MODELING MASSIVE SPATIAL DATASETS USING A CONJUGATE BAYESIAN LINEAR REGRESSION FRAMEWORK

A PREPRINT

Sudipto Banerjee[*]
UCLA Department of Biostatistics
University of California, Los Angeles
Los Angeles, CA 90095-1772
`sudipto@ucla.edu`

February 28, 2020

ABSTRACT

Geographic Information Systems (GIS) and related technologies have generated substantial interest among statisticians with regard to scalable methodologies for analyzing large spatial datasets. A variety of scalable spatial process models have been proposed that can be easily embedded within a hierarchical modeling framework to carry out Bayesian inference. While the focus of statistical research has mostly been directed toward innovative and more complex model development, relatively limited attention has been accorded to approaches for easily implementable scalable hierarchical models for the practicing scientist or spatial analyst. This article discusses how point-referenced spatial process models can be cast as a conjugate Bayesian linear regression that can rapidly deliver inference on spatial processes. The approach allows exact sampling directly (avoids iterative algorithms such as Markov chain Monte Carlo) from the joint posterior distribution of regression parameters, the latent process and the predictive random variables, and can be easily implemented on statistical programming environments such as R.

Keywords Bayesian linear regression · Exact sampling-based inference · Gaussian process · Low-rank models · Nearest-Neighbor Gaussian Processes · Sparse models.

1 Introduction

Statistical modeling and analysis for spatial and spatial-temporal data continue to receive much attention due to enhancements in computerized Geographic Information Systems (GIS) and accompanying technologies. Bayesian hierarchical spatiotemporal process models have become widely deployed statistical tools for researchers to better understand the complex nature of spatial and temporal variability See, for example, the books [1], [2], [3], [4], [5], [6] and [7] for a variety of statistical methods in diverse applications domains.

Spatial data analysis is conveniently carried out by embedding a spatial process within the familiar hierarchical modeling paradigm,

$$[\text{data} \mid \text{process}] \times [\text{process} \mid \text{parameters}] \times [\text{parameters}]\ . \tag{1}$$

Modeling for point-referenced data, which refers to data referenced by points with coordinates (latitude-longitude, Easting-Northing etc.), proceeds from a random field that introduces dependence among any finite collection of random variables. Formally, the random field is a stochastic process defined as an uncountable set of random variables, say $\{w(\ell) : \ell \in \mathcal{L}\}$, over a domain of interest $\mathcal{L}$. This uncountable set is endowed with a probability law specifying the joint distribution for any finite subset of random variables. Spatial processes are usually constructed assuming $\mathcal{L} \subseteq \Re^d$

[*] Sudipto Banerjee is Professor and Chair of the Department of Biostatistics in the University of California, Los Angeles, USA. Mailing address: 650 Charles E. Young Drive South, CHS 51-254, Los Angeles, CA 90095-1772. URL: `www.sudipto.bol.ucla.edu`

(usually $d = 2$ or 3) or, perhaps, as a subset of points on a sphere or ellipsoid. In spatiotemporal settings $\mathcal{L} = \mathcal{S} \times \mathcal{T}$, where $\mathcal{S} \subset \Re^d$ and $\mathcal{T} \subset [0, \infty)$ are the space and time domains, respectively, and $\ell = (s, t)$ is a space-time coordinate with spatial location $s \in \mathcal{S}$ and time point $t \in \mathcal{T}$ [see, e.g., 8, for details].

Probability laws over random fields are specified with a covariance function $\text{cov}\{w(\ell), w(\ell')\} = K_\theta(\ell, \ell')$ for any two points ℓ and ℓ' in $\mathcal{L}$. If $\mathcal{U}$ and $\mathcal{V}$ are finite sets comprising n and m points in $\mathcal{L}$, respectively, then $K_\theta(\mathcal{U}, \mathcal{V})$ denotes the $n \times m$ matrix whose (i, j)-th element is evaluated using the covariance function $K_\theta(\cdot, \cdot)$ between the i-th point $\mathcal{U}$ and the j-th point in $\mathcal{V}$. If $\mathcal{U}$ or $\mathcal{V}$ comprises a single point, $K_\theta(\mathcal{U}, \mathcal{V})$ is a row or column vector, respectively. A valid spatiotemporal covariance function ensures that $K_\theta(\mathcal{U}, \mathcal{U})$ is positive definite for any finite set $\mathcal{U}$, which we will denote simply as K_θ if the context is clear. A customary specification models $\{w(\ell) : \ell \in \mathcal{L}\}$ as a zero-centered Gaussian process, denoted as $w(\ell) \sim GP(0, K_\theta(\cdot, \cdot))$. For any finite collection $\mathcal{U} = \{\ell_1, \ell_2, \ldots, \ell_n\}$ in $\mathcal{L}$, the $n \times 1$ random vector $w_{\mathcal{U}} = (w(\ell_1)), w(\ell_2), \ldots, w(\ell_n))^\top$ is distributed as $N(0, K_\theta)$, where $K_\theta = K_\theta(\mathcal{U}, \mathcal{U})$. Further details on valid spatial (and spatiotemporal) covariance functions can be found in [8], [1], [2], [5], [6] and [7] and numerous references therein.

If $y(\ell)$ represents a variable of interest at point ℓ, then a customary spatial regression model at ℓ is

$$y(\ell) = x^\top(\ell)\beta + w(\ell) + \epsilon(\ell) \; , \tag{2}$$

where $x(\ell)$ is a $p \times 1$ ($p < n$) vector of spatially referenced predictors, β is the $p \times 1$ vector of slopes, and $w(\ell) \sim GP(0, K_\theta(\cdot, \cdot))$ is the spatial or spatiotemporal process and $\epsilon(\ell)$ is a white noise process modeling measurement error or fine scale variation attributed to disturbances at distances smaller than the minimum observed separations in space and/or time. We now embed (2) and the spatial process within the Bayesian hierarchical model

$$p(\theta, \beta, \tau) \times N(w \,|\, 0, K_\theta) \times N(y \,|\, X\beta + w, D_\tau) \; , \tag{3}$$

where $y = (y(\ell_1), y(\ell_2), \ldots, y(\ell_n))^\top$ is the $n \times 1$ vector of observed outcomes, X is the $n \times p$ matrix of regressors with i-th row $x^\top(\ell_i)$ and D_τ is the covariance matrix for $\epsilon(\ell)$ over $\{\ell_1, \ell_2, \ldots, \ell_n\}$. A common specification is $D_\tau = \tau^2 I_n$, where τ^2 is called the "nugget." The hierarchy is completed by assigning prior distributions to β, θ and τ.

For fitting (3) to large spatial datasets, a substantial computational expense is incurred from the size of K_θ. Since θ is unknown, each iteration of the model fitting algorithm will involve decomposing or factorizing K_θ, which typically requires $\sim n^3$ floating point operations (flops) and order of $\sim n^2$ for memory requirements. In geostatistical settings, data are almost never observed on regular grids and the configuration of points are typically highly irregular. The covariance models that have been demonstrated to be most effective for inference do not, in general, result in any computationally exploitable structure for K_θ, which makes the matrix computations prohibitive for large values of n. For Gaussian likelihoods, one can integrate out the random effects w from (3) and work with the posterior

$$p(\theta, \beta, \tau \,|\, y) \propto p(\theta, \beta, \tau) \times N(y \,|\, X\beta, K_\theta + D_\tau) \; . \tag{4}$$

This reduces the parameter space to $\{\tau^2, \theta, \beta\}$ by excluding the high-dimensional vector w, but one still needs to work with $K_\theta + D_\tau$, which is $n \times n$. These are referred to as "big-n" or "high-dimensional" problems in geostatistics.

