NNGP Models with Spatial Examples of Simulated Data and Forest Canopy Heights

Frances Lin

Dec 2022

1. Introduction

Inferences and predictions on large spatial data or data with locations $\approx 10^6$ have either been too computationally challenging or infeasible. Methods for large spatial data that are under active development. However, most of the existing methods has focused primarily on theoretical and methodological developments. These methods have not paid enough attention on the algorithmic details nor made use of high-performance computing (HPC) libraries to expedite expensive computations and delivery full Bayesian inference for large spatial data.

On the other hand, while the original NNGP (nearest neighbor Gaussian process) model (Datta et al., 2016), which is also referred to as the latent NNGP model, appears promising, the latent NNGP model is prone to high autocorrelations and slow convergence because the Gibbs sampler involves updating a high-dimensional vector of latent random effect *sequentially*.

Three alternate formulations of the NNGP model that are more efficient and practical than the latent NNGP model (2016) are proposed: (1) the collapsed NNGP model, (2) the response NNGP model, and (3) the conjugate NNGP model, and these models are accessible through the `R` package `spNNGP` (Finley et al., 2021).

Section 2 reviews Gaussian process (GP) and formulates nearest neighbor Gaussian process (NNGP) through graphical models. *Section 3* introduces the latent NNGP model, which is followed by three alternate formulations of the latent NNGP model: (1) a collapsed NNGP model, (2) a NNGP model for the response (with no latent process), and (3) a conjugate NNGP model that allows for MCMC-free exact inference. *Section 4* ..... *Section 5* includes the discussion. ..... are included in the *Appendix*.

2. Nearest Neighbor Gaussian Processes

2.1. Review of mixed-effects and GP model for spatial data

Let $(s_i, y(s_i), x(s_i))$ be a triplet, where s_i denotes the location of measurement, $y(s_i)$ denotes the response of interest and $x(s_i)$ denotes the known or observed covariates, for $i = 1, ..., n$, a spatial linear mixed-effects model is given as

$$y(s_i) = x(s_i)^T \beta + w(s_i) + \epsilon(s_i), \tag{1}$$

where β is the vector of coefficients, $w(s_i)$ is the vector of unknown or unobserved covariates or random effects, and $\epsilon(s_i) \sim^{iid} N(0, \tau^2)$ is the random noise.

Gaussian processes (GPs) are widely used in machine learning to model smooth functions for regression, classification, and other tasks (Rasmussen, 2003, as cited in Finley et al., 2021). In spatial statistics, GPs are typically used to model the latent surface $\{w(s)\}$. A GP model for the spatial surface, which is given as

$$w(s) \sim GP(0, C(\cdot, \cdot|\theta)),$$

where $C(\cdot, \cdot|\theta)$ is a covariance function, implies that the vector of random effects $w = (w(s_1), ..., w(s_n))^T$ follows a multivariate Gaussian distribution with mean zero and covariance matrix $C(\theta) = C = (c_{ij})$, where $c_{ij} = C(s_i, s_j|\theta)$ and θ is the covariance parameters of the GP. The process is completely specified by a valid covariance function $C(\cdot, \cdot|\theta)$ (Datta et al., 2016). A popular choice of θ for $C(\cdot, \cdot|\theta)$ is selected from the Matérn covariance function or Matérn kernel (Matérn, 1960). For example, let s_i and s_j be two points in $\mathcal{D}$, then the Matérn covariance function is given as

$$C(s_i, s_j; \sigma^2, \phi, \nu) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(||s_i - s_j||\phi)^\nu \mathcal{K}_\nu(||s_i - s_j||\phi), \quad \phi > 0, \nu > 0,$$

where $||s_i - s_j||$ is the Euclidean distance between locations s_i and s_j, $\theta = \{\sigma^2, \phi, \nu\}$ are respectively the marginal variance, scale (inverse of range) or decay and smoothness parameter, Γ is the gamma function, $||\cdot||$ denotes the Euclidean distance in $\mathbb{R}^d$, and $\mathcal{K}$ is the Bessel function of second kind (Stein, 1999, as cited in Datta et al., 2016; Finley et al., 2019).

The hierarchical model can be constructed by combining the mixed-effects model for the response and the GP model for the random effects and is given as

$$p(\beta, \theta, \tau^2) \times N(w|0, C(\theta)) \times N(y|X\beta + w, \tau^2 I), \tag{2}$$

where $p(\beta, \theta, \tau^2)$ is specified by assigning priors to β, θ and τ^2. Alternatively, the marginal model can be constructed by integrating w out from (2) and is given as

$$N(y|X\beta, \Sigma = C(\theta) + \tau^2 I). \tag{3}$$

In a frequentist paradigm, parameter estimation can be obtained from (3) via MLE (maximum likelihood estimation), whereas in a Bayesian framework, after assigning priors to the parameters, posterior inference can be obtained from either (2) or (3) using MCMC (Markov chain Monte Carlo; Finley et al., 2021).

