NNGP Models with Spatial Examples of Simulated Data and...

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 paid little attention on the algorithmic details and have not 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 (latent) NNGP (nearest neighbor Gaussian process) model (Datta et al., 2016) appears promising, this model is prone to high autocorrelations and slow convergence because the "sequential" Gibbs sampler involves updating a high-dimensional vector of latent random effect.

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

Section 2 introduces the NNGP (nearest neighbor Gaussian process) models, which is followed by three alternate formulations: (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 inference. Section 3. ... Section 4 includes the discussion. ... are included in the Appendix.

2. Nearest Neighbor Gaussian Processes

Review of GPs for spatial data

A spatial linear mixed effects model is given as

$$y(s_i) = x(s_i)^T\beta + w(s_i) + \epsilon(s_i),$$

where $y(s_i)$ denotes the response, $x(s_i)$ denotes the known or observed covariates, β is the vector of coefficients, $w(s_i)$ is the vector of unknown or unobserved covariates, 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

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

where $C(\cdot, \cdot|\theta)$ is a covariance function, implies that $w = (w(s_1), ..., w(s_n))^T$ 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.

A popular choice of θ for $C(\cdot, \cdot|\theta)$ is selected from the Matérn covariance function. Let s_i and s_j be two points in $\mathcal{D}$, then the Matérn covariance function is specified 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 $\theta = \{\sigma^2, \phi, \nu\}$ are respectively the marginal variance, scale (inverse of range) and smoothness parameters, Γ is the gamma function, $||\cdot||$ denotes the Euclidean distance in $\mathbb{R}^d$, and $\mathcal{K}$ is the Bessel function of second kind.

Customary Bayesian hierarchical models are constructed 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 constricted by integrating w out from (2)

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

.....

Nearest Neighbor GPs for spatial data

When n is large, evaluating (2) is computationally challenging or infeasible.

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

$$p(w) = p(w_1) \prod_{i=2}^{n} p(w_i|Pa[i]), \tag{A1}$$

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

$$p(w) = p(w_1)p(w_2|w_1)p(w_3|w_1, w_2) \cdots p(w_n|w_1, w_2, ...w_{i-1}) \tag{A2}$$

(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

$$w_1 = 0 + \eta_1,$$

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

$$w_i = a_{i1}w_1 + a_{i2}w_2 + \cdots + a_{i,i-1}w_{i-1} + \eta_i, \text{ for } i = 2, ..., n, \tag{A3}$$

or, more explicitly, as

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & \cdots & 0 & 0 \\ a_{31} & a_{32} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn_1} & 0 \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \vdots \\ \eta_n \end{bmatrix} \tag{A4}$$

(Datta et al., 2017),

or, more compactly, as

$$w = Aw + \eta, \tag{A5}$$

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.

2.0. 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 (Gaussian Process) prior for spatial random effects w in (2) with a Nearest Neighbor Gaussian Process (NNGP) prior

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

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

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

for running an MCMC (Markov chain Monte Carlo) algorithm (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 (2*) is..... However, this block update of w is not practical (Finley et al., 2021).

The MCMC implementation of the latent NNGP involves updating the n latent spatial effects w sequentially. While....., the high-dimensional MCMC model.....

2.1. Collapsed NNGP

A collapsed NNGP model (1) enjoys the frugality of a low-dimensional MCMC chain but also (2) allows for full recovery of the latent random effects w.

Consider the two-stage hierarchical specification

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

and integrate out w to avoid sampling w in the Gibb's sampler, then the collapsed NNGP model is specified as

$$y \sim N(X\beta, \Lambda = \tilde{C} + \tau^2 I), \tag{3}$$

where $\tilde{C} = \tilde{C}(\theta)$, where again $\theta = \{\sigma^2, \phi, \nu\}$ 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.

2.2. NNGP for the Response

Both the latent NNGP (sequential NNGP) and the collapsed version of it the collapsed NNGP make predication at a new location by (1) first recovering the spatial random effects w, and (2) predicting value at the new location with kriging.

However, the recovery of w is necessary if inference on the latent process is of interest. Otherwise, it becomes a computational burden. .....

.....

Consider the marginal Gaussian process for the response

$$\{y(s)\} \sim GP(x(s)^T\beta, \Sigma(\cdot, \cdot)),$$

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 specified as

$$Y \sim N(X\beta, \tilde{\Sigma}), \tag{4}$$

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 can be applied to $\tilde{\Sigma}^{-1}$.

2.3. MCMC-Free Exact Bayesian Inference Using Conjugate NNGP

3. Applications to Spatial data

3.1. Simulated data

3.2. Real data

4.

Reference

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.