximated is therefore the GP prior. The general principle of pruned recursive conditional factorization can be used on the Gaussian response $z(\cdot)$; in this case, a positive term corresponding to the nugget $\epsilon(\cdot)$ is added to the diagonal of the GP prior's covariance matrix. It also is possible to

collect both $w(\cdot)$ and $z(\cdot)$ in one vector and apply Vecchia's approximation to this joint vector; the approximation will change following how $w(\cdot)$ and $z(\cdot)$ are ordered within that vector (Katzfuss and Guinness, 2017).

Key choices of the thesis. A first choice of the thesis is to focus on NNGPs rather than on other Vecchia's approximations. While Vecchia's approximations that mix the Gaussian response and the Gaussian latent field may be very efficient, they require the observations to be Gaussian. On the other hand, NNGPs focus on the latent field $w(\cdot)$ and leave the Gaussian error $\epsilon(\cdot)$ out, which allows them to handle other types of data (binomial, Poisson, etc...) or to be used elsewhere than in the decomposition of the interest variable, to enforce some spatial coherence on a field of parameters for example. The PhD adopts the NNGP approach, and is therefore not focused on the construction of Vecchia's approximations, even though a few original results concerning KL divergence with respect to full GP are presented in the thesis.

NNGPs are used within the framework of Bayesian hierarchical models (Datta et al., 2016) and were initially fitted using MCMC. However, this approach is not immune from the usual problems of auto-correlation and poor mixing of the MCMC chains. Alternative Monte-Carlo algorithms developments (Finley et al., 2019) aim to mitigate those problems, but lose flexibility with respect to the type of data, and are sensitive to other factors such as the dimension of the geographic space (see Rue and Held, 2005, for more details). Classical yet powerful MCMC architectures (Knorr-Held and Rue, 2002) suffer from the same problems. Another solution is to use conjugate algorithms (Finley et al., 2019; Zhang et al., 2019); however, those methods are restricted to Gaussian data and to certain priors for the covariance parameters, limiting the flexibility of the model.

Here comes the second critical choice of the thesis, that stems from the first choice. Since the approach is to retain the adaptability of NNGPs, possibly at the expense of performance, the thesis focuses on improving the original MCMC algorithm of Datta et al. (2016), while retaining its flexibility.

Contributions of the thesis. The contributions of the thesis are synthesized here, but a more detailed presentation can be found in section 1.5. Much of the effort of the PhD work is focused on finding improvements for the "vanilla" Gibbs sampler. The markovian nature of NNGPs allows to transpose the parallelisability of the prior density to the sampling of the latent field $w(\cdot)$, and to make this step easier to implement using high-level languages. Chromatic samplers (Gonzalez et al., 2011) arise as a practical solution permitted by the sparsity of NNGPs. Extensive sensitivity analysis is carried out, with the conclusion that the method is "all-terrain" and behaves well with state-of-the-art NNGP settings.

Attractive but expensive algorithms that involve prior whitening of the latent field become affordable with NNGPs. Whitening consists in multiplying the

Gaussian latent field by the inverse Cholesky factor of its prior covariance matrix. If the latent field followed its prior distribution, the whitened field would be a collection of independent and identically distributed normal variables, explaining the name of the method. Applying this transformation to variables that follow the posterior distribution results in a sharp decrease of the correlation. A NNGP prediction algorithm that relies on whitening, and improved Hamiltonian sampling of the latent field are developed.

A fruitful approach is to use the interweaving methods of Yu and Meng (2011). Those methods are particularly cheap with NNGPs thanks to the matrix sparsity induced by the method. Interweaving is applied in two parts of the MCMC architecture. One basic yet very helpful application is to improve dramatically the robustness of the MCMC behavior of parameters associated to the linear regression component of the model. The second application is to couple natural and whitened parametrizations of the latent field to improve the sampling of covariance parameters (Filippone et al., 2013).

The lead of Delayed Acceptance of Christen and Fox (2005) is also explored. The thesis underlines that this method was worth a serious try because of the ease to split NNGP density and because it is compatible with the developments that use interweaving. However, after preliminary analysis, the method gave disappointing results and was not retained. The algorithmic details and the experimental results are nonetheless presented.

The end of the thesis uses the improvements brought to the stationary model to propose a nonstationary hierarchical NNGP model. This work is done in collaboration with Benoît Liquet and Sudipto Banerjee, and is presented in an article that is included in the thesis but has not been submitted yet. The simplicity and operability of the algorithms that were proposed in the first part allow to use them on complex structures, while more elaborate algorithms would be very hard to use. The whole toolbox that was devised for stationary models is brought into service: both interweaving schemes, whitened Hybrid Monte-Carlo, chromatic sampling play a role. The starting problem is that nonstationary spatial modeling is exciting and potentially rewarding, but suffers from several problems: its computational cost, the complexity and lack of interpretability of multi-layered hierarchical models, and the difficulty of model selection. The model presented in the article attempts to tackle those issues.

The model considers three extensions to a stationary model, where several parameters may vary in space: the latent process' marginal variance, the latent process' spatial range (potentially anisotropic), and the noise's variance when the observations are Gaussian. The nonstationary correlation is taken from the classical framework from Paciorek (2003). In practice, changing the marginal variance of the Gaussian process changes the amplitude of its realizations, allowing for bigger or smaller effects when needed. The range of the Gaussian process has two components. When the range is scalar, the spatial coherence of the process can vary, leading the coherence of $w(\cdot)$ to change with the region. In addition to that, an elliptic range allows to parametrize locally anisotropic fields. A spatially varying variance of the noise results in a nugget effect that

will be stronger or weaker following the location.

The first part of this work on nonstationary NNGPs is to derive analytical and empirical properties of NNGPs with nonstationary covariance. The article presents a few factorization formulas that allow efficient implementation, in particular with high-level languages such as R. Various construction heuristics of NNGPs were tested, and the results confirmed that the state-of-the-art heuristics of Guinness (2018) for stationary covariance structure also are the best in nonstationary settings.

The second part of this work is to find a flexible yet interpretable parametrization of the spatially variable parameters. In order to enforce a spatial or space-time coherence on a parameter field, the classical solution of log-Gaussian processes is used. The latent field is analyzed as:

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

Here $\theta(\cdot)$ is a parameter that can vary in space; it can be the range, the marginal variance of the latent GP $w(\cdot)$, or the variance of the noise $\epsilon(\cdot)$. The latent Gaussian process $w_\theta(\cdot)$ captures spatially coherent variations, linear regression coefficients β_θ parametrize the linear effects (including an intercept), and ζ_θ are covariance parameters for the log-Gaussian process prior. This parametrization is popular in the literature (Heinonen et al., 2016), however the thesis gives further justifications that might help to popularize the method. The second contribution as for parametrization is to extend the model to elliptical range parameters. Analogously to scalar logarithm, matrix logarithm maps positive-definite matrices into symmetric matrices by passing their eigenvalues to the logarithm. A multivariate Gaussian Process prior is then applied on the coordinates of the log-matrix in the basis of the vector space of symmetric matrices. A model with stationary range corresponds to a model with constant nonstationary scalar range. In turn, a model with elliptic range whose ellipses are circular corresponds to a model with nonstationary scalar range. Integrating the various range models within an expanding, interpretable family allows for better understanding of the parameters and easier model selection. On synthetic data sets, complex models applied to stationary data boil down to a degenerate model that induces a stationary distribution.

A third aspect is to find efficient MCMC algorithms. An important point is to sample the nonstationary parameters. After exploring the possibility of Chromatic Sampling, it appears that while this solution is conceptually possible, its usefulness is limited. A Hybrid Monte-Carlo step inspired from Heinonen et al. (2016) is used instead. This method is easy to implement for variance parameters of the latent field and the noise thanks to the properties of nonstationary NNGPs mentioned earlier. As for the range parameters, the method requires the derivative of the NNGP-induced Cholesky factor of the precision matrix with respect to each range parameter. Even though the formula is tedious to derive, it can be implemented at a reasonable cost. This sampling method is embedded within a nested interweaving algorithm, a strategy envisioned by Yu and Meng (2011) but never put in application for realistic models as far as I

know.

1.2 Space time hierarchical models for point-measurement observations

In natural and social sciences it is usual to observe some phenomena on a spatial domain and possibly at several times. Many of them, for example mineral deposits, pollution, or biomass density, can be modeled as fields with some kind of space-time coherence. A point-measurement data set is a collection of measurements of the response variable associated with their precise space-time coordinates in a domain $\mathcal{D}$ which can be $\mathbb{R}^2$, $\mathbb{R}^3$, the sphere, the Cartesian product of the sphere and $\mathbb{R}$, *et caetera*. It gives an incomplete, possibly noisy, image of the field because the number of measurements is limited while the field could be observed on infinitely many points of the space. Point-measurements are not the only possible format for such data. Sometimes only areal measurements are available, for example when the observations are aggregated on administrative regions. The development of geographical information systems allows for larger, richer point-measurement space-time data sets, impulsing research for scalable models.

Hierarchical model architecture. Let's start with the general architecture of the hierarchical model, before zooming on each of its layers. Statistical modeling of a space or space-time phenomenon $z(\cdot)$ that is observed with point-referenced measurements can be done by introducing a spatially-indexed process $w(\cdot)$ on the interest space-time domain, such that for any finite subset $\mathcal{S}$ of the spatial domain $\mathcal{D}$, there is a well-defined joint distribution of the vector $w(\mathcal{S})$. When we deal with one single spatial location, it is noted $s \in \mathcal{S}$. Many models also add linear regression on covariates $X(\cdot)$. When the observed data $z(\cdot)$ is continuous, it can be analyzed by introducing a Gaussian error $\epsilon(\cdot)$, giving a hierarchical model with Gaussian data presented in figure 1.1. There are three components: the linear regression, the spatial process, and the noise. When the data is not continuous, for example binomial or integer, the generalized spatial model of figure 1.2 is used. This model has only two components, the linear regression and the spatial effect.

First layer of the model: analysis of the interest data. The first layer of the model separates the observed variable into several components. Recall the Gaussian spatial model formula, which is used when the response is continuous:

$$z(s) = X(s)\beta^T + w(s) + \epsilon(s).$$

This space-time model differs from a simple Gaussian linear model only because of the presence of $w(\cdot)$. The analogy with linear models goes further since the response variable $z(\cdot)$ can have arbitrary distributions, and link functions

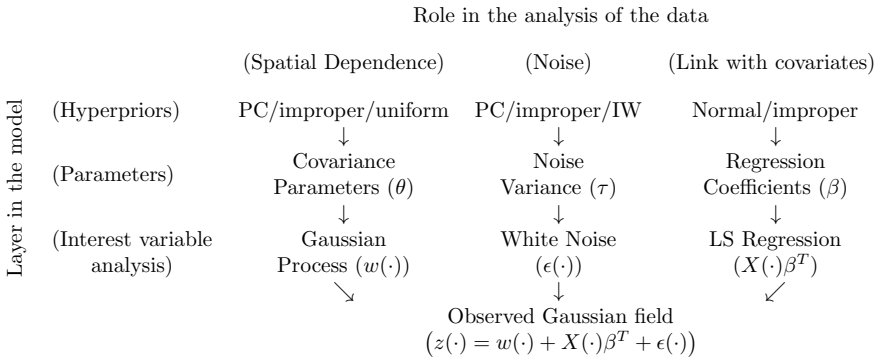


Figure 1.1: Schema of a hierarchical Gaussian space-time GP model

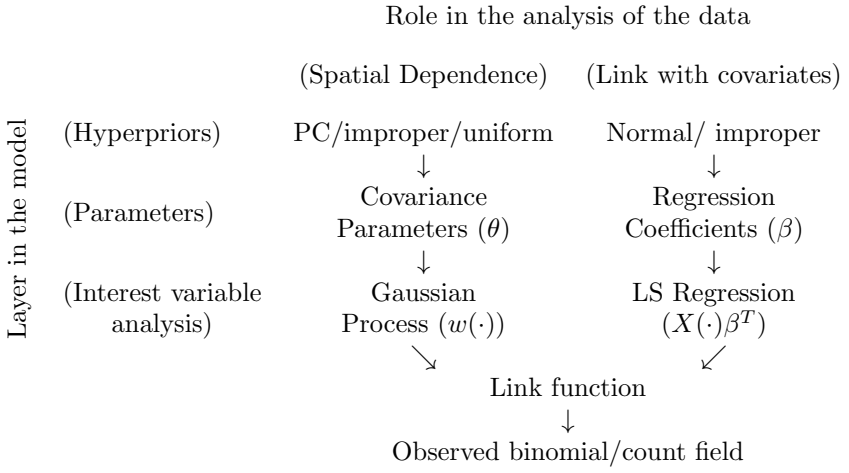


Figure 1.2: Schema of a hierarchical generalized space-time GP model

(logistic, normal cumulative density function, exponential) can make space-time models suitable for various types of data models. Take $h(\cdot)$ a link function such as the normal cumulative distribution function or the logistic function, we can formulate a space-time model for binomial observations:

$$z(s) \sim Bern\left(h\left(X(s)\beta^T + w(s)\right)\right).$$

We also can formulate a spatial count data model as

$$z(s) \sim Poisson\left(exp\left(X(s)\beta^T + w(s)\right)\right).$$

In order to keep general notations, the log-likelihood of the observed field with respect to the latent field and the regression coefficients is noted

$$l(z(\mathcal{S})|w(\mathcal{S}), \beta, \ldots).$$

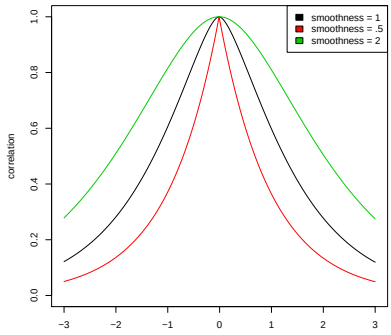
Some arguments are left un-precised because they can change with the data model: for example, in a Gaussian model, adding a parameter τ to precise the standard deviation of ϵ is necessary, while no further argument is needed for a Poisson or binomial model.

Second layer of the model: the Gaussian Process prior. Gaussian processes (GP) make an elegant prior distribution for $w(\cdot)$. The GP prior distribution of $w(\mathcal{S})$ is $\mathcal{N}(0, \Sigma(\mathcal{S}, \theta))$, where $\Sigma(\mathcal{S}, \theta)$ belongs to a family of parameter-indexed matrices $\{\Sigma(\mathcal{S}, \theta)\}$, the covariance parameters θ being unknown. The model reported here allows to infer the covariance parameters θ and to smooth the signal at the observed locations $\mathcal{S}$. Usually, the parameters of the covariance matrix are identified with the parameters of a covariance function. Using a covariance function that takes two spatial locations as arguments allows to transform the mutual dependence of all locations in a collection of pairwise links. Matérn's covariance is a popular choice. Stein (1999) gives an unambiguous guideline: "Use the Matérn model". One parametrization of the isotropic Matérn function is presented here. Take $\theta = (\sigma^2, \alpha, \nu)$, respectively the marginal variance, range and smoothness parameter, and s_1, s_2 are two points from $\mathcal{D}$. The Gamma function is noted $\Gamma(\cdot)$, and the modified Bessel function of the second kind is noted $\kappa(\cdot)$.

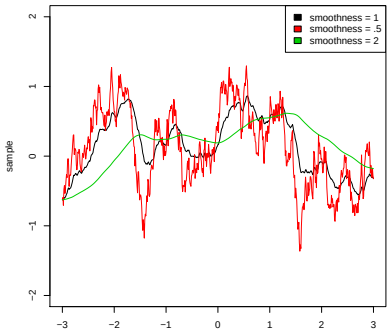
$$\mathcal{M}(s_1, s_2, \sigma^2, \alpha, \nu) = \sigma^2(2^{1-\nu}/\Gamma(\nu)) \times \left(\frac{\|s_1 - s_2\|_2}{\alpha}\right)^{\nu} \times \kappa_\nu\left(\frac{\|s_1 - s_2\|_2}{\alpha}\right).$$

I give an example of the effects of the covariance parameters in figure 1.3. Subfigures 1.3a, 1.3c and 1.3e show respectively how ν, α and σ^2 change the covariance function. Subfigures 1.3b, 1.3d, 1.3f give Gaussian process samples that correspond to the changes of covariance functions. The diversity of the samples shows that the Matérn covariance can capture a great variety of spatial behaviors.

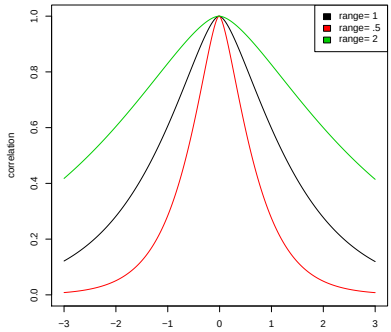
However, the use of full Matérn covariance is sometimes overkill with respect to the available data. In particular, it is sometimes hard to identify range and



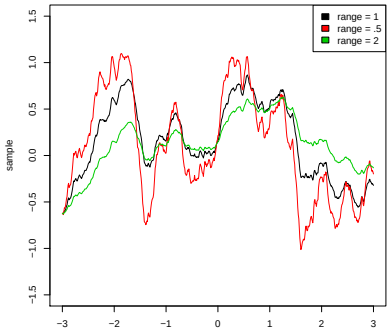
(a) Matérn covariance functions with varying smoothness



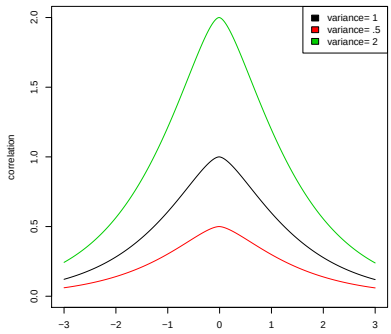
(b) Gaussian Process samples with varying smoothness



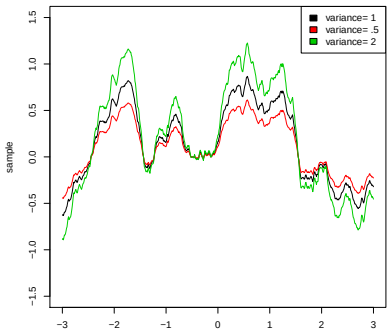
(c) Matérn covariance functions with varying range



(d) Gaussian Process samples with varying range



(e) Matérn covariance functions with varying marginal variance



(f) Gaussian Process samples with varying marginal variance

Figure 1.3: The covariance parameters of a Matérn model change the covariance function and the process samples

smoothness and in practice smoothness is often fixed. Some values of ν actually correspond to positive-definite functions with a simpler form. For $\nu = 1/2$, the covariance is exponential

$$K_{exp}(s_1, s_2, \sigma^2, \alpha) = \sigma^2 exp\left(-\frac{\|s_1 - s_2\|_2}{\alpha}\right).$$

For $\nu \to +\infty$ it is squared-exponential

$$K_{sqexp}(s_1, s_2, \sigma^2, \alpha) = \sigma^2 exp\left(-\frac{\|s_1 - s_2\|_2^2}{\alpha}\right).$$

All those functions depend only on the Euclidean distance between the two spatial locations, they are isotropic following the typology of Rasmussen and Williams (2006). Those functions also are stationary, which means that they depend only on $s_1 - s_2$ (Rasmussen and Williams, 2006). Anisotropic covariance functions can be obtained by replacing the Euclidean distance by Mahalanobis' distance. A function can be stationary but not isotropic, like those who depend on Mahalanobis' distance, but an isotropic function is always stationary since the Euclidean distance is a function of $s_1 - s_2$.

Third layer of the model: the hyperpriors. The high-level parameters are themselves subject to modeler-specified hyperpriors. Articles that present MCMC implementation of such a hierarchical model (Datta et al., 2016; Banerjee et al., 2008) prefer to use conjugate priors, such as a normal prior for the regression coefficients, a Gamma prior for the marginal variance, an Inverse Wishart distribution for noise variance. Range and smoothness have no full conditional distribution that can be easily sampled from, let alone a conjugate prior, so a constant prior on an interval is used in those articles. An interesting hyperprior is the PC (penalize complexity) distribution (Fuglstad et al., 2015b). This distribution aims to avoid identification problems between range and marginal variance when the spatial domain is too small (Zhang, 2004). From my experience, even improper constant priors on the hyper-parameters give satisfactory model behavior even though there can be a ridge-like joint distribution on the range and variance parameters.

1.3 Vecchia's approximation and Nearest Neighbor Gaussian Processes

The weakness of Gaussian processes is that they are not scalable (Datta et al., 2016; Banerjee et al., 2008). Computing the GP prior density of $w(\mathcal{S})$ involves the determinant and inverse of $\Sigma(\mathcal{S}, \theta)$, incuring a computational cost that is cubic in the size of $\mathcal{S}$. A plebiscited method, the Integrated Nested Laplace Approximation (INLA) (Lindgren et al., 2011), uses a Stochastic Partial Differential Equation representation of Gaussian Processes. Another method, that

received increasing interest the past years, is Vecchia's approximation, introduced by Vecchia (1988). One important special case is the Nearest Neighbor Gaussian Process (NNGP) (Datta et al., 2016).

1.3.1 General principle of the Nearest Neighbor Gaussian Process

Defining a NNGP starts by finding an ordering for the n locations of $\mathcal{S}$ which are noted $(s_1, \ldots, s_n)$. They can be computed with any ordering, and Datta et al. (2016) argue that it has little impact and that sorting $\mathcal{S}$ along a coordinate of $\mathcal{D}$ is a good choice. However, Guinness (2018) shows that in some cases even random ordering can be a better choice than coordinate ordering and introduces methods that show good behaviors in terms of Kullback-Leibler contrast with respect to the full GP prior (see 1.3.3 for more details). The joint latent density of $w(s_1, \ldots, s_n)$ is then written as the product of conditional densities. One realization of the process is conditioned by all of the previous realizations. Here, the non-approximated conditional form of the joint prior GP density is given, with a covariance matrix parametrized by θ, and an omitted mean equal to 0.

$$f(w(s_1, \ldots, s_n)|\theta) = f(w(s_1)|\theta) \times \Pi_{i=2}^{n} f(w(s_i)|w(s_1, \ldots, s_{i-1}), \theta).$$

Since $f(w(s_1, \ldots, s_n)|\theta))$ is a Multi-Variate Normal (MVN) distribution function, the conditional density $f(w(s_i)|w(s_1, \ldots, s_{i-1}), \theta), i \in 2, \ldots, n$ is a Normal distribution function whose mean and variance parameters are determined by

$$w(s_1, \ldots, s_{i-1}) \quad \text{and} \quad \Sigma((s_1, \ldots, s_i), \theta).$$

The approximation consists, for each conditional density, in replacing the vector $w(s_1, \ldots, s_{i-1})$ that conditions $w(s_i)$ by a much smaller subset. Denote $pa(s_i)$ the subset of the vector of locations $(s_1, \ldots, s_{i-1})$ so that, in the NNGP, $w(s_i)$ conditions on $w(pa(s_i))$: they are the parents of s_i in the Directed Acyclic Graph (DAG) that defines the NNGP (see Datta et al. (2016), Katzfuss and Guinness (2017)). A DAG is a graph where the connections are oriented, and where there is no cycle.

Usually, the size of $pa(s_i)$ is fixed $m = 5, 10$, or 15, except for the m first locations $s_1, \ldots, s_m$ that cannot have m parents. The NNGP approximation to the GP prior joint density of $w(\cdot)$ is defined as

$$\tilde{f}(w(s_1, \ldots, s_n)|\theta) = f(w(s_1)|\theta) \times \Pi_{i=2}^{n} f(w(s_i)|w(pa(s_i)), \theta) \tag{1.2}$$

This very general principle can be applied to any kind of well-defined multivariate density. Defining a joint density through factorization on a DAG always yields a valid distribution thanks to clique factorization (Lauritzen, 1996). The special case of MVN distribution has very interesting properties that are explained in 1.3.2.

I focused here on the NNGP case which approximates the latent Gaussian Process density. However, the principles of pruned recursive conditional factorization can be applied to other densities. For example, the observed variable

$z(\cdot)$ is Gaussian, the latent field $w(\cdot)$ can be easily replaced by $z(\cdot)$ in (1.2). The family of Vecchia's approximations comprises yet other cases (see 1.3.3 for more details).

1.3.2 The good computational properties of the Nearest Neighbor Gaussian Process

The NNGP defines a MVN density and allows to compute explicitly and easily a sparse right lower-triangular Cholesky factor of the precision matrix. I call this factor $\tilde{R}$. In order to show this point, let's focus again on the recursive conditional formula of Vecchia's approximation. In order to keep notations shorter, the covariance parameters are omitted: $\Sigma(\mathcal{S}, \theta)$ is noted $\Sigma(\mathcal{S})$. The notation $\Sigma(x, y)$ is used to designate the (rectangular) sub-matrix of $\Sigma(\mathcal{S})$ that corresponds to the two collections of locations $x, y \subset S$, and the square sub-matrix is abbreviated as $\Sigma(x, x) = \Sigma(x)$. In the purpose of showing that the NNGP is an unbiased approximation, the mean of the full GP is noted $\mu(\cdot)$, so that $w(s)$ has mean $\mu(s)$, even though usually $\mu = 0$ because the mean of the explained variable is explained by the linear effects $X\beta^T$. Since the full GP distribution of $w(\mathcal{S})$ is a Gaussian vector, the conditional distribution $f(w(s_i)|w(pa(s_i)), \mu(\mathcal{S}), \theta)$ will be a Normal density. Using the standard results on conditional distributions, its conditional mean parameter will be

$$\bar{\mu}_i = \mu(s_i) + \Sigma(s_i, pa(s_i))\Sigma(pa(s_i))^{-1}(w(pa(s_i)) - \mu(pa(s_i)))$$

and its conditional variance parameter will be

$$\bar{\sigma}_i^2 = \Sigma(s_i) - \Sigma(s_i, pa(s_i))\Sigma(pa(s_i))^{-1}\Sigma(pa(s_i), s_i).$$

The NNGP factor to the precision matrix $\tilde{R}$ is then constructed row by row:

- $\tilde{R}_{ii}$ receives the value $1/\bar{\sigma}_i$.
- the elements of the i^{th} row whose column indices correspond to $pa(s_i)$ receive the value $\Sigma(s_i, pa(s_i))\Sigma(pa(s_i))^{-1}/\bar{\sigma}_i$.
- the rest remains 0.

Now consider the MVN density with covariance matrix $(\tilde{R}^T\tilde{R})^{-1}$ and mean $\mu(\mathcal{S})$. Its density is

$$(2\pi)^{-n/2}(|(\tilde{R}^T\tilde{R})^{-1}|^{-1/2})exp(-(w(\mathcal{S}) - \mu(\mathcal{S}))^T\tilde{R}^T\tilde{R}(w(\mathcal{S}) - \mu(\mathcal{S}))/2).$$

Using the fact that $\tilde{R}$ is triangular, the determinant can be written

$$|(\tilde{R}^T\tilde{R})^{-1}|^{-1/2} = |\tilde{R}| = \Pi_{i=1}^n 1/\bar{\sigma}_i.$$

Using the row-by-row construction of $\tilde{R}$, the product involving $\tilde{R}$ can be written

$$-w(\mathcal{S})^T\tilde{R}^T\tilde{R}\ w(\mathcal{S})/2 = -\Sigma_{i=1}^n(w(s_i) - \bar{\mu}_i)/2\bar{\sigma}_i^2.$$

This MVN density can then be re-written as

$$\Pi_{i=1}^{n} 1/(\sqrt{2\pi}\bar{\sigma}_i) exp(-(w(s_i) - \bar{\mu}_i)^2/(2\bar{\sigma}_i^2)) \tag{1.3}$$

and therefore be identified with $\tilde{f}(w(\mathcal{S}))$.

From this construction, it clearly appears that $\tilde{R}$ is triangular and has only $m + 1$ non-null values per row at the most, making it very easy to store and manipulate using sparse matrix algebra libraries. It also appears that the NNGP leaves the mean invariant. When the Gaussian response variable $z(\cdot)$ is used instead of the latent variable $w(\cdot)$, the nugget effect $\epsilon(\cdot)$ has to be taken into account. The noise variance term τ^2 is then added to the diagonal of $\Sigma(\mathcal{S})$. The method of construction for $\tilde{R}$ is not changed, so its sparsity and triangularity are carried over; however the value of the coefficients is impacted.

1.3.3 Nota Bene of Vecchia's approximations and Nearest Neighbor Gaussian Processes

Here is a review of a few points that matter for Vecchia's approximations, with a special focus on NNGPs.

Cost. The construction of $\tilde{R}$ makes it easy to use in practice. The coefficients of a row of $\tilde{R}$ are found using Gaussian conditional expectation and variance formulas between $w(s_i)$ and $w(pa(s_i))$, for a cost that is $O(m^3)$. Then, finding $\tilde{R}$ costs $O(nm^3)$ flops, even though Guinness (2018) argues that the $O(nm^2)$ memory allocation costs more than the operations. Storing $\tilde{R}$ is also relatively cheap since a row has only m non-null coefficients, leading to a $O(nm)$ RAM requirement. Most of the time, the cost of Vecchia's approximation is linear in the size of the data and parallelizable (with the exception of some of the algorithms presented by Finley et al., 2019). The latent density $\Pi_{i=2}^{n} f(w(s_i)|w(pa(s_i)), \theta)$ can be split into $n - 1$ small jobs and dispatched to a cluster of calculators (Datta et al., 2016).

Ordering. Datta et al. (2016) argue that the ordering has little impact on the accuracy of NNGPs, and order the locations following one of their coordinates in the spatial domain. However, Guinness (2018) carries out extensive Kullbak-Leibler experiments and shows that the max-min heuristic (starting from a random site, then constructing the ordering recursively by picking the site that has the highest minimal distance with its predecessors) or even random ordering give much better performances on two dimensions. On three, four dimensions, the middle-out heuristic (ordering following the distance to the center of the cloud of observations) performs better. The max-min heuristic and random ordering have in common to quickly cover the spatial domain. Coordinate and middle-out orderings create an expanding blanket of vertices.

Parents. The choice of the conditioning sets is critical but no universal criterion exists. A popular choice is to choose $pa(s_i)$ to be s_i's nearest neighbors

among $(s_1, \ldots, s_{i-1})$, explaining the denomination "Nearest Neighbors Gaussian Process" given by Datta et al. (2016). This choice is motivated by a heuristic which argues that the nearest neighbors have the highest correlation with the considered site. Other schemes exist like mixing close and far-away observations (Datta et al., 2016; Stein et al., 2004) or multi-resolution approximation (Katzfuss and Guinness, 2017).

More advanced strategies exist such as a grouping strategy proposed by Guinness (2018), that allows to improve the Kullback-Leibler divergence between the full GP prior and Vecchia's approximation while making it cheaper to compute. This strategy is based on the "information never hurts" principle: adding more parents to an existing parent set can only lower the Kullback-Leibler divergence between Vecchia's approximation and the full GP (Guinness, 2018; Peruzzi et al., 2020).

However, I think that there are cases where the nearest neighbor heuristic is not powerful enough, or at least is very difficult to use. In the case of multivariate spatial data, the approach $m = 5, 10...$ neighbors of each p variables for good measure would lead to parent sets of size $m \times p$, and be unbearable computationally. I do not think that using only m nearest neighbors regardless of which variable is observed at those sites is a good idea. If the variables that at observed at the neighbors have little correlation with the variables that are observed at the considered point, the conditioning may be very bad. When the space where the observations are done grows in dimension, there may be anisotropy along one or several coordinates (in particular when one of the dimensions is the time) and there is no unequivocal definition of the nearest neighbors.

Shuffling the Gaussian response and the latent field. If $z(\cdot)$ is Gaussian, the random vectors obtained combining $z(\mathcal{S})$ and $w(\mathcal{S})$ are Gaussian vectors as well. The methodology that consists in pruning the recursive conditional form of the density can therefore be applied to any permutation of the joint vector $(w(\mathcal{S}), z(\mathcal{S}))$. Katzfuss and Guinness (2017) study those possibilities as the Sparse General Vecchia's approximation. Following the approach of Katzfuss and Guinness (2017), NNGPs are a special case of Vecchia's approximation where the latent field $w(\cdot)$ comes before the response variable $z(\cdot)$ in the ordering. Since the latent field comes first, its prior Gaussian density is approximated by (1.2). The Gaussian response variables are, in turn, conditioned only by the latent field at their location and Vecchia's approximation boils down to the likelihood of the Gaussian white noise. It also is possible to simply drop $w(\mathcal{S})$ and just work with $z(\mathcal{S})$ in a response NNGP (Finley et al., 2019) (note that the response NNGP is not a NNGP according to the typology of Katzfuss and Guinness (2017)).

1.3.4 Approach of the thesis with respect to Vecchia's approximations

Vocabulary. Later in the thesis, I will name "NNGP" Vecchia's approximation that is applied on the latent field's density, following the typology of Katzfuss and Guinness (2017). I am however fully satisfied with neither of those two names.

In "Nearest Neighbor Gaussian Process", I am uneasy with "Nearest Neighbor". On one hand, a latent density can be defined through recursive factorization on a DAG even if this DAG is not defined using the heuristic of nearest neighbors. Stein et al. (2004) show that a mix of nearest neighbors and farther observations improves the approximation. Conversely, general Vecchia's approximations (Katzfuss and Guinness, 2017) can be defined using the nearest neighbor heuristic, but will not qualify as NNGPs.

As for "Vecchia's approximations", I think that even though Gaussian Processes defined through recursive conditional factorization are indeed a great approximation for Gaussian processes that are defined through the covariance matrix they can be something else than approximations. For example, in my work on nonstationary NNGP, I use the covariance function of Paciorek (2003), that is defined only on the plane. Paciorek provides an extension to the sphere through truncated kernels but this function is fairly complex to implement. Instead, each conditional likelihood of the NNGP factorization is computed on the orthogonal projection of the points, from the sphere to a tangent plane. Using this formulation, a NNGP density is derived without actually defining a covariance function.

Method. I decided to focus my work on NNGPs with full data augmentation, that is implementations of the NNGP model that simulate explicitly $w(\mathcal{S})$. Because of this choice, I clearly walk in the steps of Datta et al. (2016). The first reason is that they can be applied to many types of observed data as long as a suitable link function is available. A second reason is that it is possible to build on the versatility of NNGPs with full data augmentation to tackle complex modeling. For example, latent Gaussian fields can be used for non-stationary Gaussian process modeling in one dimension and relatively few observations (Heinonen et al., 2016), and I wanted to use NNGPs to extend this approach to data sets with larger size and larger dimension.

1.4 MCMC problems in spatial NNGP models

The NNGP model initially presented by Datta et al. (2016) is fitted using a simple Gibbs sampler that loops over the latent field $w(\mathcal{S})$ (sequential update), the covariance parameters θ (Metropolis step), the regression coefficients β, and the noise variance τ (in the Gaussian case). However, this approach suffers from slow mixing (Finley et al., 2017). The "vanilla" NNGP model has a high dimension, since there are more parameters than observations. However, due to

the hierarchical nature of the model, all parameters are not on an equal footing. Some parameters such as the covariance parameters, the variance of the noise, or the regression coefficients, are high-level parameters that explain all of the observations. On the other hand, the variables from the latent field $w(\cdot)$ have a very local impact.

All the issues that I have been able to identify have in common to involve the latent field in one way or another. This implies that their answers will also have to fiddle with $w(\cdot)$. Some solutions treat the root of the problem by removing the troublesome components. One approach, the response NNGP of Finley et al. (2019) is to remove $w(\cdot)$ and just work with $z(\cdot)$. Another approach is to remove the MCMC and use a conjugate model (also in Finley et al., 2019). However, since the purpose of the thesis is to have a versatile tool thanks to plain sampling of $w(\cdot)$, the problems must be exposed, and hopefully addressed. The trouble caused by the latent field can be split into several components.

1.4.1 Spatial correlation

The first problem is that $w(\mathcal{S})$ is auto-correlated because of the NNGP prior. Sequential update of the field can mix very slowly, in particular if the observations are dense with respect to the spatial process range. A solution to this problem is to use blocked updates on $w(\mathcal{S})$.

Analytical joint sampling. A first approach is to use the analytical joint distribution in the Gaussian case (Datta et al., 2016). This approach has the inconvenient to become prohibitively costly as the size of the data augments. Due to the fact that a NNGP induces a Gaussian Markov Random Field, the cost will be roughly $\mathcal{O}(n)$, $\mathcal{O}(n^{3/2})$, $\mathcal{O}(n^2)$ following if the dimension of the spatial domain is respectively 1, 2, or 3 (Rue and Held, 2005). To address this problem, it is possible to split $w(\mathcal{S})$ into groups and update them sequentially. Those groups should be spatial clusters in order to guarantee that one group has enough freedom conditionally on the rest of the field. The problem is that the size of the groups will lower as more data is added, bringing back the original issue. Another problem is that extension to non-Gaussian data is possible but technical (Rue et al., 2004).

Hamiltonian methods. Another approach is to use Hamiltonian Monte Carlo (HMC). This method has the advantage to be easy to transpose to non-Gaussian data. The NUTS of Hoffman et al. (2014) gained much notoriety thanks to its automatic leapfrog length selection, but in the context of a hierarchical model Neal et al. (2011) report that short Hybrid Monte-Carlo hopping, potentially with momentum carry-over, is an efficient solution. My own experience with nonstationary models is consistent with the advice of Neal et al. (2011).

1.4.2 Interactions with covariance parameters

Another problem, and a much thornier one, is that the latent field impedes the mixing of the higher level parameters. A first aspect is that the covariance parameters and the latent field are correlated (Knorr-Held and Rue, 2002). The covariance parameters θ control the general aspect of the latent field, as shown in figure 1.3. New covariance parameters cannot just be sampled so easily: the latent field has to correspond to a field that could have been sampled with the proposed parameters.

Blocked update. A first solution to this problem consists in updating the field and the covariance parameters jointly, using blocked update (Knorr-Held and Rue, 2002). First, a new covariance parameter is proposed, then a new field corresponding to those parameters is sampled. The resulting Metropolis ratio is composed of:

- the ratio of proposal distributions, itself composed of:
  - the proposal distribution for the new covariance parameters.
  - the distributions to sample the latent fields knowing the covariance parameters.
- the ratio of densities, composed of:
  - the GP density of the latent field knowing the covariance parameters.
  - the likelihood of the observations knowing the latent field.
  - the hyperprior densities of the parameters.

This solution requires to use blocked sampling of $w(\mathcal{S})$ in order to avoid spatial correlation with the previous state of the latent field. It inherits the vulnerabilities of blocked field sampling concerning the scalability and the extension to non-Gaussian data. A good point is that it is possible to work on spatial clusters to sample the latent field. Another problem is that this solution relies on a random walk proposal for the covariance parameters. While the dimension of the parameters is low, which is the case in a stationary model, this solution is adequate, but it cannot be transposed as it is to a nonstationary model with more than a couple of parameters.

Collapsed Gibbs. Another solution is to avoid to sample the latent field thanks to a collapsed Gibbs (Finley et al., 2019). Like blocked sampling, the method is sensitive to the dimension of the space; moreover, unlike blocked sampling, it is not possible to work on spatial clusters. The method also is restricted to Gaussian data.

Interweaving. Yet another solution is to use interweaving (Filippone et al., 2013). This method samples covariance parameters while coupling natural and whitened parametrizations of the latent field. In general, this method scales poorly because it requires costly Cholesky factorization of the covariance matrix. The properties of NNGPs, though, make it perfectly affordable.

1.4.3 Interactions with regression coefficients

A second aspect of the interaction with high-level parameters is that the regression coefficients also cause trouble with the latent field. In particular, I showed empirically that the covariates with some spatial coherence are especially affected. I also carried over a theoretical exploration for the intercept effect, and got a result that goes in that sense even though it has some intractable terms (see chapter 3 for more details about this point). While I am not aware of some discussion of the problem for spatial models, the issue is well known for hierarchical models with random effects (Gelfand et al., 1995).

Collapsed Gibbs. The collapsed Gibbs should work well in theory, but it lacks applicability for NNGP models.

Switching parametrizations. Changing the centering of the latent field is another option (Gelfand et al., 1995). In general, it is well known that a linear recombination of the sampled variables can dramatically improve the performance of a Gibbs algorithm (Robert and Casella, 2004). In the case of the spatial model, this centering consists in replacing $w(\cdot)$ by

$$w_{center} = w + X\beta^T.$$

Of course, one is not obliged to center $w(\cdot)$ on the whole $X\beta^T$, and we can select a subset of the covariates. This method however suffers from various problems. A first potential issue is the cost of the method. With full GP, large dense matrix multiplication and inversion make the method very costly. In our case, centering is workable thanks to the good properties of NNGPs. Another issue is that centering the latent field on variables that vary within one spatial location is not possible. For example, several persons can live in the same house, being men or women, smokers or nonsmokers, etc... However, all covariates obtained through areal or gridded data are eligible for the method. The big flaw of the method is that when two parametrizations are available, they often behave like a "beauty and beast" pair (Yu and Meng, 2011). If one works greatly, the other will work terribly. It is therefore very important to chose which covariates to center and which covariates not to center.

Interweaving. Eventually, interweaving can be used (Yu and Meng, 2011). This solution mixes centered and natural parametrizations of the latent field, removing the daunting task to pick the right centering among the $2^{\text{number of covariates}}$ candidates. It inherits the rest of positive aspects and shortcomings of simple parametrization switching.

1.5 Detailed structure of the thesis

The first part of the thesis is a patchwork of ragtag developments that did not make their way into an article. It features results concerning the Kullback-Leibler precision of NNGPs, with an excursion towards meshed Gaussian Processes. Later, an algorithmic toolbox using latent field whitening is provided. It consists in a NNGP prediction algorithm adapted to the use of MCMC in high-level languages, a transposition to NNGP of the interweaving of Filippone et al. (2013), and a whitened Hybrid Monte-Carlo algorithm for the latent field (Neal et al., 2011). Eventually, the frustrated attempts to apply delayed acceptance to NNGP models are presented, with the motivation of the scheme and the results.

The second part of the thesis is an article that I wrote under the supervision of Benoît Liquet. It begins with a focus on the seemingly trivial linear regression component of NNGP models. I start with the intercept of the spatial model and show empirically that an equivalent centered parametrization works much better in terms of MCMC efficiency. I give some theoretical elements to support this point, but some intractable elements induced by GPs forbid a rigorous quantification of what is going on. I provide nonetheless some reasoning "with the hands" in order to show why the space-time coherence induced by a NNGP will make the intercept's coefficient mix poorly. After solving this problem, I had another difficulty: some covariates, having some kind of spatial coherence, may cause trouble just like the intercept. An approach consisting in picking the troublesome variables and center them manually would be tedious. Instead, I use the sparsity induced by NNGPs to implement efficiently an interweaving method (Yu and Meng, 2011) that mixes centered and non-centered parametrizations. The resulting hybrid takes advantage of the discordance of the two parametrizations.

The second part of this article aims to improve the sampling of the latent field. The Markovian nature of NNGPs allows to carry the parallelizability of NNGP density over to field sampling, leading to parallel coding for massive data sets and easy implementation in high-level languages such as R using vectorization. The groups of variables that can be sampled in parallel are identified through graph coloring. I tested three coloring algorithms, with the objective that the number of colors must be as small as possible while the time needed to color the graph must remain reasonable. Several types of graphs, including large graphs and graphs that correspond to blocked updates of the latent field, are tested. The available designs of Vecchia's approximations (ordering, number of parents) are screened. The results are compared through sensitivity analysis, and it appears that even though the number of colors changes with the properties of the graph, chromatic sampling is a viable and robust method. I also provide some clues to explain why the number of colors varies following the attributes of the graph, even though I could not go to the bottom of things, each individual subject being, I believe, able to provide enough problems for one thesis to solve. After proposing those two schemes, I put them in application. First, I tested my implementation with the package spNNGP (Finley et al., 2017) on synthetic data

sets. The proposed implementation performed well against the state-of-the-art package: in spite of its high-level coding, it does as good or better than the fine-tuned spNNGP. This implementation is used to analyze a data set of lead contamination in the mainland of the United States of America. While spNNGP had a pathological MCMC behavior, this implementation found sensible results.

The third part of the thesis concerns nonstationary NNGP modeling. It is an article resulting from a collaboration with Sudipto Banerjee and Benoît Liquet. The aim of this work is to tackle three issues of nonstationary modeling: the computational cost, the interpretability of the model, and model selection. We consider three extensions to the stationary model: a model with spatially variable range, and two heteroskedasticity models for the latent and noise processes. We use the covariance function of Paciorek (2003), embedded in the log-GP prior model of Heinonen et al. (2016). The first aspect of the work is to clarify what exactly is a non-stationary NNGP that uses the covariance from Paciorek (2003). I derive some helpful properties of the resulting NNGP, and I extend the NNGP process on the sphere.

The second contribution is not limited to NNGPs. It consists in generalizing the log-GP prior of Heinonen et al. (2016) to the elliptic covariance parameters of Paciorek (2003). As a result, a family of nonstationary models is obtained, with the complex models encompassing the simple ones, considerably simplifying model interpretation and selection.