There is already a substantial literature on high-dimensional spatial and spatiotemporal modeling and we do not attempt to undertake a comprehensive review here; see, e.g., [9] for a focused review on some popular Bayesian approaches and [10] for a comparative evaluation for a variety of contemporary statistical methods. These papers, and the references therein, offer a variety of algorithmic and model-based approaches for large data. Some published methods have scalable implementations into the millions [see, e.g., 11, 12, 13, 14, 15] but often require specialized high-performance computer architectures and libraries harnessing parallel processing or graphical processing units. Also, uncertainty quantification on the spatial process while maintaining fidelity to the underlying probability model may also be challenging. With the advent of a new generation of data products, there is a need for some simpler implementations that can be run on modest computing architectures by practicing spatial analysts. This requires new directions in thinking about high-dimensional spatial problems. Here, we will show how some elementary conjugate Bayesian linear regression models can be exploited to conduct Bayesian analysis for massive spatial datasets. While a common underlying idea is to approximate the underlying spatial process with a scalable alternative, we will ensure that such approximations will result in well-defined probability models. In this sense, these approaches can be described as model-based solutions for very large spatial datasets that can be executed on modest computing environments. One exception to the fully model-based approach will be a divide and conquer approach that we briefly review, where an approximation to the full posterior distribution for the entire data is constructed from several posterior distributions of smaller subsets of the data.

The balance of the paper proceeds as follows. The next section briefly reviews dimension reduction and sparsity inducing spatial models. Section 3 presents some standard distribution theory for Bayesian linear regression and outlines how scalable spatial process models can be cast into such frameworks. A synposis of some simulation experiments and data analysis examples are provided in Section 4. Section 5 presents an alternative approach based upon dividing and conquering the data, known as meta-kriging. The paper concludes with some further discussion in Section 6.

2 Dimension reduction and sparsity

Dimension reduction [16] is among the most conspicuous of approaches for handling large spatial datasets. This customarily proceeds by representing or approximating the spatial process in terms of the realizations of a latent process over a smaller set of points, often referred to as *knots*. Thus,

$$w(\ell) \approx \tilde{w}(\ell) = \sum_{j=1}^{r} b_\theta(\ell, \ell_j^*) z(\ell_j^*) = b_\theta^\top(\ell) z, \tag{5}$$

where $z(\ell)$ is a well-defined (usually unobserved) process and $b_\theta(\cdot,\cdot)$ is a family of basis functions or kernels, possibly depending upon some parameters θ. The collection of r locations $\{\ell_1^*, \ell_2^*, \ldots, \ell_r^*\}$ are the knots, $b_\theta(\ell)$ and z are $r \times 1$ vectors with components $b_\theta(\ell, \ell_j^*)$ and $z(\ell_j^*)$, respectively. Therefore, $\tilde{w} = B_\theta z$, where $\tilde{w} = (\tilde{w}(\ell_1), \tilde{w}(\ell_2), \ldots, \tilde{w}(\ell_n))^\top$ and B_θ is $n \times r$ with (i,j)-th element $b_\theta(\ell_i, \ell_j^*)$. We work with r (instead of n) $z(\ell_j^*)$'s and the $n \times r$ matrix B_θ. Choosing $r << n$ effectuates dimension reduction because $\tilde{w}(\ell)$, as defined in (5), spans only an r-dimensional space. When $n > r$, the joint distribution of $\tilde{w}$ is singular. Nevertheless, we construct a valid stochastic process with covariance function

$$\text{cov}(\tilde{w}(\ell), \tilde{w}(\ell')) = b_\theta^\top(\ell) V_z b_\theta(\ell') \ , \tag{6}$$

where V_z is the variance-covariance matrix (also depends upon parameter θ) for z. From (6), we see that, even if $b_\theta(\cdot,\cdot)$ is stationary, the induced covariance function is not. If the z's are Gaussian, then $\tilde{w}(\ell)$ is a Gaussian process. Every choice of basis functions yields a process and there are too many choices to enumerate here. Wikle [17] offers an excellent overview of low rank models.

Some choices of basis functions can be more computationally efficient than others depending upon the specific application. For example, [18] (also see [19]) discuss "Fixed Rank Kriging" (FRK) by constructing B_θ using very flexible families of non-stationary covariance functions to carry out high-dimensional kriging, [20] extend FRK to spatiotemporal settings calling the procedure "Fixed Rank Filtering" (FRF), [21] provide efficient constructions for B_θ for massive spatiotemporal datasets, and [22] uses spatial basis functions to capture medium to long range dependence and tapers the residual $w(\ell) - \tilde{w}(\ell)$ to capture fine scale dependence. Multiresolution basis functions [23, 24] have been shown to be effective in building computationally efficient nonstationary models. These papers amply demonstrate the versatility of low-rank approaches using different basis functions. An alternative approach specifies $z(\ell)$ itself as a spatial process. This process is called the "parent process" and one can derive a low-rank process $\tilde{w}(\ell)$ from the parent. One such derivation emerges from truncating the Karhunen-Loève (infinite) basis expansion for a Gaussian process to a finite number of terms to obtain a low-rank process [see, e.g., 25, 7]. This is equivalent to projecting the parent process on a lower-dimensional subspace determined by a partial realization of the parent over r knots of the process. This yields the *predictive process* and several variants aimed at improving the approximation [26, 27, 28, 29, 11]; also see [30] and [9] for computational details on efficiently implementing Gaussian predictive processes.

While dimension reduction methods have been applied extensively and effectively to analyze spatial data sets in the order of $n \sim 10^4$, their computational efficiency and inferential performance tend to struggle at even larger scales [9]. More recently, there has been substantial developments in full rank models that exploit sparsity. We introduce sparsity either in the covariance matrix or its inverse (the precision matrix). Covariance tapering [31, 32, 33] is in the spirit of the former by modeling var$\{w\} = K_\theta \odot K_{\text{tap},\nu}$, where $K_{\text{tap},\nu}$ is a sparse covariance matrix formed from a compactly supported, or *tapered*, covariance function with tapering parameter ν and $\odot$ denotes the element wise (or Hadamard) product of two matrices. The Hadamard product retains positive definiteness, so $K_\theta \odot K_{\text{tap},\nu}$ is positive definite. Furthermore, $K_{\text{tap},\nu}$ is sparse because a tapered covariance function is equal to 0 for all pairs of locations separated by a distance beyond a threshold ν. Covariance tapering is undoubtedly an attractive approach for constructing sparse covariance matrices, but its practical implementation for full Bayesian inference will generally require efficient sparse Cholesky decompositions, numerically stable determinant computations and, perhaps most importantly, effective memory management. These issues are yet to be tested for truly massive spatiotemporal datasets with $n \sim 10^5$ or more.

One could also devise models with sparse precision matrices. For finite-dimensional distributions conditional and simultaneous autoregressive (CAR and SAR) models [see, e.g., 1, 7, and references therein] adopt this approach for areally referenced datasets. The CAR models are special instances of Gaussian Markov random fields or GMRFs [34] that have led to the popular quadrature based Integrated Nested Laplace Approximation (INLA) algorithms [35] for

Bayesian inference and to the approximation of Gaussian processes[36]. These approaches can be computationally efficient for certain classes of covariance functions with stochastic partial differential equations (SPDE) representations (including the versatile Matérn class), but their inferential performance on spatiotemporal or multivariate Gaussian processes (perhaps specified through more general covariance or cross-covariance functions) embedded within Bayesian hierarchical models is yet to be fully developed or assessed for massive datasets.

One could also construct massively scalable sparsity-inducing Gaussian processes using essentially the techniques used in graphical Gaussian models by exploiting the relationship between the Cholesky decomposition of a positive definite matrix and conditional independence. For Gaussian processes in particular, recent developments on the Nearest Neighbor Gaussian Processes (NNGP) [37, 38, 9, 14] have proceeded from GP likelihoods using directed acyclic graphs (or DAGs) as used by Vecchia [39] and Stein et al.[40]. The NNGP is a Gaussian process whose finite-dimensional realizations will have sparse precision matrices. Other related papers using the approximation in [39] include [41], [42], [43], and [44]. Shi et al. [45] recently used the NNGP for uncertainty quantification and Ma et al. [46] used it as a part of a rich class of fused Gaussian process models.

Full Bayesian inference for low-rank and sparse Gaussian process models require iterative algorithms such as Markov chain Monte Carlo (MCMC) or INLA. Details of these implementations can be found in the aforementioned references. In the following section, we will discuss how these spatial models can be embedded within a Bayesian linear regression framework and provide some practical strategies for inference based upon direct (exact) sampling from the posterior distribution.