Alternatively, (2) is the same as

$$(\beta, \tau^2, \theta) \sim p(\beta, \tau^2, \theta)$$

$$w|\theta \sim N(0, C(\theta))$$

$$y|\beta, w, \tau^2 \sim N(X\beta + w, \tau^2 I).$$

2.2. Nearest Neighbor Gaussian Processes for large spatial data

When n is large, evaluating both (2) and (3) can be computationally challenging or infeasible. More specifically, evaluating $N(w|0, C)$ requires $\mathcal{O}(n^3)$ computations and storing the matrix C requires $\mathcal{O}(n^2)$ storage. In addition, predicting the response at new locations K requires additional $\mathcal{O}(kn^2)$ operations. Unfortunately, integrating w out from (2) does not always give computational advantages either.

One of the solutions is to replace GP prior for the spatial random effects w with a NNGP prior (Datta et al., 2016, as cited in Finley et al., 2021).

from Finley et al. (2021)

Let $\mathcal{R} = \{s_1, ..., s_n\}$ be any finite set of locations in the spatial domain $\mathcal{D}$ and $w_{\mathcal{R}} = (w(s_1), ..., w(s_n))^T$. For any location in $\mathcal{D}$, define neighbor sets as

$$\begin{gathered} N(s_1) = \{\} \text{ (empty set)}, \\ N(s_i) = \min(m, i-1) \text{ nearest neighbors of } s_i \text{ in } s_1, ..., s_{i-1}, \text{ for } i = 2, ..., n, \\ N(s) = m \text{ nearest neighbors of } s \text{ in } \mathcal{R}, \end{gathered} \tag{4}$$

then the NNGP is given as

$$w_{\mathcal{R}} \sim \prod_{i=1}^{n} p(w_i | w(N(s_i))). \tag{5}$$

from Datta et al. (2016)

Let $\mathcal{S} = \{s_1, s_2, ..., s_k\}$ be a fixed collection of distinct locations in $\mathcal{D} \subseteq \mathcal{R}^d$ ($\mathcal{S}$ is referred to as the *reference set*), then by the chain rule, joint density of $w_{\mathcal{S}}$ can be expressed as a product of conditional densities. That is,

$$\begin{aligned} p(w_{\mathcal{S}}) &= p(w(s_1))p(w(s_2)|w(s_1)) \cdots p(w(s_k)|w(s_{k-1}), ..., w(s_1)) \\ &= \prod_{i=1}^{k} p(w(s_i)| \cap_{j=1}^{k-1} w(s_j)). \end{aligned}$$

Next replace the right-hand side of... with smaller, carefully chosen, conditioning sets of size at most m, where $m \ll k$ (see, e.g., ...), then, for every $s_i \in \mathcal{S}$, a smaller conditioning sets $N(s_i) \subset \mathcal{S} \setminus \{s_i\}$ is used to construct a new density

$$\tilde{p}(w_{\mathcal{S}}) = \prod_{i=1}^{k} p(w(s_i)|w_{N(s_i)}),$$

where $w_{N(s_i)}$ is the vector of $w(s)$ over $N(s_i)$.

The pair $\{\mathcal{S}, N_{\mathcal{S}}\}$ can be viewed as a directed graph $\mathcal{G}$, where $\mathcal{S}(= \{s_1, s_2, ..., s_k\})$ is the set of nodes and $N_{\mathcal{S}}(= \{N(s_i); i = 1, 2, ..., k\})$ is the set of directed edges. $N(s_i)$ denotes the set of directed neighbors of s_i ($N(s_i)$ is referred to as the *neighbor set* for s_i). If $\mathcal{G}$ is a directed acyclic graph, then $\tilde{p}(w_{\mathcal{S}})$ is a proper multivariate joint density (see Appendix A1 of Datta et al., 2016). In addition, for a very general class of neighboring sets, $\tilde{p}(w_{\mathcal{S}})$ is a joint density of a multivariate Gaussian distribution with a sparse precision matrix $\tilde{C}_{\mathcal{S}}^{-1}$. More specifically, let $C_{N(s_i)}$ be the covariance matrix of $w_{N(s_i)}$ and $C_{s_i, N(s_i)}$ be the cross-covariance