The third contribution of this article is to find a MCMC engine able to power a complex, multi-layered hierarchical model. The algorithm relies on two pillars. The first element is a nested interweaving strategy, envisioned but not implemented by Yu and Meng (2011). This method is adapted to the fact that there are latent fields at various floors of the model. The second is a Hybrid Monte-Carlo step inspired from Heinonen et al. (2016) and adapted to NNGPs, that serves to update nonstationary covariance parameters. A potentially useful byproduct of this development is the gradient of the nonstationary NNGP density with respect to nonstationary covariance parameters: this result could be used in other approaches such as maximum *a posteriori* or maximum likelihood.

Experiments are done on synthetic data sets in order to draw empirical rules concerning model selection and identification. The results are encouraging. Analyzing a nonstationary data set with a nonstationary model gives better results following the deviance information criterion (Spiegelhalter et al., 2002). Moreover, the interpretability of the parametrization and the fact that the complex models encompass the simpler models largely helps model selection. Over-modeling can be detected from the estimates, allowing the user to downgrade the model after looking at the MCMC chains. Eventually, I use nonstationary models on real data, such as lead contamination in the United States of America (the same data set I analyzed with a stationary model in the previous part of the thesis).

The thesis ends with a short conclusion. It recaps the problems that were addressed by the thesis, and lays out the unsolved issues and open perspectives

that arise at the end of this work.

Chapter 2

Miscellaneous developments

This chapter rounds up various works that did not make their way into an article.

The first point, in section 2.1, is a couple of formulas concerning Kullback-Leibler divergence that helped me apprehend better Vecchia's approximations. In 2.1.1, I decompose the divergence of a Vecchia's approximation with respect to a full GP as a sum of conditional divergences. I also show that obtaining the smallest divergence is equivalent to a variable selection problem with an objective of conditional variance minimization. In 2.1.2, I adapt the result to another approximation of Peruzzi et al. (2020) called the Meshed Gaussian Process.

The second development started from the fact that the NNGP factor $\tilde{R}$ is sparse and triangular, allowing for fast linear solving. In particular, the whitened latent field $w^* = \tilde{R}w$ can be used *ad libitum*. MCMC steps that use this manipulation are omnipresent and instrumental in my current implementations of NNGP models. I start in 2.2.1 by exposing the surprising behavior of this parametrization, with an empirical exploration and a property that draws an unexpected bridge between the NNGPs and the predictive processes (Banerjee et al., 2008). Later, I describe an algorithmic toolbox whose elements involve this transformation in one way or another. The first item, in 2.2.2, is a prediction algorithm that is easy to use with MCMC and high-level languages. The second, in 2.2.3, is a Metropolis step that interweaves the whitened and natural parametrizations to update the covariance parameters. The third, in 2.2.4, is a Hybrid Monte-Carlo step whose efficiency is boosted thanks to approximate decorrelation of the sampled variables induced by whitening.

The third part, in 2.3, is an attempt to use delayed acceptance (Christen and Fox, 2005) with NNGP density computation. Delayed acceptance seems a promising lead for NNGP models: thank to the inherent separability of the pruned recursive conditional form, NNGP density can be sliced any way we want. I detail those points in 2.3.1. Moreover, delayed acceptance is compatible with the interweaving scheme for covariance parameters. Unfortunately, the results of delayed acceptance, which I present in 2.3.2 were disappointing.

2.1 Conditional Kullback-Leibler for Nearest Neighbor Gaussian Process

Kullback-Leibler (KL) is extensively used to assess the accuracy of Vecchia's approximations and Nearest Neighbor Gaussian Processes (Guinness, 2018; Peruzzi et al., 2020; Katzfuss et al., 2020). I report here two results that I derived and that are not in the literature, as far as I know. They can help to think about parent-picking heuristics and directed acyclic graph construction.

2.1.1 Conditional Kullback-Leibler divergence for NNGPs

Result. Conditional KL divergence is defined for two continuous distributions $p(\cdot)$ and $q(\cdot)$ and two sets of variables x and y as

$$KL(p(x|y)|q(x|y)) = \int p(x,y) log\left(\frac{p(x|y)}{q(x|y)}\right) d(x,y), \tag{2.1}$$

$d(x,y)$ denoting the Lebesgue measure on the joint sample space of x and y, in our case $\mathbb{R}^{dim(x)+dim(y)}$ since (x,y) has a Gaussian distribution. Take a set $\mathcal{S} = (s_1, \dots, s_n)$ of n ordered spatial locations, and $w(\cdot)$ the realization of the latent process at the considered site. Denote the NNGP density and the full Gaussian density respectively $\tilde{f}(\cdot)$ and $f(\cdot)$. Then, the KL divergence between the full GP joint density and its NNGP approximation writes:

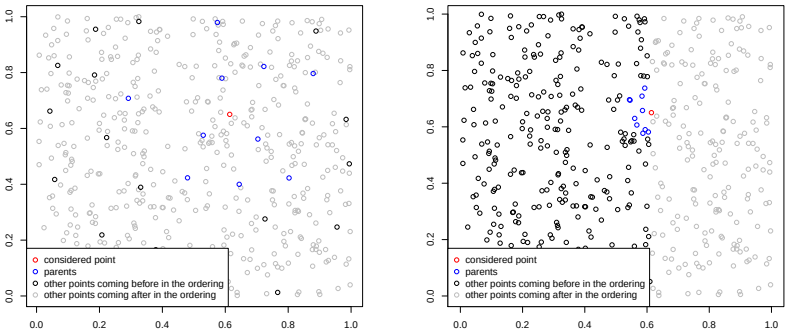
$$KL(f(w(\mathcal{S}))|\tilde{f}(w(\mathcal{S}))) = \Sigma_{i=1}^{n} KL(f(w(s_i)|w(s_1 \dots s_{i-1}))|f(w(s_i)|w(pa(s_i)))). \tag{2.2}$$

This formula allows a "divide and conquer" approach to the problem of choosing the parents sets (see 1.3.3). Given the ordering of the spatial locations, the global KL divergence of NNGP with respect to the full GP is minimized if all the conditional KL divergences are minimized. Each conditional KL divergence minimization is a variable selection problem: one looks for m parents that will minimize the divergence between the incompletely conditioned and the fully conditioned distributions. This problem can be reformulated in terms of conditional variance:

$$KL(f(w(\mathcal{S}))|\tilde{f}(w(\mathcal{S}))) = \Sigma_{i=1}^{n} log((\bar{\sigma}_i^2)_{NNGP} - (\bar{\sigma}_i^2)_{full\ GP})/2, \tag{2.3}$$

with $(\bar{\sigma}_i^2)_{NNGP} = var(w(s_i)|w(pa(s_i)))$ and $(\bar{\sigma}_i^2)_{full\ GP} = var(w(s_i)|w(s_1, \dots, s_{i-1}))$. This means that minimizing the conditional Kullbak-Leibler divergence with respect to the choice of $pa(s_i)$ among $(s_1, \dots, s_n)$ is the same thing as minimizing $var(w(s_i)|w(pa(s_i)))$.

Use and interpretation. Conditional KL divergences are difficult to compute due to the fact that the full GP conditional density becomes unaffordable as the number of conditioners augments. However, one can be interested in the subtraction of the KL divergence of two competing Vecchia's approximation



(a) Parents of a point with max-min ordering (30^{th} of 500)

(b) Parents of a point with coordinate ordering

Figure 2.1: Parents of the same point following the ordering

sharing the same ordering with respect to the full GP (Guinness, 2018). In this case, the expensive parts cancel out.

The formulas help to understand why the nearest neighbors (Datta et al., 2016) or a mix of nearest neighbors and farther observations (Stein et al., 2004) are chosen as parents sets. Indeed, the nearest neighbors of a point are the location that have the highest correlation with it, inducing a strong conditioning. However, when the covariance between those nearest neighbors is strong, they are redundant. In this case, it may be better to replace some of the nearest neighbors by farther locations that have a lower correlation with the considered point, but also have a lower correlation with each other, reducing the redundancy of the information they bring.

This decomposition gives a lead about the results of Guinness (2018), who finds that on a space with two dimensions, some orderings such as random shuffling of the locations or the max-min heuristic produce better approximations than other methods, such as ordering the points following one of their spatial coordinates or their distance from a point. The good orderings make the vertices cover quickly the spatial domain, while the bad orderings create a carpet of vertices that expands over the domain. The point is that Guinness (2018) uses the nearest neighbor heuristic to find the parents in the Directed Acyclic Graph (DAG). In the orderings that give bad results, the parents are packed on the blanket's lip and are highly correlated with each other. In the good orderings, the parents will be spread around their children and their mutual correlation is likely to be low.

Proofs. To prove the decomposition of the Kullback-Leibler contrast, do:

$$KL(f(w(\mathcal{S}))|\tilde{f}(w(\mathcal{S}))) = \int f(w(\mathcal{S}))log\left(\frac{f(w(\mathcal{S}))}{\tilde{f}(w(\mathcal{S}))}\right)d(w(\mathcal{S}))$$

(passing to conditional recursive form, equation (1.2))

$$= \int f(w(\mathcal{S}))\Sigma_{i=1}^{n} log\left(\frac{f(w(s_i)|w(s_1 \dots s_{i-1}))}{\tilde{f}(w(s_i)|w(s_1 \dots s_{i-1}))}\right)d(w(\mathcal{S}))$$

(writing integral of finite sum as sum of integrals)

$$= \Sigma_{i=1}^{n} \int f(w(\mathcal{S}))log\left(\frac{f(w(s_i)|w(s_1 \dots s_{i-1}))}{\tilde{f}(w(s_i)|w(s_1 \dots s_{i-1}))}\right)d(w(\mathcal{S}))$$

(splitting $f(w(\mathcal{S}))$ at s_i)

$$= \Sigma_{i=1}^{n} \int f(w(s_1 \dots s_i))f(w(s_{i+1} \dots s_n)|w(s_1 \dots s_i))$$
$$log\left(\frac{f(w(s_i)|w(s_1 \dots s_{i-1}))}{\tilde{f}(w(s_i)|w(s_1 \dots s_{i-1}))}\right)d(w(\mathcal{S}))$$

(applying Fubini's theorem since the KL divergence is finite because of the fact that the NNGP is non degenerate (Datta et al., 2016), and Lebesgue's measure is a product measure so $d(x, y) = dx\ dy$)

$$= \Sigma_{i=1}^{n} \int f(w(s_1 \dots s_i))log\left(\frac{f(w(s_i)|w(s_1 \dots s_{i-1})}{\tilde{f}(w(s_i)|w(s_1 \dots s_{i-1})))}\right)$$
$$\underbrace{\left(\int f(w(s_{i+1} \dots s_n)|w(s_1 \dots s_i))d(w(s_{i+1} \dots s_n))\right)}_{integrates\ to\ 1} d(w(s_1 \dots s_i))$$

(integrating out)

$$= \Sigma_{i=1}^{n} \int f(w(s_1 \dots s_i))log\left(\frac{f(w(s_i)|w(s_1 \dots s_{i-1}))}{\tilde{f}(w(s_i)|w(s_1 \dots s_{i-1}))}\right)d(w(s_1 \dots s_i))$$

(restricting NNGP parents, who are comprised in $(s_1, \dots, s_{i-1})$)

$$= \Sigma_{i=1}^{n} \int f(w(s_1 \dots s_i))log\left(\frac{f(w(s_i)|w(s_1 \dots s_{i-1}))}{f(w(s_i)|w(pa(s_i)))}\right)d(w(s_1 \dots s_i))$$

(identifying conditional KL formula)

$$= \Sigma_{i=1}^{n} KL(f(w(s_i)|w(s_1 \dots s_{i-1}))|f(w(s_i)|w(pa(s_i))))$$

In order to write the KL contrast in terms of conditional variance, start the conditional cross-entropy. The conditional KL is found by subtracting the conditional cross-entropy of the NNGP and the conditional entropy of the full GP, which is obtained by applying the formula below with $pa(s_i) = (s_1, \dots, s_{i-1})$.

$$
\begin{aligned}
H = \quad & -\int f(w(s_1,\dots,s_i)) log(f(w(s_i)|w(pa(s_i)))) dw(s_1,\dots,s_i) \\
& \text{(Integrating out, using finiteness of the cross entropy and the fact that Lebesgue's measure is a product measure)} \\
= \quad & -\int \left(\int f(w(s_1,\dots,s_i \backslash (s_i \cup pa(s_i))) | w(s_i \cup pa(s_i)) dw(s_1,\dots,s_i \backslash (s_i \cup pa(s_i))) \right) \\
& f(w(s_i \cup pa(s_i))) log(f(w(s_i)|w(pa(s_i)))) dw(s_i \cup pa(s_i)) \\
= \quad & -\int f(w(s_i \cup pa(s_i))) log(f(w(s_i)|w(pa(s_i)))) dw(s_i \cup pa(s_i)) \\
& \text{(Introducing the conditional variance and mean)} \\
= \quad & -\int f(w(s_i \cup pa(s_i))) \left(-log(\bar{\sigma_i}) - \frac{log(2\pi)}{2} - \frac{(x_i - \bar{\mu_i})^2}{2\bar{\sigma_i}^2} \right) dw(s_i \cup pa(s_i)) \\
& (\bar{\sigma_i} \text{ and } \frac{log(2\pi)}{2} \text{ are constant with respect to } w, \text{ and the density integrates to 1}) \\
= \quad & log(\bar{\sigma_i}) + \frac{log(2\pi)}{2} + \int f(w(s_i \cup pa(s_i))) \frac{(x_i - \bar{\mu_i})^2}{2\bar{\sigma_i}^2} dw(s_i \cup pa(s_i)) \\
& \text{(recognizing the formula of conditional variance)} \\
= \quad & log(\bar{\sigma_i}) + \frac{log(2\pi)}{2} + \int f(w(pa(s_i))) \int f(w(s_i)) \frac{(x_i - \bar{\mu_i})^2}{2\bar{\sigma_i}^2} dw(s_i) dw(pa(s_i)) \\
= \quad & log(\bar{\sigma_i}) + \frac{log(2\pi)}{2} + \frac{1}{2}
\end{aligned}
$$

When I found this second result, it felt inconsistent with the usual KL formula for multivariate normal distributions. After removing

$$(\mu_{NNGP} - \mu_{GP})^T \tilde{R}^T \tilde{R} (\mu_{NNGP} - \mu_{GP})$$

because the NNGP approximation does not change the mean, the Kullback-Leibler divergence between the full GP and the NNGP is

$$KL(GP|NNGP) = \frac{1}{2} \times (tr(\tilde{R}^T \tilde{R} \Sigma) - n + log(det(\tilde{R}^T \tilde{R})) - log(det(\Sigma))). \quad (2.4)$$

The "$log(\bar{\sigma_i})$" part allows to retrieve $log(det(\tilde{R}^T \tilde{R})) - log(det(\Sigma))$, but nothing seems to correspond to $tr(\tilde{R}^T \tilde{R} \Sigma)$. It felt like I forgot a term. After failing to find any error, I decided to compute $tr(\tilde{R}^T \tilde{R} \Sigma)$ on a toy example, and I obtained n, which cancels out in the KL formula. Therefore, I am quite reassured about the fact that I did not forget a component, but it would be interesting to pinpoint the exact reason why $tr(\tilde{R}^T \tilde{R} \Sigma) = n$.

2.1.2 Conditional Kullback-Leibler divergence for meshed Gaussian Processes

Peruzzi et al. (2020) introduce the meshed Gaussian Process, a hybrid between the Predictive Process (Banerjee et al., 2008) and the Nearest Neighbor Gaussian Process (Datta et al., 2016). The principle of the method is to introduce a set of auxiliary spatial locations ("knots") like in Banerjee et al. (2008). This set of p points is denoted as $\mathcal{K} = (k_1, \dots, k_p)$. A NNGP density is then defined on $\mathcal{K} \cup \mathcal{S}$, with the restriction that $\mathcal{K}$ comes before $\mathcal{S}$ in the ordering of spatial locations used to define the DAG and that $\forall s \in \mathcal{S}, pa(s) \subset \mathcal{K}$. Thanks to

the fact that $\mathcal{K}$ has a simple form, typically gridded, the computation of the covariance structure is considerably simplified.

The knots are auxiliary, so what matters in the end is to approximate best the Gaussian density at the observations and not the knots. The objective is to minimize $KL(f(w(\mathcal{S}))|\tilde{f}(w(\mathcal{S})))$, where $\tilde{f}$ is the tessellated GP density. This quantity can be controlled using:

$$\begin{aligned} KL(f(w(\mathcal{S}))|\tilde{f}(w(\mathcal{S}))) \leq \quad & \Sigma_{i=1}^{p} KL(f(w(k_i)|w(k_1,\ldots,k_{i-1})|f(w(k_i)|w(pa(k_i))) \\ & + \Sigma_{i=1}^{n} KL(f(w(s_i)|w(k_1,\ldots,k_n,s_1,\ldots,s_{i-1})|f(w(s_i)|w(pa(s_i))) \end{aligned}$$

The proof starts with a majoration of the KL divergence, relying on the positivity of conditional KL divergence:

$$\begin{aligned} & KL(f(w(\mathcal{S}))|\tilde{f}(w(\mathcal{S}))) \leq KL(f(w(\mathcal{S}))|\tilde{f}(w(\mathcal{S}))) + KL(f(w(\mathcal{K})|w(\mathcal{S}))|\tilde{f}(w(\mathcal{K})|w(\mathcal{S}))) \\ & \quad \text{(using 2 times conditional entropy property: } H(x|y) + H(y) = H(x \cup y)) \\ = \quad & KL(f(w(\mathcal{S}\cup\mathcal{K}))|\tilde{f}(w(\mathcal{S}\cup\mathcal{K}))) \\ = \quad & KL(f(w(\mathcal{K}))|\tilde{f}(w(\mathcal{K}))) + KL(f(w(\mathcal{S})|w(\mathcal{K}))|\tilde{f}(w(\mathcal{S})|w(\mathcal{K}))) \\ & \quad \text{(applying the first formula (2.2) on the two KL)} \\ = \quad & \Sigma_{i=1}^{p} KL(f(w(k_i)|w(k_1,\ldots,k_{i-1})|f(w(k_i)|w(pa(k_i))) \\ & \quad + \Sigma_{i=1}^{n} KL(f(w(s_i)|w(k_1,\ldots,k_p,s_1,\ldots,s_{i-1})|f(w(s_i)|w(pa(s_i))) \end{aligned}$$

Then the target KL can be controlled through:

- $\Sigma_{i=1}^{p} KL(f(w(k_i)|w(k_1,\ldots,k_{i-1})|f(w(k_i)|w(pa(k_i)))$, which is the Kullback-Leibler divergence between the NNGP approximation at the knots and the full GP at the knots.
- $\Sigma_{i=1}^{n} KL(f(w(s_i)|w(k_1,\ldots,k_p,s_1,\ldots,s_{i-1})|f(w(s_i)|w(pa(s_i)))$, which is the quality of the conditioning of an observation by its parent knots with respect to the full GP, which conditions by all the knots plus all the previous observations.

This formula has the advantage of "dividing and conquering" like (2.2). First, the quality of the approximation on the knots grid and the quality of the conditioning of the observations can be treated separately in order to minimize the upper bound. Second, each of those two problems can be split into smaller elemental tasks. Its big flaw is that the result is based on a majoration that may be gross.

2.2 Fast prior whitening of latent NNGP field

NNGPs allow for an useful re-parametrization of the latent field known as prior whitening. This method is known in the Gaussian process literature (Filippone et al., 2013; Heinonen et al., 2016) and MCMC good practices in general

(Neal et al., 2011), but it has not been applied to NNGP as far as I know: the recent works of Finley et al. (2019) rather develop algorithms that avoid sampling the latent field. However, $\tilde{R}$, the NNGP factor of the prior precision matrix of the latent field, is sparse and triangular, which is ideal to implement efficiently algorithms that need to whiten the latent field. From this starting point, I investigated the aspects of NNGP models that may benefit from the re-parametrization.

Subsection 2.2.1 presents the parametrization, gives a few properties, and explores empirically its behavior. Later, I apply those properties to various algorithms: latent field prediction (2.2.2), Ancillary-Sufficient Interweaving Strategy (ASIS) to update the covariance parameters (2.2.3), and prior whitening for Hamiltonian Monte-Carlo sampling of the latent field (2.2.4).

2.2.1 Presentation of latent field whitening

Prior whitening, ancillary parametrization. Prior whitening is a linear combination of the Gaussian latent field which depends on the covariance parameters. Remember that $\mathcal{S}$ is a set of n spatial coordinates. The latent field $w(\mathcal{S})$ whose prior GP density is $\mathcal{N}(0, \Sigma)$ is re-parametrized as

$$w^*(\mathcal{S}) = (\Sigma^{1/2})^{-1} w(\mathcal{S}),$$

with

$$\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^T.$$

The prior density of $w^*(\mathcal{S})$ will then be a standard multivariate normal distribution.

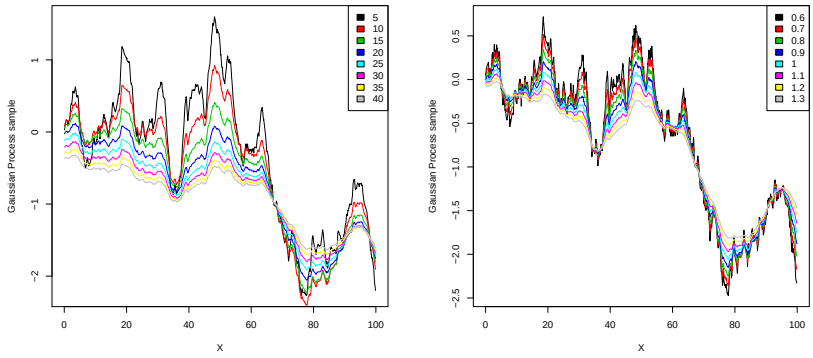
This reparametrization can be used in an Ancillary-Sufficient Interweaving Strategy (ASIS) (Yu and Meng, 2011; Filippone et al., 2013) aiming to improve the sampling of the covariance parameters θ. I call $w(\mathcal{S})$ the "natural" parametrization because it has a straightforward interpretation in the analysis of the observations in the hierarchical model (see figures 1.1 and 1.2). Following the terminology of interweaving, the natural parametrization of the latent field is called *sufficient augmentation* because conditionally on $w(\mathcal{S})$ the *a posteriori* density of the covariance parameters is independent from the observed data and the other parameters. On the other hand, the whitened latent field is an *ancillary augmentation* because it is *a priori* independent from the covariance parameters.

The properties of the whitened latent field are not the same as the natural latent field's and it is easy to get confused. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two sets of spatial points. Assume that $\Sigma^{-1/2}$ is lower triangular, like the NNGP factor $\tilde{R}$. Due to the fact that $\Sigma^{-1/2}$ is not diagonal, the whitened latent field cannot be concatenated:

$$(w^*(\mathcal{S}_1), w^*(\mathcal{S}_2)) \neq w^*(\mathcal{S}_1, \mathcal{S}_2).$$

Similarly, a permutation of the spatial locations is not equivalent to a permutation of the vector indices:

$$w^*(\mathcal{S}_{p(1,\dots,n)}) \neq w^*(\mathcal{S})_{p(1,\dots,n)},$$



(a) Changing the range (b) Changing the smoothness

Figure 2.2: Effect of a change of covariance parameters on a Matérn latent field with fixed ancillary augmentation

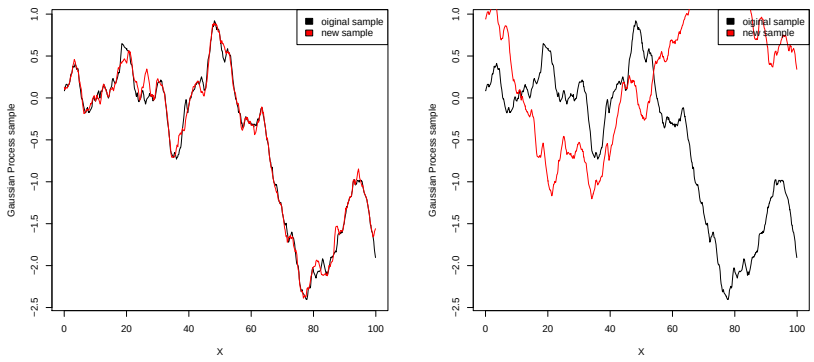
$p(1, \ldots, n)$ being a permutation of $(1, \ldots, n)$. However, $\Sigma^{-1/2}$ being lower triangular, the first elements of $w^*(\cdot)$ can be extracted. Denote $\#$ the cardinal:

$$w^*(\mathcal{S}_1 \cup \mathcal{S}_2)_{1,\ldots,\#\mathcal{S}_1} = w^*(\mathcal{S}_1).$$

Methods using prior whitening are appealing but they usually involve computing Cholesky factors and solving linear systems for large matrices. However, a NNGP model allows for direct computation of $\tilde{R}$ such that the GP prior of $w(\mathcal{S})$ has a covariance matrix given by $(\tilde{R}^T\tilde{R})^{-1}$, and $\tilde{R}$ is sparse and triangular. Using sparse matrix multiplication, it is immediate to compute $w^*(\mathcal{S}) = \tilde{R}w(\mathcal{S})$. On the other hand, sparse triangular linear solving allows to retrieve efficiently $w(\mathcal{S}) = (\tilde{R})^{-1}w^*(\mathcal{S})$. However, computing explicitly $\tilde{R}^{-1}$ quickly becomes prohibitively expensive.

Empirical exploration. Changing the covariance parameters θ and leaving $w^*(\mathcal{S})$ untouched will not alter the general profile of $w(\mathcal{S})$ but will modify its aspect locally. For example, when the range (figure 2.2a) or smoothness (figure 2.2b) parameters of a Matérn function are modified, the field will become smoother or fuzzier but will keep the same profile.

On the other hand, changing w^* does not change the way w behaves locally, but will change its global trajectory. In particular, when the ordering of the locations is such that the first locations of $\mathcal{S}$ cover the whole spatial domain rapidly (maxmin and random ordering in Guinness, 2018), the first coefficients of w^* can have an impressive leverage effect. In some cases, a handful of the first coefficients are enough to have an accurate approximation of w, and the rest of the coefficients of w^* parametrize small local variations. I illustrate this point with figure 2.3. I take an original NNGP sample $w = \tilde{R}^{-1}w^*, w^* \overset{ind.}{\sim} \mathcal{N}(0, 1)$,



(a) Changing the 900 last coeff. of w^* (b) Changing the 100 first coeff. of w^*

Figure 2.3: The first elements of the ancillary augmentation are much more important than the last.

that is represented in black in the two subfigures. In figure 2.3a, I replace the 900 last values of w^* by new draws from a standard normal. I do the same in 2.3b, but with the 100 first values. In the two cases, we obtain perturbed samples of w that are drawn in red. In this example, changing the 900 last values of w^* produces almost no effect on w, while changing the 100 first values completely transforms the sample.

Truncated whitened latent field inducing a Predictive Process approximation of the Nearest Neighbor Gaussian process. In order to present the result, I use the following notations. For a matrix M, read $M_{i\dots j,k\dots l}$ as "the sub-matrix obtained by reducing M to its rows whose indices range from i to j and its columns whose indices range from k to l". Similarly, for a vector x, read $x_{i\dots j}$ as "the sub-vector obtained by reducing x to its coefficients whose indices range from i to j".

The powerful leverage effect of the first elements of w^* comes from the fact that

$$\left(\tilde{R}^{-1}_{1\dots n,1\dots i} w^*_{1\dots i}\right)_{1\dots i} = w_{1\dots i}$$

and that

$$\left(\tilde{R}^{-1}_{1\dots n,1\dots i} w^*_{1\dots i}\right)_{i+1\dots n} = \mathbb{E}\left(w_{i+1\dots n} | w_{1\dots i}, \theta\right) \tag{2.5}$$

if the mean of $w(\cdot)$ is $(0,\dots,0)$. This means that the i first terms of $w^*(\mathcal{S})$ will allow to recover perfectly the i first terms of $w(\mathcal{S})$. The rest of $w(\mathcal{S})$ will be recovered imperfectly, but will be replaced by its expectation conditionally on the i first terms of w and the covariance parameters. If the i first observation sites are sufficiently dense in the spatial domain, the conditional expectation can be very close to the actual sample. This the principle of the predictive process

approximation to full Gaussian Processes (Banerjee et al., 2008), where the i first elements of $\mathcal{S}$ are the *knots* of the predictive process.

It is clear that the knots of a predictive process should be well spread over the spatial domain in order to guarantee a strong conditioning of the predicted locations. Similarly, the ordering of $\mathcal{S}$ that is used to define the NNGP should be such that the i first locations span all over the space. Orderings that induce such dispersion are the max-min and random ordering, which yield anyway the most accurate Vecchia's approximations on two dimensions following Guinness (2018). The result is demonstrated using block matrices. Denote:

- the upper left block as $\tilde{R}_{11} = \tilde{R}_{1\dots i,1\dots i}$.
- the lower right block as $\tilde{R}_{22} = \tilde{R}_{i+1\dots n,i+1\dots n}$.
- the lower left block as $\tilde{R}_{21} = \tilde{R}_{i+1\dots n,1\dots i}$.
- the upper right block as $\tilde{R}_{12} = \tilde{R}_{1\dots i,i+1\dots n}$.

Use the same notation for $\tilde{R}^{-1}$. For w and the other vectors of size n:

- $w_{1\dots i}$ is noted w_1.
- $w_{i+1\dots n}$ is noted w_2.

Using the fact that $\tilde{R}$ is triangular,

$$\left(\tilde{R}^{-1}\right)_{11} = \left(\tilde{R}_{11}\right)^{-1} \qquad \text{and} \qquad \left(\tilde{R}^{-1}\right)_{12} = \mathbf{0}_{\mathcal{M}_{i\times(n-i)}}.$$

It follows that

$$(\tilde{R}^{-1})_{11} w_1^* = (\tilde{R}^{-1} w^*)_1 = w_1,$$

which proves the first point. Using the usual block inversion formulas and remarking that $\tilde{R}_{12}$ has only null coefficients,

$$\left(\tilde{R}^{-1}\right)_{21} = \left(-\tilde{R}_{22}\right)^{-1} \tilde{R}_{21} \left(\tilde{R}_{11}\right)^{-1}.$$

Multiplying by w_1^* on the right and factorizing by $\left(-\tilde{R}_{22}^T \tilde{R}_{22}\right)^{-1}$ on the left, it follows that

$$\left(\tilde{R}^{-1}\right)_{21} w_1^* = \left(-\tilde{R}_{22}\right)^{-1} \tilde{R}_{21} w_1 = \left(-\tilde{R}_{22}^T \tilde{R}_{22}\right)^{-1} \left(\tilde{R}_{22}^T \tilde{R}_{21}\right) w_1.$$

Denote $Q = \tilde{R}^T \tilde{R}$ the NNGP-induced precision matrix. Using the fact that $\tilde{R}$ is triangular,

$$\tilde{R}_{22}^T \tilde{R}_{22} = \left(\tilde{R}^T \tilde{R}\right)_{22} = Q_{22} \qquad \text{and} \qquad \tilde{R}_{22}^T \tilde{R}_{21} = \left(\tilde{R}^T \tilde{R}\right)_{21} = Q_{21}.$$

Using the conditional expectation formula with a precision matrix (Rue and Held, 2005), we find the second result:

$$\left(\tilde{R}^{-1}\right)_{21} w_1^* = -\tilde{Q}_{22}^{-1} \tilde{Q}_{21} w_1 = \mathbb{E}(w_2 | w_1, \theta).$$

2.2.2 NNGP prediction using completion of the ancillary parametrization

Prediction of the latent field at unobserved locations is an important aspect of spatial modeling. There is an extensive study of Vecchia prediction in Katzfuss et al. (2020). However, the family of algorithms presented there is not well suited for MCMC implementation of a NNGP model. The method of Katzfuss et al. (2020) is to obtain the conditional mean and variance of the latent process at the predicted locations, conditionally on the response variable and the covariance parameters. But if MCMC samples of the latent field at the observed locations are available, prediction can then be done conditionally on the covariance parameters and the latent field at the observed locations, instead of the noisy response variable. Prediction algorithms already exist for MCMC implementation of NNGP models in Datta et al. (2016); Finley et al. (2019). However, those algorithms are oriented towards low-level languages and a method easy to use in R can come in handy. Moreover, those algorithms induce independence between the predicted locations conditionally on the observed locations, which leads to a decrease in the quality of the approximation following Katzfuss et al. (2020).

The R package GpGp already uses sparse forward substitution for fast NNGP sampling, see algorithm 1. This method is standard for spatial process simulation (Kroese and Botev, 2015). I apply the same idea to NNGP prediction at unobserved locations.

Algorithm 1 Algorithm used by GpGp in order to sample from $\mathcal{N}(0, (\tilde{R}^T\tilde{R})^{-1})$

input $\tilde{R}$

simulate $w^* \sim \mathcal{N}(0, I_n)$

solve $w = \tilde{R}^{-1}w^*$

return w

Denote $\mathcal{S}$ the set of spatial locations where samples of $w(\cdot)$ are available, and $\mathcal{P}$ the set of locations where prediction needs to be done. To predict at the un-observed locations, a joint NNGP prior must be defined at the observed and predicted locations. Define a DAG on $(\mathcal{S}, \mathcal{P})$, $\mathcal{S}$ coming before $\mathcal{P}$ in the ordered set. The first $\#\mathcal{S}$ nodes of the dag and the edges between those nodes have already been defined, because the prediction proceeds from a NNGP model fit using data observed at $\mathcal{S}$. The rest of the edges arrive to nodes corresponding to the locations of $\mathcal{P}$, and can come from either $\mathcal{S}$ or $\mathcal{P}$. Consider a state of the MCMC chain $(\theta^t, w^t(\mathcal{S}))$. Note $\tilde{R}((\mathcal{S}, \mathcal{P}), \theta^t)$ the joint NNGP factor computed using the covariance parameters θ^t. Its upper left corner corresponding to $\mathcal{S}$ is denoted $\tilde{R}(\mathcal{S}, \theta^t)$. Using (2.5),

$$\tilde{R}((\mathcal{S}, \mathcal{P}), \theta^t)^{-1}_{1,\cdots}\tilde{R}(\mathcal{S}, \theta^t)w^t(\mathcal{S}) = \mathbb{E}(w(\mathcal{P})|w^t(\mathcal{S}), \theta^t),$$

$\tilde{R}((\mathcal{S}, \mathcal{P}), \theta^t)^{-1}_{1,\cdots}$ being the matrix composed of the first $\#\mathcal{S}$ columns of $\tilde{R}((\mathcal{S}, \mathcal{P}), \theta^t)^{-1}$.

Using the conditional distribution formula for precision matrices (Rue and Held, 2005) and the fact that $\tilde{R}((\mathcal{S},\mathcal{P}),\theta^t)$ is lower triangular,the conditional variance matrix of $w(\mathcal{P})$ can be expressed as

$$Var(w(\mathcal{P})|w^t(\mathcal{S}),\theta^t) = \left(\tilde{R}((\mathcal{S},\mathcal{P}),\theta^t)^T_{2,2}\tilde{R}((\mathcal{S},\mathcal{P}),\theta^t)_{2,2}\right)^{-1},$$

$\tilde{R}((\mathcal{S},\mathcal{P}),\theta^t)_{2,2}$ being the lower-right square block of $\tilde{R}((\mathcal{S},\mathcal{P}),\theta^t)$ of size $\#\mathcal{P}$. With those two elements in hand, it is possible to obtain one conditional draw of $w(\mathcal{P})$ using

$$\underbrace{\tilde{R}((\mathcal{S},\mathcal{P}),\theta^t)^{-1}_{1,\cdots}\tilde{R}(\mathcal{S},\theta^t)w^t(\mathcal{S})}_{\mathbb{E}(w(\mathcal{P})|w^t(\mathcal{S}),\theta^t)} + \underbrace{\tilde{R}((\mathcal{S},\mathcal{P}),\theta^t)^{-1}_{2,2}}_{Chol(Var(w(\mathcal{P})|w^t(\mathcal{S}),\theta^t))} w^*(\mathcal{P}),$$

$$\text{with} \quad w^*(\mathcal{P}) \overset{ind.}{\sim} \mathcal{N}(0,1).$$

This draw can actually be re-written as a completion of the known elements of the whitened latent field $w^*(\mathcal{S})$ with $\#\mathcal{P}$ random elements drawn following independent standard normal distributions:

$$\tilde{R}((\mathcal{S},\mathcal{P}),\theta^t)^{-1}(w^{*\ t}(\mathcal{S}), w^*(\mathcal{P}))$$

$$\text{with} \quad w^*(\mathcal{P}) \overset{ind.}{\sim} \mathcal{N}(0,1) \quad \text{and} \quad w^{*\ t}(\mathcal{S}) = \tilde{R}(\mathcal{S},\theta^t)w^t(\mathcal{S}).$$

A pseudo code using this principle is presented in algorithm 2.

One potential limit of the algorithm is that it requires the observed locations $\mathcal{S}$ to come before the predicted locations $\mathcal{P}$ in the ordering. Indeed, in algorithm 2, the "sample" step consists in completing $w^*(\mathcal{S})$ by standard normal draws in order to get samples of $w^*(\mathcal{S},\mathcal{P})$, which is possible only if the observed locations come first (see 2.2.1). Empirical exploration of the accuracy of Vecchia's approximations by Guinness (2018) shows that some orderings of the spatial locations work better than others. On spatial domains with two dimensions, the max-min and the random ordering win; those methods guarantee that the first ordered locations will quickly span the spatial domain, contrary to other methods such as classing the points following a coordinate. The problem is that the observed locations $\mathcal{S}$ may not cover regularly the whole spatial domain where both observations and predictions are done. For example, satellite measurements may yield very dense measurements along the orbit, but no measurements between those paths (Katzfuss et al., 2020). Following the conclusions of Guinness (2018), it would be better to predict such data using a max-min or random order on $(\mathcal{S},\mathcal{P})$. This is not possible with my algorithm since shuffling or inverting $\mathcal{S}$ and $\mathcal{P}$ is forbidden. A workaround for this problem would be to change the heuristic to choose the parents of the DAG instead of changing the order of the graph. Using (2.2), we can see that if the parents of a predicted point are clumped, their mutual correlation is high and they are redundant. Prediction is done after the model is fit, so that the modeler has an idea of the actual range of the spatial process. Then, variable selection can be used in order to minimize (2.3).

Algorithm 2 NNGP prediction using ancillary representation, practical version

input $\mathcal{S}, \mathcal{P}$ ▷ Observed and Predicted spatial Locations
input $w_{1,\dots,niter}(\mathcal{S})$ ▷ MCMC samples of the latent field
input $\theta_{1,\dots,niter}$ ▷ MCMC samples of the covariance parameters
for $t \in \{1, \dots, niter\}$ **do**
  Compute $\tilde{R}((\mathcal{S}, \mathcal{P}), \theta^t)$
  Extract $\tilde{R}(\mathcal{S}, \theta^t)$ from $\tilde{R}((\mathcal{S}, \mathcal{P}), \theta^t)$ ▷ Upper left corner
  Solve $w_t^*(\mathcal{S}) = \tilde{R}(\mathcal{S}, \theta^t) w_t(\mathcal{S})$
  Sample $w_t^*((\mathcal{S}, \mathcal{P})) = (w_t^*(\mathcal{S}), \mathcal{N}(0, I_{\#\mathcal{P}}))$ ▷ Completion
  Solve $w_t((\mathcal{S}, \mathcal{P})) = \tilde{R}((\mathcal{S}, \mathcal{P}), \theta^t)^{-1} w_t^*((\mathcal{S}, \mathcal{P}))$
  Extract $w_t(\mathcal{P})$ from $w_t((\mathcal{S}, \mathcal{P}))$
end for
return $w_{1,\dots,niter}(\mathcal{P})$

2.2.3 ASIS for covariance parameters update

Covariance parameters sampling is a thorny question because the correlation between the latent field and the covariance parameters is strong. In Datta et al. (2016), the Gibbs sampler loops over the covariance parameters, the latent field, and all other parameters. This architecture often suffers from slow mixing.

The *collapsed NNGP* (Finley et al., 2019) removes the troublesome latent field and only samples from the covariance parameters and the other parameters of the model. However, various points should be considered: the algorithm is sensitive to the design of the spatial observations, and as far as I know it is difficult to apply it to latent fields with non-Gaussian likelihoods. On the other hand, the KHR architecture (Knorr-Held and Rue, 2002) blocks field and covariance parameter sampling. The method can boast a good mixing, but it involves costly Cholesky factorizations. It is possible but technical and potentially costly to extend it to fields with non-Gaussian observations (Rue et al., 2004). Moreover, the method can scale poorly like the collapsed NNGP. It is possible to keep the method feasible by splitting the field into various spatial blocks that are updated one after another, but it is at the expense of efficiency, it is complex, and the cost remains relatively high anyway.

Interweaving is another option. It consists in a mix of covariance parameters updates with two data augmentations. The general methodology is presented by Yu and Meng (2011). The particular application to covariance parameters updating is evaluated by Filippone et al. (2013). This article concludes that the method is competitive with other strategies (in particular with the KHR update). This method can also boast about the fact that it is easy to implement and that there is no *a priori* difficulty with non-Gaussian response.

The method takes advantage of the discordance between two parametriza-

tions of the data augmentation w_1 and w_2. Gibbs sampler updates of θ in models with either parametrizations are done with the relevant full conditional

$$\theta^{t+1} \sim [\theta | w_1^t, \ldots] \quad \text{or} \quad \theta^{t+1} \sim [\theta | w_2^t, \ldots],$$

where "..." represents the rest of the parameters of the model. Now suppose that there is a valid joint distribution (even degenerate) $[\theta, w_1, w_2, \ldots]$ comprising both data augmentations, that is with marginal distributions $[w_1, \theta, \ldots]$ and $[w_2, \theta, \ldots]$ corresponding to the joint distributions of the two simpler models. Interweaving consists in using the distribution that encompasses the two parametrizations, and replacing the update of θ by the following step:

$$[w_2, \theta | w_1^t, \ldots] \to [\theta^{t+1}, w_1^{t+0.5} | w_2, \ldots].$$

Two blocked updates following the full conditional distributions are combined, each preserving the target density. Usually, it is complicated to sample directly $[w_2, \theta | w_1^t, \ldots]$, and much simpler to do:

$$\underbrace{[\theta | w_1^t, \ldots] \to [w_2 | w_1^t, \theta, \ldots]}_{[\theta, w_2 | w_1^t, \ldots]} \to \underbrace{[\theta^{t+1} | w_2, \ldots] \to [w_1^{t+0.5} | w_2, \theta^{t+1}, \ldots]}_{[\theta^{t+1}, w_1^{t+0.5} | w_2, \ldots]}. \tag{2.6}$$

When the joint distribution is degenerate, $[w_2 | \theta, w_1, \ldots]$ is a deterministic transformation. It is the case with field whitening since $w^* = \tilde{R}(\theta) w$. At the end of the step w_1 is changed, but this change is not on an equal footing with a dedicated step that guarantees the irreducibility of the chain. This is the reason why "$w_1^{t+0.5}$" is written instead "w_1^{t+1}" at the end of the step. A genuine update of the data augmentation must be included in the Gibbs sampler and can be either $[w_2 | \theta, \ldots] \to [w_1 | \theta, w_2, \ldots]$ or $[w_1 | \theta, \ldots] \to [w_2 | \theta, w_1, \ldots]$, the second part of each step actually being a deterministic harmonization.

Note that the θ that is sampled in the left-hand part of (2.6) is not used in the right-hand part, so by marginalization of the joint draw

$$[\theta | w_1^t, \ldots] \to [w_2 | w_1^t, \theta, \ldots] \quad \text{is equivalent to} \quad [w_2 | w_1^t, \ldots].$$

The strategy precisely relies on the discordance between w_2 and w_1 and can even be efficient when none of the two augmentations performs well when implemented separately. A good choice for interweaving is when the two data augmentations are an ancillary-sufficient couple, giving an Ancillary-Sufficient Interweaving Strategy (ASIS). A sufficient augmentation is an *a posteriori* sufficient statistic for the parameter θ: conditioning by the data augmentation only gives the full conditional. An ancillary statistic is *a priori* independent from the parameter. In this application, the natural parametrization of the latent field is the sufficient augmentation, and the whitened latent field is the ancillary augmentation.

A change of parametrization for the latent field affects greatly the sampling procedure for the covariance parameters. When sufficient augmentation is used,

the full conditional distribution of the covariance parameters is proportional to the prior NNGP distribution $\tilde{f}(w(\mathcal{S})|\theta)h(\theta)$, $h(\cdot)$ being the hyperprior distribution and $\tilde{f}(\cdot|\cdot)$ being the NNGP prior of a latent field knowing its covariance parameters. This means that when a new covariance parameter is proposed in a Metropolis step, it is accepted or rejected following whether it explains well or not the structure of the latent field.