3 Conjugate Bayesian models for massive datasets

3.1 Conjugate Bayesian linear geostatistical models

A conjugate Bayesian linear regression model is written as

$$y \,|\, \beta, \sigma^2 \sim N(X\beta, \sigma^2 V_y) \; ; \quad \beta \,|\, \sigma^2 \sim N(\beta \,|\, \mu_\beta, \sigma^2 V_\beta) \; ; \quad \sigma^2 \sim IG(a_\sigma, b_\sigma) \; , \tag{7}$$

where y is an $n \times 1$ vector of observations of the dependent variable, X is an $n \times p$ matrix (assumed to be of rank p) of independent variables (covariates or predictors) and its first column is usually taken to be the intercept, V_y is a fixed (i.e., known) $n \times n$ positive definite matrix, μ_β, V_β, a_σ and b_σ are assumed to be fixed hyper-parameters specifying the prior distributions on the regression slopes β and the scale σ^2. This model is easily tractable and the posterior distribution is

$$p(\beta, \sigma^2 \,|\, y) = \underbrace{IG(\sigma^2 \,|\, a_\sigma^*, b_\sigma^*)}_{p(\sigma^2 \,|\, y)} \times \underbrace{N(\beta \,|\, Mm, \sigma^2 M)}_{p(\beta \,|\, \sigma^2, y)} \; , \tag{8}$$

where $a_\sigma^* = a_\sigma + n/2$, $b_\sigma^* = b_\sigma + (1/2)\left\{\mu_\beta^\top V_\beta^{-1}\mu_\beta + y^\top V_y^{-1} y - m^\top M m\right\}$, $M^{-1} = V_\beta^{-1} + X^\top V_y^{-1} X$ and $m = V_\beta^{-1}\mu_\beta + X^\top V_y^{-1} y$. Sampling from the joint posterior distribution of $\{\beta, \sigma^2\}$ is achieved by first sampling $\sigma^2 \sim IG(a_\sigma^*, b_\sigma^*)$ and then sampling $\beta \sim N(Mm, \sigma^2 M)$ for each sampled σ^2. This yields marginal posterior samples from $p(\beta \,|\, y)$, which is a non-central multivariate t distribution but we do not need to work with its complicated density function. See [47] for further details on the conjugate Bayesian linear regression model and sampling from its posterior.

We will adapt (7) to accommodate (3) or (4). Let us first consider (4) with the customary specification $D_\tau = \tau^2 I$ and let $K_\theta = \sigma^2 R(\phi)$, where $R(\phi)$ is a correlation matrix whose entries are given by a correlation function $\rho(\phi;\; \ell_i, \ell_j)$. Thus, $\theta = \{\sigma^2, \phi\}$, where σ^2 is the spatial variance component and ϕ is a spatial decay parameter controlling the rate at which the spatial correlation decays with separation between points. A simple example is $\rho(\phi;\; \ell_i, \ell_j) = \exp(-\phi\|\ell_i - \ell_j\|)$, although much richer choices are available [see, e.g., Ch 3 in 7]. Therefore, we can write $K_\theta = \sigma^2 V_y$, where $V_y = R(\phi) + \delta^2 I$ and $\delta^2 = \tau^2/\sigma^2$ is the ratio between the "noise" variance and "spatial" variance. If we assume that ϕ and δ^2 are fixed and that the prior on $\{\beta, \sigma^2\}$ are as in (7), then we have reduced (4) to (7) and direct sampling from its posterior is easily achieved as described below (8). We will return to the issue of fixing $\{\phi, \delta^2\}$ shortly.

Let us turn to accommodating (3) within (7), which would include directly sampling the spatial random effects w from their marginal posterior $p(w \,|\, y)$. Here, it is instructive to write the joint distribution of y and w in (3) as a linear model,

$$\underbrace{\begin{bmatrix} y \\ \mu_\beta \\ 0 \end{bmatrix}}_{y_*} = \underbrace{\begin{bmatrix} X & I_n \\ I_p & O \\ O & I_n \end{bmatrix}}_{X_*} \underbrace{\begin{bmatrix} \beta \\ w \end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}}_{\eta}, \; , \tag{9}$$

where $\eta \sim N(0, \sigma^2 V_{y_*})$ and $V_{y_*} = \begin{bmatrix} \delta^2 I_n & O & O \\ O & V_\beta & O \\ O & O & R(\phi) \end{bmatrix}$. If we assume that δ^2 and ϕ are fixed at known values, then V_{y_*} is fixed. We have a conjugate Bayesian linear regression model $y_* = X_* \gamma + \eta$, where γ has a flat prior and $\sigma^2 \sim IG(a_\sigma, b_\sigma)$. Thus,

$$p(\gamma, \sigma^2 \,|\, y) = \underbrace{IG(\sigma^2 \,|\, a_\sigma^*, b_\sigma^*)}_{p(\sigma^2 \,|\, y)} \times \underbrace{N(\beta \,|\, M_* m_*, \sigma^2 M_*)}_{p(\gamma \,|\, \sigma^2, y)} \,, \tag{10}$$

where $a_\sigma^* = a_\sigma + (2n+p)/2$, $b_\sigma^* = b_\sigma + (1/2)\left\{y_*^\top V_{y_*}^{-1} y_* - m_*^\top M_* m_*\right\}$, $M_*^{-1} = X_*^\top V_{y_*}^{-1} X_*$ and $m_* = X_*^\top V_{y_*}^{-1} y_*$. The posterior mean of γ is $\hat{\gamma} = M_* m_* = \left(X_*^\top V_{y_*}^{-1} X_*\right)^{-1} X_*^\top V_{y_*}^{-1} y_*$, which is the generalized least squares estimate obtained from the augmented linear system in (9). Sampling from the posterior proceeds analogous to that described below (8).

From the preceding account we see that fixing the spatial range decay parameter ϕ and the noise-to-spatial variance ratio δ^2 casts the Bayesian geostatistical model into a conjugate framework that will allow inference on $\{\beta, w, \sigma^2\}$. Note that multiplying the posterior samples of σ^2 by the fixed quantity δ^2 fetches us the posterior samples of τ^2. Therefore, we neglect uncertainty in ϕ and, partially, for one of the variance components due to fixing their ratio. This, however, provides the computational advantage that inference can be carried out without resorting to expensive iterative algorithms such as MCMC that require several iterations before sampling from the posterior distribution. This computational benefit becomes especially relevant when handling massive spatial data. Furthermore, fixing the values of δ^2 and ϕ is not entirely unreasonable given that these parameters are weakly identified by the data [48] and difficult to learn from the posterior. Nevertheless, the inference will depend upon these fixed parameters so we discuss a practical approach to fix ϕ and δ^2 at reasonable values.

3.2 Choosing ϕ and δ^2

We can set values for ϕ and δ^2 by conducting some simple spatial exploratory data analysis using the "variogram". Several practical algorithms exist for empirically calculating the variogram (or semivariogram) from observations using finite sample moments. Many of these methods for variograms are now offered in user-friendly R packages hosted by the Comprehensive R Archive Network (CRAN) (`https://cran.r-project.org`). As one example, Finley et al. [14] investigate the impact of tree cover and occurrence of forest fires on forest height. They first fit an ordinary linear regression of the form $y_{FH} = \beta_0 + \beta_1 x_{\text{tree}} + \beta_2 x_{\text{fire}} + \epsilon$ and then compute a variogram for the residuals from the ordinary linear regression.

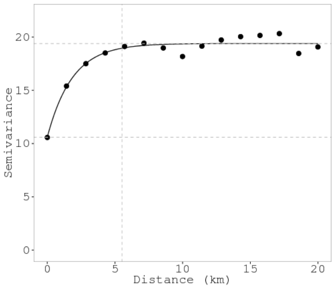


Figure 1: Variogram of the residuals from non-spatial regression indicates strong spatial pattern

Figure 1 depicts the variogram, which informs about the process parameters. The lower horizontal line represents the "nugget" or the micro-scale variation captured by the measurement error variance component τ^2. The top horizontal line represents the "sill" (or ceiling) which is the total variation captured by $\sigma^2 + \tau^2$. Therefore, the difference between the two horizontal lines is called the "partial sill" and is captured by σ^2. Finally, the vertical line represents the distance beyond which the variogram flattens or the covariance tends to zero. One can provide "eye-ball" estimates for these quantities and, in particular, fix the values of ϕ and $\delta^2 = \tau^2/\sigma^2$. Fixing these values from the variogram yields the desired highly accessible conjugate framework and the models can be estimated without resorting to Markov chain Monte Carlo (MCMC) as described earlier. Note that instead of $\{\phi, \delta^2\}$, we could also have fixed ϕ and any one of the variance components, σ^2 or τ^2, which would also yield a conjugate model with exact distribution theory. The one slight advantage of fixing δ^2 is that we will get the posterior samples of both σ^2 and τ^2, the latter obtained simply as $\sigma^2\delta^2$.