matrix between $w(s_i)$ and $w_{N(s_i)}$, then $\tilde{p}(w_{\mathcal{S}})$ is a multivariate Gaussian density with covariance matrix $\tilde{C}_{\mathcal{S}} = {B_{\mathcal{S}}}^{-1} F_{\mathcal{S}}^{-1} B_{\mathcal{S}}$ and

$$\tilde{p}(w_{\mathcal{S}}) = \prod_{i=1}^{k} N(w(s_i)|B_{s_i} w_{N(s_i)}, F_{s_i}),$$

where $B_{s_i} = C_{s_i,N(s_i)} C_{N(s_i)}^{-1}$ and $F_{s_i} = C(s_i, s_i) - C_{s_i,N(s_i)} C_{N(s_i)}^{-1} C_{N(s_i),s_i}$. This is because, by the theorem, if $p(w_{\mathcal{S}}) = N(w_{\mathcal{S}}|0, C_{\mathcal{S}})$, then $w(s_i)|w_{N(s_i)} \sim N(B_{s_i} w_{N(s_i)}, F_{s_i})$ (see Appendix A2 of Datta et al., 2016).

..... $\tilde{p}(w_{\mathcal{S}})$ is referred to as the nearest neighbor density of $w_{\mathcal{S}}$.

from Finley et al. (2019)

That is, the underlying idea of the NNGP models is similar to that of the graphical models. The joint distribution for a random vector w can be viewed as a directed acyclic graph (DAG). More specifically, $p(w) = p(w_1, w_2, ...w_n)$ can be written as

$$p(w) = p(w_1) \prod_{i=2}^{n} p(w_i|Pa[i]), \tag{4-1}$$

where $w_i \equiv w(s_i)$ and $Pa[i] = \{w_1, w_2, ...w_{i-1}\}$ is a set of parents of w_i,

or, more explicitly, as

$$p(w) = p(w_1)p(w_2|w_1)p(w_3|w_1, w_2) \cdots p(w_n|w_1, w_2, ...w_{i-1}) \tag{4-2}$$

(Datta et al., 2017). Sparse models for w can be constructed by shrinking the size of $Pa[i]$. .....

..... The multivariate Gaussian density $N(w|0, C)$ (or $w \sim N(0, C(\theta))$) in (2) can be written as a linear model

$$w_1 = 0 + \eta_1,$$

$$w_2 = a_{21} w_1 + \eta_2,$$

$$w_i = a_{i1} w_1 + a_{i2} w_2 + \cdots + a_{i,i-1} w_{i-1} + \eta_i, \text{ for } i = 2, ..., n, \tag{4-3}$$

or, more explicitly, as

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & \cdots & 0 & 0 \\ a_{31} & a_{32} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn_1} & 0 \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \vdots \\ \eta_n \end{bmatrix} \tag{4-4}$$

(Datta et al., 2017),

or, more compactly, as

$$w = Aw + \eta, \tag{4-5}$$

where A is nxn strictly lower-triangular and $\eta \sim N(0, D)$ with $D = (d_1, d_2, ..., d_n)$ is diagonal. It follows that $I - A$ is nonsingular and, by the Cholesky factorization (Cholesky decomposition), a covariance matrix C

can be factorized into a product $C = (I - A)^{-1} D (I - A)^{-T}$, where for any matrix M, M^{-T} refers to the inverse of its transpose.

However, the Cholesky factorization for the full GP covariance C does not offer any computational advantages. Instead, the sparsity was introduced through graphical models (Datta et al., 2017).

To construct a sparse precision matrix, start with a dense nxn covariance matrix C and construct a sparse strictly lower-triangular matrix A with no more than $m(\ll n)$ nonzero entries in each row and the diagonal matrix D, then the matrix $\tilde{C} = (I - A)^{-1} D (I - A)^{-T}$ is a covariance matrix and its inverse $\tilde{C}^{-1} = (I - A)^{T} D^{-1} (I - A)$ is sparse. This leads to the latent NNGP model in the section below.

NNGP can also be viewed as as a special case of a Gaussian Markov Random Field (GMRF; Rue and Held 2005, as cited in Finley et al., 2021).