When ancillary augmentation is used, the prior of w^* is a multivariate standard normal density and does not change with θ. The full conditional becomes proportional to $l(z(\mathcal{S})|\tilde{R}(\mathcal{S},\theta)^{-1}w^*, \beta, \ldots)h(\theta)$, $l(\cdot,\cdot)$ being the likelihood of the observations knowing the latent field, the linear regression, and additional parameters such as the variance of the noise if the data is Gaussian. This means that when a new covariance parameter is proposed, the latent field is modified accordingly like in figure 2.2, and the covariance parameter is accepted or rejected following whether the new latent field explains well the observed data.

Two versions of the algorithm are given. The first version (algorithm 3) underlines the sampling operations while the second (algorithm 4) is more practical.

Algorithm 3 Covariance parameters update using ASIS

input $\beta^t, \theta^t, w^t, z, X, \ldots$ ▷ More parameters such as τ can be needed

sample θ' from $q_{suff}(\cdot|\theta^t)$ ▷ Sufficient augmentation Metropolis Update

sample $a \sim \mathcal{U}([0,1])$

if $\frac{\tilde{f}(w^t|\theta')q_{suff}(\theta^t|\theta')}{\tilde{f}(w^t|\theta)q_{suff}(\theta'|\theta^t)} > a$ **then**

$\theta^{t+1/2} = \theta'$

else

$\theta^{t+1/2} = \theta^t$

end if

sample w^* conditionally on w^t and $\theta^{t+1/2}$

sample θ' from $q_{ancill}(\cdot|\theta^{t+1/2})$ ▷ Ancillary augmentation Metropolis Update

sample $a \sim \mathcal{U}([0,1])$

if $\frac{l(z|\tilde{R}(\theta')^{-1}w^*,\beta,\ldots)q_{ancill}(\theta^{t+1/2}|\theta')}{l(z|\tilde{R}(\theta^{t+1/2})^{-1}w^*,\beta,\ldots)q_{ancill}(\theta'|\theta^{t+1/2})} > a$ **then**

$\theta^{t+1} = \theta'$

else

$\theta^{t+1} = \theta^{t+1/2}$

end if

sample $w^{t+1/2}$ conditionally on w^* and θ^{t+1}

return θ^{t+1}, $w^{t+1/2}$ ▷ $w^{t+1/2}$ needs be updated later

Algorithm 4 Covariance parameters update using ASIS, practical version

input $\beta^t, \theta^t, w^t, z, X, \ldots$ ▷ More parameters such as τ can be needed

input $\tilde{R}(\theta^t)$

sample θ' from $q_{suff}(\cdot|\theta^t)$ ▷ Sufficient augmentation Metropolis Update

compute $\tilde{R}(\theta')$

$\rho_1 = \Sigma_{i=1}^n log(\tilde{R}(\theta')_{i,i}) - 1/2((w^t)^T R(\theta')^T R(\theta') w^t) - \Sigma_{i=1}^n log(\tilde{R}(\theta)_{i,i}) + 1/2((w^t)^T R(\theta)^T R(\theta) w^t)$ ▷ NNGP log-ratio

$\rho_2 = log(q_{suff}(\theta^t|\theta')) - log(q_{suff}(\theta'|\theta^t)$ ▷ Proposal distribution log-ratio

sample $a \sim \mathcal{U}([0,1])$

if $\rho_1 + \rho_2 > log(a)$ **then**

$\theta^{t+1/2} = \theta'$

$\tilde{R}(\theta^{t+1/2}) = \tilde{R}(\theta')$

else

$\theta^{t+1/2} = \theta^t$

$\tilde{R}(\theta^{t+1/2}) = \tilde{R}(\theta^t)$

end if

sample θ' from $q_{ancill}(\cdot|\theta^{t+1/2})$ ▷ Ancillary augmentation Metropolis Update

compute $\tilde{R}(\theta')$

solve $w' = \tilde{R}(\theta')^{-1} \tilde{R}(\theta^{t+1/2}) w^t$

$\rho_1 = log(l(z|w', \beta, \ldots)) - log(l(z|w^t, \beta, \ldots))$ ▷ Field likelihood log-ratio

$\rho_2 = log(q_{ancill}(\theta^t|\theta')) - log(q_{ancill}(\theta'|\theta^t))$ ▷ Proposal distribution log-ratio

sample $a \sim \mathcal{U}([0,1])$

if $\rho_1 + \rho_2 > log(a)$ **then**

$\theta^{t+1} = \theta'$

$\tilde{R}(\theta^{t+1}) = \tilde{R}(\theta')$

$w^{t+1/2} = w'$

else

$\theta^{t+1} = \theta^{t+1/2}$

$\tilde{R}(\theta^{t+1}) = \tilde{R}(\theta^{t+1/2})$

$w^{t+1/2} = w^t$

end if

return θ^{t+1}, $\tilde{R}(\theta^{t+1})$, $w^{t+1/2}$ ▷ $w^{t+1/2}$ still needs to be updated later

2.2.4 Field whitening and Hamiltonian Monte Carlo sampling

Hamiltonian Monte Carlo (HMC) (Neal et al., 2011) is an attractive method to sample from the joint density of the latent field $w(\mathcal{S})$. First, it allows to produce samples with low auto-correlation. Moreover, it is more all-terrain than analytical block sampling because it does not require the observations to be Gaussian. However, when the data size n increases, one is caught between the hammer and the anvil. One the one hand, the cost of one Leapfrog iteration augments. On the other hand, the Leapfrog integration step must be lowered in order to keep good acceptance rates, increasing the number of iterations in one HMC proposal.

In general, this problem is sensitive to multiplying the sampled variables by a matrix. It is well-known that when sampling a Gaussian vector x of covariance matrix Σ, working with $\Sigma^{-1/2}x$ can be much more efficient than using directly x (Neal et al., 2011). The problem is that either the posterior distribution of w is non-Gaussian, or such a transform is unaffordable. Prior whitening of the field (see for example Heinonen et al., 2016) doges this difficulty by multiplying the latent field w by $\tilde{R}$, and is equivalent to applying HMC on w^*. If w was following its prior distribution, w^* would be a white noise, hence the method's name; even if the components of w^* are not perfectly decorrelated *a posteriori*, they still are much less correlated than the components of w.

Replacing w by w^* in the HMC step changes the negated log-density and its gradient, see table 2.1 for a comparison. Note that in the lower right cell, the differentiation is done with respect to w and not w^*. It is because differentiating the likelihood of the observations with respect to w^* would be unaffordable. It is much better to use the Jacobian chain rule, with

$$-\frac{\partial\left(log(l(z|\tilde{R}^{-1}w^*, X, \beta, \ldots))\right)}{d(w^*)} = -(\tilde{R}^{-1})^T \frac{\partial\left(log(l(z|w = \tilde{R}^{-1}w^*, X, \beta, \ldots))\right)}{d(w)}.$$

Here the triangularity and sparsity of $\tilde{R}$ is critical for fast solving. As for the likelihood of the observations, using the conditional independence of the observations conditionally on the natural latent field allows to write it as:

$$log(l(z|w, X, \beta, \ldots)) = \Sigma_{i=1}^{n}\Sigma_{j=1}^{n_{obs}(s_i)} log(l(z_j(s_i)|w(s_i), X_j(s_i), \beta, \ldots)),$$

$n_{obs}(s)$ being the number of observations in the spatial site s, and "..." beign additional parameters (the Gaussian noise variance in the case of Gaussian observations for example). Differentiation with respect to w is therefore affordable. Using all those elements, a workable whitened HMC step can be devised. Algorithm 5 is a general outline, while hands-on instructions are given in algorithm 6.

Table 2.1: Negated log density and its gradient for natural and whitened NNGP field

	Natural	Whitened
Negated log-density	$1/2 w^T \tilde{R}^T \tilde{R} w$ $-log(l(z\|w, X, \beta, \ldots))$	$1/2 \Sigma_{i=1}^n (w_i^*)^2$ $-log(l(z\|\tilde{R}^{-1} w^*, X, \beta, \ldots))$
Negated log density gradient w.r.t w/w^*	$\tilde{R}^T \tilde{R} w$ $-\frac{\partial(log(l(z\|w,X,\beta,\ldots)))}{d(w)}$	$\Sigma_{i=1}^n w_i^*$ $-(\tilde{R}^{-1})^T \frac{\partial\left(log(l(z\|w=\tilde{R}^{-1}w^*,X,\beta,\ldots))\right)}{d(w)}$

Algorithm 5 HMC algorithm with whitening

input $w^t, \theta^t, \beta^t, z, X, \ldots$ ▷ More parameters such as τ can be needed

compute w^{*t} conditionally on w^t, θ^t

sample $p^t \sim \mathcal{N}(0, I_n)$ ▷ Momentum

sample $w^{*\prime}, p'$ using Leapfrog integration

Accept or reject $w^{*\prime}, p'$ with Metropolis step

if $w^{*\prime}$ is accepted **then**

$w^{*\ t+1} = w^{*\prime}$

else

$w^{*\ t+1} = w^{*t}$

end if

compute w^{t+1} conditionally on $w^{*\ t+1}, \theta^t$

return w^{t+1}

Algorithm 6 HMC algorithm with whitening, practical version

input $w^t, \theta^t, \beta^t, z, X, \ldots$ ▷ More parameters such as τ can be needed

input $\tilde{R}$

input ϵ ▷ Leapfrog step size

input L ▷ Number of Leapfrog steps

$w^{*t} = \tilde{R} w^t$ ▷ Current position for HMC

sample $p^t \sim \mathcal{N}(0, I_n)$ ▷ Current momentum for HMC

$w^{*\prime} = w^{*t}$ ▷ Proposed position

$p' = p^t$ ▷ Proposed momentum

$p' = p' - 1/2 \times \epsilon \times \left(\Sigma_{i=1}^n w_i^{*\prime}/2 - (\tilde{R}^{-1})^T \partial \left(log(l(z|w', \theta, \beta, \ldots)) \right) / d(w^{'}) \right)$ ▷ Half step for the momentum

for $i = 1 \ldots L$ **do** ▷ Leapfrog integration

  $w^{*\prime} = w^{*\prime} + \epsilon \times p$ ▷ Position change

  solve $w' = \tilde{R}^{-1} w^{*\prime}$ ▷ Retrieve sufficient parametrization

  if $i < L$ **then**

    $p' = p' - \epsilon \times \left(\Sigma_{i=1}^n w_i^{*\prime}/2 - \tilde{R}^{-1} \partial \left(log(l(z|w', \theta, \beta, \ldots)) \right) / d(w') \right)$ ▷ Full steps for the momentum

  end if

end for

$p' = p' - 1/2 \times \epsilon \times \left(\Sigma_{i=1}^n w_i^{*\prime}/2 - (\tilde{R}^{-1})^T \partial \left(log(l(z|w', \theta, \beta, \ldots)) \right) / d(w') \right)$ ▷ Half step for the momentum

sample $a \sim \mathcal{U}([0, 1])$

$\rho_1 = \Sigma_i^n (p_i')^2 - p_i^2$ ▷ Momentum log-density ratio

$\rho_2 = \Sigma_i^n (w_i^{*\prime})^2 - (w_i^{*t})^2$ ▷ Whitened GP prior log-density ratio

$\rho_3 = log(l(z|w', X, \beta, \ldots)) - log(l(z|w^t, X, \beta, \ldots))$ ▷ Observations likelihood log-density ratio

sample $a \sim \mathcal{U}([0, 1])$

if $\rho_1 + \rho_2 + \rho_3 > log(a)$ **then** ▷ Metropolis ratio

  $w^{t+1} = w'$ ▷ Accept

else

  $w^{t+1} = w^t$ ▷ Reject

end if

return w^{t+1}

2.3 Exploiting the separability of NNGP density with delayed acceptance

2.3.1 Delayed acceptance for NNGPs

Delayed Acceptance is a modification of the Metropolis-Hastings step that trades statistical efficiency of the MCMC chain for computational efficiency. Let $f(\cdot)$ be a target distribution and $q(\cdot|\cdot)$ be a proposal distribution. The usual Metropolis-Hastings ratio for the proposed parameter y is

$$\rho(x,y) = \underbrace{(q(x|y)/q(y|x))}_{\text{Proposal distribution}} \underbrace{\times(f(y)/f(x))}_{\text{Target density}}.$$

Delayed acceptance consists in splitting $\rho(x,y)$ as the product of k arbitrary positive functions $\rho_1(\cdot,\cdot),\dots,\rho_k(\cdot,\cdot)$, and to accept the proposed move with probability

$$\Pi_{i=1}^{k} min(\rho_i(y,x),1).$$

The cheaper ratios should be computed first: if the proposal fails to pass the first ratios, the following ratios, which are more expensive, are not computed (Banterle et al., 2014).

I chose to apply delayed acceptance to covariance parameter sampling for two reasons. First, a Metropolis-Hastings step is unavoidable since no analytical full conditional can be derived for covariance parameters (Datta et al., 2016). Second, Vecchia's approximation to a density can be split very easily since it is a product of small elements. This theoretical split is easy to apply in practice because the Cholesky factor of the Gaussian Process precision matrix is computed row by row (Guinness and Katzfuss, 2018), each row corresponding to one chunk of the recursive conditional likelihood.

My application of delayed acceptance to Vecchia's approximation is in the spirit of Christen and Fox (2005). They split the acceptance ratio as:

$$\rho_1(y,x) = \underbrace{(q(x/y)/q(y/x))}_{\text{Proposal distribution}} \times \underbrace{(f^*(y)/f^*(x))}_{\text{Target approximation}}$$

and

$$\rho_2(y,x) = \underbrace{(f^*(x)/f^*(y))}_{\text{Target approximation}} \times \underbrace{(f(y)/f(x))}_{\text{True target}}.$$

In the first ratio, $f^*(\cdot)$ is supposed to be a cheaper approximation to $f(\cdot)$. Hence $\rho_1(\cdot,\cdot)$ is supposed to be an approximation to $\rho(\cdot,\cdot)$. The first ratio can be interpreted as a pre-selection, and the costlier actual ratio is then computed only for the proposals that are the best according to the approximation. I propose to "taste" this ratio with the first n_1 observations, n_1 being much smaller than n. This is a classical Subset of Regressors approximation (Rasmussen and Williams, 2006). Another approach is to dispatch the n_1 observations among the n observations, at random for example. However, this approach is more complicated to implement and does not allow to use Ancillary Augmentation.

NNGP with sufficient augmentation. With a proposal distribution $q(\cdot|\cdot)$, a hyperprior distribution $h(\cdot)$, a NNGP density $\tilde{f}(\cdot)$, a latent field $w(\cdot)$, and respectively current and proposed covariance parameters θ and θ', the "usual" Metropolis ratio for NNGP with sufficient augmentation is:

$$\begin{aligned}\rho(\theta'|\theta) = \quad & \underbrace{(q(\theta|\theta')/q(\theta'|\theta))}_{\text{Proposal distribution}} \times \underbrace{(h(\theta')/h(\theta))}_{\text{Hyperprior}} \\ & \times \underbrace{(\tilde{f}(w(s_1,\ldots,s_n)|\theta')/\tilde{f}(w(s_1,\ldots,s_n)|\theta))}_{\text{NNGP density}}.\end{aligned}$$

The Delayed Acceptance ratios are then:

$$\begin{aligned}\rho_1(\theta'|\theta) = \quad & \underbrace{(q(\theta|\theta')/q(\theta'|\theta))}_{\text{Proposal distribution}} \times \underbrace{(h(\theta')/h(\theta))}_{\text{Hyperprior}} \\ & \times \underbrace{(\tilde{f}(w(s_1,\ldots,s_{n_1})|\theta')/\tilde{f}(w(s_1,\ldots,s_{n_1})|\theta))}_{\text{Beginning of the NNGP density}},\end{aligned}$$

and

$$\begin{aligned}\rho_2(\theta'|\theta) = \quad & \tilde{f}(w(s_{n_1+1},\ldots,s_n)|w(s_1,\ldots,s_{n_1}),\theta') \\ & \underbrace{/\tilde{f}(w(s_{n_1+1},\ldots,s_n)|w(s_1,\ldots,s_{n_1}),\theta)}_{\text{End of the NNGP density}}.\end{aligned}$$

Only the first n_1 rows of $\tilde{R}$ are needed to compute the first ratio.

Response NNGP. Response NNGP (Finley et al., 2019) has no latent field and computes directly the density of the Gaussian observations $z(\cdot)$. The covariance parameters integrate a nugget effect to account for the noise of the observations. The previous approach can be transposed *mutatis mutandis*.

NNGP with ancillary augmentation. Here, $l(\cdot)$ is the likelihood of the observations. The Cholesky factor of the precision matrix that is found using the NNGP is noted $\tilde{R}(\cdot)$. The whitened latent field is $w^* = \tilde{R}(\theta)w$. The "usual" Metropolis ratio for NNGP with ancillary augmentation is:

$$\begin{aligned}\rho(\theta',\theta) = \quad & \underbrace{(q(\theta|\theta')/q(\theta'|\theta))}_{\text{Proposal distribution}} \times \underbrace{(h(\theta')/h(\theta))}_{\text{Hyperprior}} \\ & \times \underbrace{(l(z|\tilde{R}^{-1}(\theta')w^*,\beta,\ldots)/l(z|\tilde{R}^{-1}(\theta)w^*,\beta,\ldots))}_{\text{Likelihood of the observations}}.\end{aligned}$$

The Delayed Acceptance ratios are:

$$\rho_1(\theta', \theta) = \underbrace{(q(\theta/\theta')/q(\theta'/\theta))}_{\text{Proposal distribution}} \times \underbrace{(h(\theta')/h(\theta))}_{\text{Hyperprior}} \times \underbrace{\frac{l(z_{1\dots n_1}|(\tilde{R}(\theta')^{-1}w^*)_{1\dots n_1}, \beta, \dots)}{l(z_{1\dots n_1}|(\tilde{R}(\theta)^{-1}w^*)_{1\dots n_1}, \beta, \dots)}}_{\text{Likelihood of the first } n_1 \text{ observations}}.$$

$$\rho_2(\theta', \theta) = \underbrace{\frac{l(z_{n_1+1\dots n}|(\tilde{R}(\theta')^{-1}w^*)_{n_1+1\dots n}, \beta, \dots)}{l(z_{n_1+1\dots n}|(\tilde{R}(\theta)^{-1}w^*)_{n_1+1\dots n}, \beta, \dots)}}_{\text{Likelihood of the last } n-n_1 \text{ observations}}.$$

Once again, only the n_1 first rows of $\tilde{R}$ need to be computed to solve $(\tilde{R}_{1\dots n_1,1\dots n_1})^{-1}w^*_{1\dots n_1} = (\tilde{R}^{-1}w^*)_{1\dots n_1}$.

2.3.2 Empirical exploration

Design. An experiment on synthetic data sets is set up to assess the contribution of delayed acceptance to computational efficiency. One case of the experiment is obtained by:

1. Simulating a spatial data set.
2. Running a NNGP model with a certain delayed acceptance strategy.
3. After 6000 iterations, store the chain and record interest variables about the computational efficiency of the run.

The observed fields are simple Gaussian fields given as $z(\cdot) = w(\cdot) + \epsilon(\cdot)$, the latent field w being an exponential field with range and marginal variance 1, and the error $\epsilon(\cdot)$ being a white noise with variance 1. Those synthetic data sets all have size 10000 and the spatial locations are drawn uniformly on a square of size 20×20.

A random ordering (Guinness, 2018) is used in order to improve the accuracy of the NNGP, implying that the n_1 first spatial points used in the delayed acceptance step occupy all the 20×20 spatial domain, but are less dense in the square than the original 10000 spatial points. The density of the observations (25 observations per unit of area) is high enough with respect to the range of the spatial process (1 unit of length) to guarantee that the subset-of-regressors approach is not a nonsense. Indeed, it is possible to fit a model with the same range but just 1 observation per unit of area.

Four cases are tested for the delayed acceptance: $n_1 = 500, 1000, 2500$, and no delayed acceptance. An ASIS algorithm (4) with automatic greedy proposal tuning in the first hundred iterations is used. There are one ancillary and one sufficient Metropolis steps in each iteration of the MCMC algorithm. Each of

those two steps is done with delayed acceptance, but not necessarily with the same n_1, so that there are 16 cases in total. Each of those 16 cases is replicated 50 times with different seeds.
The interest outcome variables are:

- the running time, that should be lower with delayed acceptance.
- the Effective Sample Size (ESS) is a measure of the statistical efficiency and should be lower with delayed acceptance (Christen and Fox, 2005).
- the ESS per time as a measure of computational efficiency. If delayed acceptance is beneficial, it should increase.

Results. To assess the effect of delayed acceptance on the interest variables, I treated n_1 as a factor and used linear models with interactions. A summary of the models is presented in table 2.2.

As expected, both the running time and the ESS are generally lower when using delayed acceptance. Only three cases are out of this pattern, at rows $11, 14$, and 15 of the table. Those three rows correspond to interactions between sufficient and ancillary augmentation. In those three cases, the ESS drops dramatically but the running time strongly increases. They correspond to cases where the chains mix poorly and may even stray, explaining the drop of ESS. The increase of time could be explained by a failure of the pre-acceptance step to act as a filter.

As for the ESS per time, it appears that it is not affected greatly by delayed acceptance in general, but the three pathological cases obviously cause a serious deterioration. I did not push forward on this lead whose preliminary results were not very encouraging.

The fact that I use NNGP with full data augmentation (the latent field $w(\cdot)$ is sampled within the Gibbs sampler) might explain the mediocrity of delayed acceptance's performance. A loss of statistical efficiency in the sampling of the covariance parameters actually affects the mixing of all the parameters of the model. On the other hand, the gain in time only affects the corresponding step. In a more frugal model, such as response NNGP (Finley et al., 2019), delayed acceptance might perform better than in the full NNGP model I used.

Table 2.2: Results of the Delayed Acceptance experiment.

Case		ESS/time Est.[1]	ESS/time Pr(> |t|)	ESS Est.	ESS Pr(> |t|)	time Est.	time Pr(> |t|)
(Intercept)[2]		3.31	0.00	96.50	0.00	29.14	0.00
suff. 500		-0.14	0.24	-20.18	0.00	-5.05	0.00
suff. 1000		-0.01	0.91	-19.33	0.00	-5.76	0.00
suff. 2500		-0.18	0.14	-22.48	0.00	-5.54	0.00
anc. 500		0.10	0.42	-8.98	0.00	-3.51	0.00
anc. 1000		0.37	0.00	-5.91	0.03	-4.56	0.00
anc. 2500		0.36	0.00	-8.03	0.00	-5.06	0.00
suff. 500,	anc. 500	0.40	0.02	6.17	0.10	-0.52	0.00
suff. 1000,	anc. 500	0.46	0.01	6.35	0.09	-0.52	0.00
suff. 2500,	anc. 500	0.63	0.00	11.20	0.00	-0.36	0.01
suff. 500,	anc. 1000	-2.63	0.00	-48.44	0.00	4.55	0.00
suff. 1000,	anc. 1000	0.33	0.06	0.79	0.83	-0.79	0.00
suff. 2500,	anc. 1000	0.48	0.01	5.40	0.15	-0.60	0.00
suff. 500,	anc. 2500	-2.61	0.00	-48.16	0.00	2.87	0.00
suff. 1000,	anc. 2500	-2.71	0.00	-47.96	0.00	4.05	0.00
suff. 2500,	anc. 2500	0.44	0.01	2.91	0.44	-1.01	0.00

[1] "Est." stands for "Estimate".
[2] The reference case is no delayed acceptance

Chapter 3

Chromatic sampling and fixed effect ASIS applied to NNGP models.

The chapter consists in an article that presents early attempts to improve the MCMC algorithm of Datta et al. (2016). It currently is under review (first round) in *Computational Statistics and Data Analysis*. Two methods are presented.

The first method aims to solve a problem that caused much trouble in my early implementations of NNGP models: poor mixing of the linear regression coefficients associated to the fixed effects. It was not difficult to stumble on the issue because the intercept's coefficient has an especially bad MCMC behavior. A re-parametrization of the model, where the latent Gaussian field is centered around the intercept instead of 0, is enough to solve the problem. While it is well-known that linear transformation of the variables may dramatically improve the behavior of a Gibbs sampler algorithm, the question that remains is why centering the latent field on the intercept works better. Theoretical exploration of a simplified field-intercept model reveals that the two parametrizations of the intercept will behave in opposite ways: when either does good, the other will fail. It is impossible to give a rigorous quantification of what is going on due to the fact that GP densities generate some intractable terms. However, some reasoning "with the hands" allows to link the behavior of the intercept with the fact that Gaussian Processes induce some spatial coherence.

The problem that arose with the intercept sometimes happens with other covariates. A good point is that thanks to the sparsity that is induced by NNGPs, centering the latent field on the incriminated fixed effects is computationally affordable. However, instead of proposing to pick the effects and center them, I advocate in favor of an interweaving of the two parametrizations. The first reason why is that interweaving generally is very efficient, even when none of

the two interweaved steps are. The second is that there is some incertitude about which parametrization should be chosen. In general, covariates that have some spatial coherence will cause trouble, while the others are well behaved. Worse, centering a linear effect that should not be centered will deteriorate the behavior of the associated regression coefficient. Testing the variables one by one by spatial correlation analysis or preliminary run would be tedious and negate the benefit of the approach. Instead of spending time thinking about which parametrization should be used, I prefer to round all eligible covariates (including the intercept) up in a common sampling step and forget about them.

The limit of this solution is that it cannot be applied to covariates that vary within one spatial site (e.g. people living in the same household may smoke or not, but the presence or not of asbestos in the building affects all of them). This limit must be tempered in practice because NNGP methods are applied on point-measurement data sets: therefore, all regressors obtained through grids or areas are immediately eligible.

The other method presented in the article is a Chromatic sampler that uses the sparsity induced by NNGPs in order to improve the sampling of the latent field. It is a continuation of the approach of Datta et al. (2016), in the sense that it aims to transpose the parallelisability of the computation of the NNGP density to the sampling of the latent field.

While Peruzzi et al. (2020) force a chromatic behavior on the data through a mesh of knots, the approach presented here is to adapt chromatic sampling to existing NNGPs. The two methods have their own merits. Meshed NNGPs are a great tool in cases where the Nearest Neighbor heuristic is ambiguous or difficult to apply, such as multivariate data. However, "vanilla" NNGPs are time-tested, and their accuracy is guaranteed by standard empirically tested heuristics (Guinness, 2018).

Empirical exploration showed that the method is successful in gathering the sampling operations into very few, very large groups, while remaining cheap using simple greedy coloring. Even though the heuristics of construction for NNGPs (Guinness, 2018) do affect the number of groups, it remains stable with the size of the data set, and there was no problematic case. The behavior of NNGP moral graphs with respect to coloring is surprising and arises curiosity. I did my best to explain this behavior, even though much more work would be needed to go to the bottom of things.

What chromatic samplers cannot do is to solve the problems of auto-correlation of NNGPs with full data augmentation. They must be used in the framework of efficient MCMC architectures, such as those that were benchmarked by Filippone et al. (2013), see section 2.2.3 of the thesis for the implementation details for NNGP models. On the other hand, chromatic samplers have the very positive property to be usable with any type of data and not only Gaussian data. Therefore, they perfectly fitted in the approach to find robust and all-terrain methods as basic elements for more complex methods. In particular, they were an interesting lead for complex sampling involved in non-stationary model fitting, even though in the end Hamiltonian methods worked better.

Improving performances of MCMC for Nearest Neighbor Gaussian Process models with full data augmentation

Sébastien Coube-Sisqueille[1,a] and Benoît Liquet[1,2,b]

Abstract

Even though Nearest Neighbor Gaussian Processes (NNGP) alleviate considerably MCMC implementation of Bayesian space-time models, they do not solve the convergence problems caused by high model dimension. Frugal alternatives such as response or collapsed algorithms are an answer. Our approach is to keep full data augmentation but to try and make it more efficient. We present two strategies to do so.

The first scheme is to pay a particular attention to the seemingly trivial fixed effects of the model. We show empirically that re-centering the latent field on the intercept critically improves chain behavior. We extend this approach to other fixed effects that may interfere with a coherent spatial field. We propose a simple method that requires no tuning while remaining affordable thanks to the sparsity of NNGPs.

The second scheme accelerates the sampling of the random field using Chromatic samplers. This method makes long sequential simulation boil down to group-parallelized or group-vectorized sampling. The attractive possibility to parallelize NNGP density can therefore be carried over to field sampling.

We present a R implementation of our methods for Gaussian fields in the public repository
`https://github.com/SebastienCoube/Improving_NNGP_full_augmentation`.
An extensive vignette is provided. We run our implementation on two synthetic toy examples along with the state of the art package `spNNGP`. Finally, we apply our method on a real data set of lead contamination in the United States of America mainland.

Chromatic Sampler; Interweaving; Nearest Neighbor Gaussian Process; Space-time models

---

[1]Laboratoire de Mathématiques et de leurs Applications, Université de Pau et des Pays de l'Adour, UMR CNRS 5142, E2S-UPPA, Pau, France

[2]Department of Mathematics and Statistics, Macquarie University, Sydney

[a]sebastien.coube@univ-pau.fr

[b]benoit.liquet@univ-pau.fr

Funding was provided by the Energy Environment Solutions (E2S-UPPA) consortium and the BIGCEES project from E2S-UPPA ("Big model and Big data in Computational Ecology and Environmental Sciences")

3.1 Introduction

Many social or natural phenomena happen at the scale of a territory and must be observed at various sites and possibly times. The rise of modern GPS and Geographic Information Systems made large and high-quality point-referenced data sets more and more available. Assume that, in a collections of sites $\mathcal{S}$ of the space or space-time domain $\mathcal{D}$, we have measurements $z(\cdot)$ with some kind of space or space-time coherence. This coherence can be accounted for by introducing a spatially-indexed process $w(\cdot)$ that has a well-defined joint distribution on any finite subset of the domain. We consider a Gaussian model where the observations $z(\cdot)$ have been perturbed by a Gaussian noise ϵ of standard deviation τ. Many models also add linear regression on covariates $X(\cdot)$, giving the following classical model formulation

$$z(s) = \beta_0 + X(s)\beta^T + w(s) + \epsilon(s), s \in \mathcal{S}. \tag{3.1}$$

In order to keep notations shorter, for any collection of spatial locations $\mathcal{P} \subset \mathcal{S}$, we denote the vector $\{w(s) : s \in \mathcal{P}\}$ as $w(\mathcal{P})$. Gaussian processes (GP) make an elegant prior distribution for $w(\cdot)$, see Gelfand et al. (2010). The GP prior distribution of $w(\mathcal{S})$ is $\mathcal{N}(\mu, \Sigma)$. The mean parameter of $w(\cdot)$ is usually fixed to $\mu = 0$ to avoid identification problems with the linear regression intercept β_0. The covariance matrix is computed using a positive definite function $k(\cdot)$ with covariance parameters θ, such as Matérn's covariance and its exponential and squared-exponential special cases. It can then be written as $\Sigma(\mathcal{S}, \theta)$, and its entries are $\Sigma(\mathcal{S}, \theta)_{i,j} = k(s_i, s_j, \theta)$. We denote $f(\cdot|\mu, \Sigma)$ the GP density, and we abbreviate it as $f(\cdot|\mu, \theta)$. The covariance parameters can have modeller-specified hyperpriors developed in Fuglstad et al. (2015); Datta et al. (2016).

The weakness of GPs is that computing the prior density of $w(\mathcal{S})$ involves the determinant and inverse of $\Sigma(\mathcal{S}, \theta)$, incurring a computational cost that is cubic in the size of $\mathcal{S}$. Vecchia's approximation to Gaussian likelihoods received increased attention the past years, with theoretical developments of Katzfuss and Guinness (2017); Guinness (2018); Datta et al. (2016); Finley et al. (2019) and software presented in Guinness and Katzfuss (2018); Finley et al. (2017). The Nearest Neighbor Gaussian Process (NNGP) is a special case of Vecchia's approximation that provides a surrogate of the inverse Cholesky factor of Σ and uses it to approximate GP prior density. It starts by finding an ordering for the n locations of $\mathcal{S}$ which we will denote $(s_1, \ldots, s_n)$. The ordering may have an impact on the quality of the approximation, and is discussed in Datta et al. (2016); Guinness (2018). The joint latent density of $w(s_1, \ldots, s_n)$ is then written under the recursive conditional form

$$f(w(s_1, \ldots, s_n)|\mu, \theta) = f(w(s_1)|\mu, \theta) \times \Pi_{i=2}^{n} f(w(s_i)|w(s_1, \ldots, s_{i-1}), \mu, \theta).$$

Since $f(w(s_1, \ldots, s_n)|\mu, \theta))$ is a Multi-Variate Normal (MVN) distribution function, the conditional density $f(w(s_i)|w(s_1, \ldots, s_{i-1}), \mu, \theta), i \in 2, \ldots, n$ is a Normal as well. A NNGP is obtained by replacing the vector $w(s_1, \ldots, s_{i-1})$ that conditions $w(s_i)$ by a much smaller parent subset denoted $w(pa(s_i))$ for each

conditional density. The NNGP approximation to the GP prior joint density of $w(\cdot)$ is defined as

$$\tilde{f}(w(s_1,\ldots,s_n)|\mu,\theta) = f(w(s_1)|\mu,\theta) \times \Pi_{i=2}^{n} f(w(s_i)|w(pa(s_i)),\mu,\theta). \quad (3.2)$$

This very general principle can be applied to any kind of well-defined multivariate density. However, as far as we know, MVN density approximation is the only application. This may be explained by the fact that non-Gaussian data can be handled with GP modeling using link functions. Moreover, a NNGP defines a MVN density and it is possible compute explicitly and easily the sparse Cholesky factor of the precision matrix. The choice of the parents is critical but no universal criterion exists. A popular choice is to pick s_i's nearest neighbors among $(s_1,\ldots,s_{i-1})$, explaining the denomination "Nearest Neighbors Gaussian Process" given in Datta et al. (2016). However, Datta et al. (2016); Stein et al. (2004) argue that mixing close and far-away observations can improve the approximation. This approximation is cheap and easily parallelisable. The latent density (3.2) can be split into small jobs and dispatched to a cluster of calculators (Datta et al., 2016). Its cost is linear in the number of observations under the condition that the size of each parent set is bounded. More advanced strategies exist such as grouping, proposed by Guinness (2018).

If NNGPs work around the bottleneck of GP likelihood computation, they do not solve the problem of slow MCMC convergence. In Datta et al. (2016), the Gibbs sampler loops over θ, $w(\mathcal{S})$ and β, μ is fixed to 0. The latent field $w(\mathcal{S})$ is updated sequentially or by blocks. This sampler suffers from slow mixing, in particular when n increases. Other strategies have been proposed by Finley et al. (2019) that precisely avoid to sample the field in order to reduce the dimension of the model. Yet another method (Finley et al., 2019; Zhang et al., 2019) is to use convenient conjugate distributions for models where the range of $w(\cdot)$ and the variance ratio of $w(\cdot)$ and $\epsilon(\cdot)$ is fixed, and select the fixed parameters by cross-validation. Our approach is nevertheless to improve implementations of NNGP models where the latent field is explicitly sampled. Our first reason is that there may be situations where some of the methods presented in Finley et al. (2019) perform poorly while full data augmentation works well. For example, the *collapsed NNGP* of Finley et al. (2019) enjoys low dimensionality and allows nonetheless to retrieve the latent field, but demands Cholesky factorization of large sparse matrices which may be unfeasible depending on n and the dimension of $\mathcal{D}$. The *Response NNGP* of Finley et al. (2019) retrieves the covariance parameters θ but not the latent field $w(\mathcal{S})$. Our second reason is that efficient Gibbs sampler architectures can sharply improve mixing. A NNGP defines a Markov Random Field, allowing to use the blocking methods of Knorr-Held and Rue (2002). The sparse Cholesky factor in a NNGP makes it possible to use the Ancillary-Sufficient Interweaving Strategy (AS-IS) presented in Yu and Meng (2011). The third reason is that full latent field sampling is all terrain, and can address many data models or be plugged into complex, non-stationary models like Heinonen et al. (2016), while collapsed MCMC or conjugate models are much pickier.

Here is an outline of the article. Section 3.2 focuses on the seemingly trivial linear effects of the hierarchical model. In 3.2.1 we propose a mild but efficient centering of the latent field on the least squares regression intercept. In 3.2.2, we extend centering to other linear effects, and we use interweaving from Yu and Meng (2011) to propose a robust, tuning-less application. Section 3.3 targets the simulation of the random field. In 3.3.1, we propose to use the chromatic samplers developed by Gonzalez et al. (2011) in order to carry the attractive parallelizability of NNGP density over to field sampling. In 3.3.2, we analyze the sensitivity of NNGP graph coloring and we benchmark coloring algorithms. We apply our methods in section 3.4. We present our implementation (available at `https://github.com/SebastienCoube/Improving_NNGP_full_augmentation`) in 3.4.1. We test our implementation along with the state of the art package `spNNGP` presented in Finley et al. (2017) on synthetic toy examples in 3.4.2. In 3.4.3, we present an application on lead contamination in the mainland of the United States of America. The article ends by a discussion in Section 3.5.

3.2 Latent field centering

3.2.1 Centering the latent field on the intercept

The mean parameter μ of the prior density for the latent field $w(\cdot)$ is usually set to 0 in order to avoid identification problems with the intercept β_0. We call this formulation standard, since it is found in state of the art papers such as Datta et al. (2016); Finley et al. (2019). We name samples of the standard formulation $w_s(\cdot)$. Our proposal is to replace $w_s(\mathcal{S})$ by a centered $w_c(\mathcal{S}) = w_s(\mathcal{S}) + \beta_0$ in the Gibbs Sampler. This substitution is a non degenerate linear transform that keeps the model valid, while keeping the possibility to transform the samples back to standard parametrization if needed. The centered parametrization can also be seen as a slightly different model, with (3.1) becoming

$$z(s) = X(s)\beta^T + w_c(s) + \epsilon(s), s \in \mathcal{S}, \tag{3.3}$$

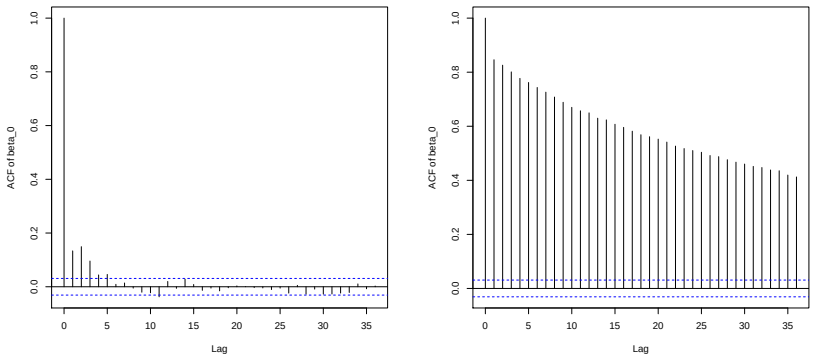
and the prior density of $w_c(\mathcal{S})$ becoming

$$\tilde{f}(w_c(\mathcal{S})|\mu = \beta_0, \theta).$$

Those changes impact the full conditional distributions. Table 3.1 summarizes the changes in a Gibbs sampler for a Gaussian model found in Datta et al. (2016). We denote $f(\cdot|\cdot, \cdot)$ the normal density function, and $\tilde{Q}$ the latent field's precision matrix defined by the NNGP. We abbreviate the interest variables $X(\mathcal{S})$ as X. We denote the vector made of n times 1 as $\mathbf{1}$. The matrix obtained by adding $\mathbf{1}$ to the left side of X is named $[\mathbf{1}|X]$. We did not feature prior distributions on the high-level parameters like θ, τ or β: their full conditionals would not be affected since centering changes only the NNGP prior and the observed data likelihood. Even if the modification is minor, the improvement

Table 3.1: Changes in the full conditional distributions

Variable	Standard	Centered
β_0		$f(\beta_0, (\mathbf{1}^T\tilde{Q}\mathbf{1})^{-1}(\mathbf{1}^T\tilde{Q}w_c), (\mathbf{1}^T\tilde{Q}\mathbf{1})^{-1})$
β		$f(\beta, (X^TX)^{-1}(X^T(z-w_c))), \tau^2(X^TX)^{-1})$
(β_0, β)	$f(\beta, ([\mathbf{1}\|X]^T[\mathbf{1}\|X])^{-1}([\mathbf{1}\|X]^T(z-w_s))),$ $\tau^2([\mathbf{1}\|X]^T[\mathbf{1}\|X])^{-1})$	
θ	$\tilde{f}(w_s(\mathcal{S})\|0, \theta)$	$\tilde{f}(w_c(\mathcal{S})\|\beta_0, \theta)$
τ	$\Pi_{s\in\mathcal{S}} f(z(s)\|w_s(s) + \beta_0 + X(s)\beta^T, \tau)$	$\Pi_{s\in\mathcal{S}} f(z(s)\|w_c(s) + X\beta^T, \tau)$
$w(s),$	$\tilde{f}(w_s(s)\|w_s(\mathcal{S}\backslash s), 0, \theta)$	$\tilde{f}(w_c(s)\|w_c(\mathcal{S}\backslash s), \beta_0, \theta)$
$s \in \mathcal{S}$	$f(z(s)\|w_s(s) + \beta_0 + X(s)\beta^T, \tau)$	$f(z(s)\|w_c(s) + X(s)\beta^T, \tau)$

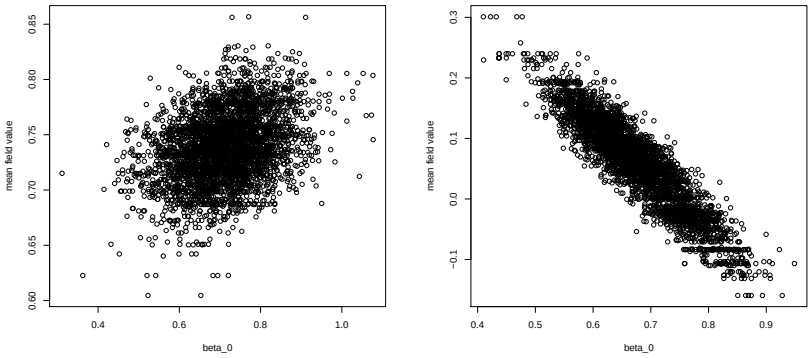


(a) Centered model (b) Standard model

Figure 3.1: The ACF of β_0 drops much faster in the centered model than in the standard model

in the mixing of the intercept is clear. We simulated a little toy example with 1000 observations and we ran the two Gibbs samplers. The autocorrelation plots (Figure 3.1) are clearly in favor of the centered formulation. Even though for this toy example the standard model mixes after a few hundred iterations, this is not the case for larger data sets. We observed empirically that there is much more correlation between $w(\mathcal{S})$ and β_0 in the standard implementation. Plotting $\frac{1}{n}\Sigma w(\mathcal{S})$ against β_0 (Figure 3.2) displays a clear ridge in the case of the standard model (Figure 3.2b). This means that the whole latent field has to shift upwards and downwards for the intercept to explore its posterior distribution. Ridge-like densities are a well known plague of Gibbs samplers, and linear recombination is one of the tools to get rid of it, see Gelfand et al. (1995) and Robert and Casella (2004).

The behavior of the toy example arises from the fact that the fraction of β_0^t that is carried over in w^{t+1} and β_0^{t+1} changes following the model. Take a simpler spatial model where only an intercept and the Gaussian latent field are estimated, while the Gaussian noise variance τ^2 and the NNGP precision $\tilde{Q}$ are known. The intercept coefficient has an improper constant prior. Assume that



(a) Centered model (b) Standard model

Figure 3.2: When plotting $\frac{1}{n}\Sigma w(\mathcal{S})$ against β_0, the standard model exhibits a ridge-shaped point cloud

the latent field is sampled in one step (which is usually not the case unless the data size is very small).
Denote the diagonalization $\tilde{Q} = V^T \lambda V$, V being a square matrix of eigenvectors and λ being a diagonal matrix of eigenvalues. The eigenvalues are positive since $\tilde{Q}$ is defined as the cross-product of two NNGP factors (Datta et al., 2016; Katzfuss and Guinness, 2017). Note α_i, $i = 1, \ldots, n$ the coordinates of the vector $\mathbf{1} = (1, \ldots, 1)$ in the orthonormal basis V. The following results are proved in 3.6.
The first result is that a fraction of β_0 is carried over in the empirical mean of the latent field. Note $\rho = \Sigma_{i=1}^n \alpha_i^2 (\tau^2 \lambda + I_n)_{i,i}^{-1} / n$. Then,

$$\overline{w_s^{t+1}} = \mu_s - \rho\beta_0^t + \epsilon_s^{t+1} \quad \text{and} \quad \overline{w_c^{t+1}} = \mu_c + (1-\rho)\beta_0^t + \epsilon_c^{t+1},$$

μ_s and μ_c being fixed, ϵ_s and ϵ_c being stochastic innovations, and t being the index of the iteration. Moreover, we have .

$$0 \leq \rho \leq 1$$

Using this result, we can see that if a high fraction of β_0 is carried over in the mean of the standard latent field, then a low fraction of β_0 will be carried over in the mean of the centered latent field, and conversely. This is clearly what can be seen in figure 3.2.