The above crude estimates can be improved using a K-fold cross-validation. We split the data randomly into K different folds. Let $S[k]$ be the k-th folder of observed points and let $S[-k]$ denote the observed points outside of $S[k]$. For each k, we compute the predictive mean $E[y(S[k]) \,|\, y(S[-k])]$. We then compute the "Root Mean Square Predictive Error" (RMSPE) [49] and choose the value of $\{\phi, \delta^2\}$ corresponding to the smallest RMSPE from a grid of candidate values. The range of the grid is based on interpretation of the hyper-parameters. We suggest a reasonably wide range for δ^2 (e.g., $[0.001, 1000]$), which accommodates one variance component substantially dominating the other in either direction. For the spatial decay ϕ we suggest a lower bound of $\frac{3}{\text{maximum inter-site distance}}$, which, based upon the exponential covariance function, indicates that the spatial correlation drops below 0.05 at the maximum inter-site distance, and an upper bound that can be initially set as 100 times of the lower bound. Functions like `variofit` in the R package `geoR` [50] can provide empirical estimates for $\{\phi, \delta^2\}$ from an empirical variogram. After initial fitting, we can shrink the range and refine the grid of the candidate values for more precise estimators.

3.3 Conjugate Bayesian geostatistical models for massive spatial data

Conjugate models can be estimated by sampling directly from their joint posterior density and, therefore, completely obviate problems associated with MCMC convergence. This is a major computational benefit. However, the challenges in analyzing massive spatial data do not quite end here. When the number of spatial locations providing measurements are in the order of millions as in [14], then the matrices K_θ, V_y or V_{y_*} that we encountered earlier in different model parametrizations will be too massive to be efficiently loaded on to the machine's memory, let alone be computed with. This precludes efficient likelihood computations and has led several researchers to propose models specifically adapted for spatial analysis. We briefly present adaptations of (9) using two different classes of models for massive spatial data: (i) low-rank process models and (ii) NNGP models.

As discussed in Section 2, in low rank models the $n \times 1$ spatial effect w in (3) is replaced by $B_\theta z$, where B_θ is the $n \times r$ matrix whose i-th row is $b_\theta^\top(\ell_i)$. Dimension reduction is achieved by fixing r to be much smaller than n so that we only deal with r random effects instead of n. The framework in (9) can be easily adapted to this situation as below:

$$\underbrace{\begin{bmatrix} y \\ \mu_\beta \\ 0 \end{bmatrix}}_{y_*} \underset{=}{=} \underbrace{\begin{bmatrix} X & B_\theta \\ I_p & O \\ O & I_r \end{bmatrix}}_{X_*} \underbrace{\begin{bmatrix} \beta \\ z \end{bmatrix}}_{\gamma} \underset{+}{+} \underbrace{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}}_{\eta}, \quad , \tag{11}$$

where $\eta \sim N(0, \sigma^2 V_{y_*})$ and $V_{y_*} = \begin{bmatrix} \delta^2 I_n & O & O \\ O & V_\beta & O \\ O & O & V_z \end{bmatrix}$ is $(n+p+r) \times (n+p+r)$ and fixed, and V_z is now $r \times r$ instead of the $n \times n$ matrix $R(\phi)$ in (9). Computations for (11) proceed analogous to those for (8), but benefits accrue in terms of storage and the number of floating point operations (flops) when conducting the exact conjugate Bayesian analysis for this model.

We now outline the construction of sparse NNGP models. These can be regarded as a special case of Gaussian Markov Random Fields (GMRFs) with a neighborhood structure specified using a directed acyclic graph (DAG). The computational benefits for NNGP models accrue from the ease of inverting sparse matrices. This is immediate from noting that the expense to obtain $V_{y_*}^{-1}$ in (10) is dominated by $R(\phi)^{-1}$. Therefore, if $R(\phi)^{-1}$ is easily available then the inference for $\gamma = \{\beta, w\}$ will be inexpensive. Modeling sparse $R(\phi)^{-1}$ can be easily achieved as follows. Writing $N(w \,|\, 0, \sigma^2 R_\phi)$ as $p(w_1)\prod_{i=2}^{n} p(w_i \,|\, w_1, w_2, \ldots, w_{i-1})$ is equivalent to the following set of linear models,

$$w_1 = 0 + \eta_1 \quad \text{and} \quad w_i = a_{i1}w_1 + a_{i2}w_2 + \cdots + a_{i,i-1}w_{i-1} + \eta_i \;\text{ for } i = 2, \ldots, n \;,$$

or, more compactly, simply $w = Aw + \eta$, where A is $n \times n$ strictly lower-triangular with elements $a_{ij} = 0$ whenever $j \geq i$ and $\eta \sim N(0, D)$ and D is diagonal with diagonal entries $d_{11} = \text{var}\{w_1\}$ and $d_{ii} = \text{var}\{w_i \,|\, w_j : j < i\}$ for $i = 2, \ldots, n$. From the structure of A it is evident that $I - A$ is unit lower-triangular, hence nonsingular, and $R_\phi = (I - A)^{-1} D (I - A)^{-\top}$.

We now introduce sparsity in $R_\phi^{-1} = (I - A)^\top D (I - A)$ by letting $a_{ij} = 0$ whenever $j \geq i$ (since A is strictly lower-triangular) and also whenever ℓ_j is not among the m nearest neighbors of ℓ_i, where m is fixed by the user to be a small number. It turns out that a very effective approximation emerges by recognizing that the lower-triangular elements of A are precisely the coefficients of a linear combination of $w(\ell_j)$'s equating to the conditional expectation $\text{E}[w(\ell_i) \,|\, \{w(\ell_j) : j < i\}]$. Thus, the $m \times 1$ vector $\tilde{a}_i$ of non-zero entries in the i-th row of A are obtained by solving the $m \times m$ linear system $\tilde{R}_{\phi,N_i,N_i}\tilde{a}_i = R_{\phi,N_i,i}$, where $\tilde{R}_{\phi,N_i,N_i}$ is the $m \times m$ principal submatrix extracted from R_ϕ corresponding to the m neighbors of i (indexed by elements of a neighbor set N_i) and $R_{\phi,N_i,i}$ is the $m \times 1$ vector extracted by choosing the m indices in N_i from the i-th column of R_ϕ. Once $\tilde{a}_i$ is obtained, the i-th diagonal entry of D is obtained as $d_{ii} = R_\phi[i,i] - \tilde{a}_i^\top R_{\phi,N_i,i}$. These computations need to be carried out for each $i = 2,\ldots,n$ (note that for $i=1$, $d_{11} = \sigma^2$ and $a_{11} = 0$), but m can be kept very small (say 5 or 10 even if n 10^7) so that the expense is $O(nm^3)$ and still feasible. The details can be found in [9]. This notion is familiar in Gaussian Graphical models and have been used in [39] and, more recently, in [37] and [14] to tackle massive amounts of spatial locations.