3. NNGP Models

3.0. Latent NNGP

The original NNGP model proposed by Datta et al. (2016) constructed the neighbor sets based on m nearest neighbors and replaced the GP prior for spatial random effects w in (2) with a NNGP prior

$$w \sim N(0, \tilde{C}(\theta)). \tag{6}$$

This model is referred to as the latent NNGP model, which uses a fully Bayesian hierarchical specification

$$p(\beta, \theta, \tau^2) \times N(w|0, \tilde{C}(\theta)) \times N(y|X\beta + w, \tau^2 I), \tag{7}$$

for running an MCMC (Markov chain Monte Carlo) algorithm, and the parameters $\{w, \beta, \theta, \tau^2\}$ are updated in a Gibb's sampler (Finley et al., 2021).

Normal priors for β and inverse Gamma priors for the variance components τ^2 ensure that they yield conjugate full conditionals in the Gibbs sampler (Finley et al., 2021). The remaining covariance parameters θ are updated using random-walk Metropolis steps for their respective full conditionals (Finley et al., 2021).

The full conditional distribution for w in (7) is

$$w|\cdot \sim (B(y - X\beta)/\tau^2, B),$$

where $B = \tilde{C}^{-1}(\theta) + I/\tau^2$ is the full conditional precision matrix. However, this block update of w is not practical. This is because even though B is as sparse as $\tilde{C}^{-1}(\theta)$ is, unlike $\tilde{C}^{-1}(\theta)$, the determinant of B cannot be calculated in $\mathcal{O}(n)$ FLOPs (Finley et al., 2021). Instead, the MCMC implementation of the latent NNGP model involves updating the n full conditions (or parameters) $w_i|\cdot$ *sequentially* (Finley et al., 2021). But, MCMC convergence for high-dimensional model like this is difficult to study to prove reliable (Finley et al., 2019) and can also imply slow convergence (Finley et al., 2019; Finley et al., 2021).

Alternatively, (7) is the same as

$$(\beta, \tau^2, \theta) \sim p(\beta, \tau^2, \theta)$$

$$w|\theta \sim N(0, \tilde{C}(\theta))$$

$$y|\beta, w, \tau^2 \sim N(X\beta + w, \tau^2 I).$$

Three alternate variants of the latent NNGP model are proposed. To reduce the parameter dimensionality of the latent NNGP model, these models consider marginalizing over the entire vector of spatial random effects (Finley et al., 2019).

3.1. Collapsed NNGP

A collapsed NNGP model not only enjoys the frugality of a low-dimensional MCMC chain but also allows for full recovery of the latent random effects w (Finley et al., 2019).

Consider the two-stage hierarchical specification $N(w|0, \tilde{C}(\theta)) \times N(y|X\beta + w, \tau^2 I)$ and integrate out w to avoid sampling w in the Gibb's sampler, then the collapsed NNGP model is specified as

$$y \sim N(X\beta, \Sigma = \tilde{C}(\theta) + \tau^2 I), \tag{7*}$$

where $\theta = \{\sigma^2, \phi, \nu\}$ for Matérn covariance function.

A normal prior $N(\mu_\beta, V_\beta)$ is used for β, inverse-Gamma priors are used for the spatial and noise variances σ^2 and τ^2, and uniform priors are used for the range and smoothness parameters $1/\phi$ and ν.

3.2. NNGP for the Response

Both the latent NNGP and collapsed NNGP model (the collapsed version of the latent NNGP model) make predication at a new location by first recovering the spatial random effects w and predicting value at the new location with kriging. However, if inference on the latent process is of interest, the recovery of w is necessary. Otherwise, it is often a computational burden (Finley et al., 2019).

Instead of using NNGP for the latent Gaussian process w, the response NNGP model applies the marginal Gaussian process for the response (Finley et al., 2019, as cited in Finley et al., 2021).

Consider the GP marginal model for the response

$$Y \sim N(X\beta, \Sigma),$$

where Σ is the marginalized covariance function Σ and is specified as $\Sigma(s_i, s_j) = C(s_i, s_j|\theta) + \tau^2\delta(s_i, s_j)$, where δ is the Kronecker delta (Finley et al., 2021).

Since the covariance function of an NNGP can be derived from any parent GP, next replace the full GP covariance Σ with its NNGP analogue $\tilde{\Sigma}$, then the response NNGP marginal model is given as

$$Y \sim N(X\beta, \tilde{\Sigma}), \tag{8}$$

where $\tilde{\Sigma}$ is the NNGP covariance matrix derived from $\Sigma = C(\theta) + \tau^2 I$. (Finley et al., 2021). The sparsity in Section 2.2 can be applied to $\tilde{\Sigma}^{-1}$.

The dimension of the parameter space is reduced from $\mathcal{O}(n)$ to $\mathcal{O}(1)$, and the lower dimensional NNGP tends to have improved MCMC convergence (Finley et al., 2019, as cited in Finely et al., 2021).