The next question is why ρ is closer to 1 than to 0. This point is difficult to clarify because there is no analytic expression for terms where $\tilde{Q}$ is involved. For example, the range parameters are updated through a Metropolis step in Datta et al. (2016) because a full conditional draw is challenging. However, we can start from $\rho = \Sigma_{i=1}^n \alpha_i^2 (\tau^2 \lambda + I_n)_{i,i}^{-1} / n$ and make a deduction: if the sum is high, then $\lambda_{i,i}$ is small when α_i is big; and conversely $\lambda_{i,i}$ is big when α_i is small. We

can re-formulate: $\lambda_{i,i}^{-1}$ is big when α_i is big. Now, remark that V and λ^{-1} respectively are the spatial basis and coefficients of the Karhunen-Loève decomposition of the NNGP prior. This means that α_i is high for the first components of the decomposition, where λ^{-1} is the highest. In other terms, $\mathbf{1}$ "resonates" with the first spatial basis functions of the Karhunen-Loève decomposition. This conclusion is consistent with the fact that $\tilde{Q}$ parametrizes a spatially coherent latent field, inducing that a few spatial basis functions are enough to describe most of w. For example, in the Predictive Process model of Banerjee et al. (2008), w is approximated by a degenerate process with a low-rank covariance matrix. Similarly, Gelfand et al. (2010) report that "it is generally the case that the Empirical Orthogonal Functions (EOF) associated with the largest eigenvalues [of the Karhunen-Loève decomposition] represent larger-scale spatial variation, and conversely, the EOFs associated with the smallest eigenvalues correspond to smaller-scale spatial variation".

Now, let's focus on the expressions of $[\beta_0^{t+1}|\beta_0^t]$. Like the mean of the latent field, they can be expressed with a fixed part, a geometric carry-over, and an innovation. In the standard model, a fraction ρ of β_0^t is carried over. We already discussed this quantity. As for the centered model, the fraction of β_0^t which is conserved in β^{t+1} is

$$\Sigma_{i=1}^n \left((\alpha_i^2\lambda_{i,i})(\tau^2\lambda_{i,i})/(\tau^2\lambda_{i,i}+1)\right) / \Sigma_{i=1}^n \alpha_i^2\lambda_{i,i}.$$

Once again, the geometric term is between 0 and 1 since

$$0 \leq (\tau^2\lambda_{i,i})/(\tau^2\lambda_{i,i}+1) \leq 1.$$

Like before, suppose that α is big when λ is small. Then, when α_i is the largest, $(\tau^2\lambda_{i,i})/(\tau^2\lambda_{i,i}+1)$ will be much smaller than 1, resulting in the geometric carry-over term $\Sigma_{i=1}^n \left((\alpha_i^2\lambda_{i,i})(\tau^2\lambda_{i,i})/(\tau^2\lambda_{i,i}+1)\right)$ being much smaller than $\Sigma_{i=1}^n \alpha_i^2\lambda_{i,i}$. Therefore, we can expect a small proportion of β_0^t to remain in β_0^{t+1}.

3.2.2 Adaptation to other fixed effects through interweaving

Field centering can be extended to other fixed effects. In most cases it is unnecessary because centering and scaling $X(\mathcal{S})$ is enough to considerably improve chain behavior. Even worse, the Gibbs sampler usually behaves very bad if the random field is centered on other fixed effects than the intercept. There are nonetheless cases where bad mixing of the regression coefficients happens again. In this case, it is often useful to try and center $w(\cdot)$ not only on the intercept but also on the troublesome covariates' fixed effect. However, doing preliminary runs and picking manually which fixed effects the field needs to be centered on would be tedious.

Interweaving, introduced by Yu and Meng (2011), combines the advantages of the two strategies and removes the need to choose. The method takes advantage of the discordance between two parametrizations to construct the following

step:

$$[w_2|w_1] \to [\theta^{t+1}|w_2],$$

w_1 and w_2 being two data augmentations and θ the parameter. Usually, it is complicated to sample directly $[w_2|w_1]$. Drawing an intermediary $\theta^{t+0.5}$ gives

$$[w_1|\theta^t] \to [\theta^{t+0.5}|w_1] \to [w_2|\theta^{t+0.5}, w_1] \to [\theta^{t+1}|w_2].$$

It is possible that $[w_2|\theta, w_1]$ is a deterministic transformation, giving a degenerate joint distribution. Note that interweaving is not alternating: an alternating scheme would be $[w_1|\theta^t] \to [\theta^{t+0.5}|w_1] \to [w_2|\theta^{t+0.5}] \to [\theta^{t+1}|w_2]$. The strategy is usually very efficient if the two parametrizations are an ancillary-sufficient couple, giving an Ancillary-Sufficient Interweaving Strategy (ASIS), and can even be efficient when none of the two augmentations performs well when implemented separately. Algorithm 7 presents the steps to update the regression coefficients with interweaving. The two parameterizations are w which is uncentered and corresponds to the standard parametrization w_s, and v, which is centered on all the fixed effects and is a generalization of w_c. For the sake of simplicity, we suppose that there is only one measurement of X per spatial location and we use an improper constant joint hyperprior on (β_0, β). The parameters that depend on the state in the Gibbs sampler are indexed by t. If the observations were not Gaussian, the second "simulate" step would be left unchanged while the first "simulate" step would be adapted just like in any generalized NNGP model Datta et al. (2016).

There are two limitations to this approach. The first is the case where several measurements of the interest variable $z(\cdot)$ and the regressors $X(\cdot)$ are done at the same spatial location. The model must be extended as

$$z(s,i) = X(s,i)\beta^T + w(s) + \epsilon(s,i), s \in \mathcal{S}, 1 \leq i \leq n_{obs}(s),$$

$n_{obs}(s) \geq 1$ being the number of observations in the site s. In this setting, some variables vary within one spatial locations while other do not. For example, the presence of asbestos in buildings may be considered as a location-wise regressor while smoking is an observation-wise regressor. If the regressors vary within one location, it is impossible to center the field on the corresponding fixed effects. This would mean that the normal random variable $w(s)$ has several mean parameters at the same time. However, it is still possible to restrict interweaving to the regression coefficients associated to the location-wise variables. Our implementation allows one to specify which regressors are associated to spatial location and which are associated to individual measurements. A NNGP being a point-measurement model, regressors obtained through gridded and areal data are immediately eligible for this method.

The second limitation is the computational cost. With improper constant prior, the centered regression coefficients follow a MVN distribution whose mean and variance need to be computed at each update of θ. The sparsity induced by Vecchia's approximation is critical for the feasibility of the method because it ensures that matrix multiplications involving $\tilde{Q}$ are affordable. Using a sparse

matrix formulation for X could further alleviate this operation if X has dummy variables or null measurements.

Algorithm 7 Regression coefficient updating with interweaving

input $\tilde{Q}^t, w^t, X, \beta^t, \beta_0^t, \tau^t$
simulate $\beta_0^{t+.5}, \beta^{t+.5}$ following $\mathcal{N}(([\mathbf{1}|X]^T[\mathbf{1}|X])^{-1}([\mathbf{1}|X]^T z), \tau^2([\mathbf{1}|X]^T[\mathbf{1}|X])^{-1})$
$v = w^t + X(\beta^{t+.5})^T$
simulate $\beta_0^{t+1}, \beta^{t+1}$ following $\mathcal{N}(([\mathbf{1}|X]^T\tilde{Q}^t[\mathbf{1}|X])^{-1}([\mathbf{1}|X]^T\tilde{Q}^t v), ([\mathbf{1}|X]^T\tilde{Q}^t[\mathbf{1}|X])^{-1})$
$w^{t+0.5} = v - X(\beta^{t+1})^T$
return $\beta_0^{t+1}\beta^{t+1}, w^{t+0.5}$

3.3 Chromatic sampler for Nearest Neighbor Gaussian Process

3.3.1 Chromatic samplers and how to apply them to NNGP

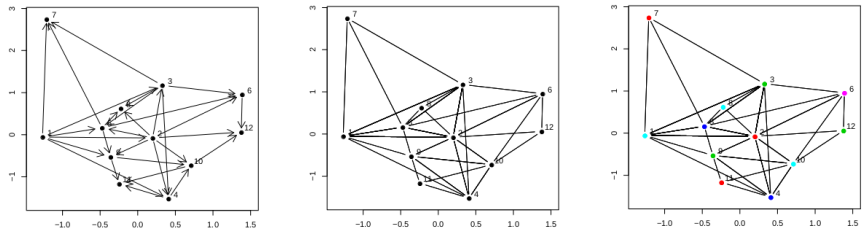
In a Gibbs sampler, the parameters of a model are updated sequentially. If a set of variables happens to be mutually independent conditionally on the other variables of the model and are updated consecutively by the Gibbs algorithm, their sampling can be parallelized. Let's consider a Gibbs sampler or a Metropolis-Within-Gibbs aiming to sample from a joint multivariate distribution $f(x_1, \ldots, x_n)$.

$$\begin{aligned}
x_1^{t+1} &\sim f(x_1|x_2^t, \ldots, x_n^t) \\
&\ldots \\
x_i^{t+1} &\sim f(x_i|x_1^{t+1}, \ldots, x_{i-1}^{t+1}, x_{i+1}^t, \ldots, x_n^t) \\
&\ldots \\
x_n^{t+1} &\sim f(x_n|x_1^{t+1}, \ldots, x_{n-1}^{t+1}).
\end{aligned}$$

Let's introduce $p \leq n$ vectors $X_1, \ldots, X_p$ so that $(x_1, \ldots, x_n) = (X_1, \ldots, X_p)$, and suppose that $\forall X \in X_1, \ldots, X_p$, either X has only one element or the elements of X are conditionally independent given the other variables. The Gibbs sampler can then be re-written

$$\begin{aligned}
x_i^{t+1} \in X_1 &\sim f(x_i|X_2^t, \ldots, X_p^t) \\
&\ldots \\
x_i^{t+1} \in X_j &\sim f(x_i|X_1^{t+1}, \ldots, X_{j-1}^{t+1}, X_{j+1}^t, \ldots, X_p^t) \\
&\ldots \\
x_i^{t+1} \in X_p &\sim f(x_i|X_1^{t+1}, \ldots, X_{p-1}^{t+1}).
\end{aligned}$$

Since all elements from X_j are simulated from independent densities, it is possible to parallelize their sampling.



(a) Directed Acyclic Graph (b) Moral Graph (c) Colored Moral Graph

Figure 3.3: Moralization and coloring of a DAG

A NNGP is defined on a Directed Acyclic Graph (DAG) by Datta et al. (2016), see Katzfuss and Guinness (2017) for discussion about other Vecchia approximations. Then, using the argument of recursive kernel factorization given in Lauritzen (1996), it has the Markov properties on the moral graph obtained by un-directing the edges and "marrying" the parents in the DAG (Figure 3.3). Graph vertex coloring associates one color to each node of a graph while forbidding that two connected nodes have the same color, just like coloring a map while forbidding that two countries that share a border have the same color. Using inductively the Global Markov property, it is possible to guarantee mutual conditional independence for the variables or the blocks that correspond to vertices sharing the same color. Chromatic sampling can be applied straightforwardly to the Gibbs sampler presented in Datta et al. (2016). It also allows to compute normalizing constants and can be combined with the covariance parameter blocking proposed by Knorr-Held and Rue (2002).

Chromatic samplers can be applied to blocked sampling as well. This method consists in updating the latent field in various locations at once. Chromatic sampling is a special case of blocked sampling, because in general there is no conditional independence within one block. Precisely, sampling the latent field jointly in a region of the domain reduces the negative impact of spatial autocorrelation on the behavior of MCMC chains. Blocked sampling may be applied to the latent field alone (Datta et al., 2016) or improve both covariance parameters and field sampling in Knorr-Held and Rue (2002). Even though there is no conditional independence within one block, there is some conditional independence between the blocks as long as there is no edge between any pair of their respective vertices, allowing for chromatic sampling. The matrix $B^T AB$ indicates the connections between the blocks, A being the adjacency matrix of the NNGP latent field's Markov graph, and B a vertex-block indicator matrix ($B_{i,j} = 1$ if vertex i belongs to block j).

3.3.2 Coloring of NNGP moral graphs: sensitivity analysis and benchmark of the algorithms

Coloring the moral graph $\mathcal{G}^m$ is a critical step in chromatic sampling and determines the attractiveness of the method with respect to the "vanilla" versions of the algorithms (one-site sequential sampling or blocked sampling with several blocks). We focus on two variables to summarize the efficiency of chromatic sampling:

- The number of colors: the smaller this number, the fewer the number of steps in the chromatic sampler.
- The time needed for coloring, that must be small with respect to the running time of the MCMC chains.

This section has two objectives. The first is to test the sensitivity of those two interest variables to the properties of $\mathcal{G}^m$ and the coloring algorithm using variance-based sensitivity analysis. The second objective is to benchmark various coloring algorithms and find a rule to choose the algorithm. We test various factors that may change the structure of $\mathcal{G}^m$:

- Size n.
- Number of parents in the DAG m.
- Spatial domain dimension d.
- Ordering of the points.

We also test 3 coloring algorithms, given in detail in 3.7.1:

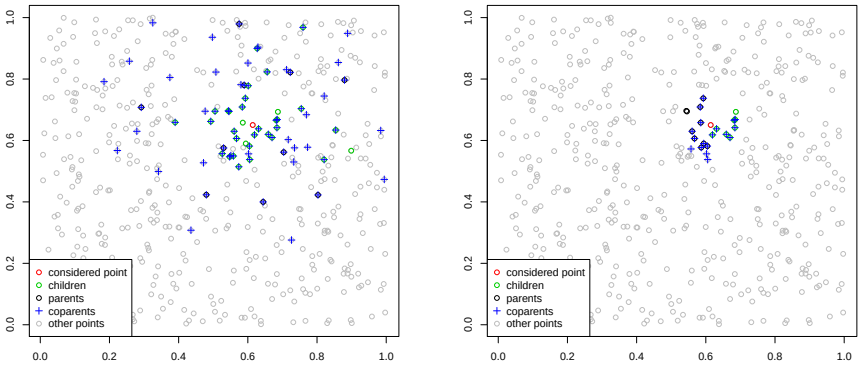
- Naive greedy coloring: coloring each vertex with the smallest available color.
- Degree greedy coloring: reorder the vertices following their number of neighbors, and apply naive greedy coloring.
- DSATUR heuristic: color the node that has the highest number of distinct colors among its neighbors (*Degree of SATURation*), and break ties using the number of neighbors.

The full results of the experiments are given in 3.7.2, and the sensitivity analyses are summarized in table 3.2.

Pilot experiment

Design. The objective is to do preliminary sensitivity analysis and benchmark on small graphs. We test the three coloring algorithms and graphs with the following attributes, each case being replicated 10 times:

- Graph size $n = 500, 1000, 2000$.



(a) Max-min ordering
(the considered point is the 30^{th} of 500)

(b) Coordinate ordering

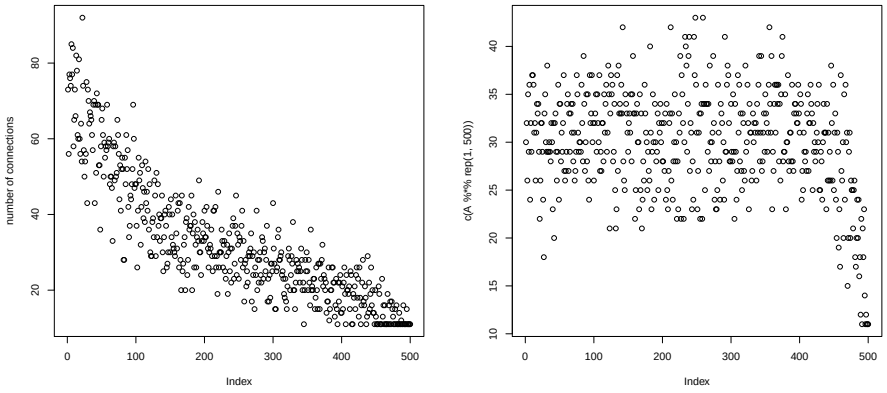
Figure 3.4: Connections of the same point, with two different orderings.

- Number of parents $m = 5, 10, 20$.
- Dimension $d = 2, 3$.
- Ordering following the first coordinate from Datta et al. (2016), at random, or using MaxMin heuristic from Guinness (2018).

Sensitivity. The color count is overwhelmingly driven by the number of parents, the ordering, and interactions between them. The role of the parents is not surprising: the larger the parent sets, the more edges in the graph, the more colors needed. As for the ordering, it does not change the density but rather the distribution of the edges, which may explain why the number of colors is much smaller in the coordinate ordering. In a graph obtained through Coordinate ordering and the Nearest Neighbor heuristic, a vertex tends to be connected with its immediate spatial surroundings. Indeed its parents in the DAG will be its predecessors along the coordinate used for ordering, its children will be its successors, and its co-parents will mostly be a mix of the closest parents and children. In a graph obtained using random or max-min ordering, the connections can be much longer, in particular for points coming early in the ordering. This results in some vertices being connected to many other vertices, leading to a denser graph. This point is illustrated in figures 3.4 and 3.5.

The number of colors is robust with respect to the graph size because n and its interactions have very low percentages in table 3.2. Since the numbers of colors are small with respect to n (45 colors at the most), this makes chromatic sampling a good candidate for large data sets.

This point is counter-intuitive because a bigger graph is more complex than



(a) Max-min ordering (b) Coordinate ordering

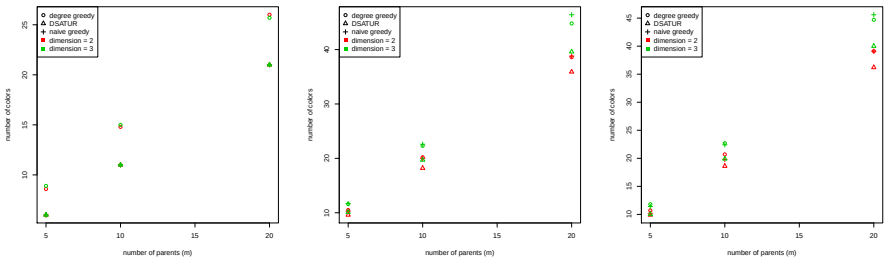
Figure 3.5: Number of connections of a point given its place in the ordering (same graphs as in figure 3.4)

a smaller graph and should therefore be more difficult to color. The markovian nature of NNGPs may be a lead to explain this point, because several sets of vertices that are not linked by a direct edge can be colored independently. While new vertices are added in the graph, older vertices can be forgotten and their colors can be re-used.

The dimension of the spatial domain d plays almost no role in the sensitivity analysis, and the choice of the coloring algorithm has a very marginal effect on the color count. Closer examination of the means reveals nonetheless that their effect is not nonexistent but rather dwarfed by the prominent role of the ordering of the vertices and the number of parents. For graphs obtained with max-min or random ordering (3.6b and 3.6c), the number of colors increases if $d = 3$.

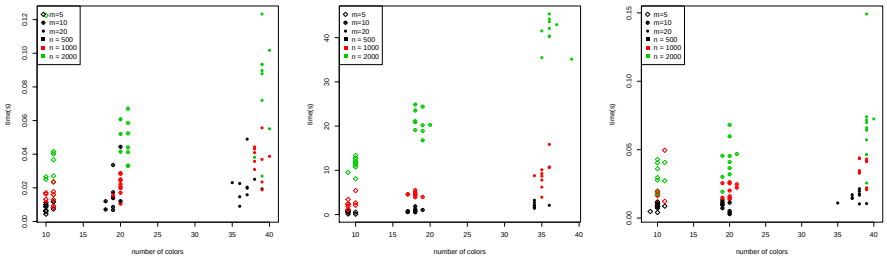
The running time is affected by n, as expected. However, it is mostly explained by the coloring algorithm and its interactions with n and m. In Figure 3.7, we see the results of the experiment when the ordering of the spatial points is random and $d = 2$. The number of parents m defines well-separated vertical clouds of points, showing a clear, positive impact on the number of colors. It also increases the running time: the clouds of points on the right are stretched higher along the ordinates axis. The graph size n affects the running time positively. The other cases with different ordering and dimension all show this clear, chromatography-like profile.

Benchmark. In order to see if one coloring algorithm has the better of the others, we compare the average number of colors for each case of the experiment in table 3.5. Regardless of the ordering, m, and n, the number of colors favors



(a) Coordinate ordering (b) Max-min ordering (c) Naive

Figure 3.6: Impact of the spatial domain dimension and the coloring algorithm on the mean number of colors, for graphs of size $n = 2000$.



(a) Degree (b) DSATUR (c) Naive

Figure 3.7: Repartition of the number of colors and running time with random ordering and d=2.

Table 3.2: Sensitivity analysis*

	Pilot		Large		Blocked	
	colors	time	colors	time	colors	time
ordering	15.4	3.4	10.6	8.7	40.7	1.7
algo	0.8	24.7	0.5	1.3	1.0	8.6
d	0.6	0.0	1.1	0.7	0.4	0.0
m	76.4	2.8	80.6	26.0	17.2	1.0
n/n blocks	0.2	11.8	0.1	40.7	15.2	16.2
ordering:algo	0.3	6.8	0.1	0.3	0.3	3.3
ordering:d	0.3	0.0	0.5	0.4	0.2	0.0
ordering:m	5.3	0.8	5.2	5.4	7.9	0.5
ordering:n	0.0	3.3	0.0	3.0	6.4	6.1
algo:d	0.0	0.1	0.0	0.0	0.0	0.1
algo:m	0.2	5.5	0.2	0.6	0.1	1.9
algo:n	0.0	23.3	0.0	0.7	0.9	31.3
d:m	0.2	0.0	0.4	0.5	0.0	0.0
d:n	0.0	0.0	0.0	0.2	0.1	0.1
m:n	0.0	2.4	0.0	8.0	6.3	3.5
total	99.7	84.8	99.5	96.6	96.6	74.6

* Read: "In the pilot experiment, the ordering of the spatial points explained 15.4 percents of the variance of the number of colors"

systematically but slightly DSATUR over the two simpler algorithms. In the case of coordinate ordering, naive greedy coloring reaches the performances of DSATUR. While the two simple methods are very economical, the running time becomes high in DSATUR when the graph size augments (Figure 3.7b). We conclude that regardless of the structure of the graph, DSATUR must be chosen for smaller graphs. The two other methods must be chosen for larger graphs because DSATUR will become prohibitively expensive.

Coloring for large graphs

Design. The objective is to test the sensitivity of the two interest variables and to benchmark coloring algorithms when the graphs are bigger. The experiment is the same as before, with two differences:

- Only naive greedy and degree greedy coloring algorithms are tested.
- The graph size $n = 50000, 100000, 200000$.

Each case is replicated 10 times.

Sensitivity. For the number of colors, the results are the same as before (table 3.2). It is mostly determined by the ordering and the number of parents. The

robustness of the number of colors with respect to n is confirmed. The running time is affected mostly by n, but the ordering and m also play a role.

Benchmark. In table 3.6, we can see that naive coloring systematically has a lower mean number of colors than degree coloring. It is also slightly faster due to the fact that the vertices are not sorted. Anyway, the running times are short in both cases and are never bigger than 15 seconds. We conclude that naive greedy coloring is the better option for large data sets.

Coloring blocked graphs

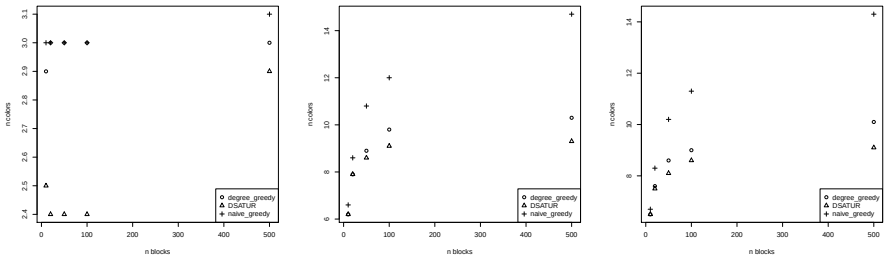
Design. The objective is to carry out sensitivity analysis and benchmark to graphs that correspond to spatial blocks used for block-update of the latent field. Spatial clusters of vertices are found using a K-means algorithm on $n = 10000$ spatial locations, and coloring is applied to the Markov graph between the blocks. The orderings, the numbers of parents, and the dimensions remain the same as in the previous experiments. The parameters that change are:

- The graph size $n_{blocks} = 10, 20, 50, 100, 500$.
- All three algorithms (DSATUR, naive greedy, and degree greedy) are tested

Each case is replicated 10 times.

Sensitivity. The sensitivity of the number of colors (table 3.2) differs from the previous experiments. Even though m still matters, it is the ordering that becomes the most important variable.

This loss of importance of m can be explained by the fact that one edge is enough to connect two blocks, and once two blocks are connected adding new edges between them is redundant. On the other hand, the disposition of the edges in the space, which is induced by the ordering, keeps all its importance. Short edges induced by a coordinate ordering (3.4b) will connect adjacent spatial blocks, while the long edges induced by the max-min heuristic (3.4a) will connect distant regions. The important interaction between m and the ordering is well explained by this hypothesis. In table 3.7 we see that m barely plays any role for coordinate ordering, while it keeps having an important impact for the other two orderings. Indeed, when $\mathcal{S}$ is ordered following a coordinate, adding more short connections between contiguous spatial blocks does not change anything: those blocks already are connected. For max-min and random ordering, though, adding long edges may link distant regions that were not connected yet. After m and the ordering, the number of blocks is the third most important variable. As expected, the more blocks in the graph, the more colors are needed. However, we remark that Max-Min and Random orderings perform poorly for graphs with few blocks, and actually need almost one color per block. Once the graphs get bigger, the number of colors stabilizes. Therefore, the observed sensitivity with respect to the number of blocks is mostly induced by the bad coloring of graphs with few blocks. The point can be visualized in figure 3.8 for $m = 5$ and $d = 2$.



(a) Coordinate ordering (b) Max-min ordering (c) Naive

Figure 3.8: Number of colors following the number of blocks, for spatial domain dimension $d = 2$ and number of parents $m = 5$.

Benchmark. Incontestably, DSATUR has the smallest number of colors, as seen in figure 3.8 and table 3.7. Interestingly, degree greedy coloring has the second smallest number of colors. If we assume that the number of blocks will always be smaller than 1000, we can discard the running time from our criteria and say that DSATUR is the best option for blocked graphs. However, in the cases with random and Max-Min orderings and low numbers of blocks, chromatic sampling does not greatly reduce the number of steps with respect to vanilla block sampling.

3.4 Implementation, testing and application

3.4.1 About our implementation

We tested our implementation along with the state of the art package spNNGP presented by Finley et al. (2017), that uses the Gibbs sampler architecture given by Datta et al. (2016). spNNGP uses Rcpp (Eddelbuettel et al., 2011) and parallelizes the computation of NNGP density. In order to monitor convergence using the diagnostics from Gelman et al. (1992) and Brooks and Gelman (1998), various chains need to be run one after the other. Our implementation is available at `https://github.com/SebastienCoube/Improving_NNGP_full_augmentation`. The code is done in R (see R Core Team (2018a)), with the AS-IS Gibbs sampler architecture of Yu and Meng (2011). Chromatic sampling is implemented for individual locations. We used the package GpGp (Guinness and Katzfuss, 2018) for Vecchia's approximation factor computation. Our implementation runs several chains in parallel using the package parallel (R Core Team, 2018b), but GpGp does not implement parallel Vecchia's approximation factor computation within each chain like spNNGP. We emphasized the ease of use, with real-time Gelman-Rubin diagnostics and chains plotting, greedy MCMC tuning in the first hundred iterations, and the possibility to start, stop, and run again easily. For some data sets, our implementation has an advantage

over spNNGP because multiple measurements at the same spatial site are allowed. However, unlike spNNGP, we have only implemented a Gaussian model so far.

3.4.2 Toy examples

We present two toy examples in order to test our implementation, with the latent field NNGP implementation of spNNGP as a reference. For both implementations, 5 nearest neighbors were used for NNGP. The toy examples are Gaussian. We compare the MCMC behavior using the number of iterations and the time needed before the chains have mixed following the Gelman-Rubin-Brooks $\hat{R}$. We also compare the estimated covariance parameters with the values that were used to simulate the toy example. The covariance parameters are reported individually, and in the second toy example we report the Mean Square Error (MSE) of the fitted fixed effects with respect to their true value. Eventually we compare the quality of the denoising using the MSE of the denoised field predicted by the model with respect to the simulated latent field. The first toy example is a simple Gaussian field simulated as follows.

1. Simulate spatial locations $\mathcal{S} \sim \mathcal{U}([0, 50] \times [0, 50])$.
2. Simulate latent field $w(\mathcal{S}) \sim \mathcal{N}(0, \Sigma(\mathcal{S})), \Sigma(\mathcal{S})_{i,j} = exp(-0.5\|s_i, s_j\|)$.
3. Simulate observed variable $z(\mathcal{S}) = w(\mathcal{S}) + \epsilon(\mathcal{S}), \epsilon(\mathcal{S}) \sim \mathcal{N}(0, 5I_n)$.

The second toy example intends to highlight the positive effect of our architecture when covariates have some spatial coherence. We integrate covariates that are areal indicators, and others that are white noise.

1. Simulate spatial locations $\mathcal{S} \sim \mathcal{U}([0, 50] \times [0, 50])$ and note $\mathcal{S}_1$ the first coordinates of the locations.
2. Simulate latent field $w(\mathcal{S}) \sim \mathcal{N}(0, \Sigma(\mathcal{S})), \Sigma(\mathcal{S})_{i,j} = exp(-0.5\|s_i - s_j\|)$.
3. Simulate regressors $X = [X_1|X_2]$ with $X_1 = [\mathbb{1}_{1 \leq \mathcal{S}_1 < 2}|\mathbb{1}_{2 \leq \mathcal{S}_1 < 3} \dots |\mathbb{1}_{49 \leq \mathcal{S}_1 \leq 50}]$ and X_2 a matrix of side $n \times 49$ with coefficients drawn following independent $\mathcal{N}(0, 1)$.
4. Simulate regression coefficients $\beta \sim \mathcal{N}(0, I_{98})$.
5. Simulate observed variable $z(\mathcal{S}) = w(\mathcal{S}) + X\beta^T + \epsilon(\mathcal{S}), \epsilon(\mathcal{S}) \sim \mathcal{N}(0, 5I_n)$.

The results of the runs on the toy examples are presented in table 3.3. The estimates are close to the target and there is no clear gap between the methods. Due to the fast-mixing AS-IS architecture from Yu and Meng (2011) and Filippone et al. (2013), our implementation needed much less iterations than the latent model of spNNGP (Even taking into account the fact that one AS-IS iteration needs two covariance parameters updates): our model takes thousands iterations to converge, while spNNGP needs tens of thousands. The response model,

Table 3.3: Summary of the toy examples runs

(a) Summary of the first toy example

method	n iter.	time (min)	MSE	σ^2	τ^2	α
spNNGP	15000	28	0.40	1.07	4.99	1.10
Our code	3000	28	0.38	1.08	5.00	1.01
spNNGP res.	8000	13		1.06	5.00	0.99
true values				1.00	5.00	1.00

σ^2: marginal variance of w; τ^2: variance of ϵ; α: range of w

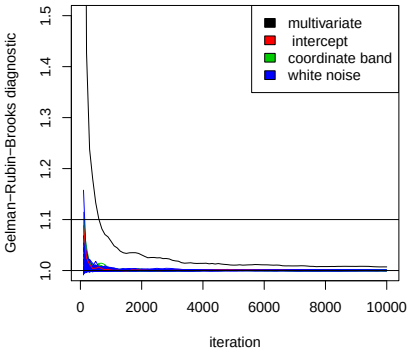
(b) Summary of the second toy example

method	n iter.	time (min)	MSE	$\beta-$MSE	σ^2	τ^2	α
spNNGP	25000	74	0.45	0.053	0.91	5.08	0.90
Our code	3000	36	0.42	0.057	0.98	5.06	1.13
spNNGP res.	10000	50		0.047	0.88	5.10	0.72
true values					1.00	5.00	1.00

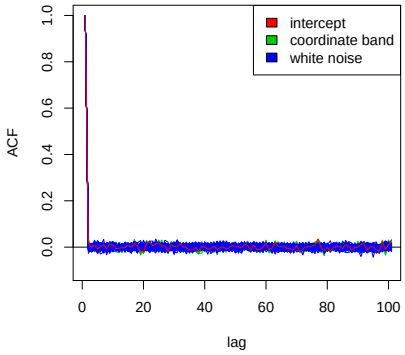
σ^2: marginal variance of w; τ^2: variance of ϵ; α: range of w

in spite of its frugality, needed a few thousands iterations to converge, like our implementation. The running times end up being of the same order, due to the efficient multi-process implementation of spNNGP which compensates the number of iterations.

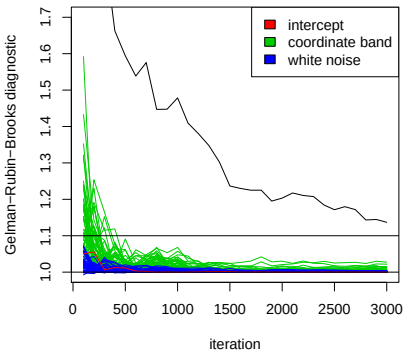
Let's now focus on the behavior of the regression coefficients in the second toy example (Figure 3.9). The best model regarding the mixing of the regression coefficients is incontestably the response model (3.9a, 3.9b). However, the covariance parameters needed more time to mix than the regression coefficients, explaining why 10000 iterations were needed. Moreover, the response model cannot retrieve the latent field, explaining why its MSE could not be computed. Except for the response model, we can see that the coherent regression coefficients of X_1, in green in 3.9, mix slower than the fuzzy coefficients of X_2, in blue. Nonetheless, for our implementation, the $\hat{R}$ diagnostics dropped to 1 in a few hundred iterations (figure 3.9c), against the tenths of thousands needed for spNNGP (figure 3.9e). For our implementation, the autocorrelations dropped to 0 after a few dozen iterations (figure 3.9d). The auto-correlations of spNNGP for the regression coefficients of X_1 were still between 0.4 and 0.6 after 100 iterations (figure 3.9f), while the autocorrelation for the coefficients of X_2 remain stuck slightly above 0. It is then clear that the chains behave much better in our implementation than in spNNGP. Moreover, the good behavior of our implementation could not be reproduced if we did not indicate that interweaving could be used, see figures 3.9g, 3.9h.



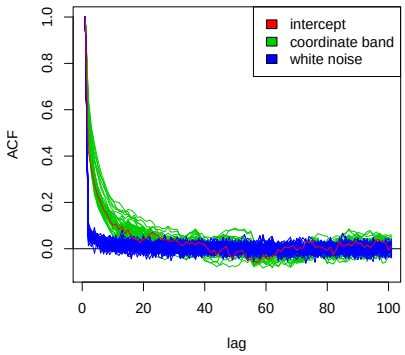
(a) $\hat{R}$ with response spNNGP



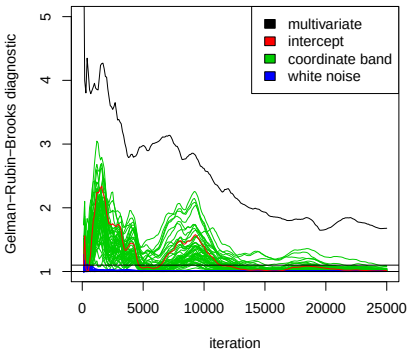
(b) Autocorrelations with response spNNGP



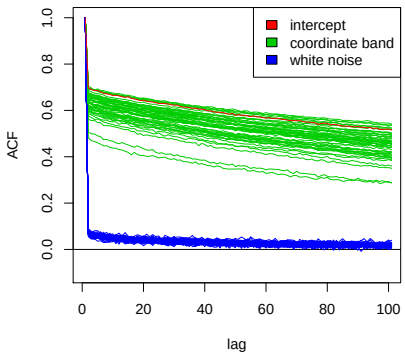
(c) $\hat{R}$ of our implementation



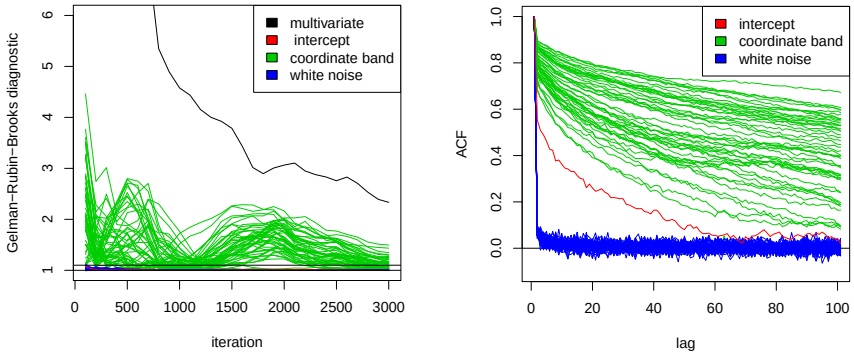
(d) Autocorrelations of our implementation



(e) $\hat{R}$ with spNNGP



(f) Autocorrelations with spNNGP



(g) $\hat{R}$ of our implementation without interweaving

(h) Autocorrelations of our implementation without interweaving

Figure 3.9: Behavior of the regression coefficients with spNNGP and with our implementation

3.4.3 Application to lead contamination analysis

We used our implementation to study a heavy metal contamination data set proposed by Hengl (2009)[1]. The dataset gathers measurements made by the United States Geological Survey of Grossman et al. (2004) and several covariates, including geophysical and environmental information about the sampling site, and potential contamination sources nearby. We added the predominant subsoil rock type given by the USGS study presented in Horton (2017)[2]. We scaled the quantitative regressors. After removing missing data, there was 64274 observations. We assumed the model

$$log(z(s)) = w(s) + X(s)\beta^T + \epsilon(s),$$

s being the sampling location, $X(\cdot)$ being the aforementioned covariates, $w(\cdot)$ being a latent Gaussian field with exponential covariance on the sphere, and ϵ being a white noise.

The model converged in 4000 iterations, and 1 hour and 38 minutes were needed. We tried to analyze the real data set with spNNGP in order to compare the results and the running time. Surprisingly, spNNGP had a pathological behavior in spite of its good performances on simulated data. The scale parameter kept straying towards values several orders of magnitude above the variance of the observed field, even with starting points corresponding to our estimates. This behavior was observed with both latent and response model, and various orderings of the locations.

[1] https://spatial-analyst.net/book/NGS8HMC
[2] https://mrdata.usgs.gov/geology/state/

Table 3.4: Summary of the covariance parameters and a subset of the fixed effects

	mean	qtile 0.025	median	qtile 0.975	st dev
Scale	0.198	0.188	0.198	0.209	0.0053
Noise variance	0.178	0.175	0.178	0.181	0.0017
Range (Km)	35.6	33.3	35.5	38.3	1.2700
(Intercept)	2.83	2.79	2.83	2.87	0.019
Air pollution dsty	0.0403	0.0195	0.0404	0.0605	0.0106
Mineral op[1] dsty	0.0180	0.0044	0.0182	0.0312	0.0069
Toxic reject dsty	0.0641	0.0433	0.0639	0.0852	0.0107
Carbon biomass dsty	-0.0543	-0.0673	-0.0541	-0.0412	0.0066
Population dsty	0.1360	0.1100	0.1360	0.1640	0.0138
Night light	0.0384	0.0303	0.0384	0.0466	0.0042
Roads dsty	0.0193	0.0139	0.0193	0.0248	0.0028

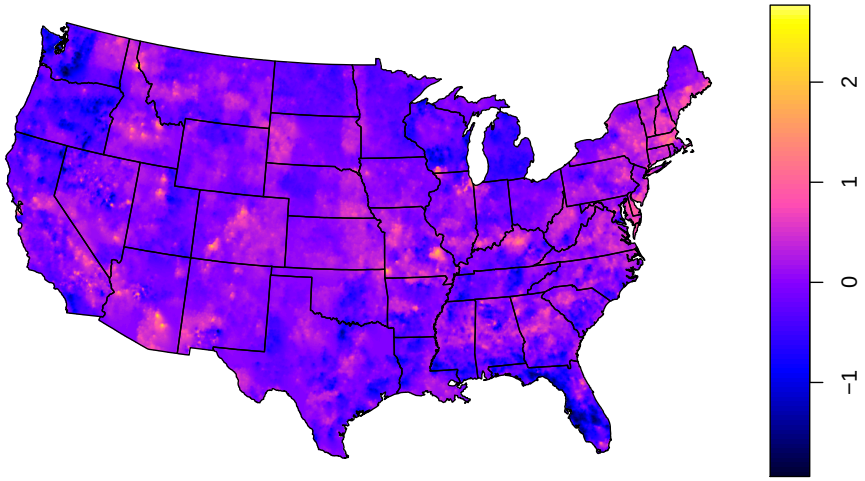
1: "operations"

We present our implementation's estimates of the covariance parameters and some of the fixed effects in table 3.4. We left out some regressors such as the geological classification, indications about nearby mineral observations, the geophysical characteristics of the sampling site. The variances of the latent field and the noise have equivalent order ($\sigma^2 = 0.20, \tau^2 = 0.18$). The spatial range is 30 Km. With a rule of the thumb, this means that the correlation drops to 10% of the scale for locations separated by 60 Km. The regressors behave as expected: the urbanization level and contamination indicators have a positive, certain effect on lead concentration. However, the values of the regression coefficients remain modest with respect to the scale of the latent field.

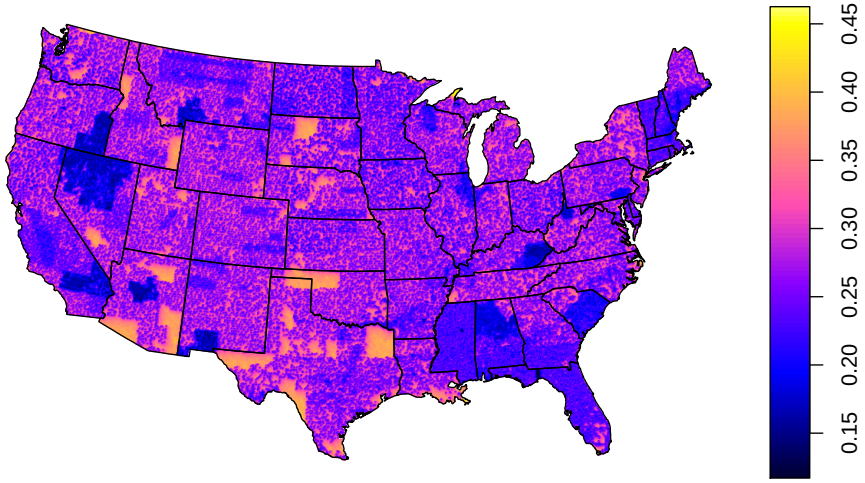
We also provide prediction of the latent field on a 5-Km grid on the territory of the USA mainland. Predictions at un-observed locations are done using the MCMC samples of the covariance parameters θ and $w(\mathcal{S})$, see for example Finley et al. (2019). We report the predicted latent mean and standard deviation in figures 3.10a and 3.10b. The standard deviation map must be put in relation with the sampling sites map (Figure 3.11). The patches with high standard deviation correspond to zones with no measurement, while territories with dense sampling, such as Florida, will have low predicted standard deviation.

3.5 Discussion

We presented two ways to improve the behavior of NNGPs with full data augmentation, that can be simply applied to previous implementations. What's more, while we assumed a Gaussian data model throughout the article, the two methods we proposed can be easily applied to other models.
While our article focused on a basic NNGP model, our field centering may have applications in complex models. Space-varying regression coefficients are an



(a) Predicted latent mean



(b) Predicted latent standard deviation (closely related to sampling density)

Figure 3.10: Latent field predictions

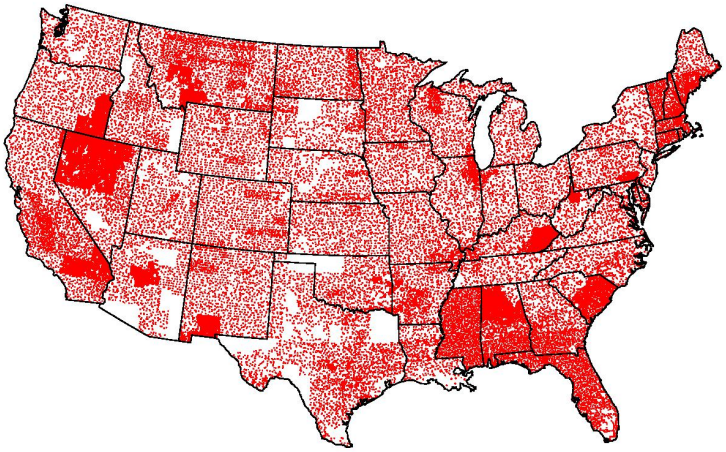


Figure 3.11: Sampling sites

extension to GP models (Datta et al. (2016); Banerjee et al. (2008)). If we consider the latent field $w(\cdot)$ as a space-varying intercept, it seems natural to try to center a spatially variable parameter on the corresponding fixed effect. The extension to other fixed effects we presented could prove valuable in the case in which the regressor with a spatially variable β is correlated with other variables from X. Another possible extension could be a GP defined as the sum of two or more GPs. It could have an interest in various applications, such as: modeling seasonality in a space-time process, modeling a process with short-range and long-range interactions, defining one non-separable space-time process as a sum of two separable processes. The equivalent of standard parametrization would be $z(\cdot) = \beta_0 + w_1(\cdot) + w_2(\cdot) + \epsilon$, $w_1(\cdot)$ and $w_2(\cdot)$ being GPs of mean 0. One could try out a Russian doll centering: $z(\cdot) = v_1(\cdot) + \epsilon$ where $v_1(\cdot)$ has mean $v_2(\cdot)$, and $v_2(\cdot)$ has mean β_0. In this case it might be necessary to find an *ad hoc* interweaving scheme.