The framework in (9) now assumes the form

$$\underbrace{\begin{bmatrix} y \\ \mu_\beta \\ 0 \end{bmatrix}}_{y_*} = \underbrace{\begin{bmatrix} X & I_n \\ I_p & O \\ O & D^{-1/2}(I-A) \end{bmatrix}}_{X_*} \underbrace{\begin{bmatrix} \beta \\ w \end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}}_{\eta}, \quad , \tag{12}$$

where $\eta \sim N(0, \sigma^2 V_{y_*})$ and $V_{y_*} = \begin{bmatrix} \delta^2 I_n & O & O \\ O & V_\beta & O \\ O & O & I_n \end{bmatrix}$ is $(2n+p) \times (2n+p)$ and fixed with much greater sparsity. While this approach can also be subsumed into the framework of (9), its efficient implementation on standard computing architectures needs careful consideration and involves solving a large linear system with $(n+p) \times (n+p)$ coefficient matrix $X_*^\top X_*$. This matrix is large, but is sparse because of $(I-A)^\top D^{-1}(I-A)$. Since $(I-A)$ has at most $m+1$ nonzero entries in each row, an upper bound of nonzero entries in $(I-A)$ is $n(m+1)$ and, therefore, the upper bound in $(I-A)^\top D^{-1}(I-A)$ is $n(m+1)^2$. This sparsity can be exploited by sparse linear solvers such as conjugate gradient methods that can be implemented on modest computing environments.

Sampling from the joint posterior distribution $p(\gamma, \sigma^2 \,|\, y_*)$ is achieved in the following manner. First, the least-squares estimate $\hat{\gamma}$ is obtained using a sparse least-square solver using a preconditioned conjugate gradient algorithm. Subsequently, σ^2 is sampled from its marginal posterior density $IG(a_*, b_*)$, where $a_* = a_\sigma + n/2$ and $b_* = b_\sigma + (1/2)(y_* - X_*\hat{\gamma})^\top (y_* - X_*\hat{\gamma})$, and then for each sampled σ^2, γ is sampled from $N\left(\hat{\gamma}, \sigma^2 \left(X_*^\top V_{y_*}^{-1} X_*\right)^{-1}\right)$. In general, solving $X_*^\top X_* \hat{\gamma} = X_*^\top y_*$ requires $\mathcal{O}(\frac{1}{3}(n+p)^3)$ flops, but when $p \ll n$, the structure of X_* and $X_*^\top X_*$ ensures memory requirements in the order of $n(m+1)^2$ and the computational complexity in the order of $nm + n(m+1)^2$ flops. Details on such implementations on modest computing platforms can be found in [15].

3.4 Spatial prediction

Let $\tilde{\mathcal{L}} = \{\tilde{\ell}_1, \tilde{\ell}_2, \ldots, \tilde{\ell}_{\tilde{n}}\}$ be a set of $\tilde{n}$ locations where we wish to predict the outcome $y(\ell)$. Let $\tilde{Y}$ be an $\tilde{n} \times 1$ vector with i-th element $\tilde{Y}(\tilde{\ell}_i)$ and let $\tilde{w}$ be the $\tilde{n} \times 1$ vector with elements $w(\tilde{\ell}_i)$. The predictive model augments the joint distribution $p(\theta, w, \beta, \tau, y)$ to

$$p(\theta, \tau, \beta, w, y, \tilde{w}, \tilde{Y}) = p(\theta, \tau, \beta) \times p(w \,|\, \theta) \times p(\tilde{w} \,|\, w, \theta) \times p(y \,|\, \beta, w, \tau) \times p(\tilde{Y} \,|\, \beta, \tilde{w}, \tau)\ . \tag{13}$$

The factorization in (13) also implies that $\tilde{Y}$ and w are conditionally independent of each other given $\tilde{w}$ and β. Predictive inference for spatial data evaluates the posterior predictive distribution $p(\tilde{Y}, \tilde{w} \,|\, y)$. This is the joint posterior distribution for the outcomes and the spatial effects at locations in $\tilde{\mathcal{L}}$. This distribution is easily derived from (13) as

$$p(\tilde{Y}, \tilde{w}, \beta, w, \theta, \tau \,|\, y) \propto p(\beta, w, \theta, \tau \,|\, y) \times p(\tilde{w} \,|\, w, \theta) \times p(\tilde{Y} \,|\, \beta, \tilde{w}, \tau)\ . \tag{14}$$

Sampling from (14) is achieved by first sampling $\{\beta, w, \theta, \tau\}$ from $p(\beta, w, \theta, \tau \,|\, y)$. For each drawn sample, we make one draw of the $\tilde{n} \times 1$ vector $\tilde{w}$ from $p(\tilde{w} \,|\, w, \theta)$ and then, using this sampled $\tilde{w}$, we make one draw of $\tilde{Y}$ from $p(\tilde{Y} \,|\, \beta, \tilde{w}, \tau)$. The resulting samples of $\tilde{w}$ and $\tilde{Y}$ will be draws from the desired posterior predictive distribution $p(\tilde{w}, \tilde{Y} \,|\, y)$. This delivers inference on both the latent spatial random effect $\tilde{w}$ and the outcome $\tilde{Y}$ at arbitrary locations since $\mathcal{L}$ can be any finite collection of samples. Summarizing these distributions by computing their sample means, standard errors, and the 2.5-th and 97.5-th quantiles (to produce a 95% credible interval) yields point estimates with associated uncertainty quantification.

It is instructive to see how the entire inference for Gaussian outcomes can be cast into an augmented linear regression model. The predictive model for $\tilde{Y}$ can be written as a spatial regression

$$\tilde{Y} = \tilde{X}\beta + \tilde{w} + \tilde{\epsilon}\ ; \quad \tilde{w} = Cw + \omega\ , \tag{15}$$

where $\tilde{X}$ is the $\tilde{n} \times p$ matrix of predictors observed at locations in $\tilde{\mathcal{L}}$ and $\tilde{\epsilon} \sim N(0, \tilde{D}_\tau)$, where $\tilde{\epsilon}$ is the $\tilde{n} \times 1$ vector with elements $\epsilon(\tilde{\ell}_i)$. The second equation in (15) expresses the relationship between the spatial effects $\tilde{w}$ across the unobserved locations in $\tilde{\mathcal{L}}$ and the spatial effects across the observed locations in $\mathcal{L}$. Since there is one underlying random field over the entire domain, the covariance function for the random field specifies the $\tilde{n} \times n$ coefficient matrix C. In particular, if $w \sim N(0, K_\theta)$, then $C = K_\theta(\tilde{\mathcal{L}}, \mathcal{L})K_\theta^{-1}$ and $\omega \sim N(0, F_\theta)$, where $F_\theta = K_\theta(\tilde{\mathcal{L}}, \tilde{\mathcal{L}}) - K_\theta(\tilde{\mathcal{L}}, \mathcal{L})K_\theta^{-1}K_\theta(\mathcal{L}, \tilde{\mathcal{L}})$. The model for the data and the predictions is combined into

$$\underbrace{\begin{bmatrix} y \\ \mu_\beta \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{y_*} = \underbrace{\begin{bmatrix} X & I_n & O & O \\ I_p & O & O & O \\ O & C & -I_{\tilde{n}} & O \\ \tilde{X} & O & I_{\tilde{n}} & -I_{\tilde{n}} \end{bmatrix}}_{X_*} \underbrace{\begin{bmatrix} \beta \\ w \\ \tilde{w} \\ \tilde{Y} \end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \end{bmatrix}}_{\eta}, \quad , \tag{16}$$

where $\eta \sim N\left(0, \begin{bmatrix} D_\tau & O & O & O & O \\ O & V_\beta & O & O & O \\ O & O & K_\theta & O & O \\ O & O & O & F_\theta & O \\ O & O & O & O & \tilde{D}_\tau \end{bmatrix}\right)$. If locations where predictions are sought are fixed by study design, then fitting (16) using the Bayesian conjugate framework can be beneficial. On the other hand, one can first estimate $\{\beta, w, \sigma^2\}$ and store samples from their posterior distribution. Then, for any arbitrary set of points in $\tilde{\mathcal{L}}$, for each stored sample of the parameters we draw one sample of $\tilde{w} \sim N(Cw, F_\theta)$ followed by one draw of $\tilde{Y} \sim N(\tilde{X}\beta + \tilde{w}, \tilde{D}_\tau)$. The resulting $\{\tilde{w}, \tilde{Y}\}$ will be the desired posterior predictive samples for the latent spatial process and the unobserved outcomes. Again, the advantage of this formulation is that an efficient least squares algorithm to solve (16) that can exploit the sparsity of the design matrix X_* will immediately deliver inference on the regression slopes (β), the spatial process (w) at observed points, the interpolated process ($\tilde{w}$) at unobserved points, and the predicted response ($\tilde{Y}$) all at once.