3.3. MCMC-Free Exact Bayesian Inference Using Conjugate NNGP

MCMC methods are commonly used to obtain approximate Bayesian inference since the normalizing constant often involves high-dimensional integrals and is therefore hard to compute (Salakhutdinov, 2010). However, running MCMC methods such as the Gibbs' sampler for several thousand iterations may still be very slow (Finley et al., 2019).

The conjugate NNGP model offers exact (MCMC-free) posterior inference by fixing certain covariance parameters (i.e., ϕ and α) in the response NNGP model (Finley et al., 2021).

Recall that Σ of the GP marginal model in Section 2.2 is given as

$$\Sigma = \Sigma(s_i, s_j) = C(s_i, s_j|\theta) + \tau^2\delta(s_i, s_j),$$

express the covariance function $C(\cdot,\cdot|\theta)$ as $\sigma^2 R(\cdot,\cdot|\phi)$, where σ^2 is the marginal variance and R is the correlation function parameterized by ϕ, i.e., $\theta = \{\sigma^2, \phi\}$, and rewrite $\tau^2 = \alpha\sigma^2$, then

$$\Sigma = \Sigma(s_i, s_j) = \sigma^2 R(s_i, s_j|\phi) + \alpha\sigma^2\delta(s_i, s_j)$$

$$= \sigma^2(R(s_i, s_j|\phi) + \alpha\delta(s_i, s_j)).$$

This implies that the MCMC-free conjugate NNGP marginal model is

$$Y \sim N(X\beta, \sigma^2\tilde{M}), \tag{9}$$

where $\tilde{M} = \tilde{M}(\phi, \alpha)$ is a known covariance matrix once ϕ and α are fixed (Finley et al., 2021). The fixed values of ϕ and α are either chosen based on a variogram or can be selected more formally using K-fold cross-validation on hold-out data (Finley et al., 2021). This leaves β and σ^2 the only unknown parameters (Finley et al., 2021).

Normal-Inverse-Gamma prior for (β, σ^2) leads to conjugate Normal-Inverse-Gamma posterior distributions, and other summary quantities of β and σ^2 can easily and exactly be obtained (Finley et al., 2021). That is, for fixed ϕ and α, the conjugate Bayesian linear regression model can be constructed as

$$IG(\sigma^2|a_\sigma, b_\sigma) \times N(\beta|\mu_\beta, \sigma^2 V_\beta) \times N(y|X\beta, \sigma^2\tilde{M})$$

with joint posterior distribution

$$p(\beta, \sigma^2|y) \propto IG(\sigma^2|a^*_\sigma, b^*_\sigma) \times N(\beta|B^{-1}b, \sigma^2 B^{-1})$$

$$= p(\sigma^2|y) \times p(\beta|\sigma^2, y),$$

where

$$a^*_\sigma = a_\sigma + n/2, \quad b^*_\sigma = b_\sigma + \frac{1}{2}(\mu_\beta^T V_\beta^{-1}\mu_\beta + y^T\tilde{M}y - b^T B^{-1} b)$$

and

$$B = V_\beta^{-1} + X^T\tilde{M}^{-1}X, \quad b = V_\beta^{-1}\mu_\beta + X^T\tilde{M}^{-1}y.$$

Marginal posterior distributions for β and σ^2 are respectively

$$\beta|y \sim MVS_{t_{2a^*_\sigma}}(B^{-1}b, \frac{b^*_\sigma}{a^*_\sigma}B^{-1})$$

and

$$\sigma^2|y \sim IG(a^*_\sigma, b^*_\sigma),$$

where $MVS_{t_\kappa}(B^{-1}b, (b/a)B^{-1})$ denotes the multivariate noncentral Student's t distribution with degrees of freedom κ, mean $B^{-1}b$ and variance $bB^{-1}/(a-1)$. The marginal posterior mean and variance for σ^2 are $b^*_\sigma/(a^*_\sigma - 1)$ and $b^{2*}_\sigma/(a^*_\sigma - 1)^2(a^*_\sigma - 2)$, respectively.

4. Applications to Spatial data

4.1. Simulated data

4.2. Real data

5. Discussion

Reference

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

Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Appendix A1 of Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. Journal of the American Statistical Association, 111(514), 800-812.

Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Appendix A2 of Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. Journal of the American Statistical Association, 111(514), 800-812.

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

Salakhutdinov, R. (2010) Approximate Inference using MCMC. MIT. https://www.mit.edu/~9.520/spring10/Classes/class21_mcmc_2010.pdf