Beyond the improvements of chromatic sampling in the NNGP algorithm, exploration of the moralized graph could be an interesting approach to study Vecchia's approximation and evaluate heuristics concerning ordering and picking parents. For example, Guinness (2018) has explored how various ordering and grouping strategies affected the Kullback-Leibler divergence of Vecchia's approximation with respect to the full GP density. Those strategies have a graphical translation. Grouping takes an existing graph and adds new edges, making it closer to the full GP's graph (i.e. the saturated DAG and moralized

graph). Ordering modifies the structure of the graph and the length of the edges, just like the mixing of observations explored in Stein et al. (2004). For example, a coordinate or a middle-out ordering with Nearest Neighbor heuristic will make graph where each vertex connected to its closest neighbors, while we could use a classical concept of Geography and say that a random or a max-min ordering will generate graphs not unlike a Christallerian system. Focusing on the neighbor-picking heuristics gives one a close-up shot of what is going on and has a direct algorithmic translation, but some descriptive statistics about the moralized graphs could give a more general view.

Bibliography

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008, September). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics 7*(4), 434–455.

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

Eddelbuettel, D., R. François, J. Allaire, K. Ushey, Q. Kou, N. Russel, J. Chambers, and D. Bates (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software 40*(8), 1–18.

Filippone, M., M. Zhong, and M. Girolami (2013). A comparative evaluation of stochastic-based inference methods for gaussian process models. *Machine Learning 93*(1), 93–114.

Finley, A., A. Datta, and S. Banerjee (2017). spnngp: spatial regression models for large datasets using nearest neighbor gaussian processes. *R package version 0.1 1*.

Finley, A. O., A. Datta, B. D. Cook, D. C. Morton, H. E. Andersen, and S. Banerjee (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics 28*(2), 401–414.

Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue (2015). Interpretable priors for hyperparameters for gaussian random fields. *arXiv preprint arXiv:1503.00256*.

Gelfand, A. E., P. Diggle, P. Guttorp, and M. Fuentes (2010). *Handbook of Spatial Statistics (Chapman Hall CRC Handbooks of Modern Statistical Methods)*. Chapman Hall CRC Handbooks of Modern Statistical Methods. Taylor and Francis.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parametrisations for normal linear mixed models. *Biometrika 82*(3), 479–488.

Gelman, A., D. B. Rubin, et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.

Gonzalez, J., Y. Low, A. Gretton, and C. Guestrin (2011). Parallel gibbs sampling: From colored fields to thin junction trees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 324–332. JMLR Workshop and Conference Proceedings.

Grossman, J. N. et al. (2004). *The National Geochemical Survey-database and documentation*.

Guinness (2018). Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics 60*(4), 415–429.

Guinness and Katzfuss (2018). *GpGp: Fast Gaussian Process Computation Using Vecchia's Approximation*.

Heinonen, M., H. Mannerström, J. Rousu, S. Kaski, and H. Lähdesmäki (2016). Non-stationary gaussian process regression with hamiltonian monte carlo. In *Artificial Intelligence and Statistics*, pp. 732–740.

Hengl, T. (2009). *A practical guide to geostatistical mapping*. Hengl Amsterdam.

Horton, J. D. (2017). The state geologic map compilation (sgmc) geodatabase of the conterminous united states.

Katzfuss, M. and J. Guinness (2017, Aug). A general framework for Vecchia approximations of Gaussian processes. *arXiv e-prints*, arXiv:1708.06302.

Knorr-Held, L. and H. Rue (2002). On block updating in markov random field models for disease mapping. *Scandinavian Journal of Statistics 29*(4), 597–614.

Lauritzen, S. L. (1996). *Graphical models*. Oxford Statistical Science Series. OUP.

Peruzzi, M., S. Banerjee, and A. O. Finley (2020). Highly scalable bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *Journal of the American Statistical Association*, 1–14.

R Core Team (2018a). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team (2018b). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods* (2nd ed.). Springer Texts in Statistics. Springer.

Rue, H. and L. Held (2005). *Gaussian Markov random fields: theory and applications* (1 ed.). Monographs on statistics and applied probability 104. Chapman Hall/CRC.

Stein, M. L., Z. Chi, and L. J. Welty (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(2), 275–296.

Yu, Y. and X.-L. Meng (2011). 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.

Zhang, L., A. Datta, and S. Banerjee (2019). Practical bayesian modeling and inference for massive spatial data sets on modest computing environments. *Statistical Analysis and Data Mining: The ASA Data Science Journal 12*(3), 197–209.

APPENDIX

3.6 Appendix: stochastic form of the intercept-field model

We obtain the full conditionals of w_c and w_s using the conditional expectation and variance formulas with precision matrices of Rue and Held (2005), the joint precision of both (w_s, z) or (w_c, z) being $\begin{bmatrix} \tilde{Q}+\tau^2 I_n & -\tau^2 I_n \\ -\tau^2 I_n & \tau^2 I_n \end{bmatrix}$. The expectation of w_s is 0, the expectation of w_c is β_0, and the expectation of z is always β_0.

Distributions with the standard model. The full conditional distributions of β_0 and w_s are:

$$[\beta_0|w_s] \sim \mathcal{N}(\overline{z-w_s}, \tau^2/n), [w_s|\beta_0] \sim \mathcal{N}(-(\tilde{Q}+I_n/\tau^2)^{-1}(-I_n/\tau^2)(z-\beta_0), (\tilde{Q}+I_n/\tau^2)^{-1}).$$

Note $\mathbf{1}$ the vector of length n and filled with ones. From the second full conditional, we have a formula for the mean of w_s, which is obtained with $\overline{w_s} = \mathbf{1}^t w_s/n$. It has 3 terms: one is fix, the second is a geometric carry-over of β_0, and the third is stochastic:

$$[\overline{w_s}|\beta_0] \sim \underbrace{\mathbf{1}^T(\tau^2\tilde{Q}+I_n)^{-1}(z)/n}_{\text{fixed}} - \underbrace{(\mathbf{1}^T(\tau^2\tilde{Q}+I_n)^{-1}\mathbf{1}/n)(\beta_0)}_{\text{carry-over}} + \underbrace{\mathcal{N}((0, \mathbf{1}^T(\tilde{Q}+I_n/\tau^2)^{-1}\mathbf{1}/n^2)}_{\text{innovation}}.$$

Injecting the full conditional of $\bar{w}_s$ into β_0's, we identify an expression with 3 terms like before:

$$\begin{aligned}[\beta_0^{t+1}|\beta_0^t] \quad \sim \quad & \underbrace{\overline{z} - \mathbf{1}^T(\tau^2\tilde{Q}+I_n)^{-1}(z)/n}_{\text{fixed}} + \underbrace{(\mathbf{1}^T(\tau^2\tilde{Q}+I_n)^{-1}\mathbf{1}/n)\beta_0^t}_{\text{carry-over}} \\ & + \underbrace{\mathcal{N}(0, \mathbf{1}^T(\tilde{Q}+I_n/\tau^2)^{-1}\mathbf{1}/n^2 + \tau^2/n)}_{\text{innovation}}.\end{aligned}$$

Distributions with the centered model. The full conditional of β_0 and w_c are:

$$[\beta_0|w_c] \sim \mathcal{N}(\mathbf{1}^T\tilde{Q}w_c/\mathbf{1}^T\tilde{Q}\mathbf{1}, 1/\mathbf{1}^T\tilde{Q}\mathbf{1}), [w_c|\beta_0] \sim \mathcal{N}(\beta_0+(\tilde{Q}+I_n/\tau^2)^{-1}(z-\beta_0)/\tau^2, (\tilde{Q}+I_n/\tau^2)^{-1}).$$

The mean of w_c behaves like the mean of w_s except for the term that depends on β_0:

$$\begin{aligned}[\overline{w_c}|\beta_0] \quad \sim \quad & \underbrace{\mathbf{1}^T(\tau^2\tilde{Q}+I_n)^{-1}(z)/n}_{\text{fixed}} - \underbrace{(1-(\mathbf{1}^T(\tau^2\tilde{Q}+I_n)^{-1}\mathbf{1}/n))\beta_0}_{\text{carry-over}} \\ & + \underbrace{\mathcal{N}(0, \mathbf{1}^T(\tilde{Q}+I_n/\tau^2)^{-1}\mathbf{1}/n^2)}_{\text{innovation}}.\end{aligned} \tag{3.4}$$

Injecting the full conditional of w_c into the full conditional of β_0, we have

$$[\beta_0^{t+1}|\beta_0^t] \sim \underbrace{\mathbf{1}^T\tilde{Q}(\tilde{Q}+I_n/\tau^2)^{-1}z/\tau^2\mathbf{1}^T\tilde{Q}\mathbf{1}}_{\text{fixed}} + \underbrace{\mathbf{1}^T\tilde{Q}(I_n-(\tau^2\tilde{Q}+I_n)^{-1})\mathbf{1}\beta_0/\mathbf{1}^T\tilde{Q}\mathbf{1}}_{\text{carry-over}}$$
$$+\underbrace{\mathcal{N}(0,\mathbf{1}^T\tilde{Q}(\tilde{Q}+I_n/\tau^2)^{-1}\tilde{Q}\mathbf{1}/(\mathbf{1}^T\tilde{Q}\mathbf{1})^2+1/\mathbf{1}^T\tilde{Q}\mathbf{1})}_{\text{innovation}}$$

Passing to the SVD. Let's compare first the expressions of $[\overline{w_s}|\beta_0]$ and $[\overline{w_c}|\beta_0]$. Denote the diagonalization $\tilde{Q} = V^T\lambda V$, V being a square matrix of eigenvectors and λ being a diagonal matrix of eigenvalues. The eigenvalues are positive since $\tilde{Q} = \tilde{R}^T\tilde{R}$. Using the fact that adding I_n adds 1 to every eigenvalue without affecting the eigenvectors,

$$(\tau^2\tilde{Q}+I_n)^{-1}/n = V^T(\tau^2\lambda+I_n)^{-1}V/n.$$

Let $\alpha = (\alpha_1,\ldots,\alpha_n)$ be the coordinates of $\mathbf{1}$ in the orthonormal basis defined by V.

$$\mathbf{1}^T(\tau^2\tilde{Q}+I_n)^{-1}\mathbf{1}/n = \Sigma_{i=1}^n\alpha_i^2(\tau^2\lambda+I_n)_{i,i}^{-1}/n.$$

Using that $\tilde{Q}$ is positive-definite on the left and that $\Sigma_1^n\alpha_i =< \mathbf{1},\mathbf{1} >= n$ on the right, we have

$$0 \leq \mathbf{1}^T(\tau^2\tilde{Q}+I_n)^{-1}\mathbf{1}/n \leq 1.$$

As for the centered model, we re-write:

$$\tilde{Q}(I_n-(\tau^2\tilde{Q}+I_n)^{-1}) = V^T(\lambda(I_n-(\tau^2\lambda+I_n)^{-1}))V = V^T(\tau^2\lambda^2(\tau^2\lambda+I_n)^{-1})V.$$

Once this is done, we can express the fraction of β_0^t which is conserved in β^{t+1} in the centered model as

$$\mathbf{1}^T\tilde{Q}(I_n-(\tau^2\tilde{Q}+I_n)^{-1})\mathbf{1}/\mathbf{1}^T\tilde{Q}\mathbf{1} = \Sigma_{i=1}^n\left((\alpha_i^2\lambda_{i,i})(\tau^2\lambda_{i,i})/(\tau^2\lambda_{i,i}+1)\right)/\Sigma_{i=1}^n\alpha_i^2\lambda_{i,i}.$$

Like before, using the fact that the eigenvalues of $\tilde{Q}$ are positive,

$$0 \leq (\tau^2\lambda_{i,i})/(\tau^2\lambda_{i,i}+1) \leq 1$$

.

3.7 Appendix: coloring

3.7.1 Details about the coloring algorithms

Algorithm 8 Naive greedy coloring

input A ▷ Input adjacency matrix

$(c_1, \ldots, c_n) = (0, \ldots, 0)$ ▷ Initialize colors

for $c_i \in (c_1, \ldots, c_n)$ **do** ▷ Coloration loop

$c_i = min((1, \ldots, n) \backslash (c_J))$ with $J = \{J / A_{i,j} = 1\}$ ▷ Using smallest available color

end for

return $(c_1, \ldots, c_n)$

Algorithm 9 Degree greedy coloring

input A ▷ Input adjacency matrix

$(c_1, \ldots, c_n) = (0, \ldots, 0)$ ▷ Initialize colors

find $(nd_1, \ldots, nd_n) = (1, \ldots, 1) \cdot A$ ▷ Compute connection degrees of nodes

find $(o(1), \ldots, o(n))$, a permutation of $1, \ldots, n$ such that $i < j \Rightarrow nd_{o(i)} \leq nd_{o(j)}$ ▷ order nodes by decreasing connection degree

for $c_i \in (c_{o(1)}, \ldots, c_{o(n)})$ **do** ▷ Coloration loop

$c_i = min((1, \ldots, n) \backslash (c_J))$ with $J = \{J / A_{i,j} = 1\}$ ▷ Using smallest available color

end for

return $(c_1, \ldots, c_n)$

Algorithm 10 DSATUR

input A ▷ Input adjacency matrix
$(c_1, \ldots, c_n) = (0, \ldots, 0)$ ▷ Initialize colors
$(sd_1, \ldots, sd_n) = (0, \ldots, 0)$ ▷ Initialize saturation degrees
$(nd_1, \ldots, nd_n) = (1, \ldots, 1) \cdot A$ ▷ Compute connection degrees of nodes
while $0 \in (c_1, \ldots, c_n)$ **do** ▷ Coloration loop
    $j = \{i / c_i = 0\}$
    $j = \{i \in j / sd_i == max_{i \in j}(sd_i)\}$ ▷ Saturation degree selection rule
    if $\#j > 1$ **then**
        $j = \{i \in j / nd_i == max_{i \in j}(nd_i)\}$ ▷ Node degree tiebreaking rule
    end if
    if $\#j > 1$ **then**
        Reduce j to its first element ▷ lexicographical tiebreaking rule
    end if
    $c_j = min((1, \ldots, n) \backslash (c_{i/A_{i,j}=1}))$ ▷ Using smallest available color
    $sd_{i/A_{i,j}=1} = sd_{i/A_{i,j}=1} + 1$ ▷ Updating saturation degrees
end while
return $(c_1, \ldots, c_n)$

3.7.2 Results of coloring experiments

Table 3.5: Case-by-case mean number of colors in the pilot experiment.

			Coordinate ordering			Max-min ordering			Random ordering		
m	n	d	degree	DSATUR	naive	degree	DSATUR	naive	degree	DSATUR	naive
5	500	2	7.4	6.0	6.0	9.8	9.1	10.1	10.3	9.2	10.0
		3	7.9	6.0	6.0	10.8	9.6	11.0	10.6	10.0	11.1
	1000	2	8.2	6.0	6.0	10.1	9.4	10.3	10.6	9.3	10.2
		3	8.0	6.0	6.0	11.8	10.0	11.1	11.3	10.1	11.1
	2000	2	8.6	6.0	6.0	10.5	9.6	10.3	10.7	9.9	10.2
		3	8.9	6.0	6.0	11.6	10.1	11.7	11.8	10.0	11.4
10	500	2	12.9	11.0	11.0	18.7	17.4	18.9	19.0	17.8	19.4
		3	13.0	11.0	11.0	20.9	19.1	21.2	21.0	18.9	21.3
	1000	2	13.7	11.0	11.0	19.2	17.8	19.5	19.9	17.9	20.0
		3	13.6	11.0	11.0	21.3	19.4	22.3	21.6	19.4	22.0
	2000	2	14.8	11.0	11.0	20.2	18.2	20.0	20.7	18.6	19.8
		3	15.0	11.0	11.0	22.3	19.7	22.6	22.7	19.9	22.4
20	500	2	23.0	21.0	21.0	36.5	33.9	37.4	36.8	34.2	37.5
		3	23.3	21.0	21.0	40.8	37.1	44.1	40.5	37.5	43.4
	1000	2	24.9	21.0	21.0	37.6	35.0	38.2	38.6	35.2	38.5
		3	23.9	21.0	21.0	42.9	38.4	45.3	43.1	38.8	44.7
	2000	2	26.0	21.0	21.0	38.6	35.9	38.8	39.1	36.2	39.1
		3	25.7	21.0	21.0	44.8	39.6	46.4	44.7	40.0	45.6

Table 3.6: Case-by-case mean number of colors for large graphs.

			Coordinate ordering		Max-min ordering		Random ordering	
m	n	d	degree	naive	degree	naive	degree	naive
5	50000	2	10.0	8.8	11.3	11.0	12.3	11.1
		3	10.0	8.7	13.1	12.6	13.1	12.3
	100000	2	10.1	9.0	11.7	11.0	12.1	11.0
		3	10.0	9.1	13.1	13.0	13.2	12.3
	200000	2	10.1	9.3	11.9	11.2	12.1	11.4
		3	10.3	9.7	13.3	13.0	13.5	12.8
10	50000	2	17.1	13.8	21.5	21.0	22.7	20.8
		3	17.0	14.0	24.5	24.1	25.5	23.7
	100000	2	17.0	15.6	22.0	21.2	22.9	21.0
		3	17.0	15.5	24.9	24.2	25.6	23.9
	200000	2	17.9	16.7	22.2	21.3	22.9	21.1
		3	18.0	16.8	25.2	24.2	26.4	24.3
20	50000	2	31.4	21.1	41.4	40.1	43.0	40.4
		3	31.2	21.6	49.7	48.2	50.2	47.8
	100000	2	30.8	25.1	41.9	40.7	44.3	40.8
		3	31.0	24.6	50.0	48.5	50.8	47.6
	200000	2	30.3	28.6	41.7	40.7	44.3	40.8
		3	30.2	28.7	50.5	48.9	51.6	48.2

Table 3.7: Case-by-case mean number of colors for blocked graphs.

			Coordinate ordering			Max-min ordering			Random ordering		
m	blocks	d	degree	DSATUR	naive	degree	DSATUR	naive	degree	DSATUR	naive
5	10	2	2.9	2.5	3.0	6.2	6.2	6.6	6.5	6.5	6.7
		3	2.7	2.3	2.7	7.5	7.5	7.5	7.3	7.3	7.5
	20	2	3.0	2.4	3.0	7.9	7.9	8.6	7.6	7.5	8.3
		3	3.0	2.6	3.0	9.0	8.5	9.6	8.6	8.3	9.5
	50	2	3.0	2.4	3.0	8.9	8.6	10.8	8.6	8.1	10.2
		3	3.0	2.4	3.0	10.9	10.2	12.7	10.3	9.4	12.0
	100	2	3.0	2.4	3.0	9.8	9.1	12.0	9.0	8.6	11.3
		3	3.0	2.2	3.0	11.8	11.1	14.0	11.0	10.2	13.0
	500	2	3.0	2.9	3.1	10.3	9.3	14.7	10.1	9.1	14.3
		3	3.0	3.0	3.0	13.2	11.7	17.3	12.4	11.1	16.1
10	10	2	2.9	2.5	3.0	8.1	8.1	8.2	8.6	8.6	8.6
		3	2.7	2.3	2.7	9.3	9.3	9.3	9.3	9.3	9.3
	20	2	3.0	2.4	3.0	11.4	11.4	12.2	11.9	11.9	12.5
		3	3.0	2.6	3.0	13.5	13.4	13.8	12.5	12.3	13.0
	50	2	3.0	2.4	3.0	14.8	14.2	17.1	14.0	13.6	16.0
		3	3.0	2.4	3.0	17.5	16.6	19.0	16.2	15.5	18.2
	100	2	3.0	2.4	3.0	16.7	16.0	20.6	15.7	15.0	19.2
		3	3.0	2.2	3.0	19.8	18.5	22.9	18.0	17.1	21.2
	500	2	3.6	3.1	4.1	19.6	18.2	26.5	18.6	17.0	24.5
		3	3.8	3.4	4.3	22.9	21.0	30.9	20.9	18.8	28.0
20	10	2	2.9	2.5	3.0	9.8	9.8	9.8	9.9	9.9	9.9
		3	2.7	2.3	2.7	10.0	10.0	10.0	10.0	10.0	10.0
	20	2	3.0	2.4	3.0	16.2	16.2	16.7	16.3	16.3	16.5
		3	3.0	2.6	3.0	17.7	17.7	17.7	17.1	17.1	17.2
	50	2	3.0	2.4	3.0	25.1	24.5	27.0	23.0	22.9	25.4
		3	3.0	2.4	3.0	27.9	27.8	30.4	24.8	24.3	27.4
	100	2	3.0	2.4	3.0	30.4	29.1	34.6	26.7	25.7	31.0
		3	3.0	2.2	3.0	33.7	32.4	38.2	29.8	28.9	33.9
	500	2	5.2	4.7	5.6	37.2	35.2	47.6	34.0	31.2	42.6
		3	5.2	4.8	5.8	42.0	39.0	55.0	38.9	35.1	49.1

Chapter 4

Nonstationary spatial modeling using Nearest Neighbor Gaussian Processes

Pourquoi faire simple quand on peut faire compliqué ? [1]
-The Shadoks

This chapter consists in an article redacted in collaboration with my advisor Benoît Liquet and Sudipto Banerjee.

After finding techniques to improve NNGP models with explicit sampling of the latent field, I felt that I had to do something with those methods; preferably, something that cannot do without those methods. I started from the work of Heinonen et al. (2016), who propose an elegant framework for nonstationary Gaussian Processes using Hybrid Monte-Carlo implementation. This article presents a model that allows covariance parameters to vary in the space using Gaussian Process priors, inducing a nonstationary behavior of the response. Here, the Gaussian Processes are replaced by Nearest Neighbor Gaussian Processes and the algorithmic toolbox that has been devised in the first part of the thesis is used to power the model. This transposition of methods results in a computational scale-up: Heinonen et al. (2016) work with small data sets, indexed in spaces of one dimension; the article transposes their approach to larger data sets with more spatial dimensions.

More broadly, the article also aims to tackle three thorny and interdependent problems of nonstationary modeling: (i) parameter interpretation, (ii) model selection, and (iii) computational efficiency.

Like in Heinonen et al. (2016), understanding the model is made easier by

[1] Why make it simple when we can make it complicated ?

the use of logarithm transformation. This transformation and the corresponding prior are extended to models with locally elliptic covariance from Paciorek (2003) through the original *matrix log NNGP prior*. Moreover, more parameters are estimated in our architectures than in Heinonen et al. (2016), providing as a consequence an expanding family where the nonstationary models encompass the stationary models.

The issue of model selection and overfitting is investigated through extensive experimentation on synthetic data sets. It appears that thanks to the fact that the simple models are embedded in the complex ones, a nonstationary model is able to degenerate back to a stationary model when the data is stationary. As a consequence, over-modeling can be detected from the parameters of the model and no overfitting was observed on the synthetic data sets. On real data applications, nonstationary models were selected using the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002).

The problem of computation is addressed through a strategy with two components. The first method is a Hybrid Monte-Carlo algorithm invented by Heinonen et al. (2016) but transposed here to NNGP settings, that aims to sample from fields of latent parameters with spatial auto-correlation. In order to use HMC, I had to find the gradients of NNGP densities with respect to the nonstationary parameters. This task was fairly tedious, but now that it is done I believe that the formulas could be used in other applications than in HMC. Thanks to the Markovian nature of the NNGP, a bounded amount of operations is needed to differentiate the NNGP density; the method is therefore affordable, even though it is costlier than stationary NNGP. The second is a Nested Interweaving strategy. This method, envisioned by Yu and Meng (2011) but not applied to realistic models as far as I know, answers to the fact that the model has several hierarchical layers with NNGP latent fields. While the MCMC strategy presented in the model is operational on synthetic data sets, I still encounter difficulties to model real data sets with the full NNGP prior. Alternative computational strategies are discussed in the article and in the conclusion of the thesis, section 5.2.

Nonetheless, the model works smoothly when the behavior of the spatial effect is explained only by covariates (and, ironically, spatial patterns can be captured by putting spatial basis functions in the covariates, see the conclusion of the thesis for further discussion). The model is applied to analyze the lead contamination data set. The properties of the spatial process are explained using several environmental and anthropic variables that may impact the emission and/or the diffusion of the lead in the environment. Following the DIC, the nonstationary model is a clear improvement with respect to the stationary model. The nonstationary model brings a little bit of change to the predicted values; in contrast, the variance of the predictions is heavily affected. Some variables clearly impact the covariance structure, allowing to interpret their role in the emission and diffusion of the lead.

Nonstationary Nearest Neighbor Gaussian Process: hierarchical model architecture and MCMC sampling

Sébastien Coube-Sisqueille[1,a] Sudipto Banerjee[2,b] and Benoît Liquet[1,3,c]

Abstract

Nonstationary spatial modeling is exciting and potentially rewarding, but suffers from several problems: its computational cost, the complexity and lack of interpretability of multi-layered hierarchical models, and the difficulty of model selection. We tackle those problems by introducing a nonstationary Nearest Neighbor Gaussian Process (NNGP) model. NNGPs are a good starting point to address the problem of the computational cost because of their accuracy and affordability. We study the behavior of NNGPs that use a nonstationary covariance function, deriving some algebraic properties and exploring the impact of ordering on the effective covariance induced by NNGPs. To simplify results analysis and model selection, we introduce a readable hierarchical model architecture. In particular, we make parameter interpretation and model selection easier by integrating stationary range, nonstationary range with circular parameters, and nonstationary range with elliptic parameters in a consistent framework. Given the NNGP approximation and the model architecture, we propose a MCMC implementation based on Hybrid Monte-Carlo and nested interweaving of parametrizations, available at `https://github.com/SebastienCoube/Nonstat-NNGP`. We carry out experiments on synthetic data sets to find empirical practical rules concerning MCMC algorithm choice, hyperparameter tuning, and model selection. Finally, we use those guidelines to analyze a data set of lead contamination in the United States of America.

Keywords: Hybrid Monte-Carlo; Interweaving; Spatial Model

---

[1]Laboratoire de Mathématiques et de leurs Applications, Université de Pau et des Pays de l'Adour, UMR CNRS 5142, E2S-UPPA, Pau, France

[2]Fielding School of Public Health, University of California, Los Angeles

[3]Department of Mathematics and Statistics, Macquarie University, Sydney

[a]sebastien.coube@univ-pau.fr

[b]sudipto@ucla.edu

[c]benoit.liquet@univ-pau.fr

Funding was provided by the Energy Environment Solutions (E2S-UPPA) consortium and the BIGCEES project from E2S-UPPA ("Big model and Big data in Computational Ecology and Environmental Sciences")

4.1 Introduction

Bayesian hierarchical models for analyzing spatially and temporally oriented data continue to be widely deployed in diverse scientific and technological applications in the physical, environmental and health sciences (Cressie and Wikle, 2015; Banerjee et al., 2014; Gelfand et al., 2019). Such models are constructed by embedding a spatial process within a hierarchical structure,

$$[\text{data} \,|\, \text{process},\ \text{parameters}] \times [\text{process} \,|\, \text{parameters}] \times [\text{parameters}]\ , \qquad (4.1)$$

which specifies the joint probability law of the data, an underlying spatial process and the parameters. The process in (4.1) is a crucial inferential component that introduces spatial and/or temporal dependence, allows us to infer about the underlying data generating mechanism, and to carry out predictions over entire spatial-temporal domains.

Point-referenced spatial data, which will be our focus here, refer to measurements over a set of locations with fixed coordinates. These measurements are assumed to arise as a partial realization of a spatial process over the finite set of locations. A stationary Gaussian process is a conspicuous specification in spatial process models. Stationarity imposes a simplifying assumption on the dependence structure of the process such as the association between measurements at any two points being a function of the separation between the two points. While this assumption is unlikely to hold in most scientific applications, stationary Gaussian process models are easier to compute. Also, they can effectively capture spatial variation and substantially improve predictive inference that are widely sought in environmental data sets. The aforementioned references provide several examples of stationary Gaussian process models and their effectiveness.

Nonstationary spatial models attempt to relax assumptions of stationarity and can enhance wide-ranging benefits to inference. For example, in situations where variability in the data is a complex function of space composed of multiple locally varying processes, the customary stationary covariance kernels may be inadequate. Here, the richer and more informative covariance structures in nonstationary processes, while adding complexity, may be more desirable by improving smoothing, goodness of fit and predictive inference. Nonstationary spatial models have been addressed by a number of authors (Higdon, 1998; Fuentes, 2002; Paciorek, 2003; Banerjee et al., 2008; Cressie and Johannesson, 2008; Yang and Bradley, 2021; Risser and Calder, 2015; Risser, 2016; Fuglstad et al., 2015a; Gelfand et al., 2010, chapter 9)

The richness sought in nonstationary models have been exemplified in a number of the above references. Paciorek (2003) and Kleiber and Nychka (2012) introduce nonstationarity by allowing the parameters of the Matérn class to vary with location, yielding local variances, local ranges, local geometric anisotropies and local smoothness. Such ideas have been extended and further developed in a number of different directions but have not been devised for implementation on massive data sets in the order of 10^5+. For example, recent works have addressed data sets in the order of hundreds (Risser and Calder, 2015;

Ingebrigtsen et al., 2015; Heinonen et al., 2016) or thousands (Fuglstad et al., 2015a) of locations, but this is modest with respect to the size of commonly encountered spatial data (see the examples in Datta et al., 2016; Heaton et al., 2019; Katzfuss and Guinness, 2017).

A second challenge with nonstationary models is overparametrization arising from complex space-varying covariance kernels. This can lead to weakly identifiable models that are challenging to interpret and difficult to estimate. This issue also complicates model evaluation and selection as the inference becomes very sensitive to the specifications of the model.

We devise a new class of nonstationary spatial models for massive data sets that build upon Bayesian hierarchical models based upon directed acyclic graphs (DAGs) such as the Nearest Neighbor Gaussian Process models (Datta et al., 2016) and, more generally, the family of Vecchia's approximations (Katzfuss and Guinness, 2017) to nonstationarity, which allow us to exploit their attractive computational and inferential properties (Katzfuss and Guinness, 2017; Finley et al., 2019; Guinness, 2018). The idea is to endow the nonstationary process model from Paciorek (2003) with NNGP specifications on the processes defining the parameters. Our approach relies upon matrix logarithms to specify processes for the elliptic covariance parameters of Paciorek (2003). The resulting parametrization is sparser than Paciorek (2003) or Risser and Calder (2015) and is a natural extension of the usual logarithmic prior for positive parameters such as the marginal variance, the noise variance, and the range when it is not elliptic. We embed this nonstationary NNGP in a coherent and interpretable hierarchical Bayesian model framework as in Heinonen et al. (2016), but differ from Heinonen et al. (2016) in our focus on modeling large spatial data sets.

A key challenge is learning about the nonstationary covariance processes. Our approach is a Hamiltonian Monte Carlo (HMC) algorithm directly derived from Heinonen et al. (2016). Here, we draw distinctions from Heinonen et al. (2016) who used full GP and classical matrix calculus that are impracticable for handling massive data and, specifically, for NNGP or other DAG-based models. We devise such algorithms specifically for NNGP models to achieve computational efficiency. We also differ from Heinonen et al. (2016) in that we pursue hierarchical latent processmodelling. Estimating the latent field (Finley et al., 2019) allows us to model non-Gaussian responses as well. In order to obtain an efficient algorithm, we hybridize the approach of Heinonen et al. (2016) with interweaving strategies of Yu and Meng (2011); Filippone et al. (2013). We implement a nested interweaving strategy that was envisioned by Yu and Meng (2011), but not applied to realistic models as far as we know. Our Gibbs sampler otherwise closely follows Coube and Liquet (2020), which is itself a tuned version of Datta et al. (2016) using elements from Yu and Meng (2011) and Gonzalez et al. (2011) to improve the computational efficiency to improve MCMC behavior. We answer to the problem of interpretability by a parsimonious and readable parametrization of the nonstationary covariance structure, allowing to integrate random and fixed effects. We construct a nested family of models, where the simpler models are merely special states of the complex models. While we do not develop automatic model selection of the

nonstationarity, we observe through experiments on synthetic data sets that a complex model that is unduly used on simple data will not overfit but rather degenerate towards a space corresponding to a simpler model. This behavior allows to detect over-modeling from the MCMC samples without waiting for full convergence.

The balance of the article proceeds as follows. Section 4.2 outlines the nonstationary models we develop: covariance and data models, the properties of NNGP density and the model architecture. Section 4.3 details the MCMC implementation of the model, with two pillars: the Gibbs sampler architecture using interweaving of parametrizations in section 4.3.1, and the use of HMC in section 4.3.2. In section 4.4 we focus on application: we use experiments on synthetic data to find guidelines, and we apply them to analyze a data set of lead contamination in the US mainland.

4.2 Nonstationary nearest neighbor space time model

4.2.1 Process and response models

Let $\mathcal{S} = (s_1, s_2, \ldots, s_n)$ be a collection of n spatial locations indexed in a spatial domain $\mathcal{D}$. This domain can be $\mathbb{R}$, when the observations are referenced along time for example; in $\mathbb{R}^2$ or the sphere, for spatial data, which will be our scope here; or in spaces of higher dimensions, that allow to add a depth or a temporal component to spatial locations using Cartesian product. We introduce nonstationarity through the following extensions: (i) spatially-varying marginal variance $\sigma^2(\mathcal{S})$; (ii) spatially-variable, possibly anisotropic range; and (iii) a spatially variable Gaussian noise variance $\tau^2(\mathcal{S})$. All three models are encompassed in equations (4.2)—(4.5). For Gaussian data, we envision a regression equation

$$z(s) = X(s)\beta^T + w(s) + \epsilon(s) \,, \tag{4.2}$$

where the heteroskedastic noise's prior distribution is

$$\epsilon(s_i) \overset{ind}{\sim} \mathcal{N}\left(0, \tau^2(s_i)\right). \tag{4.3}$$

The GP prior for the latent field is

$$w(\mathcal{S}) \sim \mathcal{N}(0, \Sigma). \tag{4.4}$$

The covariance function between two locations, used in order to compute the latent NNGP prior on $w(\mathcal{S})$, is

$$\Sigma_{i,j} = K(s_i, s_j) = \sigma(s_i)\sigma(s_j)K_0(s_i, s_j, \alpha(s_i), \alpha(s_j)), \tag{4.5}$$

where $\sigma(s_1 \ldots s_n)$ is a collection of (positive) spatially-variable marginal standard deviations, K_0 is a correlation function, and $\alpha(s_1, \ldots, s_n)$ is a collection of

spatially variable range parameters. Those parameters can be positive-definite matrices, giving a locally anisotropic nonstationary covariance structure, or positive numbers, giving a locally isotropic nonstationary range. The first case is given by Paciorek (2003).

$$K_0(s,t,A(s),A(t)) = \frac{2^{d/2}|A(s)|^{1/4}|A(t)|^{1/4}}{|A(s)+A(t)|^{1/2}} K_i\left(d_M\left(s,t,(A(s)+A(t))/2\right)\right), \tag{4.6}$$

$A(s)$ and $A(t)$ being the range matrices, d being the dimension of the space-time domain, $d_M(\cdot,\cdot,\cdot)$ being the Mahalanobis distance, and K_i being an isotropic correlation function. Note that when the range matrices $A(\cdot)$ are constant, the covariance structure is anisotropic but stationary. The nonstationary correlation with isotropic range parameter is obtained by setting $A = \alpha I_d$ and identifying the Mahalanobis distance with matrix I_d and the Euclidean distance $d_E(\cdot,\cdot)$:

$$K_0(s,t,\alpha(s),\alpha(t)) = \left(\frac{\sqrt{2}\alpha(s)^{1/4}\alpha(t)^{1/4}}{(\alpha(s)+\alpha(t))^{1/2}}\right)^d K_i\left(d_E(s,t)/\left((\alpha(s)+\alpha(t))/2\right)\right). \tag{4.7}$$

Nonstationary models are more complex than their stationary counterparts and should, therefore, have more issues. We try to anticipate them in this subsection and try to propose solutions. Whether or not those problems actually occur and our proposals are efficient is to be tested.

1. Combining a nonstationary marginal variance and range models sounds attractive, however we have concerns about the possibility to identify the two parameters. Identification is a problem for stationary models when the spatial domain is not large enough (Zhang, 2004). The problem can be addressed using PC priors in order to reduce the ridge of equivalent range-marginal variance combinations to one of its points (Fuglstad et al., 2015b). Due to the fact that covariance functions quickly drop to 0, the locations that will have a non-null covariance with respect to one site are concentrated around it. The observations that will effectively allow to infer the covariance parameters at this site will then be reduced to a cluster of points around the site, a situation that reminds of the fill-in asymptotic of Zhang (2004).

2. We also suspect that a non-stationary model may overfit when the observations are not dense enough with respect to the spatial process range. Consider a situation where the observations are dense enough to tell $w(\cdot)$ apart from $\epsilon(\cdot)$ but not to have precise estimates of the latent field. The samples of $w(\cdot)$ will vary and give broad *a posteriori* confidence intervals. In the case of a nonstationary model, this variability could be explained by the nonstationary marginal variance and/or range, leading to a poor identification between the latent field value and those parameters.

3. Another point is to tell spatially variable process variance $\sigma^2(\cdot)$ apart from spatially variable noise variance $\tau^2(\cdot)$. The samples of the latent

field $w(\mathcal{S})$ can be quite fuzzy, in particular when the correlation function K_0 has low smoothness (for example: exponential kernel, that is Matérn covariance with smoothness $\nu = 0.5$). A combination of sample fuzziness and spatially variable marginal variance could be difficult to distinguish from a heteroskedastic noise.

4. Eventually, it is difficult to identify range and smoothness when a Matérn model is used, even for stationary models. It may be wise to leave smoothness as a hyperparameter and use special cases of the Matérn function such as the exponential kernel ($\nu = 1/2$) or the squared exponential kernel ($\nu = 1$) as isotropic correlation $K_0(\cdot)$.

Solutions for problems 1, 2 and 3 would be:

1. Not to use full nonstationary model if the identification problems are confirmed

2. To use priors to guarantee that the spatially variable parameters will have a strong, smooth, large-scale spatial cohesion. For example, in problem 2, a short-scale prior for $\sigma(\cdot)$ will allow $\sigma(s)$ to go along $w(s)$, while a large-scale prior will bound it to nearby realizations of $\sigma(\cdot)$, giving a restoring force that will prevent $\sigma(s)$ from moving around freely. The extreme of this approach would be a prior that is so stiff that it is practically equivalent to a stationary model.

4.2.2 Nonstationary NNGP

Useful formulas about nonstationary NNGP. The stationary NNGP proceeds from the pruned recursive conditional form

$$\tilde{f}(w(s_i) \,|\, w(s_1, \ldots, s_{i-1}), \theta) = f(w(s_i) \,|\, w(pa(s_i)), \theta), \tag{4.8}$$

where $pa(s_i)$ is the set of parents for location s_i defined through the Directed Acyclic Graph (DAG), $\tilde{f}(\cdot)$ is the NNGP density, and $f(\cdot)$ is the non-approximated GP with mean 0 and covariance parameters (range and marginal variance) θ. The NNGP density is normal and the Cholesky factor of its precision matrix is extremely sparse (Katzfuss and Guinness, 2017). We note this factor $\tilde{R}$, such that the precision of the NNGP is $\tilde{R}^T\tilde{R}$ and its covariance is $\left(\tilde{R}^T\tilde{R}\right)^{-1}$.

The non-stationary NNGP density replaces the set of uniform covariance parameters θ by the spatially-variable $\theta(\mathcal{S})$. A critical point is that if the covariance model is given by (4.6) or (4.7) we have:

$$\tilde{f}(w(s_i) \,|\, w(s_1) \ldots, w(s_{i-1}), \theta(\mathcal{S})) = f(w(s_i) \,|\, w(pa(s_i)), \theta(s_i \cup pa(s_i))). \tag{4.9}$$

This means that instead of conditioning by $\theta(\mathcal{S})$ in the recursive conditional form we can condition only by $\theta(s_i \cup pa(s_i))$. We illustrate in annex 4.6.1

Another useful property of NNGP using the covariance model of (4.5) concerns the marginal variance. Note $\tilde{R}_0$ the NNGP factor obtained using the

correlation function $K_0(\cdot)$ instead of the covariance function $K(\cdot)$ and $\sigma(\mathcal{S})$ the nonstationary standard deviations taken at all spatial locations. The NNGP Cholesky factor can be written as

$$\tilde{R} = \tilde{R}_0 \; diag(\sigma(\mathcal{S}))^{-1}. \tag{4.10}$$

So the nonstationary NNGP behaves like full GP in the sense that, w_0 being the normalized NNGP with marginal variance 1, we have $diag(\sigma(\mathcal{S}))w_0(\mathcal{S}) = w(\mathcal{S})$. Demonstration is provided in annex 4.6.2. Therefore, the parts that intervene in NNGP density can be re-written as:

$$|(\tilde{R}^T\tilde{R})^{-1}|^{-1/2} = |\tilde{R}| = |\tilde{R}_0 \; diag(\sigma(\mathcal{S}))^{-1}| = \Pi_{i=1}^{n}(\tilde{R}_0)_{i,i}/\sigma(s_i) \tag{4.11}$$

in virtue of the triangularity of $\tilde{R}$, and

$$(w)^T\tilde{R}^T\tilde{R}(w) = \sigma^{-1}(\mathcal{S})^T \; diag(w)\tilde{R}_0^T\tilde{R}_0 \; diag(w)\sigma^{-1}(\mathcal{S}). \tag{4.12}$$

When evaluated in $\sigma^{-1}(\mathcal{S})$, (4.12) is proportional to MVN log-density with mean 0 and precision matrix $diag(w)\tilde{R}_0^T\tilde{R}_0 \; diag(w)$.

Nonstationary NNGP on the sphere. Paciorek (2003) gives a general method to construct a nonstationary function on the sphere using truncated kernels. This approach seemed tedious to transpose to NNGP, so we took advantage of the fact that the NNGP is defined locally to define nonstationary NNGPs on the sphere without defining a nonstationary covariance on the sphere. If the ordering of the points (Guinness, 2018) guarantees that the parents of a point s_i are close (within a few hundred kilometers), they can be projected on the tangent plane intersecting the sphere in s_i with little deformation. The nonstationary conditional Gaussian distribution can then be computed on the tangent plane, and a NNGP distribution arises from the local behaviors. Note that even though NNGP is widely used as an approximation of a full Gaussian process, here we are defining a NNGP without knowing the actual covariance function.
This approach is straightforward to apply in the case of (4.7), since the Euclidean distance on the tangent plane is not affected by a rotation of the plane's basis. However, in the case of (4.6), the Mahalanobis distance is used and rotation of the plane's basis matters. For regions that exclude the poles (not necessarily the actual magnetic poles but any couple of opposed points on earth), the tangent plane can be parametrized using the North and East directions as a basis. We did not find an approach that allows to work on the whole sphere.

Kullback-Leibler divergence between nonstationary NNGP and full nonstationary GP. The Kullback-Leibler (KL) contrast between two multivariate normal distributions $\mathcal{N}_0$ and $\mathcal{N}_1$ with respective means μ_0 and μ_1 and covariances Σ_0 and Σ_1 is

$$KL\left(\mathcal{N}_0 \parallel \mathcal{N}_1\right) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + \left(\mu_1 - \mu_0\right)^{\mathsf{T}}\Sigma_1^{-1}\left(\mu_1 - \mu_0\right) - k + \ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right)\right).$$

Consider a full GP and a NNGP with nonstationary marginal variance $\sigma(\mathcal{S})$, the same mean, and respective covariance matrices written thanks to (4.10) as a sandwich of spatial correlation and marginal variance matrices:

$$diag(\sigma(\mathcal{S}))\ \Sigma_0\ diag(\sigma(\mathcal{S})) \quad \text{and} \quad diag(\sigma(\mathcal{S}))\ (\tilde{R}_0^T \tilde{R}_0)^{-1}\ diag(\sigma(\mathcal{S})).$$

A first point is that if the marginal variance of the NNGP and the full GP are identical, the KL contrast boils down to the divergence between Σ_0 and $(\tilde{R}_0^T \tilde{R}_0)^{-1}$. Indeed,

- in the left-hand term, which can be re-written as

  $$tr\left(diag(\sigma(\mathcal{S}))\ \Sigma_0\ diag(\sigma(\mathcal{S}))\ diag(1/\sigma(\mathcal{S}))\ (\tilde{R}_0^T \tilde{R}_0)\ diag(1/\sigma(\mathcal{S}))\right)$$

  the inner variance matrices trivially cancel each other out, while the fact that the trace only uses the diagonal terms suppresses the outer variance matrices.