4 Illustrative examples

We present a part of some simulation experiments conducted in [15], where we generated data using the spatial regression model in (2) over a set of $n = 1200$ spatial locations within a unit square and using an exponential covariance function to specify the spatial process. While 1200 spatial locations may seem too modest, we use this to draw comparisons with a full GP model that will be too expensive for large datasets. The model included an intercept and a single predictor generated from a standard normal distribution.

We fit a full Gaussian process based model (labeled as full GP in Table 1) using the `spBayes` package in `R`, a latent NNGP model with $m = 10$ neighbors using the sequential MCMC algorithm described in [37] (using the `spNNGP` package), and the conjugate latent NNGP model described in the preceding section with $m = 10$ neighbors. We will refer to the latent NNGP model fitted using MCMC (with all process parameters unknown) as simply the NNGP or latent NNGP model, while we will explicitly use "conjugate" to describe the conjugate latent NNGP model.

These models were trained using $n = 1000$ observations, while the remaining 200 observations were withheld to assess predictive performance. The fixed parameters$\{\phi, \delta^2\}$ for the conjugate latent NNGP model were picked through the K-fold cross-validation algorithm described in Section 3.2. The intercept and slope parameters in β were assigned improper flat priors and an $IG(2, b)$ (mean b) prior was used for σ^2. For the latent NNGP and full GP models, the spatial decay ϕ was modeled using a fairly wide uniform prior $U(2.2, 220)$ prior and Inverse-Gamma priors $IG(2, b)$ (mean b) were used for the nugget (τ^2) and the partial sill (σ^2) in order to compare the conjugate Bayesian models with other models. The shape parameter was fixed at 2 and the scale parameter was set from the empirical estimate provided by the variogram using the `geoR` package [50]. The parameter estimates and performance metrics are provided in Table 1. Table 1 presents parameter estimates and performance metrics for the candidate models. The inference for β is almost indistinguishable across the three models. The full GP and the NNGP fully estimate $\{\sigma^2, \tau^2, \phi\}$ using MCMC and yield very similar results. The conjugate NNGP does not estimate ϕ and estimates $\{\sigma^2, \tau^2\}$ subject to the constraint that their ratio δ^2 is fixed. This results, expectedly, in slightly narrower credible intervals for σ^2 and τ^2. Overall, the parameter estimates are very comparable across the models.

Turning to model comparisons, Zhang et al.[15] computed the posterior distribution of the Kullback-Leibler divergence (KL-D) by computing it between each candidate model and the full GP for each posterior sample. The KL-D values

Table 1: Simulation study summary table: posterior mean (2.5%, 97.5%) percentiles

	True	Full GP	NNGP	Conj NNGP
β_0	1	1.07(0.72, 1.42)	1.10 (0.74, 1.43)	1.06 (0.76, 1.46)
β_1	-5	-4.97 (-5.02, -4.91)	-4.97 (-5.02, -4.91)	-4.97 (-5.02, -4.91)
σ^2	2	1.94 (1.63, 2.42)	1.95 (1.63, 2.41)	1.94 (1.77, 2.12)
τ^2	0.2	0.14 (0.07, 0.23)	0.15 (0.06, 0.24)	0.17 (0.16, 0.19)
ϕ	16	19.00 (13.92, 23.66)	18.53 (14.12, 24.17)	17.65
KL-D	–	4.45(1.16, 9.95)	5.13(1.66, 11.39)	3.58(1.27, 8.56)
MSE(w)	–	297.45(231.62, 444.79)	303.38(228.18, 429.54)	313.28 (258.96, 483.75)
RMSPE	–	0.94	0.94	0.94
time(s)	–	2499 + 23147	109.5	12 + 0.6

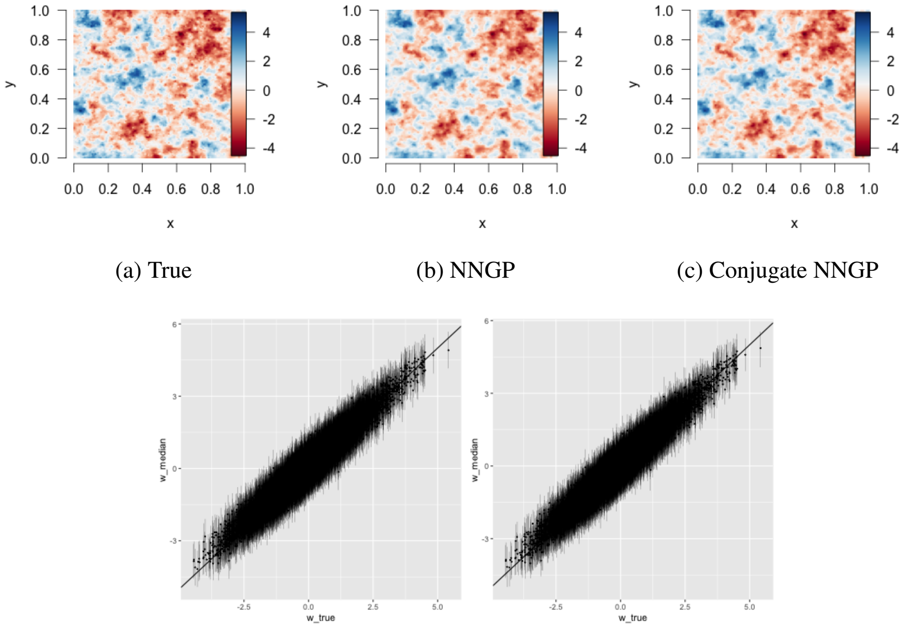
presented in Table 1 show no significant differences between the three models in their separation from the true full GP model. The root mean-squared prediction error (RMSPE) values (computed from the hold-out set of 200 locations) across all three models are also similar, further corroborating the comparable predictive performance of the conjugate model with the full Gaussian process.

In terms of timing (presented in seconds in Table 1), the recorded time of the conjugate models includes the time for choosing hyper-parameters through cross-validation and (“+") the time for sampling from the posterior distribution. The recorded time of the full GP model consists of the time for MCMC sampling and (“+") the time for recovering the regression coefficients and predictions. The full latent NNGP model is 200 times faster than the full Gaussian process based model, while the conjugate latent NNGP model uses one tenth of the time required by the latent NNGP model to obtain similar inference on the regression coefficients and latent process. Further simulation experiments conducted by Zhang et al. [15] also show that interpolation of the latent process is almost indistinguishable between the conjugate and full models.

Next, we present a second simulation example using exactly the same setup as in the preceding example, but with $n = 12,000$ spatial locations. Here, we fit a latent NNGP model using the MCMC algorithm in [37] and the conjugate latent NNGP model. We used $10,000$ locations for training the models while the remaining 2000 locations were used for predictive assessment. We summarized the results from the latent NNGP model using a post burn-in posterior sample for $10,000$ iterations. This was deemed adequate based upon the customary convergence diagnostics available in the `coda` and `mcse` packages within the `R` computing environment [51, 52]. The inference from the conjugate latent NNGP model were based on 300 samples. This is sufficient for the conjugate latent NNGP model since the conjugate model provides independent samples from the exact posterior distribution. The full MCMC-based NNGP model took about 1268 seconds to deliver full Bayesian inference, while the conjugate model took only $99 + 14 = 113$ seconds (99 seconds for the cross-validation to fix $\{\phi, \delta^2\}$ and 13 seconds for sampling from the posterior distribution). We found that the RMSPE values for the full latent NNGP and the conjugate model computed using the 2000 hold-out locations were almost identical (0.67 up to 2 decimal places).

The parameter estimates from the full NNGP and conjugate NNGP models in this larger simulation experiment reveal essentially the same story as in Table 1 so we do not repeat them here. Instead, we focus on the estimation of the latent process and the predictive performance for the two models. Figure 2 shows interpolated surfaces from the simulation example: 2(a) shows an interpolated map of the “true” spatial latent process w, while 2(b) and 2(c) present the posterior means of $w(s)$ over the entire domain obtained from the full latent NNGP model and the conjugate latent NNGP model, respectively. The recovered spatial residual surfaces are almost indistinguishable, and are comparable to the true interpolated surface of $w(s)$. Figure 2(d)–(e) present the 95% credible intervals for the spatial effects w from the latent NNGP model and the conjugate latent NNGP model. These intervals are plotted against the true values of w from the generated model. We found that 9567 out of 10000 credible intervals successfully included the true value for the conjugate model, while the corresponding number was a very comparable 9584 for the full NNGP model.