- the NNGP does not affect the mean, so $\mu_1 = \mu_0$ and the middle term vanishes.

- in the right-hand term, $|AB| = |A|\ |B|$ makes the determinants that involve $\sigma(\mathcal{S})$ cancel out.

This means that in a model where only the marginal variance is nonstationary, the KL contrast will be the same as in a full stationary model. One can then refer to the study of Guinness (2018).

In order to get an understanding of what happens when the range is nonstationary, we simulated credible nonstationary range parameter fields on two dimensions following the log-GP prior (see section 4.2.3). We computed the KL divergence between full GP and NNGP while checking for the number of nearest neighbors, the ordering, and the intensity of nonstationarity through the marginal variance of the log-GP. In the case of scalar range parameters, the impact of the intensity of nonstationarity is weak, but it becomes non-negligible when anisotropy is added. Regardless of the covariance type, the most important factors are the number of parents and the ordering. Like Guinness (2018) we found that the random and max-min orderings induce the lowest KL divergence. Details are provided in 4.7.

4.2.3 Log-Gaussian Process priors for spatially variable covariance parameters

Definition of the log-Gaussian Process prior. An attractive option to enforce some kind of coherence in the latent fields of positive parameters is to use a log-Gaussian Process (log-GP) prior (Heinonen et al., 2016). The latent field $\theta(\mathcal{S})$ is analyzed as

$$log(\theta(s)) = w_\theta(s) + X_\theta(s)\beta_\theta^T \quad \forall\ s \in \mathcal{S} \quad \text{and} \quad w_\theta(\mathcal{S}) \sim \mathcal{N}(0, \zeta_\theta). \tag{4.13}$$

Some regression coefficients β_θ parametrize linear effects (that may in particular include an intercept), and ζ_θ is a set of high-level covariance parameters for the log-GP prior.
The linear effects can answer two types of problems: first, a Gaussian Process is not the perfect tool to capture all spatial patterns. Some regressors such as the latitude, the time, the easting, or some factors, powers, sinusoidals derived from them can be of high interest. Also, one can be interested in explaining the parameter field using some regressors. By construction of the space-time hierarchical model, there is only one realization of $w(\cdot)$ by spatial location, so there cannot be more than two range or marginal variance parameters at the same location. Therefore, only regressors that do not change within one spatial location are allowed.
However, there are more than one realization of the Gaussian error process $\epsilon(\cdot)$ if several observations are done at the same spatial site. The Gaussian data model at site s becomes

$$z(s,i) = X(s,i)\beta^T + w(s) + \epsilon(s,i), s \in \mathcal{S}, 1 \leq i \leq (n_{obs}(s)), \tag{4.14}$$

$n_{obs}(s) \geq 1$ being the number of observations in the site s. While the regressors $X_{\tau^2}(s,i)$ may change within the spatial site s, the latent field $w_{\tau^2} \sim GP(\zeta_{\tau^2})$ is fixed, giving

$$\epsilon(s,i) \sim \mathcal{N}\left(0, \tau^2(s,i)\right) \quad \text{and} \quad log(\tau^2(s,i)) = w_{\tau^2}(s) + X_{\tau^2}(s,i)\beta_{\tau^2}^T. \tag{4.15}$$

A convenient and intuitive tool to compare covariance parameters. We favor log-GP priors for various reasons. The first is that a Gaussian variable is defined on the real numbers, while isotropic range parameters and the variance of a Gaussian noise are positive numbers. With respect to truncated distributions or other workarounds, simply passing the parameters to the logarithm seems less trouble. Moreover, the covariance parameters are essentially sizes: a variance is the size of a distribution, a range is the width or the volume of a covariance function. When comparing sizes, using the logarithm feels more natural: a random variable with variance 1.1 differs from a variable with variance 1 as much as a variable with variance 11 differs from a variable with variance 10, and not 10 times less. Following the typology of measurements of Stevens et al. (1946), the logarithm maps the ratio scale into the interval scale and allows for a good interpretability of the value of w_θ and β_θ.

Using the logarithm also wipes some parametrization problems out. For example, it is legitimate to wonder if it is better to use the variance, the precision, or the standard deviation in order to parametrize the heteroskedastic noise variance $\tau^2(\cdot)$ and the marginal variance $\sigma^2(\cdot)$. Similarly, is it better to compare the scalar range parameters $\alpha(\cdot)$ or some power such as $\alpha^d(\cdot)$? In a d-dimensional space-time domain, the covariance function's radius varies proportionally to α, but its volume varies proportionally to α^d. Once passed to the logarithm, those parametrizations only differ by a multiplicative constant. The parametrization problem is then turned into a parameter problem and can be solved fitting (or tuning) the intercept and the log-GP variance σ_θ^2.

4.2.4 Extension of the log-GP prior to positive-definite matrices for anisotropic range parameters

The previous log-GP distribution is not straightforwardly extended to nonstationary covariance functions with anisotropic range parameters such as those of Paciorek (2003). However, using the matrix generalization of scalar logarithm allows to find a consistent generalization. We proceed by analogy with 3 guidelines: the operations and objects involved in the matrix log-GP must be generalizations of their log-GP counterparts; the good properties of the log-GP prior must be carried over; the log-GP prior must be a special case of the matrix log-GP prior.

Matrix logarithm. Let A be a positive definite matrix of size $d \times d$. Let $(\lambda_1, \ldots, \lambda_d)$ and $(v_1, \ldots, v_d)$ be its eigenvalues and eigenvectors. The logarithm $log(A)$ is defined as the matrix whose eigenvalues are $log(\lambda_1), \ldots, log(\lambda_d)$, and the eigenvectors are $(v_1, \ldots, v_d)$. It is clear that $log(\cdot)$ maps the positive definite matrices into the symmetric matrices and that $log(A^{-1}) = -log(A)$, which removes parametrization problems like logarithm transformation.

Matrix log-GP prior. Analogously to (4.13), the logarithm of the range matrices is analyzed as

$$log(A(s)) = W(s) + \Sigma_{i=1}^{n_{X_A}} X_i(s) \times B_i, \tag{4.16}$$

n_{X_A} being the number of covariates, $X_i(\cdot),\ 1 \leq i \leq n_{X_A}$, the i^{th} regressor and $B_i, 1 \leq i \leq n_{X_A}$ being a $d \times d$ symmetric matrix. Since B_i does not depend on s, $\Sigma_{i=1}^{k} X_i(s) \times B_i$ must be seen as a linear effect.
On the other hand $W(s)$ is a $d \times d$ random symmetric matrix. Consider an orthonormal basis of the symmetric matrices $(e_1, \ldots, e_{d(d+1)/2})$. Denote $(w_1(s), \ldots, w_{d(d+1)/2}(s))$ the coordinates of $W(s)$ in this basis. Analogously to (4.13), the matrix log-GP prior is defined as:

$$\big(w_1(s_1, \ldots, s_n), \ldots, w_{d(d+1)/2}(s_1, \ldots, s_n)\big) \sim \mathcal{N}(0, S \otimes \Sigma_0), \tag{4.17}$$

$\otimes$ being the Kronecker product, Σ_0 being a $n \times n$ hyper-parameter correlation matrix and S being a positive-definite matrix with $d(d+1)/2$ rows and columns. Σ_0 is a spatial correlation matrix and accounts for correlation between sites, while S is the matrix analogous of a marginal variance parameter and accounts for the multivariate correlation within one site. Like the marginal variance of the log-GP prior, S is estimated by the model.

Denoting β_{A_i} with $1 \leq i \leq d(d+1)/2$ the vector obtained with the projections of $(B_1, \ldots, B_{n_{X_A}})$ on the i^th element of the basis of symmetric matrices, and $log(A)_i$ the projection of $log(A)$ on the same element, we have

$$(log(A(\mathcal{S}))_1, \ldots, log(A(\mathcal{S}))_{d(d+1)/2}) \sim \mathcal{N}(\mu, S \otimes \Sigma_0), \tag{4.18}$$

the mean μ being obtained by stacking vertically $X_A\beta_{A_1}, \ldots, X_A\beta_{A_{d(d+1)/2}}$. The stationary model and the nonstationary range model with scalar range are included in a model with matrix log-GP. If S is null and that all B_is are null

except for the pseudo-intercept, then the induced correlation is stationary. As for the scalar range case, denote v the coordinates of the matrix $I_d/\sqrt{d(d+1)/2}$ in the chosen basis of symmetric matrices. If $S = \sigma_\alpha^2 \times v^T v$, the random effect $w_1(s), \ldots, w_{d(d+1)/2}(s)$ is degenerate and its support is restricted to the matrices that are proportional to I_d. If in addition $B_i = \beta_i \Sigma_{j=1}^{d(d+1)/2} v_j e_j = \beta_i I_d/\sqrt{d(d+1)/2}$, we recognize the nonstationary correlation with scalar range parameters of (4.7).

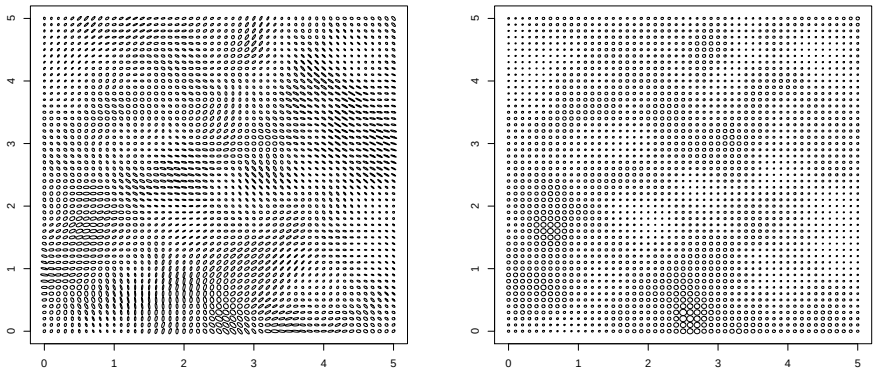
This point is illustrated in figure 4.1. On the left hand side, we see range ellipses generated with the matrix log NNGP prior (see below section 4.2.5), and a NNGP sample corresponding to those parameters that has been obtained using $w = \tilde{R}^{-1} w^*$, with $w^* \sim \mathcal{N}(0, I)$ and $\tilde{R}$ obtained from the elliptic range parameters using (4.6). On the right hand side, we removed the anisotropy components of the ellipses (the matrices are projected on $I_d/\sqrt{d(d+1)/2}$) and we are left with circles. While $\tilde{R}$ was obtained using (4.7) and the circular range parameters, w^* was not changed. The resulting NNGP sample is obtained like before as $w = \tilde{R}^{-1} w^*$. We can see that the circles of figure 4.1b and the ellipses of figure 4.1a have similar sizes even if they differ in shape. The NNGP samples paths are very similar too, and figure 4.1c looks like figure 4.1d "plus some local anisotropy".

Evaluation and sampling. Sampling is easy using (4.17). Simulation of the matrix log-GP prior on a grid with no linear effects and null intercept shows a coherent field of ellipses as seen in Figure 4.1a. Predictions of $A(\cdot)$ can be done applying standard Gaussian conditional distribution formulas to (4.17).

Interpretation. The coordinates of the log-matrix are not on an equal footing. On the one hand, the coordinate of the matrix $I_d/\sqrt{d(d+1)/2}$, noted v, controls the determinant of the log-matrix. Moving this coordinate of the log-matrix will inflate or deflate the corresponding range ellipse homothetically but will not change its orientation or the intensity of the anisotropy. On the other hand, the rest of the coordinates of the basis for the symmetric matrices do not affect the determinant. They change the shape of the ellipse, either by modifying its direction or the ratio of its axes (an increase of either axis being compensated by an inverse decrease of the other, the determinant does not move). Those coordinates cannot be separated in groups parametrizing the rotation or the elongation alone, they all do both.

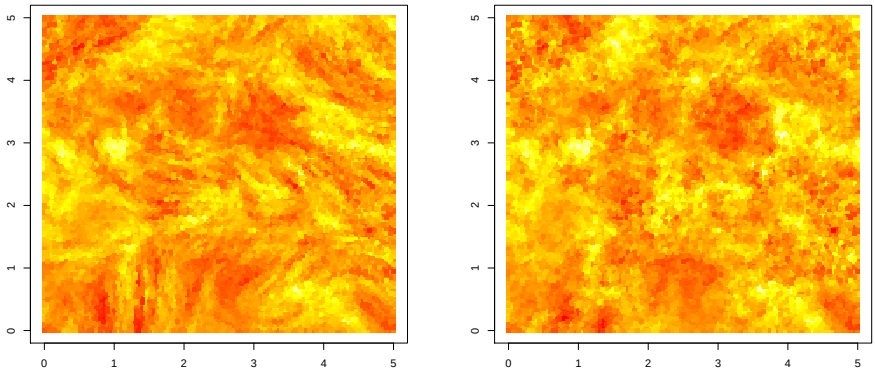
4.2.5 Hierarchical architecture using NNGPs

In view of the good computational properties and accuracy of NNGP approximation (Guinness, 2018; Katzfuss and Guinness, 2017), we use log-NNGP and Matrix log-NNGP as priors for the spatially variable parameters. They are obtained replacing the covariance matrix $\Sigma(\mathcal{S}, \theta)$ by a NNGP approximation $(\tilde{R}_\theta^T \tilde{R}_\theta)^{-1}$ in the log-GP and matrix log-GP priors.



(a) Ellipses obtained with matrix log NNGP

(b) Circles obtained with scalar log NNGP



(c) NNGP samples corresponding to the ellipses

(d) NNGP samples corresponding to the circles

Figure 4.1: Example of range ellipses and GP samples induced by the log-NNGP and matrix log-NNGP priors

Estimating or not the covariance structure of log-NNGP priors. The log-NNGP and matrix log-NNGP priors for the latent covariance parameters fields $\theta(\cdot)$ are themselves parametrized by at least a covariance matrix $(\tilde{R}_\theta^T \tilde{R}_\theta)^{-1}$ and an intercept that is integrated in β_θ. In Heinonen et al. (2016), the covariance parameters and the intercept are treated as hyper-parameters. We choose to leave only the hyperprior range as a user-chosen parameter, while estimating σ_θ^2 (or its counterpart S for elliptic range) and β_θ. On the other hand, we chose not to sample α_θ. First, it is a costly operation since it involves to compute $\tilde{R}_\theta$. Moreover, in view of the identification problems that could occur in nonstationary models, we advocated for a prior with a high spatial coherence. In the case of a log-GP prior, this means that the range α_θ should be high with respect to the domain size. Given the identification problem that occurs between range and variance in fill-in asymptotic (Zhang, 2004), estimating the range would bring a very marginal improvement.

Prior distributions on high-level parameters. The high-level parameters that are estimated by the model are the linear regression coefficients β_θ and the variance parameter σ_θ^2 or S. We put an improper prior on β_θ. This choice of an improper prior on the logarithm of a positive parameter is quite standard, but the literature of stationary spatial models generally advocates for stronger priors (Fuglstad et al., 2015b; Datta et al., 2016). As for the variance parameter, we put a uniform prior on a $[-8; 3]$ window for $log(\sigma_\theta^2)$, and for each log-eigenvalue of S in the matrix case. The bottom of the interval induces a model that is practically stationary since the variance of the field of parameters will be very close to 0. We chose not to let the variance fall any lower in order to avoid straying and numerical problems. On the other hand, $exp(3) \approx 20$, which means that the latent field can have a high variance too. We did not choose to allow the field to go any higher because of numerical problems, and because there would be no sensible interpretation of an extremely variable field of parameters.

4.3 MCMC strategy

Log-NNGP and matrix log-NNGP induce a hierarchical model on four levels

$$\begin{aligned} &[\text{data} \,|\, \text{process, parameter process, parameters}] \quad \times \quad [\text{process} \,|\, \text{parameter process, parameters}] \\ &\qquad [\text{parameter process} \,|\, \text{parameters }] \quad \times \quad [\text{parameters}]. \end{aligned}$$

The implementation of a Gibbs sampler for such a model is a difficult task. Generally, the bigger and the more complex the model, the more trouble should be expected. In addition, one very problematic point stands out: the high correlation between fields and parameters. The problem is well-known in stationary models and various articles address it through tactics such as blocking (Knorr-Held and Rue, 2002), collapsing (Finley et al., 2019) or interweaving (Yu and Meng, 2011; Filippone et al., 2013). In our model, this behavior is likely to occur in both [process | parameter process, parameters] and [parameter process | parameters]. Moreover, spatial auto-correlation makes it difficult to sample from the latent

fields, in particular the parameter fields that have a long range because of the model construction. We decided to address those problems with a MCMC strategy based on two pillars:

1. A nested interweaving scheme, well adapted to the structure of our model.
2. Hybrid Monte-Carlo steps to deal with spatial auto-correlation.

The lead of chromatic sampling was investigated, but it become clear that HMC was a better option. We outline the approach in annex 4.8.

4.3.1 Gibbs sampler architecture using Nested Interweaving

Interweaving. Interweaving is a method introduced by Yu and Meng (2011), which improves the convergence speed of models relying on data augmentation. Usually various parametrizations of the data augmentation are available. For example, in the context of our NNGP model, the latent field

$$w \sim \mathcal{N}(0, (\tilde{R}^T \tilde{R})^{-1})$$

can be re-parametrized as

$$w_{center} = X_{locs}\beta^T + w \sim \mathcal{N}(X_{locs}\beta^T, (\tilde{R}^T \tilde{R})^{-1}),$$

X_{locs} being the covariates that do not change within one spatial location. Another parmetrization is the prior whitening:

$$w^* = \tilde{R}w \overset{a\ priori}{\sim} \mathcal{N}(0, I_n).$$

The component-wise interweaving strategy of Yu and Meng (2011) can be applied when two data augmentations w_1 and w_2 have a joint distribution $[\theta, w_1, w_2]$ (even if it is degenerate) such that its marginals $[\theta, w_1]$ and $[\theta, w_2]$ correspond to the two models with the different data augmentations. It takes advantage of the discordance between the two parametrizations to construct the following step in order to sample θ^{t+1}:

$$[\theta, w_2 | w_1^t, \ldots] \rightarrow [\theta^{t+1}, w_1^{t+0.5} | w_2, \ldots],$$

"..." being the other parameters of the model. Since all the draws are done from full conditional distributions, the target joint distribution is always preserved. Joint sampling of the parameter and the data augmentation is much easier to implement when decomposed as:

$$\underbrace{[\theta | w_1^t, \ldots] \rightarrow [w_2 | w_1^t, \theta, \ldots]}_{[\theta, w_2 | w_1^t, \ldots]} \rightarrow \underbrace{[\theta^{t+1} | w_2, \ldots] \rightarrow [w_1^{t+0.5} | w_2, \theta^{t+1}, \ldots]}_{[\theta^{t+1}, w_1^{t+0.5} | w_2, \ldots]}.$$

It is possible that the joint distribution is degenerate as long as it is well-defined, so that $[w_2 | \theta, w_1]$ and $[w_1^{t+0.5} | w_2, \theta^{t+1}, \ldots]$ are often deterministic transformations (in our application they are). For this reason even though the data augmentation is changed at the end of the sampling of θ, w_1 still has to be updated

in a separate step in order to have an irreducible chain: that is why we indexed it by $t + 0.5$.

The strategy being based on the discordance between two parametrizations, it is a good choice to pick an ancillary-sufficient couple, giving an Ancillary-Sufficient Interweaving Strategy (ASIS). Interweaving can work very well even though none of the two augmentations performs well when implemented separately. Following the terminology of Yu and Meng (2011), w is sufficient when *a posteriori* $(\theta|w, z) = (\theta|w)$, z being the observed data. It is sufficient when it is *a priori* independent from θ. ASIS already proved its worth for GP models: Filippone et al. (2013) show empirically that updating covariance parameters in a Gaussian Process model benefits from interweaving w (sufficient) and w^* (ancillary), while Coube and Liquet (2020) show that interweaving w (ancillary) and w_{center} (sufficient) improves the sampling of the fixed effect coefficients.

Nested interweaving for high-level parameters. The problem here is that there are some latent fields on various layers of the model. Nested ASIS is envisioned by Yu and Meng (2011) for such models, even though the authors do not provide application to realistic models.

Consider a high-level parameter concerning the log-NNGP distributions of the covariance parameters. Those high-level parameters may be the marginal variance of a log-NNGP distribution or the regression coefficients from (4.13), or (4.16) and (4.17). This parameter is noted ζ. Note $(w_\theta)_1$ and $(w_\theta)_2$ the two parametrizations for the corresponding log-NNGP field of covariance parameters. Eventually, note w_1 and w_2 the two parametrizations of the NNGP latent field from (4.2). If we aim to sample regression coefficients β_θ, the centered and non-centered parametrizations of w_θ will be used, while if we aim to sample the marginal variance S or σ^2_θ, the whitened parametrization will be used. A nested interweaving step aiming to update ζ can be devised as

$$\underbrace{\left.\begin{array}{c}[\zeta, (w_\theta)_2, w_2|(w_\theta)_1, w_1, \ldots] \to [\zeta, (w_\theta)_1, w_2|(w_\theta)_2, w_1, \ldots] \\ \swarrow \\ [\zeta, (w_\theta)_2, w_1|(w_\theta)_1, w_2, \ldots] \to [\zeta, (w_\theta)_1, w_1|(w_\theta)_2, w_2, \ldots]\end{array}\right\} \text{interweaving } w}_{\text{interweaving } w_\theta} \tag{4.19}$$

Like before, it is much easier to sample sequentially, for example the blocked draw $[\zeta, (w_\theta)_2, w_2|(w_\theta)_1, w_1, \ldots]$ writes as

$$[\zeta|(w_\theta)_1, w_1, \ldots] \to \underbrace{[(w_\theta)_2|\zeta, (w_\theta)_1, w_1, \ldots]}_{\text{deterministic}} \to \underbrace{[w_2|\zeta, (w_\theta)_1, (w_\theta)_2, w_1, \ldots]}_{\text{deterministic}}.$$

Like before too, w_θ needs to be updated later, using an interweaving of parametrizations of w.

Centering-upon-whitening nested interweaving for the log-NNGP regression coefficients. Coube and Liquet (2020) show that updating the regression coefficients of (4.2) using an interweaving of w and $w_{center} = w + X\beta^T$

considerably improves the behavior of the chains, in particular when some covariates have spatial coherence. The limit is that it cannot be applied for the covariates that vary within one spatial location. We apply this strategy to update β_α, β_{τ^2} and β_{σ^2}. The covariates explaining the range and marginal variance cannot vary within one spatial location anyway. In the case of the scalar range and the latent field's marginal variance, we are using nested interweaving. The two relevant parametrizations for the NNGP latent field of (4.2) are the natural parametrization and the whitened latent field $w^* = \tilde{R}w$. So, for β_α and β_{σ^2}, the sampling step derived from (4.19) is

$$[\beta_\theta|w_\theta, w, \ldots] \to [\beta_\theta|(w_\theta)_{center}, w, \ldots] \to [\beta_\theta|w_\theta, w^*, \ldots] \to [\beta_\theta|(w_\theta)_{center}, w^*, \ldots].$$

For the sake of simplicity we do not write the implicit updates of the latent fields at each sampling of β_θ. $(w_\theta)_{center}$ being a sufficient augmentation, sampling from $[\beta_\theta|(w_\theta)_{center}, w, \ldots]$ is the same as sampling from $[\beta_\theta|(w_\theta)_{center}]$. The procedure is described in Coube and Liquet (2020). As for the updates conditionally on w_θ, they can be done with an usual Metropolis-within-Gibbs sweep over the components of β_θ or with a Hybrid Monte-Carlo step detailed in the following section 4.3.2. When the model has elliptic range parameters, the approach is essentially the same but the "Vec trick" is needed. Details are given in annex 4.9.1.

We have not found a satisfactory couple of parametrizations for w when the matter is to update β_{τ^2}. We use simple (non-nested) interweaving:

$$[\beta_{\tau^2}|w_{\tau^2}, w, \ldots] \to [\beta_{\tau^2}|(w_{\tau^2})_{centered}, w, \ldots].$$

Whitening-upon-whitening nested interweaving for the log-NNGP variance. In the case of the marginal variance σ^2_θ of a log-NNGP prior, two parametrizations of w_θ are available. The sufficient parametrization is the natural parametrization, while the ancillary parametrization is the whitened $w^*_\theta = \tilde{R}_{0_\theta} w_\theta / \sigma_\theta$, $\tilde{R}_{0_\theta}$ being the hyperprior correlation NNGP factor. Like before, for the latent field, we use w and w^*. For the marginal variance σ^2, the circular range α, and the elliptic range A, the step writes:

$$[\sigma^2_\theta|w_\theta, w, \ldots] \to [\sigma^2_\theta|w^*_\theta, w, \ldots] \to [\sigma^2_\theta|w_\theta, w^*, \ldots] \to [\sigma^2_\theta|w^*_\theta, w^*, \ldots].$$

Since w_θ is a sufficient statistic for σ^2_θ, $[\sigma^2_\theta|w_\theta, w, \ldots]$ or $[\sigma^2_\theta|w_\theta, w^*, \ldots]$ are equivalent to $[\sigma^2_\theta|w_\theta]$. The procedure to update a marginal variance with such a parametrization is well-known (Banerjee et al., 2008; Datta et al., 2016). When the ancillary parametrization w^*_θ is used, a Metropolis-Hastings step or a HMC step detailed below can be used. Contrary to the case of fixed effects, it is straightforward to generalize to the matrix log NNGP prior.

Like before, only the sufficient parametrization of w is used for the noise variance. The step is:

$$[\sigma^2_{\tau^2}|w_{\tau^2}, w, \ldots] \to [\sigma^2_{\tau^2}|w^*_{\tau^2}, w, \ldots].$$

4.3.2 Hybrid Monte-Carlo to sample parameter fields with log-GP priors

Hybrid Monte-Carlo (HMC) (Neal et al., 2011) has already been implemented successfully by Heinonen et al. (2016) for nonstationary Gaussian processes, but in our case the gradient of the model negated log-density with respect to $w_\theta(\mathcal{S})$ must be found since we are using NNGP and not full GP, and the applicability of HMC must be questioned for large spatial data sets. The other divergence with respect to the methodology of Heinonen et al. (2016) is that we sample the latent field $w(\cdot)$, while Heinonen et al. (2016) use a response model. Given the fact that we use a nested ASIS strategy with the natural and the whitened parametrizations of the latent NNGP field, we need the gradients of the density with both model formulations.

Generic form for the gradients. We work with a log-NNGP or —matrix log NNGP prior for the three spatially variable parameters (range, scale, noise variance). The negated log likelihood with respect to the field $w_\theta(\mathcal{S})$ will then be

$$H = -log(\tilde{f}_\theta(w_\theta(\mathcal{S})) - g_\theta(w_\theta(\mathcal{S}))$$

$\tilde{f}_\theta(\cdot)$ being a NNGP density with covariance $(\tilde{R}_\theta^T \tilde{R}_\theta)^{-1}$ involved in the log-NNGP prior and $g_\theta(\cdot)$ depending on the role of θ in the model. The gradient of this potential is written as:

$$\nabla_{w_\theta} H = \tilde{R}_\theta^T \tilde{R}_\theta w_\theta - \nabla_{w_\theta} g_\theta(w_\theta(\mathcal{S})).$$

However, this potential leads to inefficient HMC. According to Heinonen et al. (2016) prior whitening may lead to improvements of "several orders of magnitude". In order to do a whitened HMC step, we search the gradient of H with respect to $w_\theta^*(\mathcal{S}) = \tilde{R}_\theta w_\theta(\mathcal{S})$. This gradient is given as

$$\nabla_{w_\theta^*} H = \tilde{R}_\theta w_\theta - (\tilde{R}_\theta^{-1})^T \nabla_{w_\theta} g_\theta(w_\theta(\mathcal{S})). \tag{4.20}$$

Details are provided in annex 4.10.1. Solving or multiplication involving $\tilde{R}_\theta$ is not an issue thanks to the fact that $\tilde{R}_\theta$ is sparse and triangular. In the matrix log NNGP prior case, where $\tilde{R}_\theta = S^{-1/2} \otimes \tilde{R}_0$, the "Vec trick" $(A \otimes B)\, C = B\, Vec(C)\, A^T$ comes in handy. The problem that remains is to compute $\nabla_{w_\theta} g_\theta(w_\theta(\mathcal{S}))$. In the following paragraphs, we derive this gradient for the various parameters of the model (range, marginal variance, noise variance).

Marginal variance latent field. When sufficient augmentation is used, the marginal variance intervenes in the NNGP density of the latent field. Using (4.11) and (4.12), the gradient writes

$$-\nabla_{w_{\sigma^2}} g_{\sigma^2}(w_{\sigma^2}) = 1/2 + \sigma^{-1}(\mathcal{S}) \;\circ\; \Big(diag(w)\tilde{R}_0^T \tilde{R}_0\; diag(w) \quad \sigma^{-1}(\mathcal{S})\Big)\,/2, \tag{4.21}$$

with $\circ$ the Hadamard product.
When ancillary augmentation is used, the marginal variance has an impact on the observed field likelihood with respect to the latent field. The gradient with respect to $w_{\sigma^2}(\mathcal{S})$ will write

$$-\nabla_{w_{\sigma^2}} g_{\sigma^2}(w_{\sigma^2}) = \nabla_w \; l(z(\mathcal{S})|\tilde{R}^{-1}w^*, \beta, \ldots) \; \circ (\tilde{R}^{-1}w^*/2), \tag{4.22}$$

$l(\cdot)$ being the likelihood of the observations knowing the latent field, the fixed effect, and the noise variance in the case of a Gaussian model. The demonstrations are given in the annex 4.10.2. Given the fact that only sparse matrix multiplications and sparse triangular system solving are involved, computing those two gradients is affordable.

Range latent field. In the following, $w_\alpha(s_j)$ can be the latent field for a model with isotropic range, (4.13), or one of the $d(d+1)/2$ latent fields for a model with anisotropic range (4.17). We find the gradient of the negated log-density with respect to w_α using a two-step method.
The first part is to compute the derivatives of $\tilde{R}$ with respect to w_α. The details are presented in annex 4.10.3. Even if the formulas seem daunting, their computational cost actually is affordable thanks to the sparsity induced by the covariance functions (4.6) and (4.7). In annex 4.10.4, we estimate that both the number of flops and the RAM needed respectively to compute and store the derivatives of $\tilde{R}$ are proportional to $n\ m^2$, n being the size of the data set and m being the size of the parents sets used in the NNGP approximation.
The second step is to express the gradient of H using the derivatives of $\tilde{R}$. In the case of sufficient augmentation, we have

$$-\frac{\partial g_\alpha(w_\alpha)}{\partial w_\alpha(s_i)} = \left(w^T\tilde{R}^T\right)\frac{\partial \tilde{R}}{\partial w_\alpha(s_i)}w + \Sigma_{j/s_j\in\{s_i\cup ch(s_i)\}}\frac{\partial \tilde{R}_{j,j}}{\partial w_\alpha(s_i)}/\tilde{R}_{j,j}, \tag{4.23}$$

$ch(s_i)$ being s_i's children in the DAG used to define the NNGP and $\tilde{R}_{j,j}$ being the j^{th} diagonal term of $\tilde{R}$. When ancillary parametrization is used,

$$-\frac{\partial g_\alpha(w_\alpha)}{\partial w_\alpha(s_i)} = \nabla_w log(l(z(s_i)|\tilde{R}^{-1}w^*, X, \beta, \ldots))^T\tilde{R}^{-1}\frac{\partial \tilde{R}}{\partial(w_\alpha(s_i))}\tilde{R}^{-1}w^*, \tag{4.24}$$

where $l(\cdot)$ is the likelihood of the observations. Note that we use the gradient ∇_w of the likelihood with respect to w, evaluated at $w = \tilde{R}^{-1}w^*$. Details about the gradients are found in annex 4.10.5. Due to the fact that $\partial\tilde{R}/\partial(w_\alpha(s_j))$ is extremely sparse, the gradients can be computed for an affordable computational cost discussed in annex 4.10.6.

Noise variance latent field. The noise variance intervenes directly in the Gaussian likelihood of the observed field. The gradient of the negated log-density writes:

$$-\nabla_{w_{\tau^2}} g_{\tau^2}(w_{\tau^2}) = 1/2 - \tau^2(\mathcal{S}) \circ (z(\mathcal{S}) - w(\mathcal{S}) - X(\mathcal{S})\beta^T)^2/2. \tag{4.25}$$

Details are given in annex 4.10.7. Note that this method should be affordable for other link functions as long as they are cheap to evaluate and differentiate.

Regression coefficients for the covariance parameters. This paragraph shows how to update β_θ with a HMC step. This method is especially useful for the range parameters since it avoids an unaffordable Metropolis-within-Gibbs sweep over β_α. Using the Jacobian chain rule,

$$\nabla_{\beta_\theta} H = J_{\beta_\theta}^T log(\theta) \cdot \nabla_{log(\theta)} H = X_\theta^T \cdot \nabla_{log(\theta)} H.$$

In the case of the log-range and log-variance, there is a one-to-one correspondence between $log(\theta)$ and w_θ, so that it is possible to replace $\nabla_{log(\theta)} H$ by $\nabla_{w_\theta} H$. In the case of the noise variance, $\nabla_{log(\tau^2)} H$ is straightforward to derive using $\frac{\partial l(z_i(s)|\tau_i^2(s), w(s), X_i(s), \beta)}{\partial \tau_i^2(s)}$ (i being the index of the observation at site s).

Marginal variance for the covariance parameters. When the ancillary parametrization $w_\theta^* = \tilde{R}_\theta w_\theta / \sigma_\theta$ is used, changing σ_θ^2 (or the within site multivariate covariance S in the case of the elliptic range) has an impact on w_θ. Using the Jacobian chain rule,

$$\nabla_{\sigma_\theta^2} H = J_{\sigma_\theta^2}^T w_\theta \cdot \nabla_{w_\theta} H = J_{\sigma_\theta^2}^T (\sigma_\theta \tilde{R}_\theta^{-1} w_\theta^*) \cdot \nabla_{w_\theta} H = (\tilde{R}_\theta^{-1} w_\theta^* / 2\sigma_\theta)^T \cdot \nabla_{w_\theta} H.$$

In the case of elliptic range parameters, we have in virtue of the "Vec trick":

$$\nabla_S H = J_S^T w_A \cdot \nabla_{w_A} H = J_S^T (S^{1/2} \otimes \tilde{R}_A^{-1} w_A^*) \cdot \nabla_{w_A} H = J_S^T (Vec(\tilde{R}_A^{-1} w_A^* (S^{1/2})^T)) \cdot \nabla_{w_A} H.$$

In order to get the Jacobian, the derivatives of $S^{1/2}$ with respect to S are obtained by finite differences. The derivatives of $\tilde{R}_A^{-1} w_A^* (S^{1/2})^T$ are in turn obtained by matrix multiplication, and plugged into the $Vec(\cdot)$ operator.

4.4 Data analysis

While the raw results of the experiments and applications are too heavy to be available online, our code can be found at `https://github.com/SebastienCoube/Nonstat-NNGP`. Extensive vignettes and code enabling to reproduce the applications are provided.

4.4.1 Empirical guidelines

This subsection aims to give some answers to practical problems that may arise with our model.

Over-modeling. For the choice of the model, the results of an experiment presented in annex 4.11.2 and summarized in figure 4.8 and 4.9 tell us that over-modeling does not hurt in terms of Deviance Information Criterion (DIC) (Spiegelhalter et al., 1998). However, the problem of wasting time and resources fitting a complex and costlier model remains.

If over-modeling does not affect the performance of the model, it is because the non-stationary model encompasses the stationary model, and boils down

to stationarity when confronted with stationary data. When stationary data is analyzed with a non-stationary model, the marginal variance parameter of the log-NNGP prior sticks to 0, inducing a degenerate distribution. The parameter latent field ends up being constant, effectively inducing a stationary model.

Model selection. An useful corollary is that over-modeling can be detected just by looking at the MCMC chains, without needing to wait for full convergence. For example, in figure 4.2, we can see the 2000 first states, for 3 separate chains, of the log-variance parameter for a range log-NNGP prior. On the left, the data is stationary, and the log-variance is very low. On the right, the data is non-stationary, and the log-variance is high enough to allow the parameter to move in the space.

As for the model with anisotropic range parameters, it is also possible to detect over-modeling from the estimates. In order to do so, we look at the matrix logarithm of S from (4.17). Recall that if $S \approx v^T \sigma^2 v$, v being the projection of $I_d/\sqrt{d(d+1)/2}$ in the chosen basis of symmetric matrices, then the model is effectively a nonstationary scalar range model. If S is null, the model is stationary. We monitor three indicators:

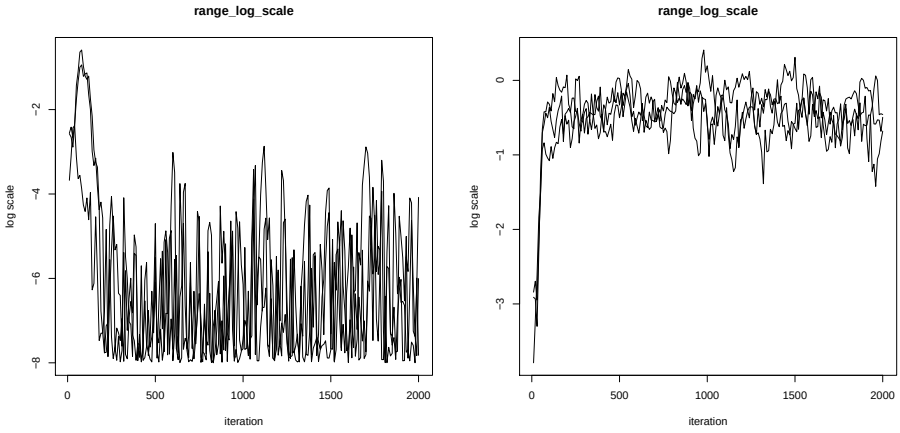
$$v\ log(S)\ v^T, \qquad u\ log(S)\ u^T, \qquad x\ log(S)\ x^T$$

with u, x being a completion of v in the basis of the symmetric matrices.

In figure 4.3, we show the behavior of the indicators for three data sets: a stationary data set (figure 4.3a), a nonstationary data set with scalar range (figure 4.3b), and a nonstationary data set with elliptic range (figure 4.3c). We can see that all three components are very low in figure 4.3a, implying $S \approx \mathbf{0}_{3\times 3}$, which makes in turn w_A constant, eventually inducing a stationary prior for w. When the range is nonstationary with scalar parameters (4.3b), vSv^T (in black) raises while the two other indicators are low. Eventually, when the data is nonstationary with elliptic range parameters (4.3c), all three indicators are high.

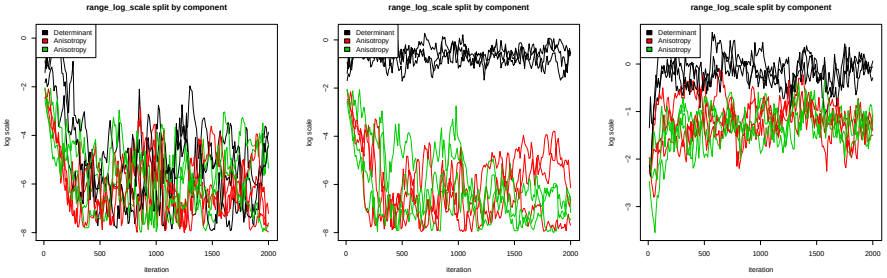
Identification of the parameters. A first approach to tell the identification of parameters is to use model comparison criteria such as the DIC. If the parameters are not well-identified, then a change in the chosen model, for example replacing a model with nonstationary range by a model with nonstationary noise variance, should not affect the chosen criterion. From the experiment presented in 4.11.2, it is clear that nonstationary noise variance is well-identified and that forgetting it in relevant cases leads to under-fitting.

Omitting both scalar range and marginal variance of the latent NNGP process leads to under-fitting as well, but the identification of those two parameters is less clear. On the one hand, on data with nonstationary range, a model with nonstationary variance does not do as good as a model with nonstationary range (see figure 4.8d in annex). On the other hand, the converse is not true for data with nonstationary variance (figure 4.8f); and on data with both non-stationary range and marginal variance, models with only either nonstationary range or



(a) Stationary data (b) Data with nonstationary range

Figure 4.2: Log-variance of the log-NNGP prior of the range parameter



(a) Stationary data (b) Data with nonstationary scalar range (c) Data with nonstationary elliptic range

Figure 4.3: Log-scale analysis of the matrix log-NNGP prior of the range parameter

variance do as good as the model with both (figure 4.8b). This problem is not surprising: on small domains, range and variance are difficult to identify for stationary models (Zhang, 2004).

However, a troubling observation shows that there is some kind of identification: when given the possibility, our model is able to make the right choice between the two parameters. In figure 4.10, we used boxplots to summarize results of the models that estimate both nonstationary marginal variance and range. On the left (4.10a), we can see estimates for the log-variance of w_α's log-NNGP prior. On the right (4.10b), we see its counterpart for w_{σ^2}. In both subfigures, the boxplots are separated following the type of the data, ($\emptyset$) being stationary data, (α) being data with nonstationary range, (σ^2) being data with nonstationary variance, and ($\alpha + \sigma^2$) being data with both nonstationarities (annex 4.11.1 presents the naming system in detail). Recall that when the log-variance is low, the corresponding field is practically stationary. Then we can see that the right kind of nonstationarity is detected for all four configurations: when data is stationary, both log variances are very low, when the data is (σ^2), then only the log-variance of w_{σ^2} is high, etc.

4.4.2 Case study: lead concentration in the United States of America mainland

About the data set. The lead data set presented by Hengl (2009) features various heavy metals measurements, including lead concentration. Various anthropic (density of air pollution, mining operations, toxic release, night lights, roads) and environmental (density of earthquakes, indices of green biomass, elevation, wetness, visible sky, wind effect) covariates are provided. Those variables may impact the emission of the lead, its diffusion, or both. The lead concentration and the covariates have been observed on 58097 locations, with a total of 64274 observations. As we can see in figure 4.4, the measures are irregular, with gaps and denser patches. The observations were passed to the log.

About the analysis. We used a NNGP with 5 neighbors and the max-min order. We tested 3 models: a model with non-stationary circular range, scale and noise ($\alpha + \sigma^2 + \tau^2$), a model with just scale and noise ($\sigma^2 + \tau^2$), and a stationary model ($\emptyset$). With the full log-NNGP prior, the chains had a pathological behavior, forcing us to integrate only a linear regression in the covariance structure. The model ($\alpha + \sigma^2 + \tau^2$) needed 5000 iterations and 3 hours to converge following the Gelman-Rubin-Brooks diagnostics (Neal, 2011). The model ($\sigma^2 + \tau^2$) was tested in order to see whether the "over-modeling does not hurt" rule that we deduced from synthetic experiments held in practice. The DIC of the three models are compared in table 4.1. The model ($\alpha+\sigma^2+\tau^2$) is therefore selected. Even though the DIC clearly tells that a nonstationary model must be chosen, it is legitimate to worry about problems of identification between all those effects. To investigate the problem, we present the correlation matrix of the MCMC samples for $(\beta, \beta_\alpha, \beta_{\sigma^2}, \beta_{\tau^2})$ in annex, figure 4.12. Problems of

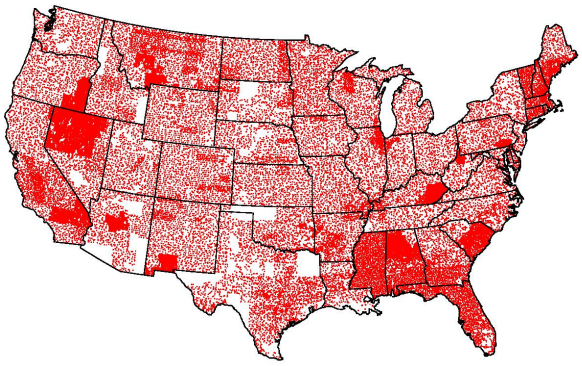


Figure 4.4: Measure sites for lead concentration

Table 4.1: DIC performances of various models on the lead data set

Model	$(\alpha + \sigma^2 + \tau^2)$	$(\sigma^2 + \tau^2)$	$(\emptyset)$
DIC	76758	78415	86158

identification should lead to high correlation between the samples. Our conclusion is that there is some correlation, but not any worse than usual. Indeed, some sub-diagonals are standing out, denoting that there is some correlation between the components of $\beta_\alpha, \beta_{\sigma^2}$ and β_{τ^2} associated to the same covariate. However, the strongest correlation occurs between the intercept coefficients of β_α and β_{σ^2}, which corresponds to the well-known problem of stationary models (Zhang, 2004).

Results. After using a PCA on the table of regression coefficients $(\beta_\alpha|\beta_{\sigma^2}|\beta_{\tau^2})$ obtained by aggregating the columns "mean" of table 4.2 and excluding the intercepts, it appears that the role of the covariates can be summarized by their role in the incoherence of the observed signal. A variable that increases the incoherence will lower the range for the latent process and increase the variance for both the noise and the latent process. Details are given in annex, figure 4.11. The wetness and the elevation seem to augment the coherence, while the presence of green biomass and the density of lights make the signal fuzzier. This incoherence component is plotted in figure 4.5. The predicted means of lead contamination are quite similar between the stationary and nonstationary model (figure 4.6). However, the prediction of the nonstationary model appears sharper, pointier in Colorado or Arizona, where the spatial effects are more incoherent. On the other hand, we can see in figure 4.5 that the spatial effects

are more coherent in Missouri except for a few spots corresponding to cities (Kansas City in the West, Saint Louis in the East, Springfield in the South-West) and mining counties (Viburnum in the South-East), who respectively affect the coherence of the spatial effect through the road and night light densities, and the density of mining operations. If we squint at the predicted means in Missouri, we can indeed see that the nonstationary predictions are smoother.

The predicted standard deviations are very different following whether the model is stationary or nonstationary (figure 4.7). In the stationary model, the only thing that imports is the spatial density of the observations (figure 4.4). In the nonstationary model, regions with high spatial coherence such as the Midwest will have lower standard deviation, and other regions such as the West Coast will have high standard deviation even if the measurements are dense there.

Table 4.2: Summary of the *A Posteriori* samples of β_α, β_{σ^2}, and β_{τ^2} in the model $(\alpha + \sigma^2 + \tau^2)$

	β_α			β_{σ^2}			β_{τ^2}		
	mean	q0.025	q0.975	mean	q0.025	q0.975	mean	q0.025	q0.975
(Intercept)	-4.838	-4.923	-4.752	-1.568	-1.626	-1.503	-1.878	-1.894	-1.860
air pollution.	0.104	0.031	0.187	-0.071	-0.122	-0.020	-0.023	-0.046	-0.001
mining dens.	-0.017	-0.064	0.029	0.102	0.072	0.134	0.058	0.040	0.075
earthquake dens.	-0.011	-0.067	0.043	-0.056	-0.098	-0.012	0.002	-0.016	0.019
toxic release	-0.104	-0.186	-0.016	0.094	0.041	0.146	-0.036	-0.063	-0.004
green biomass	-0.406	-0.468	-0.331	0.016	-0.025	0.056	0.184	0.166	0.202
elevation	-0.172	-0.254	-0.094	-0.139	-0.194	-0.080	-0.507	-0.537	-0.475
night lights dens.	-0.269	-0.338	-0.206	0.115	0.067	0.163	0.171	0.140	0.203
road dens.	-0.030	-0.064	0.005	0.037	0.008	0.064	0.040	0.018	0.061
wetness idx.	0.515	0.422	0.602	-0.306	-0.367	-0.244	-0.300	-0.332	-0.267
visible sky idx.	-0.185	-0.233	-0.139	-0.007	-0.045	0.033	-0.051	-0.077	-0.025
wind effect idx.	-0.020	-0.072	0.027	0.142	0.107	0.178	0.075	0.053	0.096

4.5 Summary and open problems

This paper undertook to generalize the NNGP model to nonstationary covariance structures. We delivered a solution that takes into account the problematic aspects of computational cost, model selection, and interpretation of the parameters. Along the way, we developed various tools that could be useful in other contexts:

1. We found a flexible and interpretable parametrization for local anisotropy, embedding the nonstationary models in a coherent family *à la* Russian doll. Thanks to the logarithmic transform, the user can easily interpret the parameters. This family of models seems quite resilient with respect to overfitting, and could be useful in models that do not use NNGP.

2. The derivatives of nonstationary NNGP density can be used elsewhere than in HMC, in MAP or maximum likelihood approaches for example.

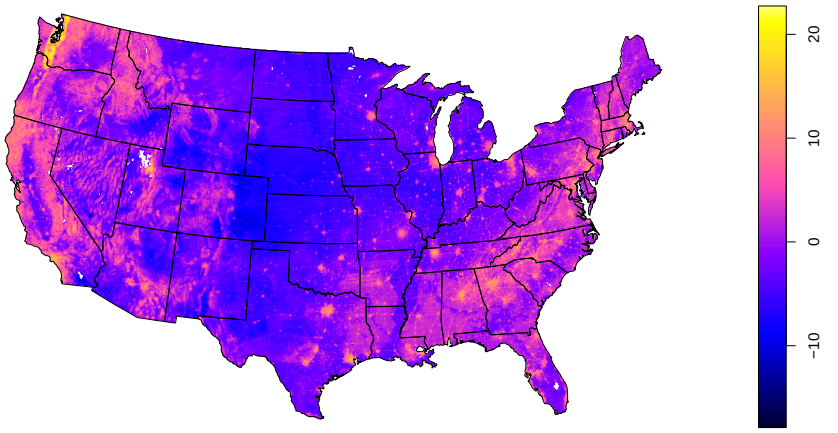


Figure 4.5: Visualization of the spatial incoherence of the lead measurements

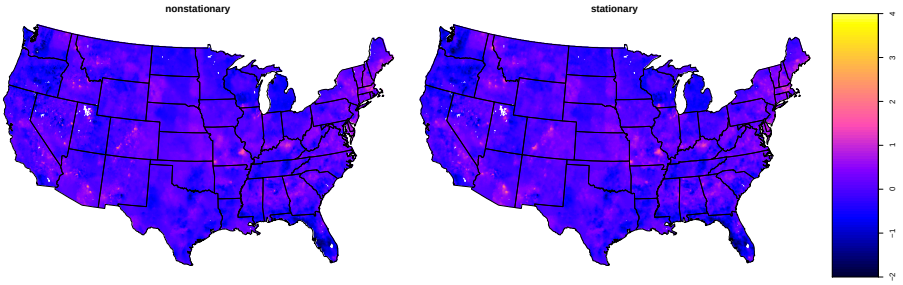


Figure 4.6: Predicted latent mean of the lead concentration

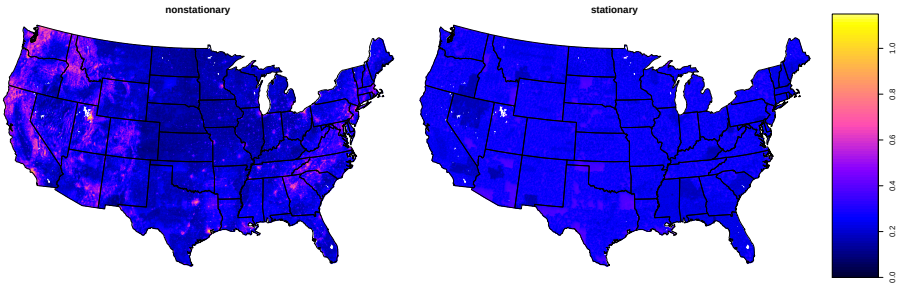


Figure 4.7: Predicted latent standard deviation of the lead concentration

3. We proved that nested ASIS (Yu and Meng, 2011) can be a viable strategy for multi-layered hierarchical models with large data augmentations.

A problem that we leave unsolved is the definition of a NNGP prior with anisotropic covariance parameters on the whole sphere. Working locally and using the projection of the range ellipses on tangent planes might allow to define both the nonstationary NNGP and matrix log NNGP prior. The matter of defining fixed effects for the matrix-valued in the absence of a common basis remains open.

A possible extension is an implementation of the model in more than two dimensions. In particular, elliptic covariances in 3 dimensions might prove useful to quantify drifts, for example rain moving across a territory. The matter to keep in mind is that ellipses in higher dimensions incur more differentiation, since the matrix logarithm of the range parameters will have 6 coordinates instead of 3. A computational scale up, discussed below, may be necessary.

An exciting but complex subject is whether we can extend our nonstationary spatial model to multivariate data. For example, the lead data set of Hengl (2009) features other heavy metals than lead, and those variables may share common sources and diffusion mechanisms.

A first axis would be to tackle "multivariate nonstationarity", that is a multivariate spatial model where spatial covariance is nonstationary. *A priori*, nothing forbids us to use Paciorek (2003)'s nonstationary covariance as a nonstationary cross-covariance, with

$$K_{multi}(s_i, s_j, v(\cdot), w(\cdot)) = K(s_i, s_j, \theta_v(s_i), \theta_w(s_j)) \times C(v(\cdot), w(\cdot)),$$

$K(\cdot)$ being a nonstationary covariance function taking a couple of spatial locations (s_i, s_j) and a couple of covariance parameters $(\theta_v(s_i), \theta_w(s_j))$ observed at those locations as arguments, and $C(v(\cdot), w(\cdot))$ being a measure of the association between $v(\cdot)$ and $w(\cdot)$. $C(v(\cdot), w(\cdot))$ is not exactly a within-site correlation because if the ranges of $v(\cdot)$ and $w(\cdot)$ differ in the site s, then

$$K(s, s, \theta_v(s), \theta_w(s)) \leq 1.$$

If the association matrix obtained using $C(\cdot)$ is positive-definite, then thanks to Schur's theorem $K(\cdot)$ is positive-definite as well. The covariance parameters θ_w and θ_v are allowed to vary in the spatial domain, inducing a nonstationary cross-covariance whose marginals correspond to the univariate nonstationary case.

Another point would be "nonstationary multivariateness", where the relationship between the variables changes in space and/or depending on some regressors. Since $C(\cdot)$ has to be a positive-definite matrix, this point seems right up the alley for our matrix log-GP prior, expanding the model while keeping the same framework. What remains to be done is to find a combination of $C(s_i, v)$ and $C(s_j, w)$ into $C(s_i, s_j, v, w)$ that preserves the positive-definiteness.

The problem is that doing multivariate NNGP is not so easy, even in the stationary case. Defining a relevant DAG for multivariate data is not an elucidated point. Using m parents of each variable for good measure will quickly become unaffordable. As far as we know, heuristics like mixing m_1 parents of the same variables and m_2 parents of the rest have not been tested. Peruzzi et al. (2020) use tessellated Gaussian processes, a modified NNGP that centralizes the variables in auxiliary spatial locations. The auxiliary spatial locations having a simple layout, typically gridded, many elements can be reused and the computation ends up being much more economical. However, the fact that some elements can be re-used precisely comes from the stationarity of the function. Moreover, Peruzzi et al. (2020) induces conditional independence between the observed locations. In prediction, this behavior leads to a degradation of model performances (Katzfuss et al., 2020). In one way or another, multivariate nonstationary NNGP would mean more parents, so more NNGP density differentiation. A computational scale-up is required.

This, and the perspective to have coordinate spaces of dimension 3 or more, lead us to the third point. Given the fact that we have found the gradients of the model density and that the *a posteriori* distributions of the model are well-behaved thanks to the logarithmic parametrization (actually, in the lead data set, the high-level parameters passed the test of Henze and Wagner (1997) for multivariate normality), the option of *Maximum A Posteriori* (MAP) estimation should be considered seriously. While we will lose some nuances of the *a posteriori* distribution, there is a trade-off between MCMC and NNGP. The computational effort in flops and RAM that is not spent on doing MCMC could be re-invested in doing NNGP with a richer Vecchia's approximation.

Bibliography

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008, September). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

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

Cressie, N. and G. Johannesson (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 209–226.

Cressie, N. and C. K. Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

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

Filippone, M., M. Zhong, and M. Girolami (2013). A comparative evaluation of stochastic-based inference methods for gaussian process models. *Machine Learning 93*(1), 93–114.

Finley, A. O., A. Datta, B. D. Cook, D. C. Morton, H. E. Andersen, and S. Banerjee (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics 28*(2), 401–414.

Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika 89*(1), 197–210.

Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue (2015a). Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics 14*, 505–531.

Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue (2015b). Interpretable priors for hyperparameters for gaussian random fields. *arXiv preprint arXiv:1503.00256*.

Gelfand, A. E., P. Diggle, P. Guttorp, and M. Fuentes (2010). *Handbook of Spatial Statistics (Chapman Hall CRC Handbooks of Modern Statistical Methods)*. Chapman Hall CRC Handbooks of Modern Statistical Methods. Taylor and Francis.

Gelfand, A. E., M. Fuentes, J. A. Hoeting, and R. L. Smith (2019). *Handbook of environmental and ecological statistics*. CRC Press.

Gonzalez, J., Y. Low, A. Gretton, and C. Guestrin (2011). Parallel gibbs sampling: From colored fields to thin junction trees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 324–332. JMLR Workshop and Conference Proceedings.

Guinness (2018). Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics 60*(4), 415–429.

Heaton, M. J., A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics 24*(3), 398–425.

Heinonen, M., H. Mannerström, J. Rousu, S. Kaski, and H. Lähdesmäki (2016). Non-stationary gaussian process regression with hamiltonian monte carlo. In *Artificial Intelligence and Statistics*, pp. 732–740.

Hengl, T. (2009). *A practical guide to geostatistical mapping*. Hengl Amsterdam.

Henze, N. and T. Wagner (1997). A new approach to the bhep tests for multivariate normality. *Journal of Multivariate Analysis 62*(1), 1–23.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics 5*(2), 173–190.

Ingebrigtsen, R., F. Lindgren, I. Steinsland, and S. Martino (2015). Estimation of a non-stationary model for annual precipitation in southern norway using replicates of the spatial field. *Spatial Statistics 14*, 338–364.

Katzfuss, M. and J. Guinness (2017, Aug). A general framework for Vecchia approximations of Gaussian processes. *arXiv e-prints*, arXiv:1708.06302.

Katzfuss, M., J. Guinness, W. Gong, and D. Zilber (2020). Vecchia approximations of gaussian-process predictions. *Journal of Agricultural, Biological and Environmental Statistics 25*(3), 383–414.

Kleiber, W. and D. Nychka (2012). Nonstationary modeling for multivariate spatial processes. *Journal of Multivariate Analysis 112*, 76–91.

Knorr-Held, L. and H. Rue (2002). On block updating in markov random field models for disease mapping. *Scandinavian Journal of Statistics 29*(4), 597–614.

Lauritzen, S. L. (1996). *Graphical models*. Oxford Statistical Science Series. OUP.

Neal, R. (2011, May). MCMC Using Hamiltonian Dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, Volume 20116022. Chapman and Hall/CRC.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo 2*(11), 2.

Paciorek, C. J. (2003). *Nonstationary Gaussian processes for regression and spatial modelling*. Ph. D. thesis, Citeseer.

Peruzzi, M., S. Banerjee, and A. O. Finley (2020). Highly scalable bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *Journal of the American Statistical Association*, 1–14.

Risser, M. D. (2016). Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches. *arXiv preprint arXiv:1610.02447*.

Risser, M. D. and C. A. Calder (2015). Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics 26*(4), 284–297.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van der Linde (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical report, Research Report, 98-009.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology) 64*(4), 583–639.

Stevens, S. S. et al. (1946). On the theory of scales of measurement.

Yang, H.-C. and J. R. Bradley (2021). Bayesian inference for big spatial data using non-stationary spectral simulation. *Spatial Statistics 43*, 100507.

Yu, Y. and X.-L. Meng (2011). 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.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association 99*(465), 250–261.

APPENDIX

4.6 Appendix: demonstrations

4.6.1 Recursive conditional form of nonstationary NNGP

Start with:

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

Then re-write

$$f(w(s_i)|w(pa(s_i)),\theta(\mathcal{S})) = f(w(s_i \cup pa(s_i))|\theta(\mathcal{S}))/f(w(pa(s_i))|\theta(\mathcal{S})).$$

The joint distributions

$$f(w(s_i \cup pa(s_i))|\theta(\mathcal{S})) \quad \text{and} \quad f(w(pa(s_i))|\theta(\mathcal{S}))$$

are fully parametrized respectively by

$$\Sigma(s_i \cup pa(s_i),\theta(\mathcal{S})) \quad \text{and} \quad \Sigma(pa(s_i),\theta(\mathcal{S})).$$

Since the covariance functions given by equations (4.6) or (4.7) allow to compute $\Sigma_{i,j}$ using only $\theta(s_i)$ and $\theta(s_j)$ instead of $\theta(\mathcal{S})$,

$$f(w(s_i \cup pa(s_i))|\theta(\mathcal{S})) = f(w(s_i \cup pa(s_i))|\theta(s_i \cup pa(s_i)))$$

and

$$f(w(pa(s_i))|\theta(\mathcal{S})) = f(w(pa(s_i))|\theta(pa(s_i))) = f(w(pa(s_i))|\theta(s_i \cup pa(s_i))).$$

Conclude with

$$f(w(s_i)|w(pa(s_i)),\theta(\mathcal{S})) = f(w(s_i \cup pa(s_i))|\theta(s_i \cup pa(s_i)))/f(w(pa(s_i))|\theta(s_i \cup pa(s_i))).$$

4.6.2 Marginal variance of nonstationary NNGP

Let $\tilde{R}$ be the NNGP factor using the nonstationary covariance $K(\cdot)$ and let $\tilde{R}_0$ be the NNGP factor using the nonstationary correlation $K_0(\cdot)$ from (4.5) and either (4.6) or (4.7) respectively. From the construction of $\tilde{R}$, we introduce

$$\bar{\sigma}_i = \left(\Sigma(s_i,s_i) - \Sigma(s_i,pa(s_i))\Sigma(pa(s_i),pa(s_i))^{-1}\Sigma(pa(s_i),s_i)\right)^{1/2}$$

the standard deviation of $w(s_i)$ conditionally on $w(pa(s_i))$, and its counterpart $(\bar{\sigma_0})_i$ obtained with Σ_0 instead of Σ. The i^{th} row will have:

$$\begin{cases} -\Sigma(s_i, pa(s_i))\Sigma(pa(s_i), pa(s_i))^{-1}/\bar{\sigma}_i \text{ at the column indices that correspond to } pa(s_i) \\ 1/\sigma_i \text{ at column index } i \\ 0 \text{ elsewhere} \end{cases}$$

Introducing Σ_0 and $\sigma(\mathcal{S})$, we find that

$$\begin{aligned} \bar{\sigma}_i &= \left(\Sigma(s_i, s_i) - \Sigma(s_i, pa(s_i))\Sigma(pa(s_i), pa(s_i))^{-1}\Sigma(pa(s_i), s_i)\right)^{1/2} \\ &= (\sigma(s_i)^2\Sigma_0(s_i, s_i) - \sigma(s_i)\Sigma_0(s_i, pa(s_i))\ diag(\sigma(pa(s_i))) \\ &\quad diag(\sigma(pa(s_i)))^{-1}\Sigma_0(pa(s_i), pa(s_i))^{-1}\ diag(\sigma(pa(s_i)))^{-1} \\ &\quad diag(\sigma(pa(s_i)))\Sigma_0(pa(s_i), s_i)\sigma(s_i))^{1/2} \\ &= \sigma(s_i)(\bar{\sigma}_0)_i \end{aligned}$$

The coefficients of row i become:

$$\begin{cases} \left(-\Sigma_0(s_i, pa(s_i))\Sigma_0(pa(s_i), pa(s_i))^{-1}/(\bar{\sigma_0})_i\right) diag(\sigma(pa(s_i)))^{-1} \text{at the indices of } pa(s_i) \\ 1/(\sigma(s_i)(\bar{\sigma_0})_i) \text{ at index } i \\ 0 \text{ elsewhere} \end{cases}$$

It follows that

$$\tilde{R}_{i,j} = \tilde{R}_{0_{i,j}}/\sigma(s_j),$$

which proves the result.

4.7 Appendix: details about KL divergence

4.7.1 Scalar range case

Synthetic data sets with 10000 observations were simulated on a domain with size 5×5. The spatially variable log-range had mean $log(0.1)$. Three factors were tested:

- the intensity of nonstationarity, by letting the log range's variance take different values (0.1, 0.3, and 0.5).
- the ordering (coordinate, max-min, random, middleout).
- the number of parents (5, 10, 20).

Using a linear model with interactions shows that the intensity of nonstationarity has almost no role. The most important factor is the number of parents. Eventually, the NNGP approximation can be improved using the max-min and random order, joining Guinness (2018)'s conclusions for stationary models in 2 dimensions. See table 4.3 for more details about the effects of the factors.

4.7.2 Elliptic range case

Synthetic data sets with 10000 observations were simulated on a domain with size 5×5. The spatially variable log-matrix range had mean $log(.1) \times I_2/\sqrt{2}$. Three factors were tested:

- the intensity of nonstationarity, by letting the variance of the coordinates of the log-range matrix take different values: ($0.1 \times I_3$, $0.3 \times I_3$, and $0.5 \times I_3$).
- the ordering (coordinate, max-min, random, middleout).
- the number of parents (5, 10, 20).

The outcome is treated with a linear model, whose summary is presented in table 4.4. Contrary to the first experiment, the intensity of the nonstationarity does play a role.

Table 4.3: Summary of linear regression of the KL divergence, in the scalar range case.*

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	186.5156	0.4517	412.96	0.0000
nonstat.intensity 0.3	1.4745	0.2957	4.99	0.0000
nonstat.intensity 0.5	3.2509	0.2957	10.99	0.0000
ordering max min	-47.6966	0.5914	-80.66	0.0000
ordering middle out	-4.7491	0.5914	-8.03	0.0000
ordering random	-47.2462	0.5914	-79.89	0.0000
10 nearest neighbors	-135.9274	0.5914	-229.86	0.0000
20 nearest neighbors	-176.4046	0.5914	-298.31	0.0000
max min: 10	25.9771	0.8363	31.06	0.0000
middle out: 10	1.9064	0.8363	2.28	0.0228
random: 10	25.7530	0.8363	30.79	0.0000
max min: 20	41.4488	0.8363	49.56	0.0000
middle out: 20	3.3930	0.8363	4.06	0.0001
random: 20	40.9979	0.8363	49.02	0.0000

*The reference case has coordinate ordering, 5 nearest neighbors, and a log-range variance of 0.1

Table 4.4: Summary of linear regression of the KL divergence, in the elliptic range case*.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	243.6011	1.8448	132.05	0.0000
nonstat.intensity 0.3	23.8570	1.2077	19.75	0.0000
nonstat.intensity 0.5	50.5490	1.2077	41.86	0.0000
ordering max min	-50.5886	2.4154	-20.94	0.0000
ordering middle out	-0.0311	2.4154	-0.01	0.9897
ordering random	-50.6093	2.4154	-20.95	0.0000
10 nearest neighbors	-176.3113	2.4154	-72.99	0.0000
20 nearest neighbors	-238.4647	2.4154	-98.73	0.0000
max min: 10	19.4831	3.4159	5.70	0.0000
middle out: 10	-1.6965	3.4159	-0.50	0.6195
random: 10	19.5230	3.4159	5.72	0.0000
max min: 20	38.3413	3.4159	11.22	0.0000
middle out: 20	-1.6284	3.4159	-0.48	0.6337
random: 20	38.3553	3.4159	11.23	0.0000

The reference case has coordinate ordering, 5 nearest neighbors, and a log-range variance of 0.1

4.8 Appendix: chromatic samplers for parameter fields with log-NNGP priors

Coube and Liquet (2020) show that chromatic samplers are an operational solution to sample the latent field $w(\mathcal{S})$. We tried to adapt this method to update $w_\theta(\mathcal{S})$, but it was less efficient than HMC. We report our result here in the eventuality of someone finding a better use for them.

Conditional independence in the full conditional of w_α and w_{σ^2} when sufficient augmentation is used. In order to use chromatic sampler, the Markov graph of the full conditional distribution of $w_\theta(\mathcal{S})$ must be found. Assume a log-NNGP or matrix log-NNGP prior on the covariance parameter field $\theta(\mathcal{S})$ and the NNGP prior for the latent field $w(\mathcal{S})$. Let $\mathcal{G}_\theta$ and $\mathcal{G}_w$ be the moralized graphs induced by their respective NNGP DAGs. Both graphs have the same vertices, identified with the spatial locations $\mathcal{S}$. Let $\mathcal{E}_\theta, \mathcal{E}_w \subset \mathcal{S} \otimes \mathcal{S}$ be their respective edges. Then, the full conditional distribution of $w_\theta(\mathcal{S})$ has the global Markov property on the graph

$$\mathcal{G}_{w+\theta} = (\mathcal{S}, \mathcal{E}_w \cup \mathcal{E}_\theta)$$

To prove this point, remark that (4.13) and (4.9) imply that the distribution of $w_\theta(\mathcal{S})$ conditionally on the rest of the parameters is proportional to

$$\tilde{f}_\theta(w_\theta(\mathcal{S})|\zeta_\theta)\tilde{f}(w(\mathcal{S})|\theta(\mathcal{S})).$$

The left-hand part is the parameters' log-NNGP prior and the right-hand part is the latent field's NNGP prior. We call here $pa(s_i)$ the parents of s_i in the DAG used to define the latent field density $\tilde{f}(\cdot)$. Using (4.9), we can write

$$\begin{aligned}\tilde{f}(w(\mathcal{S})|\theta(\mathcal{S})) &= f(w(s_1)|\theta(s_1))\Pi_{i=2}^n f(w(s_i)|w(pa(s_i)),\theta(s_i \cup pa(s_i))) \\ &= \psi_1(\theta_1)\Pi_{i=2}^n \psi_i(\theta(s_i \cup pa(i))), \quad \psi_1,\ldots,\psi_n \text{ being nonnegative.}\end{aligned}$$

By definition, $(s_i \cup pa(s_i))$ are complete sets or cliques on $\mathcal{G}_w$, so the NNGP density evaluated in $\theta(\mathcal{S})$ factorizes on $\mathcal{G}_w$. On the other hand, using recursive conditional factorization, the log-NNGP factorizes on $\mathcal{G}_\theta$.
Since the NNGP density and the log-NNGP prior respectively factorize on $\mathcal{G}_w$ and $\mathcal{G}_\theta$, they *a fortiori* factorize on $(\mathcal{S}, \mathcal{E}_w \cup \mathcal{E}_\theta)$. Their product then factorizes on $(\mathcal{S}, \mathcal{E}_w \cup \mathcal{E}_\theta)$. Therefore, it has the Global Markov Property on $(\mathcal{S}, \mathcal{E}_w \cup \mathcal{E}_p)$ (Lauritzen, 1996).

Distribution of w_{σ^2} with ancillary augmentation. Using (4.10) and the independence of the observations conditionally on the linear effects and the latent field, the likelihood can be factorized as

$$l\left(z(\mathcal{S})|\left(\tilde{R}_0\ diag(\sigma(\mathcal{S}))^{-1}\right)^{-1} w^*,\beta,\ldots\right) = \Pi_{i=1}^n l\left(z(s_i)|\sigma(s_i)\left(\tilde{R}_0^{-1}w^*\right)_i,\beta,\ldots\right). \tag{4.26}$$

When this likelihood is evaluated with respect to $\sigma(\mathcal{S})$, it can be factorized on the edge-less graph. So the full conditional distribution of $w_\sigma(\mathcal{S})$ can be factorized on $\mathcal{E}_\sigma = \mathcal{E}_\sigma \cup \emptyset$. This is consistent with the fact that $w_\sigma(s_i)$ will move only $w(s_i)$, in $]0,+\infty[$ if $w(s_i) > 0$, and in $]-\infty,0[$ if $w(s_i) < 0$.

4.9 Appendix: details about interweaving

4.9.1 Centered update of the regression coefficients for the matrix log NNGP

Start from (4.17) and (4.18), centering the latent field on the fixed effects gives

$$w_{A_{centered}} = \left(w_1(\mathcal{S}),\ldots,w_{d(d+1)/2}(\mathcal{S})\right) + \mu.$$

The problem is that μ is obtained stacking various fixed effects vertically, so that the approach of Coube and Liquet (2020) cannot be done directly. Write the Cholesky factorization of the multivariate prior matrix of the log-NNGP prior:

$$(S \otimes \Sigma_0) = (S^{1/2} \otimes \Sigma_0^{1/2})(S^{1/2} \otimes \Sigma_0^{1/2})^T.$$

Note that $\Sigma_0^{-1/2}$ is the NNGP Cholesky factor that parametrizes the spatial correlation in the matrix log-NNGP prior. Introduce the whitened matrix log-GP field

$$w_A^* = \left(w_1^*(\mathcal{S}),\ldots,w_{d(d+1)/2}^*(\mathcal{S})\right) = (S^{1/2}\otimes\Sigma_0^{1/2})^{-1/2}\left(w_1(\mathcal{S}),\ldots,w_{d(d+1)/2}(\mathcal{S})\right)$$

Using the "Vec trick", we can re-write:

$$w_{A_{centered}} = Vec(\mu) + \Sigma_0^{1/2}(w_A^*)(S^{1/2})^T.$$

Recall that μ is obtained by by stacking vertically $X_A\beta_{A_1}, \ldots, X_A\beta_{A_{d(d+1)/2}}$, so that

$$Vec(\mu) = X_A(\beta_{A_1}|\ldots|\beta_{A_{d(d+1)/2}}),$$

where $(\beta_{A_1}|\ldots|\beta_{A_{d(d+1)/2}})$ is obtained by stacking horizontally the vectors of regression coefficients. Multiply on the right by $(S^{1/2})^{-T}$ in order to do an inter-component whitening (the spatial correlation remains):

$$w_{A_{centered}}(S^{1/2})^{-T} = X_A(\beta_{A_1}|\ldots|\beta_{A_{d(d+1)/2}})(S^{1/2})^{-T} + \Sigma_0^{1/2}(w_A^*).$$

Since the $d(d+1)/2$ columns of $\Sigma_0^{1/2}(w_A^*)$ are independent, the method of Coube and Liquet (2020) can be used to update each component of $(\beta_{A_1}|\ldots|\beta_{A_{d(d+1)/2}})(S^{1/2})^{-T}$, which are then transformed back to $(\beta_{A_1}|\ldots|\beta_{A_{d(d+1)/2}})$.

4.10 Appendix: gradients for HMC updates of the covariance parameters

4.10.1 General form of the gradient with respect to w_θ^*

Start from

$$H = \ -log(\tilde{f}_\theta(w_\theta(\mathcal{S})|\zeta_\theta)) - g_\theta(w_\theta(\mathcal{S})) \quad \propto \quad w_\theta^T\tilde{R}_\theta^T\tilde{R}_\theta w_\theta/2 \ - g_\theta(w_\theta(\mathcal{S})),$$

ζ_θ being the covariance parameters of the log-NNGP prior. Find the gradient of H with respect to w_θ:

$$\nabla_{w_\theta} H = \tilde{R}_\theta^T\tilde{R}_\theta w_\theta - \nabla_{w_\theta} g_\theta(w_\theta(\mathcal{S})),$$

$\tilde{R}_\theta$ being the NNGP factor informed by the covariance parameters ζ_θ. Then, apply the Jacobian (J) chain rule $\nabla\psi \circ \phi(x) = (J^T\phi)(x) \cdot (\nabla\psi)(\phi(x))$ with $\psi = H$ et $\phi(w_\theta^*) = \tilde{R}_\theta^{-1}w_\theta^*$. With $J^T\left(\tilde{R}_\theta^{-1}w_\theta^*(\mathcal{S})\right) = (\tilde{R}_\theta^{-1})^T$, we obtain

$$\nabla_{w_\theta^*} H = w_\theta^* - (\tilde{R}_\theta^{-1})^T\nabla_{w_\theta} g_\theta(\tilde{R}_\theta^{-1}w_\theta^*(\mathcal{S}))$$

4.10.2 Gradient of the negated log-density with respect to w_{σ^2}

Sufficient augmentation. When sufficient augmentation is used, the marginal variance intervenes in the NNGP density of the latent field. Using (4.12) and (4.11),

$$\tilde{f}(w(\mathcal{S})|\alpha, \sigma^2(\mathcal{S})) = exp\left(-\sigma^{-1}(\mathcal{S})^T \ diag(w)\tilde{R}_0^T\tilde{R}_0 \ diag(w)\sigma^{-1}(\mathcal{S})/2\right)\Pi_{i=1}^n(\tilde{R}_0)_{i,i}/\sigma(s_i).$$

Passing to the negated log-density

$$-g_{\sigma^2} = -log\left(\tilde{f}(w(\mathcal{S})|\alpha, \sigma^2(\mathcal{S}))\right) \quad = \quad cst + \Sigma_{i=1}^n log(\sigma(s_i)) \;\; + \\ \sigma^{-1}(\mathcal{S})^T \; diag(w)\tilde{R}_0^T\tilde{R}_0 \; diag(w)\sigma^{-1}(\mathcal{S})/2$$

One the one hand,

$$\nabla_{w_{\sigma^2}} \Sigma_{i=1}^n log(\sigma(s_i)) = \nabla_{w_{\sigma^2}} \Sigma_{i=1}^n log((\sigma^2(s_i))^{1/2}) = \nabla_{w_{\sigma^2}} \Sigma_{i=1}^n log(\sigma^2(s_i))/2 = (1/2, \dots, 1/2)$$

On the other hand, using $\sigma^{-1}(s) = (\sigma^2(s))^{-1/2} = exp(-(w_{\sigma^2}(s) + X_{\sigma^2}(s)\beta_{\sigma^2}^T)/2)$,
we can write the Jacobian of σ^{-1} with respect to w_{σ^2}:

$$J_{w_{\sigma^2}}\sigma^{-1}(\mathcal{S}) = J_{w_{\sigma^2}} exp(-(w_{\sigma^2}(\mathcal{S}) + X_{\sigma^2}(\mathcal{S})\beta_{\sigma^2}^T)/2) = -diag(\sigma^{-1}(\mathcal{S})/2).$$

We also find the following gradient:

$$\nabla_{\sigma^{-1}}\sigma^{-1}(\mathcal{S})^T \; diag(w)\tilde{R}_0^T\tilde{R}_0 \; diag(w)\sigma^{-1}(\mathcal{S})/2 = diag(w)\tilde{R}_0^T\tilde{R}_0 \; diag(w)\sigma^{-1}(\mathcal{S}).$$

With the Jacobian chain rule, we combine the two previous formulas to find

$$-\nabla_{w_{\sigma^2}}\sigma^{-1}(\mathcal{S})^T \; diag(w)\tilde{R}_0^T\tilde{R}_0 \; diag(w)\sigma^{-1}(\mathcal{S})/2 = \sigma^{-1}(\mathcal{S}) \circ \left(diag(w)\tilde{R}_0^T\tilde{R}_0 \; diag(w) \quad \sigma^{-1}(\mathcal{S})\right)/2,$$

with $\circ$ the Hadamard product.
Combining the two terms, we have

$$-\nabla_{w_{\sigma^2}} g_{\sigma^2} = 1/2 - \sigma^{-1}(\mathcal{S}) \;\circ\; \left(diag(w)\tilde{R}_0^T\tilde{R}_0 \; diag(w) \quad \sigma^{-1}(\mathcal{S})\right)/2. \quad (4.27)$$

Ancillary augmentation. When ancillary augmentation is used, the marginal variance has an impact on the observed field likelihood with respect to the latent field. Like in (4.26), the likelihood can be written as

$$l(z(\mathcal{S})|(\tilde{R}(\sigma^2(\mathcal{S}), \alpha)^{-1}w^*), \beta, \dots) = \Pi_{i=1}^n l\left(z(s_i)|\sigma(s_i)\left(\tilde{R}_0^{-1}w^*\right)_i, \beta, \dots\right)$$

Passing to the negated log-density and introducing $\sigma^2(s) = exp(w_{\sigma^2}(s) + X_{\sigma^2}(s)\beta_{\sigma^2}^T)$,
the negated log-likelihood can be written as

$$-g_{\sigma^2} = -\Sigma_{i=1}^n log\left(l\left(z(s_i)|exp\left(\left(w_{\sigma^2}(s_i) + X_{\sigma^2}(s_i)\beta_{\sigma^2}^T\right)/2\right)\left(\tilde{R}_0^{-1}w^*\right)_i, \beta, \dots\right)\right)$$

Using the chain rule on

$$w(s_i) = exp\left(\left(w_{\sigma^2}(s_i) + X_{\sigma^2}(s_i)\beta_{\sigma^2}^T\right)/2\right)\left(\tilde{R}_0^{-1}w^*\right)_i,$$

the gradient with respect to $w_{\sigma^2}(\mathcal{S})$ will write

$$-\nabla_{w_{\sigma^2}} g_{\sigma^2} = \nabla_w \; l(z(\mathcal{S})|\tilde{R}^{-1}w^*, \beta, \dots) \; \circ (\tilde{R}^{-1}w^*/2), \quad (4.28)$$

with $\circ$ being the Hadamard product.

4.10.3 General derivative of $\tilde{R}$ with respect to nonstationary range parameters

The aim is to find $\partial\tilde{R}/\partial w_\alpha(s_j)$ with $j \in 1,\dots,n$. Let's focus on the i^{th} row of $\tilde{R}$, noted $\tilde{R}_{i,.}$. The index of the row i can be different from j. To find the derivative of $\tilde{R}_{i,.}$ with respect to $w_\alpha(s_j)$, we need to use the covariance matrix between s_i and its parents $pa(s_i)$. Let's note Σ^i the covariance matrix corresponding to $(pa(s_i), s_i)$, and let's block it as $\Sigma^i = \left[\begin{array}{c|c} \Sigma^i_{11} & \Sigma^i_{12} \\ \hline \Sigma^i_{21} & \Sigma^i_{22} \end{array}\right]$ Σ^i_{11} being a $m \times m$ covariance matrix corresponding to $pa(s_i)$, and Σ^i_{22} being a 1×1 matrix corresponding to s_i. From its construction, $\tilde{R}_{i,.}$ has non-null coefficients only for the column entries that correspond to s_i and its parents $pa(s_i)$. Therefore there is no need to compute the gradient but for those coefficients. The diagonal element $\tilde{R}_{i,i}$ has value $1/\bar{\sigma}_i$, $\bar{\sigma}_i$ being the standard deviation of $w(s_i)$ conditionally on $w(pa(s_i))$. The elements that correspond to $pa(s_i)$ have value $-\Sigma^i_{21}(\Sigma^i_{11})^{-1}/\bar{\sigma}_i$.
Let's start by the diagonal coefficient $\tilde{R}_{i,i}$:

$$
\begin{aligned}
\partial(\tilde{R}_{ii})/\partial w_\alpha(s_j) &= \partial((\bar{\sigma}_i^2)^{-1/2})/\partial w_\alpha(s_j) \\
&\quad \text{(chain rule)} \\
&= -(\bar{\sigma}_i^{-3}/2) \times \partial(\bar{\sigma}_i^2)/\partial w_\alpha(s_j) \\
&\quad \text{(using conditional variance formula)} \\
&= -(\bar{\sigma}_i^{-3}/2) \times \partial(\Sigma^i_{22} - \Sigma^i_{21}(\Sigma^i_{11})^{-1}\Sigma^i_{12})/\partial w_\alpha(s_j) \\
&\quad \text{(product rule)} \\
&= -(\bar{\sigma}_i^{-3}/2) \times \partial(\Sigma^i_{22})/\partial w_\alpha(s_j) \\
&\quad +(\bar{\sigma}_i^{-3}) \times \partial(\Sigma^i_{21})/\partial w_\alpha(s_j)(\Sigma^i_{11})^{-1}\Sigma^i_{12} \\
&\quad +(\bar{\sigma}_i^{-3}/2) \times \Sigma^i_{21}\partial\left((\Sigma^i_{11})^{-1}\right)/\partial w_\alpha(s_j)\Sigma^i_{12} \\
&\quad \text{(derivative of inverse)} \\
&= -(\bar{\sigma}_i^{-3}/2) \times \partial(\Sigma^i_{22})/\partial w_\alpha(s_j) \qquad (a) \\
&\quad +(\bar{\sigma}_i^{-3}) \times \partial(\Sigma^i_{21})/\partial w_\alpha(s_j)(\Sigma^i_{11})^{-1}\Sigma^i_{12} \qquad (b) \\
&\quad -(\bar{\sigma}_i^{-3}/2) \times \Sigma^i_{21}(\Sigma^i_{11})^{-1}\left(\partial(\Sigma^i_{11})/\partial w_\alpha(s_j)\right)(\Sigma^i_{11})^{-1}\Sigma^i_{12} \qquad (c)
\end{aligned}
$$

Let's now differentiate the coefficients that correspond to $pa(s_i)$, located on row $\tilde{R}_{i,}$ at the left of the diagonal:

$$
\begin{aligned}
\partial(-\Sigma^i_{21}(\Sigma^i_{11})^{-1}/\bar{\sigma}_i)/\partial w_\alpha(s_j) &= -\partial(\Sigma^i_{21}(\Sigma^i_{11})^{-1} \times \tilde{R}_{ii})/\partial w_\alpha(s_j) \\
&\quad \text{(product rule)} \\
&= -\left(\partial\Sigma^i_{21}/\partial w_\alpha(s_j)\right)(\Sigma^i_{11})^{-1} \times \tilde{R}_{ii} \\
&\quad -\Sigma^i_{21}\left(\partial\left((\Sigma^i_{11})^{-1}\right)/\partial w_\alpha(s_j)\right) \times \tilde{R}_{ii} \\
&\quad -\Sigma^i_{21}(\Sigma^i_{11})^{-1}\partial\tilde{R}_{ii}/\partial w_\alpha(s_j) \\
&\quad \text{(derivative of inverse)} \\
&= -\left(\partial\Sigma^i_{21}/\partial w_\alpha(s_j)\right)(\Sigma^i_{11})^{-1} \times \tilde{R}_{ii} \qquad (d) \\
&\quad +\Sigma^i_{21}(\Sigma^i_{11})^{-1}\left(\partial\Sigma^i_{11}/\partial w_\alpha(s_j)\right)(\Sigma^i_{11})^{-1} \times \tilde{R}_{ii} \qquad (e) \\
&\quad -\Sigma^i_{21}(\Sigma^i_{11})^{-1} \times \underbrace{\partial\tilde{R}_{ii}/\partial w_\alpha(s_j)}_{\text{already known}} \qquad (f)
\end{aligned}
$$

From those derivatives, it appears that the elements that are needed to get the derivative of $\tilde{R}_{i,\cdot}$ are Σ^i and $\partial\Sigma^i/\partial w_\alpha(s_j)$ (with $s_j \in s_i \cup pa(s_i)$). The former needs anyway to be computed in order to obtain $\tilde{R}$. The latter can be approximated using finite differences:

$$\partial\Sigma^i/\partial w_\alpha(s_j) \approx \left(\Sigma^i(w_\alpha(s_j) + dw_\alpha(s_j)) - \Sigma^i(w_\alpha(s_j))\right)/dw_\alpha(s_j).$$

4.10.4 Computational cost of the derivative of $\tilde{R}$ with respect to nonstationary range parameters

We can see that the differentiation of $\tilde{R}_{i,\cdot}$ is non-null only for $\alpha(pa(s_i) \cup s_i)$ because the entries of Σ^i are given by $K(s,t,\alpha(s),\alpha(t))$ with $s,t \in s_i \cup pa(s_i)$. Conversely, if $w_\alpha(s_j)$ moves, only the rows of $\tilde{R}$ that correspond to s_i and its children on the DAG move as well. This means that in order to compute the derivative of $\tilde{R}$ with respect to α_j, the row differentiation operation must actually be done $|ch(s_j)| + 1$ times and not n times. Knowing the fact that $\Sigma_{j=1}^n |ch(s_j)| = \Sigma_{j=1}^n |pa(s_j)| = m \times n$ (m being the number of nearest neighbors used in Vecchia's approximation), we can see that row differentiation must be done $(m+1) \times n$ times in order to get all the derivatives of $\tilde{R}$ with respect to $\alpha(s_1, \ldots, s_n)$.
Given the fact that one row has $m+1$ non-null terms and that $(m+1) \times n$ rows are differentiated, the cost in RAM to store the differentiation of $\tilde{R}$ will be $O(m+1)^2 n$, which remains acceptable for $m = 5, 10$.
On the other hand, the flop cost of differentiation itself may seem daunting. However, the fact that spatially-variable covariance parameters affect pairwise covariances considerably simplifies the problem. In the derivatives, there are only 3 terms that depend on (s_j), they are $\partial(\Sigma^i_{22})/\partial w_\alpha(s_j)$, $\partial(\Sigma^i_{12})/\partial w_\alpha(s_j)$, and $\partial(\Sigma^i_{11})/\partial w_\alpha(s_j)$. Let's separate the cases:

1. When $i \neq j$
   (a) $\partial(\Sigma^i_{12})/\partial w_\alpha(s_j)$ has only one non-null coefficient.

(b) $\partial(\Sigma^i_{11})/\partial w_\alpha(s_j)$ is a $m \times m$ matrix with cross structure (non-null coefficients only for the row and the column corresponding to s_j).

(c) $\partial(\Sigma^i_{22})/\partial w_\alpha(s_j)$ is a null 1×1 matrix.

2. When $i = j$

(a) $\partial(\Sigma^i_{12})/\partial w_\alpha(s_j)$ is a dense vector of length m.

(b) $\partial(\Sigma^i_{11})/\partial w_\alpha(s_j)$ is null.

(c) $\partial(\Sigma^i_{22})/\partial w_\alpha(s_j)$ is null because a change in $w_\alpha(s_i)$ does not affect the marginal variance of $w(s_i)$ (a change in $w_{\sigma^2(s_i)}$ does).