Turning to a real example, we present a synopsis of the analysis in [15] of a spatial dataset from NASA comprising sea surface temperature (in degrees Centigrade) observations over 2,827,252 spatial locations of which approximately 90% (2,544,527) were used for model fitting and the rest were withheld for cross-validatory predictive assessment. Details of the dataset can be found in `http://modis-atmos.gsfc.nasa.gov/index.html` and details on the analysis can be found in [15]. The salient feature of the analysis is that a conjugate Bayesian framework for the NNGP model as in (12) was able to deliver full inference including the estimation of the spatial latent effects in about 2387 seconds. Sampling from the posterior distribution was achieved using direct sampling as described below (12). Since this algorithm is fast and directly samples from the posterior, hence there is no burn-in period for convergence, it was run over a grid of values of $\{\delta^2, \phi\}$. For each such value, a posterior predictive assessment over the cross-validatory hold-out set was



(a) True

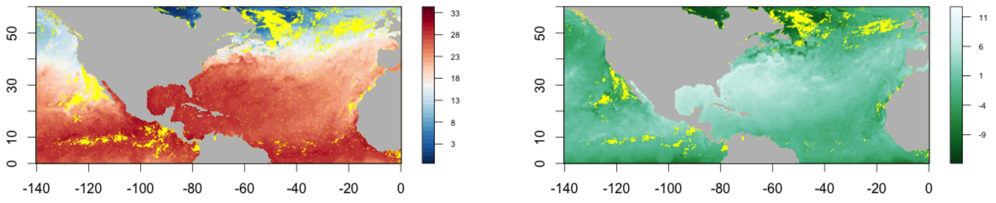
(b) NNGP

(c) Conjugate NNGP

(d) CIs of w from NNGP

(e) CIs of w from conjugate NNGP

Figure 2: Interpolated maps of (a) the true generated surface, (b) the posterior means of the spatial latent process $w(s)$ for the NNGP and (c) posterior means of $w(s)$ for the conjugate latent NNGP. The 95% confidence intervals for the spatial effects w from (d) the NNGP and (e) the conjugate NNGP. The NNGP models were all fit using $m = 10$ nearest neighbors.



(a) Posterior mean of sea-surface temperature

(b) Posterior predictive mean of latent spatial effects

Figure 3: Posterior predictive maps of sea-surface temperature (in degree centigrade) and latent spatial effects. The land is colored in gray, locations in the ocean without observations are indicated in yellow.

carried out and the value of $\{\delta^2, \phi\}$ producing the least RMSPE was selected as optimal inputs for which the estimates of $\{\gamma, \sigma^2\}$ were presented.

5 Spatial Meta-Kriging

A different approach toward BIG DATA problems relies upon divide and conquer methods. The idea here is divide and conquer (or map and reduce) by pooling posterior inference across a partition of data subsets. Once again consider the

Bayesian linear regression model

$$p(\beta, \sigma^2 \,|\, y) \propto IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\beta \,|\, \mu_\beta, \sigma^2 V_\beta) \times N(y \,|\, X\beta, \sigma^2 V_y) \;, \tag{17}$$

where y is $N \times 1$, X is $N \times p$, β is $p \times 1$, V_y is a fixed $N \times N$ covariance matrix, μ_β is a fixed $p \times 1$ vector and V_β is a fixed $p \times p$ matrix. The joint posterior density $p(\beta, \sigma^2 \,|\, y)$ is available in closed form as $p(\beta, \sigma^2 \,|\, y) = p(\sigma^2 \,|\, y) \times p(\beta \,|\, \sigma^2, y)$, where the marginal posterior density $p(\sigma^2 \,|\, y) = IG(\sigma^2 \,|\, a^*, b^*)$ and the conditional posterior density $p(\beta \,|\, \sigma^2, y) = N(\beta \,|\, Mm, \sigma^2 M)$ with $a^* = a_\sigma + N/2$, $b^* = b_\sigma + c/2$, $m = V_\beta^{-1}\mu_\beta + X^\top V_y^{-1} y$, $M^{-1} = V_\beta^{-1} + X^\top V_y^{-1} X$ and $c = \mu_\beta^\top V_\beta^{-1}\mu_\beta + y^\top V_y^{-1} y - m^\top M m$. Therefore, exact posterior inference can be carried out by first sampling σ^2 from $IG(a^*, b^*)$ and then sampling β from $N(Mm, \sigma^2 M)$ for each sampled value of σ^2. This results in samples from $p(\beta, \sigma^2 \,|\, y)$. Besides the fixed hyperparameters in the prior distributions, this exercise requires computing m, M and c.

Now consider a situation where N is large enough so that memory requirements for computing (17) is unfeasible. One possible resolution is to replace the likelihood in (17) with a composite likelihood that assumes independence across blocks formed by partitioning the data into K subsets. We partition the $N \times 1$ vector y into K subvectors with y_k as the $n_k \times 1$ subvector forming the k-th subvector, where $\sum_{k=1}^{K} n_k = N$. The size of the k-th subset is n_k. These sizes need not be the same across k, but will be chosen in a manner so that each of the subsets can be fitted easily with the computational resources available. Also, let X_k be the $n_k \times p$ matrix of predictors corresponding to y_k and let V_{y_k} be the marginal correlation matrix for y_k. The conjugate Bayesian model with a block-independent composite likelihood assumes that $y_k = X_k\beta + \epsilon_k$, where $\epsilon_k \overset{ind}{\sim} N(0, \sigma^2 V_{y_k})$. The Bayesian specification is completed by assigning priors to σ^2 and β as in (17). If we distribute the analysis to K different computing cores, where the k-th core fits the above model but only with the likelihood $N(y_k \,|\, X_k\beta, \sigma^2 V_{y_k})$, then the quantities needed for sampling from the full $p(\beta, \sigma^2 \,|\, y)$ can be computed entirely using quantities obtained from the individual subsets of the data. For each $k = 1, 2, \ldots, K$ we independently compute $m_k = V_\beta^{-1}\mu_\beta + X_k^\top V_{y_k}^{-1} y_k$ and $M_k^{-1} = V_\beta^{-1} + X_k^\top V_{y_k}^{-1} X_k$ based upon the k-th subset of the data. We then combine them to obtain $m = \sum_{k=1}^{K}(m_k - (1 - 1/K)V_\beta^{-1}\mu_\beta)$ and $M^{-1} = \sum_{k=1}^{K}(M_k^{-1} - (1 - 1/K)V_\beta^{-1})$. Subsequently, we compute $c = \mu_\beta^\top V_\beta^{-1}\mu_\beta + \sum_{k=1}^{K} y_k^\top V_{y_k}^{-1} y_k - m^\top M m$. Therefore, sampling from the posterior distribution of β and σ^2 given the entire dataset can be achieved using quantities computed independently from each of the K smaller subsets of the data. There is no need to interact between the subsets and one does not require to store or compute with large objects based upon the entire dataset. This computation can also be done sequentially. We first obtain the posterior distribution $p(\beta, \sigma^2 \,|\, y_1)$ based only upon the first data set. This posterior becomes the prior for the next step and we obtain $p(\beta, \sigma^2 \,|\, y_1, y_2) \propto p(\beta, \sigma^2 \,|\, y_1) \times p(y_2 \,|\, \beta, \sigma^2)$ and so on until we arrive at $p(\beta, \sigma^2 \,|\, y_1, y_2, \ldots, y_K) \propto p(\beta, \sigma^2 \,|\, y_1, y_2, \ldots, y_{K-1}) \times p(y_K \,|\, \beta, \sigma^2)$.