The costliest part of the formulas is to compute $(\Sigma^i_{11})^{-1}$. However, this part needs only to be computed one time since it is not affected by differentiation. Even better, $(\Sigma^i_{11})^{-1}$ and $\Sigma^i_{21}(\Sigma^i_{11})^{-1}$ can be used to compute $\tilde{R}$ and then recycled on the fly to compute the derivatives. The computational effort needed to get them can then be removed from the cost of the derivative and remain in the cost of $\tilde{R}$.
Applying all those remarks gives table 4.5.

Table 4.5: costs to compute $\partial \tilde{R}_{i,.}/\partial(w_\alpha(s_j))$

	(a)	(b)	(c)	(d)	(e)	(f)
i = j	O(1)	O(m)	0	$O(m^2)$	0	0
$s_i \in ch(s_j)$	0	O(1)	O(m)	O(m)	O(m)	0

Using table 4.5 and again $\Sigma_{j=1}^n |ch(s_j)| = \Sigma_{j=1}^n |pa(s_j)| = m \times n$, we can see that the matrix operations should have a total cost of $O(m^2 \times n)$.
The cost of the finite difference approximation to $\partial\Sigma^i/\partial w_\alpha(s_j)$ must be added to this. The cost of computing the finite differences in one coefficient of Σ^i depends on whether isotropic or anisotropic range parameters are used. In the case of isotropic range parameters, only a recomputation of the covariance function (4.7) with range $exp(log(\alpha(s) + dw))$ instead of $exp(log(\alpha(s)))$ will be needed. In the other case, the SVD of $log(A)$ must be computed again. What's more, the covariance function (4.6) involves the Mahalanobis distance instead of the Euclidean distance. The cost will then depend on d, and be higher than in the case with isotropic covariance parameters.
However, due to (4.5), it appears that if $w_\alpha(s_j)$ moves, only the row and column of Σ^i that correspond to s_j will be affected. Moreover, due to the symmetry of Σ^i, the row and the column will be changed exactly the same way. Therefore, computing $\partial\Sigma^i/\partial w_\alpha(s_j)$ involves only $m+1$ finite differences since Σ^i is of size $(m+1) \times (m+1)$.
The finite difference $\partial\Sigma^i/\partial w_\alpha(s_j)$ must be computed $m+1$ time for each row of $\tilde{R}$, and there is n rows. Therefore, the total cost of the finite differences should be $O(m+1)^2 n$
Therefore, we can hope that careful implementation of the derivative of $\frac{\partial \tilde{R}}{\partial(\alpha(s_1,\ldots,s_n))}$

will cost $O(n(m+1)^2)$ operations, in the same order as computing $\tilde{R}$ itself (Guinness, 2018).

4.10.5 Gradient of the negated log-density with respect to w_α

Sufficient augmentation. The negated log density of $\alpha(\mathcal{S})$ with sufficient augmentation writes

$$log\left(|\tilde{R}\left(\mathcal{S},\alpha(\mathcal{S})\right)|\right)+w^T\tilde{R}\left(\mathcal{S},\alpha(\mathcal{S})\right)^T\tilde{R}\left(\mathcal{S},\alpha(\mathcal{S})\right)w\times 1/2.$$

Let's write the derivative of the log-determinant $log(|\tilde{R}|)$:

$$\begin{aligned}
\partial log(|\tilde{R}|)/\partial w_\alpha(s_j) &= \partial(\Sigma_{i=1}^n log(\tilde{R}_{i,i}))/\partial w_\alpha(s_j)(\text{because } \tilde{R} \text{ is triangular})\\
&= \Sigma_{i=1}^n \partial log(\tilde{R}_{i,i})/\partial w_\alpha(s_j)\\
&\quad (\text{only the rows corresponding to } s_j \text{ and its children are affected})\\
&= \Sigma_{i/s_i\in\{s_j\cup ch(s_j)\}}\partial log(\tilde{R}_{i,i})/\partial w_\alpha(s_j)\\
&\quad (\text{log-function derivative})\\
&= \Sigma_{i/s_i\in\{s_j\cup ch(s_j)\}}\left(\partial\tilde{R}_{i,i}/\partial w_\alpha(s_j)\right)/R_{i,i}
\end{aligned}$$

Let's write the derivative of $w^T\tilde{R}^T\tilde{R}w\times 1/2$:

$$\begin{aligned}
\partial\left(w^T\tilde{R}^T\tilde{R}w\times 1/2\right)/\partial w_\alpha(s_j) &= \partial\left((w^T\tilde{R}^T)(\tilde{R}w)\times 1/2\right)\partial w_\alpha(s_j)\\
&= \partial(w^T\tilde{R}^T)/\partial w_\alpha(s_j)(\tilde{R}w)\times 1/2+\\
&\quad (w^T\tilde{R}^T)\partial(\tilde{R}w)/\partial w_\alpha(s_j)\times 1/2\\
&= (w^T\tilde{R}^T)\partial(\tilde{R}w)/\partial w_\alpha(s_j)\\
&= (w^T\tilde{R}^T)(\partial\tilde{R}/\partial w_\alpha(s_j)w)
\end{aligned}$$

Ancillary Augmentation. The negated log density of $\theta(\mathcal{S})$ with ancillary augmentation writes

$$-log(l(z|\tilde{R}^{-1}w^*,X,\beta,\ldots)).$$

Here, we do not write the white Gaussian prior on w^* because it is not affected by $\alpha(\mathcal{S})$. We leave empty slots in the data likelihood function $l(\cdot)$ because additional parameters can be needed, such as noise standard deviation τ^2 in a Gaussian model. Applying differentiation, we get

$$\partial \left(-log(l(z|\tilde{R}^{-1}w^*, X, \beta, \ldots)) \right) / \partial(w_\alpha(s_j))$$

(Conditional independence)

$$= \Sigma_{i=1}^{n} - \partial \left(log(l(z(s_i)| \left(\tilde{R}^{-1}w^* \right)_i, X, \beta, \ldots)) \right) / \partial(w_\alpha(s_j))$$

(Chain rule)

$$= \Sigma_{i=1}^{n} - \partial \left(\tilde{R}^{-1}w^* \right)_i / \partial(w_\alpha(s_j)) \times$$
$$\partial \left(log(l(z(s_i)|w(s_i) = \left(\tilde{R}^{-1}w^* \right)_i, X, \beta, \ldots)) \right) / \partial(w(s_i))$$

(w^* is not changed by θ)

$$= \Sigma_{i=1}^{n} - \left(\partial \tilde{R}^{-1} / \partial(w_\alpha(s_j)) w^* \right)_i \times$$
$$\partial \left(log(l(z(s_i)|w(s_i) = \left(\tilde{R}^{-1}w^* \right)_i, X, \beta, \ldots)) \right) / \partial(w(s_i))$$

(Differentiation of inverse)

$$= \Sigma_{i=1}^{n} \left(\tilde{R}^{-1} \partial \tilde{R} / \partial(w_\alpha(s_j)) \tilde{R}^{-1} w^* \right)_i \times$$
$$\partial \left(log(l(z(s_i)|w(s_i) = \left(\tilde{R}^{-1}w^* \right)_i, X, \beta, \ldots)) \right) / \partial(w(s_i))$$

(Taking gradient of $log(l(\cdot))$ in w)

$$= \nabla_w log(l(z(s_i)|\tilde{R}^{-1}w^*, X, \beta, \ldots)) \tilde{R}^{-1} \partial \tilde{R} / \partial(w_\alpha(s_j)) \tilde{R}^{-1} w^*$$

4.10.6 Computational cost of the gradient of the negated log-density with respect to w_α

Both sufficient and ancillary formulations have a partial derivative with a term under the shape:

$$u^T \partial \tilde{R} / \partial(w_\alpha(s_j)) v,$$

with u an v two vectors that do not depend on $w_\alpha(s_j)$ and with affordable cost. Due to its construction, $\partial \tilde{R} / \partial(w_\alpha(s_j))$ has non-null rows only at the rows that correspond to s_j and $ch(s_j)$, and each of those rows has itself at most $m+1$ non-null coefficients. Sparse matrix-vector multiplication $(\partial \tilde{R} / \partial(w_\alpha(s_j)))v$ therefore costs $O((m+1) \times (1 + |ch(s_j)|))$ operations. Given the fact that $\Sigma_{j=1}^{n} |ch(s_j)| = \Sigma_{j=1}^{n} |pa(s_j)| = n \times m$, we can expect that the computational cost needed to compute $(\partial \tilde{R} / \partial(w_\alpha(s_j)))v$ for $j \in 1, \ldots, n$ will be $O(n \times (m+1)^2)$ operations, which is affordable.

Moreover, due to the fact that $\partial \tilde{R} / \partial(w_\alpha(s_j))$ has non-null rows only at the rows that correspond to s_j and $ch(s_j)$, we can deduce that $(\partial \tilde{R} / \partial(w_\alpha(s_j)))v$ has non-null terms only on the slots that correspond to s_i and its children. Computing $u^T (\partial \tilde{R} / \partial(w_\alpha(s_j)))v$ will then cost $O(ch(s_j) + 1)$ operations. Using again $\Sigma_{j=1}^{n} |ch(s_j)| = \Sigma_{j=1}^{n} |pa(s_j)| = n \times m$, we can deduce that (if we know

already $(\partial\tilde{R}/\partial(w_\alpha(s_j)))v)$ computing $u^T(\partial\tilde{R}/\partial(w_\alpha(s_j)))v$ for $j \in 1,\ldots,n$ will cost $O(n(m+1))$.

4.10.7 Gradient of the negated log-density with respect to w_{τ^2}

The noise variance intervenes directly in the Gaussian likelihood (noted $f(\cdot)$) of the observed field:

$$f(z(\mathcal{S})|w(\mathcal{S}) + X(\mathcal{S})\beta^T, diag(\tau^2(\mathcal{S}))) = \Pi_{i=1}^n f(z(s_i)|w(s_i) + X(s_i)\beta^T, \tau^2(s_i)).$$

Passing to the negated log-density and introducing $\tau^2(s) = exp(w_{\tau^2}(s)+X_{\tau^2}(s)\beta_{\tau^2}^T)$, we have (within an additive constant)

$$\Sigma_{i=1}^n w_{\tau^2}(s_i)/2 + exp(-w_{\tau^2}(s) - X_{\tau^2}(s)\beta_{\tau^2}^T)(z(s_i) - w(s_i) - X(s_i)\beta^T)^2/2.$$

Differentiating with respect to w_{τ^2} brings

$$\nabla_{w_{\tau^2}} f(z(\mathcal{S})|w(\mathcal{S})+X(\mathcal{S})\beta^T, diag(\tau^2(\mathcal{S}))) = 1/2-\tau^2(\mathcal{S})\circ(z(\mathcal{S})-w(\mathcal{S})-X(\mathcal{S})\beta^T)^2/2.$$

4.11 Appendix: experiments on synthetic data sets

4.11.1 Objectives of the experiments

We would like to investigate

1. the improvements caused by using nonstationary models when it is relevant.
2. the problems caused by using nonstationary models when it is irrelevant.
3. the potential identification / overfitting problems of nonstationary models.

Our general approach to find answers to those questions is to run our implementation on synthetic data sets and analyze their results. Following the nonstationary process and data model we defined using (4.5), (4.3), and (4.6)/(4.7), there is 12 possible configurations counting the full stationary case: 2 marginal variance models, 2 noise variance models, 3 range models. In order to keep the section readable, we use the following notation for the different models:

- $(\emptyset)$ is the stationary model.
- (σ^2) is a model with nonstationary marginal variance.
- (τ^2) is a model with heteroskedastic noise variance.
- (α) is a model with nonstationary range and isotropic range parameters.

- (A) is a model with nonstationary range and elliptic range parameters.
- Complex models are noted using "+". For example, a model with non-stationary marginal variance and heteroskedastic noise variance is noted $(\sigma^2 + \tau^2)$.

4.11.2 Under-modeling, over-modeling, and identification

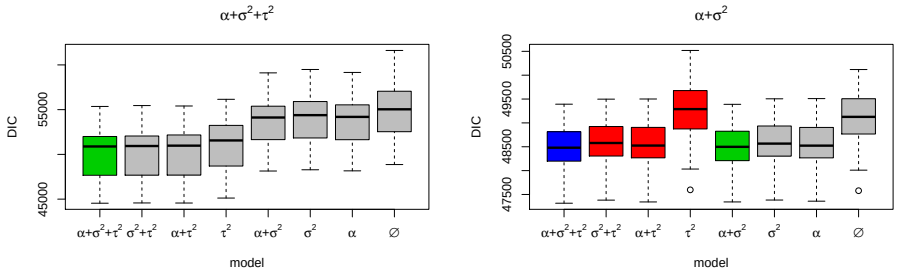
Model-data cases of interest. This experiment aims to answer to problems 1, 2, 3. Our approach here is to use a possibly misspecified model and see what happens. Four cases are possible:

- The "right" model, in the sense it matches perfectly the process used to generate the data (however, potential identification and overfitting problems may cause it to be a bad model in practice).
- "Wrong" models, where some parameters that are stationary in the data are non-stationary in the model, and some parameters that are stationary in the model are non-stationary in the data.
- Under-modeling, where some parameters that are stationary in the model are non-stationary in the data, but all parameters that are stationary in the data are stationary in the model.
- Over-modeling, where some parameters that are stationary in the data are non-stationary in the model, but all parameters that are stationary in the model are stationary in the data.

If a nonstationary model actually helps to analyze nonstationary data, we should see if the "right" model does better than under-modeling. The problem of overfitting will be assessed by comparing over-modeling, under-modeling, and the "right" model. If there is some overfitting, over-modeling or even "right"modeling would have worse performances than simpler models. Identification problems will be monitored by looking at the "wrong" models and under-modeling. If some model formulations are interchangeable, then some of the "wrong" models should perform as good as the "right" model. Also, if two parametrizations are equivalent, then using either parametrization should do as good as using both, therefore under-modeling should do as good as the "true" model. The models are compared using the DIC.

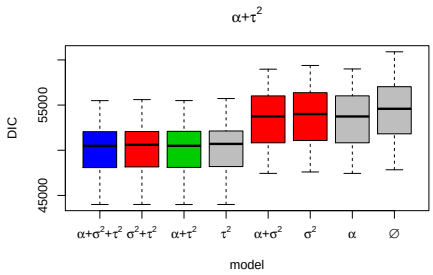
Models with $(\sigma), (\alpha)$, and (τ). We started with the eight models obtained by combining $(\sigma), (\alpha)$, and (τ), giving us $(\emptyset)$, (σ^2), (τ^2), (α), $(\sigma^2+\tau^2)$, $(\tau^2+\alpha)$, $(\sigma^2+\alpha)$, and $(\sigma^2+\tau^2+\alpha)$. We tested each data-model configuration, yielding 64 situations in total. Each case was replicated 30 times. The results are summarized by box-plots in figure 4.8.

Models with (α) and (A). We focused on the case of elliptic range parameters with the three models obtained by combining (α) and (A), giving us $(\emptyset)$, (α), and (A). Like before, we tested the 9 data-model configurations 30 times each. The results are summarized by box-plots in figure 4.9.

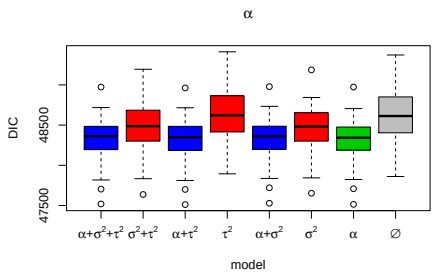


(a) $(\sigma^2 + \tau^2 + \alpha)$ data

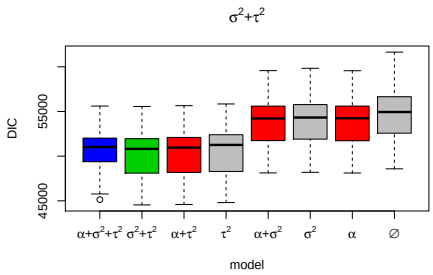
(b) $(\sigma^2 + \alpha)$ data



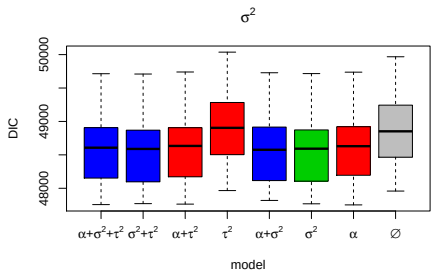
(c) $(\tau^2 + \alpha)$ data



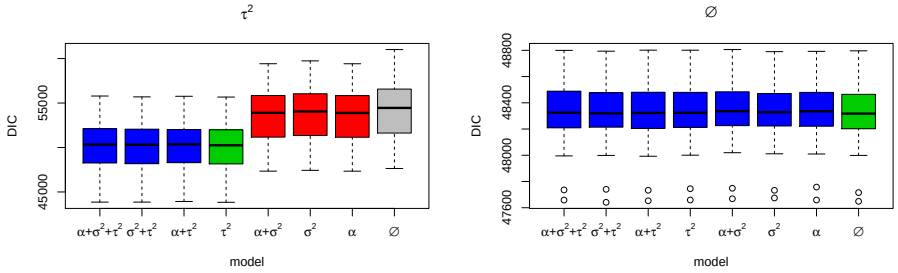
(d) (α) data



(e) $(\sigma^2 + \tau^2)$ data

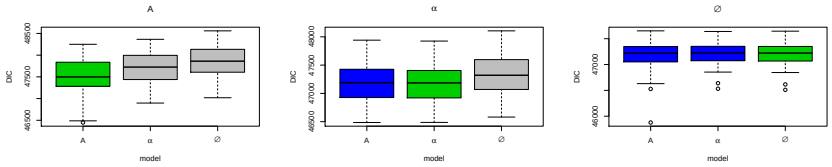


(f) (σ^2) data



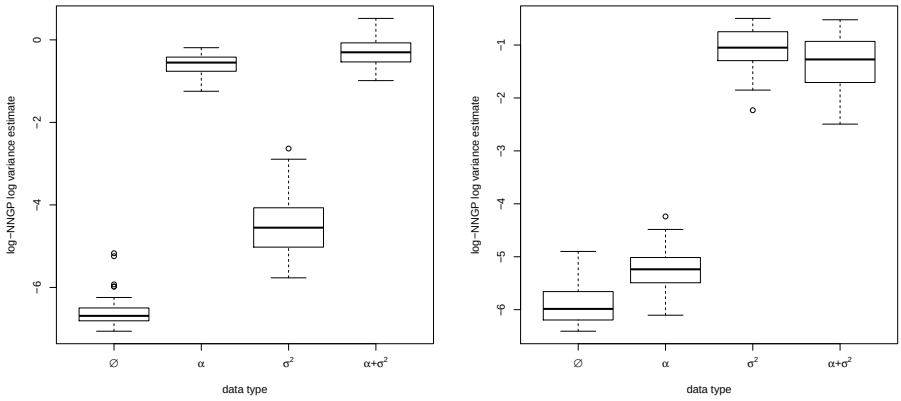
(g) (τ^2) data (h) ($\emptyset$) data

Figure 4.8: DIC of the models following the data
Legend: "right model" ■; "wrong model" ■; "over-modeling" ■; "under-modeling" ■



(a) (A) data (b) (α) data (c) ($\emptyset$) data

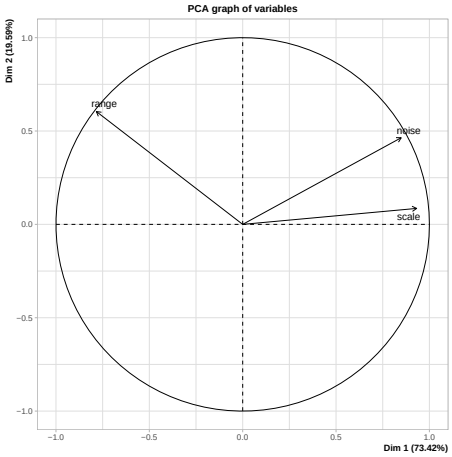
Figure 4.9: DIC of the models following the data, in the anisotropy model
Legend: "right model" ■; "over-modeling" ■; "under-modeling" ■



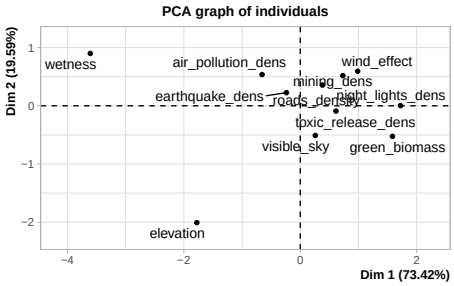
(a) Estimates of the log variance for w_α (b) Estimates of the log variance for w_σ

Figure 4.10: Estimates of the log-variance of w_α and w_σ in the model $(\alpha + \sigma^2)$ following the type of the data

4.12 Appendix: case study of lead concentration



(a) Circle correlation plot



(b) Coordinates of the Projection of the individuals

Figure 4.11: PCA of $(\beta_\alpha|\beta_{\sigma^2}|\beta_{\tau^2})$

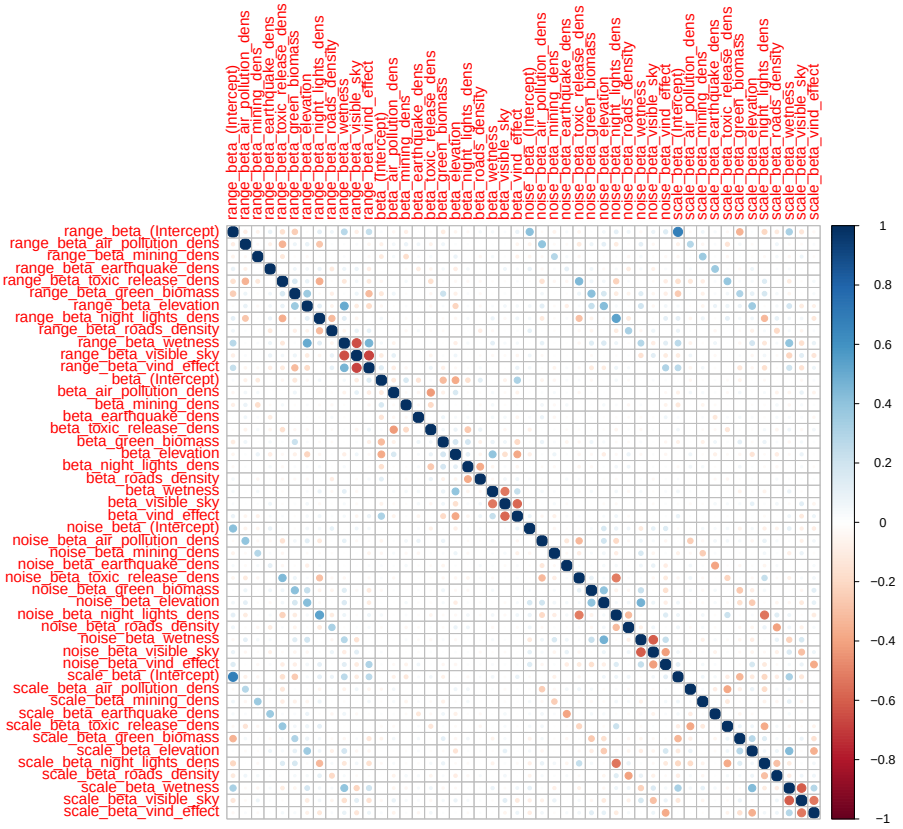


Figure 4.12: Correlation plot of the high level parameters for the lead data set analysis

Four groups are present: "range_beta_..." parametrizes the range, "beta_..." the linear regression, "scale_beta_..." the marginal variance, and "noise_beta_..." the noise variance.

Chapter 5

Conclusion

5.1 Contribution with respect to the initial objectives

The objective of this work was to improve the MCMC sampling of Nearest Neighbors Gaussian Process models with full data augmentation, in order to enjoy their possibilities while repressing their computational misbehavior. The first part of the work was to find upgrades for the original model presented by Datta et al. (2016). Conditionally on that, the second part was to exploit the specificities of NNGPs with full augmentation in order to propose new model architectures that require to sample explicitly a NNGP latent field. In accordance with an open science objective, the implementations of the methods are provided, with extensive vignettes giving hands-on examples and scripts allowing to reproduce the applications.

In this perspective, the thesis proposes and tests various modifications of the basic MCMC algorithm to tackle its mixing and convergence problems, while retaining its versatility.

In the first part of the dissertation, I present various methods that did not make their way into an article. The first of those developments stems from the fact that fast forward solving is extremely cheap and scalable with NNGPs. From this point, all I had to do was to connect the dots and look for strategies that rely on this transformation. The interweaving of Filippone et al. (2013) improves the sampling of the covariance parameters. Whitened HMC (Neal et al., 2011) allows for efficient Hybrid Monte-Carlo, even with high spatial autocorrelation. I also propose a prediction algorithm relying on this transformation, that permits to work with MCMC samples while remaining high-level. Another lead I investigated is delayed acceptance (Christen and Fox, 2005), but its results were disappointing in spite of its promising appearances.

The second part of the thesis presents an article promoting two other modifications of the basic algorithm. The first proposal is yet another interweaving

strategy. This original method targets the regression coefficients of the linear part of the NNGP model. Its practical efficiency also relies on the properties of NNGPs. Experiments on synthetic data sets show that it greatly improves the sampling of coefficients associated with covariates having some spatial coherence. This was also the occasion for me to take a closer look at the linear part of the NNGP model, and to expose that it is much less trivial that it may seem at first sight because of its interactions with the latent field.

The second method is the use of Chromatic Samplers to obtain a parallelizable and high-level step to update the latent NNGP field. Empirical exploration proved that this method is widely and easily applicable to NNGPs with usual settings. Incidentally, it was an opportunity to explore how the heuristics used in the construction of the DAG affected the structure of the NNGP's Markov graph.

The proposed algorithms were implemented and tested against the spNNGP package. The implementation is openly available at `https://github.com/SebastienCoube/Improving_NNGP_full_augmentation`. (Finley et al., 2017) on synthetic data sets. The results were promising given the fact that the proposed implementation was not low-level and fine-tuned, contrary to Finley et al. (2017). On simple data sets, the methods are equivalent; on "tricky" data sets with correlated covariates, spNNGP was clearly outperformed both in terms of MCMC behavior quality and computational time. The implementation allowed to study a data set of lead contamination in the mainland of the United States of America, while spNNGP gave erratic results on this application.

The third part of the dissertation builds on the first two parts in order to propose an application that would not be possible without explicit sampling of the latent field.

A nonstationary NNGP model is proposed, with a hierarchical architecture that integrates latent fields of spatially variable covariance parameters, allowing to capture spatial variations of the covariance structure. This architecture aims to answer to three thorny problems of nonstationary modeling. The first problem is the interpretability of the parametrization for a complex, multi-layer model. The second is the model selection when several nonstationary models are available. The third problem is the computational complexity.

The problem of the parametrization has been tackled by embedding the models in an expanding and coherent family, so that the simpler models are comprised in the complex ones. The use of logarithm transformations allows for intuitive interpretation of the parameters. Another good point is that complex nonstationary models can be laid out fairly simply under a probabilistic formulation.

Model selection is eased by the structure of the model family. On synthetic data sets, over-modeling does not lead to over-fitting; instead, the model degenerates and is practically equivalent to a simpler model. In those cases, over-modeling can be detected easily by looking at the MCMC samples. Those results on synthetic data sets must be taken with a pinch of salt. Nonetheless, on real data sets such as the lead contamination, nonstationary modeling

performed better than simpler models in terms of DIC.

As for the problem of computation, a MCMC strategy based on Hybrid Monte-Carlo and nested interweaving is proposed, its implementation being freely available at `https://github.com/SebastienCoube/Nonstat-NNGP`. Hybrid Monte-Carlo (using the whitening method presented in the beginning of the thesis) is used in order to sample spatially correlated parameter fields. Nested interweaving is needed because latent fields are present at various levels of the model. This strategy is envisioned by the inventors of interweaving (Yu and Meng, 2011), but as far as I know nested interweaving has not been put in application for large models. While the method allows to work on data sets in the order of a few tens of thousands observations, which is respectable given the size of data treated with state-of-the-art nonstationary methods (Fuglstad et al., 2015a), it is clear that a scale-up is still required to work with large modern data sets such as Datta et al. (2016)'s.

5.2 Perspectives

The interweaving scheme used to improve the behavior of the regression coefficients has one shortcoming, which is that it cannot be applied to covariates that do not change within one spatial location. I did not put too much effort in looking for a solution because there was no need to: indeed, the NNGP model being applied to point-measurement data, all regressors obtained through grids or areas are invariant in a point. The experiments on synthetic data sets show that the regressors with some spatial coherence cause trouble when the "vanilla" algorithm is used. A problematic behavior should then occur when a covariate has some spatial coherence and some within-site variability. What one would like to do in that case is to replace the troublesome variable by its site per site average, so that interweaving can be applied. If it is absolutely necessary to work with the exact covariate, the averaged covariate can still serve as an instrument to approach the actual density. The lead of using this approximation as a proposal density in a Metropolis step might be worth investigating.

This work concerning the trouble caused by the interference between the linear effect and the latent field in the NNGP model might be re-used in other cases where various effects that have some spatial coherence cohabit within the same model. One can imagine a model with two latent fields, one seasonal effect and one temporal drift for example, or one long-scale correlation and some local variations. A "Russian doll" parametrization where one field is centered on the other might be useful in this case. However, it remains to find whether there is an efficient formulation for such a model using NNGPs.

While the chromatic sampler has its limits such as spatial auto-correlation, I cannot think of ways to make it better in the Gaussian data case. What has to be done is to find an efficient implementation for non-Gaussian data, since the exact Gaussian draw has to be replaced. The problem is that we are spoiled for choice and several methods must be benchmarked. In the case of binomial data,

Polya-Gamma variables are an option (Polson et al., 2013; Finley et al., 2017). More generally, another method is to use a Metropolis step within the chromatic sampler, but this requires to tune the proposal distributions and there are many of them. Other options such as the Slice Sampler need no tuning, but they are more expensive. I think that a good lead is to take advantage of the one thing that does not change, whatever the data model: the NNGP prior, whose full conditional distribution may be used as a proposal distribution in a Metropolis step.

Beyond the sampling of the latent field, the exploration of NNGP moral graphs from the perspective of coloring highlighted how the heuristics used to build the NNGP DAG affect the induced graphs. While Guinness (2018) carried out systematic exploration of how the heuristics affect the quality of Vecchia's approximation, I have not heard of an analogous study focused on the properties of the DAG and/or the Markov graph. For example, degree distribution of the moral graph varies greatly following the rank of the vertex in the DAG ordering, with several modalities following the ordering heuristic. The edge lengths and directions in the DAG change with the ordering heuristic.

Having insight on the relationship between the graph and the quality of the approximation would help to look for better heuristics since it should be easier to imagine the impact of a heuristic on the graph than on the induced Vecchia's approximation. An example to support this method is that I remarked that ordering the locations following a coordinate introduces some anisotropy along that coordinate in the NNGP samples (incidentally, I suspect that this happens in Datta et al. (2016), where the predictions of the biomass of the US are stretched), while ordering following the distance to the center provokes some radial anisotropy. Those two cases have in common that the undue anisotropy follows the general direction of the edges in the DAG. I think that it is no coincidence if Guinness (2018) finds that for two-dimensional geographic spaces coordinate and middle-out ordering are the least accurate. So, if I had to find a new ordering heuristic, at least I would know that I should not come up with one that produces a DAG with combed edges.

As for the nonstationary NNGP model, I am very satisfied with the interpretability and the behavior of the model with respect to over-fitting. However, there is still work to do in order to improve its scalability and robustness. An interesting workaround might arise from a result that I presented in the "catch-all" part of the thesis, equation (2.5). In this development I show that a truncated ancillary augmentation of the latent field induces a Predictive Process (PP) (Banerjee et al., 2008). Out of curiosity, I treated cases that cause problems to the log-NNGP nonstationary model using a formulation where the log-NNGP random effects are suppressed and replaced by a PP basis of size $k = 50$ or 100 elements integrated in the fixed effects:

$$log(\theta(s)) = X_\theta(s)\beta_\theta^T + w_\theta(s) \quad \text{becomes} \quad log(\theta(s)) = X_\theta(s)\beta_\theta^T + X_{\theta PP}(s)\beta_{\theta PP}^T,$$

where $\theta(s)$ is a nonstationary covariance parameter in the spatial site s, $X_\theta(s)$ are the "usual" covariates (for example, elevation, wetness, etc, in the lead

application), $w_\theta(s)$ is the log-NNGP latent field, and $X_{\theta PP}(s)$ is the spatial basis. The spatial basis is obtained by solving $X_{\theta PP} = \tilde{R}_{0_\theta}^{-1} M$, $\tilde{R}_{0_\theta}$ being the prior NNGP Cholesky factor used in the log-NNGP prior, and M being a $n \times k$ matrix with only null coefficients except a diagonal $M_{i,i} = 1$. In other terms, the basis is composed of the k first columns of the NNGP correlation matrix Cholesky factor $\tilde{R}_{0_\theta}^{-1}$. The MCMC behavior of the model was satisfying and allowed to retrieve spatial nonstationarity patterns. Indeed, the model with spatial basis is much smaller than the full log-NNGP model: instead of the tenths of thousands parameters of the latent log-NNGP field $w_\theta(\cdot)$, there is only a few dozen additional regression coefficients needing to be estimated. Moreover, the hyperprior range of the log-NNGP prior being high by construction of the model, only a few elements of the spatial basis are enough to outline a coherent field, like in figure 2.3. The approach has the merit to work, but it lacks the interpretability of the log-NNGP prior. The interesting point is that adding a Normal prior

$$\beta_{\theta PP} \overset{a\ priori}{\sim} \mathcal{N}(\mathbf{0}_k, \sigma_\theta I_k)$$

is enough to turn this linear model component into a degenerate NNGP, where the ancillary augmentation has been truncated:

$$w_\theta = \sigma_\theta \tilde{R}_{0_\theta}^{-1} w_\theta^* \quad \text{is replaced by} \quad w_{\theta_{PP}} = \tilde{R}_{0_\theta}^{-1} M \beta_{\theta PP}^T = \sigma_\theta \tilde{R}_{0_\theta}^{-1} \underbrace{M \beta_{\theta PP}^T / \sigma_\theta}_{\text{truncated } w_\theta^*},$$

where $w_\theta^* \overset{a\ priori}{\sim} \mathcal{N}(0_n, I_n)$ while $\beta_{\theta PP}^T / \sigma_\theta \overset{a\ priori}{\sim} \mathcal{N}(0_k, I_k)$. The inconvenient of a PP model is its over-smoothing (Banerjee et al., 2008; Datta et al., 2016). In this specific application, I do not think that it is a big problem because of the fact that the log-NNGP prior already is a smooth, large-scale prior. The gain of the approach would be a reduction in the dimension of the high-level layers of the model, inducing a much simpler and economical MCMC architecture.

Aside of the practical issues, many open problems and possibilities arise from the developments on nonstationary modeling. First, I did not find a solution to nonstationary anisotropic modeling on the whole sphere because of the impossibility to find a common parametrization for the elliptic range parameters. Transposing and deepening the recursive tangent projection method that was used to define a nonstationary NNGP in the absence of a nonstationary covariance function on the sphere might be a good research direction. This method defines a global behavior from a collection of local specifications, which sounds like a good start in a problem caused by the absence of a common parametrization.

Another point, that arises naturally, is to try and work on multivariate nonstationary models. Multivariate spatial modeling also is a thorny subject in itself, even if recent works related to NNGP (Peruzzi et al., 2020; Taylor-Rodriguez et al., 2019) tackle the subject. But it is clear that some data, such as the lead contamination, are best modeled with a nonstationary framework;

but on the other hand, they come from multivariate data sets and there may be some interest in joint modeling.

In addition to that, I wonder if it is possible to work on some "nonstationary multivariateness", a model where the association between the interest variables changes following some covariates or the spatial location. If it was possible to define such a model, I think that my work on matrix logarithms would be reusable to model variable association matrices.

A lead that should be considered seriously is the option of the Maximum *A Posteriori* (MAP) estimation for NNGP models. Indeed, there is a trade-off between MCMC and NNGP. The computational power that is not invested in the exploration of the posterior distribution can be re-used in the construction of richer Vecchia's approximations.

Given the fact that I spent three years doing MCMC, switching to gradient-based methods might sound like a confession of failure. A very partial one then. It is because of the acquaintance with the full distributions that has been gained thanks to MCMC sampling that I know that the *a posteriori* distributions of the parameters are well-behaved, in the sense that they are "always" unimodal (I never stumbled upon a multimodal case) and generally symmetric tanks to the logarithmic parametrization. For example, in the case study of nonstationary modeling of lead contamination, the MCMC samples of the regression coefficients for the nonstationary covariance parameters passed the joint multivariate normality test of Henze and Wagner (1997). Therefore, even if summarizing the *a posteriori* distribution by its mode causes a loss of information, we can expect this summary to be reasonable, and compensated by the possibility to do better NNGPs.

Moreover, the present work on MCMC leaves us with a toolbox that may come in handy for a MAP approach. First, this PhD showed for the purpose of implementing Hamiltonian methods that the gradient of the NNGP density is affordable, even in complex cases such as a nonstationary covariance function. Methods such as the gradient descent and the coordinate descent are known to be sensitive to the parametrization. Thanks to the developments on interweaving, several parametrizations of the gradient are available - I wonder if some lowbrow transposition of interweaving to gradient algorithms would give an interesting result. And of course, the methods that were developed for a Gibbs sampler can be readily salvaged for use in a coordinate descent algorithm. For example, chromatic sampling updates the latent field by computing its conditional mean and then adding a noise whose intensity is proportional to the conditional variance. One can remove the noise and remain with the mean, which is the maximum of the conditional distribution. Yet another possibility is to draw from the stochastic EM algorithm and apply gradient methods only on the upper levels of the model while keeping the lower parameters stochastic.

Bibliography

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008, September). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

Banterle, M., C. Grazian, and C. P. Robert (2014, June). Accelerating Metropolis-Hastings algorithms: Delayed acceptance with prefetching. *arXiv:1406.2660 [stat]*. arXiv: 1406.2660.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics 7*(4), 434–455.

Christen, J. A. and C. Fox (2005, December). Markov chain Monte Carlo Using an Approximation. *Journal of Computational and Graphical Statistics 14*(4), 795–810.

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

Cressie, N. and G. Johannesson (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 209–226.

Cressie, N. and C. K. Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

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

Eddelbuettel, D., R. François, J. Allaire, K. Ushey, Q. Kou, N. Russel, J. Chambers, and D. Bates (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software 40*(8), 1–18.

Filippone, M., M. Zhong, and M. Girolami (2013). A comparative evaluation of stochastic-based inference methods for gaussian process models. *Machine Learning 93*(1), 93–114.

Finley, A., A. Datta, and S. Banerjee (2017). spnngp: spatial regression models for large datasets using nearest neighbor gaussian processes. *R package version 0.1 1*.

Finley, A. O., A. Datta, B. D. Cook, D. C. Morton, H. E. Andersen, and S. Banerjee (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics 28*(2), 401–414.

Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika 89*(1), 197–210.

Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue (2015a). Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics 14*, 505–531.

Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue (2015b). Interpretable priors for hyperparameters for gaussian random fields. *arXiv preprint arXiv:1503.00256*.

Gelfand, A. E., P. Diggle, P. Guttorp, and M. Fuentes (2010). *Handbook of Spatial Statistics (Chapman Hall CRC Handbooks of Modern Statistical Methods)*. Chapman Hall CRC Handbooks of Modern Statistical Methods. Taylor and Francis.

Gelfand, A. E., M. Fuentes, J. A. Hoeting, and R. L. Smith (2019). *Handbook of environmental and ecological statistics*. CRC Press.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parametrisations for normal linear mixed models. *Biometrika 82*(3), 479–488.

Gelman, A., D. B. Rubin, et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.

Gonzalez, J., Y. Low, A. Gretton, and C. Guestrin (2011). Parallel gibbs sampling: From colored fields to thin junction trees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 324–332. JMLR Workshop and Conference Proceedings.

Grossman, J. N. et al. (2004). *The National Geochemical Survey-database and documentation*.

Guinness (2018). Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics 60*(4), 415–429.

Guinness and Katzfuss (2018). *GpGp: Fast Gaussian Process Computation Using Vecchia's Approximation*.

Heaton, M. J., A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics 24*(3), 398–425.

Heinonen, M., H. Mannerström, J. Rousu, S. Kaski, and H. Lähdesmäki (2016). Non-stationary gaussian process regression with hamiltonian monte carlo. In *Artificial Intelligence and Statistics*, pp. 732–740.

Hengl, T. (2009). *A practical guide to geostatistical mapping*. Hengl Amsterdam.

Henze, N. and T. Wagner (1997). A new approach to the bhep tests for multivariate normality. *Journal of Multivariate Analysis 62*(1), 1–23.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics 5*(2), 173–190.

Hoffman, M. D., A. Gelman, et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res. 15*(1), 1593–1623.

Horton, J. D. (2017). The state geologic map compilation (sgmc) geodatabase of the conterminous united states.

Ingebrigtsen, R., F. Lindgren, I. Steinsland, and S. Martino (2015). Estimation of a non-stationary model for annual precipitation in southern norway using replicates of the spatial field. *Spatial Statistics 14*, 338–364.

Katzfuss, M. and J. Guinness (2017, Aug). A general framework for Vecchia approximations of Gaussian processes. *arXiv e-prints*, arXiv:1708.06302.

Katzfuss, M., J. Guinness, W. Gong, and D. Zilber (2020). Vecchia approximations of gaussian-process predictions. *Journal of Agricultural, Biological and Environmental Statistics 25*(3), 383–414.

Katzfuss, M., J. Guinness, and E. Lawrence (2020). Scaled vecchia approximation for fast computer-model emulation. *arXiv preprint arXiv:2005.00386*.

Katzfuss, M., M. Jurek, D. Zilber, W. Gong, J. Guinness, J. Zhang, and F. Schäfer (2020). Gpvecchia: Fast gaussian-process inference using vecchia approximations. *R package version 0.1 3*.

Kleiber, W. and D. Nychka (2012). Nonstationary modeling for multivariate spatial processes. *Journal of Multivariate Analysis 112*, 76–91.

Knorr-Held, L. and H. Rue (2002). On block updating in markov random field models for disease mapping. *Scandinavian Journal of Statistics 29*(4), 597–614.

Kroese, D. P. and Z. I. Botev (2015). Spatial process simulation. In *Stochastic geometry, spatial statistics and random fields*, pp. 369–404. Springer.

Lauritzen, S. L. (1996). *Graphical models*. Oxford Statistical Science Series. OUP.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.

Neal, R. (2011, May). MCMC Using Hamiltonian Dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, Volume 20116022. Chapman and Hall/CRC.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo 2*(11), 2.

Paciorek, C. J. (2003). *Nonstationary Gaussian processes for regression and spatial modelling*. Ph. D. thesis, Citeseer.

Peruzzi, M., S. Banerjee, and A. O. Finley (2020). Highly scalable bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *Journal of the American Statistical Association*, 1–14.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association 108*(504), 1339–1349.

R Core Team (2018a). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team (2018b). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*, 266.

Risser, M. D. (2016). Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches. *arXiv preprint arXiv:1610.02447*.

Risser, M. D. and C. A. Calder (2015). Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics 26*(4), 284–297.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods* (2nd ed.). Springer Texts in Statistics. Springer.

Rue, H. and L. Held (2005). *Gaussian Markov random fields: theory and applications* (1 ed.). Monographs on statistics and applied probability 104. Chapman Hall/CRC.

Rue, H., I. Steinsland, and S. Erland (2004). Approximating hidden gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(4), 877–892.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van der Linde (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical report, Research Report, 98-009.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology) 64*(4), 583–639.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging* (1 ed.). Springer Series in Statistics. Springer-Verlag New York.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stein, M. L., Z. Chi, and L. J. Welty (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(2), 275–296.

Stevens, S. S. et al. (1946). On the theory of scales of measurement.

Taylor-Rodriguez, D., A. O. Finley, A. Datta, C. Babcock, H.-E. Andersen, B. D. Cook, D. C. Morton, and S. Banerjee (2019). Spatial factor models for high-dimensional and large spatial data: An application in forest variable mapping. *Statistica Sinica 29*, 1155.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological) 50*(2), 297–312.

Yang, H.-C. and J. R. Bradley (2021). Bayesian inference for big spatial data using non-stationary spectral simulation. *Spatial Statistics 43*, 100507.

Yu, Y. and X.-L. Meng (2011). 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.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association 99*(465), 250–261.

Zhang, L., A. Datta, and S. Banerjee (2019). Practical bayesian modeling and inference for massive spatial data sets on modest computing environments. *Statistical Analysis and Data Mining: The ASA Data Science Journal 12*(3), 197–209.

Zilber, D. and M. Katzfuss (2021). Vecchia–laplace approximations of generalized gaussian processes for big non-gaussian spatial data. *Computational Statistics & Data Analysis 153*, 107081.