Clearly such exact recovery of the full posterior crucially depends on the conditional independence across the different data blocks (e.g., $p(y_k \,|\, \beta, \sigma^2, y_1, \ldots, y_{k-1}) = p(y_k \,|\, \beta, \sigma^2)$ for each $k = 2, \ldots, K$). While this works for uncorrelated outcomes, as in standard linear regression, such recovery is precluded for spatial and spatiotemporal process models and, more generally, for correlated data. Nevertheless, we can develop a general approximation framework for obtaining the full posterior from posterior densities calculated over smaller subsets. One general way to pool information across these individual posteriors is to use the unique *Geometric Median* (GM) of the subset posteriors, as developed by Minsker [53]. Assume that the individual posterior densities $p_k \equiv p(\Omega \,|\, y_k)$ reside on a Banach space $\mathcal{H}$ equipped with norm $\|\cdot\|$. The GM is defined as $\pi^*(\cdot \,|\, y) = \arg\min\limits_{\pi \in \mathcal{H}} \sum\limits_{k=1}^{K} \|p_k - \pi\|_\rho$, where $y = (y_1^\top, y_2^\top, \ldots, y_K^\top)^\top$. The norm quantifies the distance between any two posterior densities $\pi_1(\cdot)$ and $\pi_2(\cdot)$ as $\|\pi_1 - \pi_2\|_\rho = \left\| \displaystyle\int \rho(\Omega, \cdot) d(\pi_1 - \pi_2)(\Omega) \right\|$, where $\rho(\cdot)$ is a positive-definite kernel function. Assume $\rho(z_1, z_2) = \exp(-\|z_1 - z_2\|^2)$. The GM is unique and lies in the convex hull of the individual posteriors, so $\pi^*(\Omega \,|\, y)$ is a legitimate probability density. Specifically, $\pi^*(\Omega \,|\, y) = \sum_{k=1}^{K} \alpha_{\rho,k}(y) p_k$,$\sum_{k=1}^{K} \alpha_{\rho,k}(y) = 1$, each $\alpha_{\rho,k}(y)$ being a function of ρ, y, so that $\int_\Omega \pi^*(\Omega \,|\, y) d\Omega = 1$. Computing the GM $\pi^* \equiv \pi^*(\Omega \,|\, y)$ is achieved by an iterative algorithm that estimates $\alpha_{\rho,k}(y)$ from the subset posteriors p_k for each $k = 1, 2, \ldots, K$. To further elucidate, we use a well known result that the GM π^* satisfies $\pi^* = \frac{\sum_{k=1}^{K} \|p_k - \pi^*\|_\rho^{-1} p_k}{\sum_{k=1}^{K} \|p_k - \pi^*\|_\rho^{-1}}$, so that $\alpha_{\rho,k}(y) = \frac{\|p_k - \pi^*\|_\rho^{-1}}{\sum_{j=1}^{K} \|p_k - \pi^*\|_\rho^{-1}}$. There is no apparent closed-form solution for $\alpha_{\rho,k}(y)$ satisfying this equation, so Weiszfeld's algorithm [53] is used to estimate these functions.

This approach has been extended to spatial process settings by Guhaniyogi and Banerjee [54, 55]. The advantage here is that one can use existing Bayesian geostatistical software to sample from the posterior distributions of the different subsets. This can be performed either in parallel over multiple cores or across different machines altogether. One then needs to save only the post burn-in samples and execute Weiszfeld's algorithm to these samples. Weiszfeld's algorithm is extremely fast and easy to program.

6 Discussion

This article has attempted to provide a brief overview of how some Bayesian geostatistical models designed for large spatial and/or spatiotemporal datasets can be further scaled up to analyze massive datasets with observed locations in the order of 10^6 or more by exploiting the familiar theory of conjugate Bayesian linear regression models and adapting them to incorporate latent spatial processes. The resulting distribution theory is available in closed form, thereby circumventing the need for iterative algorithms such as MCMC or INLA. We have also provided a brief overview of a distributed approach (spatial meta-kriging) that relies upon analyzing exclusive subsets of the data and combining them to approximate the full posterior in the spirit of a spatial meta-analysis.

Of course, this requires some compromise in terms of full Bayesian inference. Some parameters need to be provided as fixed inputs for the distribution theory to be available in closed form. Learning about these input parameters will be done using exploratory data analysis and cross-validation methods. A practical approach that seems to be quite effective for analyzing massive datasets in modest computing environments is to choose the optimal value of the process parameters based upon the minimum RMSPE over hold-out locations. While such approaches may produce slightly shrunk credible and prediction intervals due to the effect of fixing a parameter, the effect is seen to be moderate in practical spatial analysis and the approach could form a useful tool for quick spatial analysis within the Bayesian paradigm for massive spatial datasets. However, the method of learning about these parameters is still ad-hoc and can possibly be improved with more sophisticated optimization methods. Nevertheless, the approach outlined here can be a useful tool in the spatial analyst's toolbox for exploring Bayesian spatial regression at massive scales. We also point out that the conjugate Bayesian linear regression framework can accommodate almost all of the model-based GP approximations for dimension reduction or sparsity induction. Any spatial covariance structure that leads to efficient computations can, in principle, be used.

While the article has focused on the NNGP as a choice for introducing sparsity in the model, more general GMRF specifications are also admissible here. In fact, there has been much recent activity within the framework of Vecchia approximations [see, e.g., 43, 44], where the models are being derived using DAGs over the expanded set of observations and process realizations. While certainly promising, their benefits and improvements over GMRFs are yet to be demonstrated in large scale case studies. For Vecchia type of likelihoods, there is also interest in choosing the number of neighbors. First, it should be intuitively clear that DAGs constructed using shrunk neighbor sets will yield probability models farther away from the full model as the neighbor sets get smaller. To see this, consider a random vector $w = (w_A^\top, w_B^\top)^\top$, where A and B are mutually exclusive sets containing indices for the elements of w, and let $p(w) = p(w_A)p(w_B \mid w_A)$ denote the joint probability density for w. Consider two submodels $p_1(w) = p(w_A) \times p(w_B \mid w_{N_{1B}})$ and $p_2(w) = p(w_A) \times p(w_B \mid w_{N_{2B}})$, where $N_{2B} \subset N_{1B} \subset A$. The model p_2 will be farther than p_1 from p in the terms of the Kullback-Leibler divergence:

$$\begin{aligned} KL(p\|\|p_2) - KL(p\|\|p_1) &= \int \left\{ \log\left(\frac{p(w)}{p_2(w)}\right) - \log\left(\frac{p(w)}{p_1(w)}\right) \right\} p(w)dw \\ &= \int \log\left(\frac{p_1(w)}{p_2(w)}\right) p(w)dw = \int \log\left(\frac{p(w_B \mid w_{N_{1B}})}{p(w_B \mid w_{N_{2B}})}\right) p(w)dw \\ &= \int \log\left(\frac{p(w_B \mid w_{N_{1B}})}{p(w_B \mid w_{N_{2B}})}\right) p(w_B \mid w_{N_{1B}}) p(w_{N_{1B}}) dw_B dw_{N_{1B}} \\ &= \int \left\{ \int \log\left(\frac{p(w_B \mid w_{N_{1B}})}{p(w_B \mid w_{N_{2B}})}\right) p(w_B \mid w_{N_{1B}}) dw_B \right\} p(w_{N_{1B}}) dw_{N_{1B}} \geq 0 \;, \end{aligned} \tag{18}$$

where we have used the fact that $A \setminus N_{1B}$ is mutually exclusive of N_{1B} and, crucially, also of N_{2B} (since $N_{2B} \subset N_{1B}$) to legitimately integrate out $w_{A \setminus N_{1B}}$. The final conclusion follows from a customary application of Jensen's inequality to show that the inner integral in the last equation is non-negative. Equation (18) provides an alternate distribution-free proof of a result for Gaussian likelihoods by Guinness (Theorem 1 in [42]). These results also indicate that the ordering of the variables to construct the approximation can affect model performance and certain designs to determine the ordering can produce improved results (as demonstrated in [42]). Datta et al. [38] argued against fixing the neighborhoods in spatiotemporal contexts (since neighbors in space and neighbors in time may not align) and demonstrate a computationally efficient method to learn about neighbors in spatiotemporal domains.

Finally, we point toward a few future directions of research in this domain. Much of the spatial literature on modeling massive spatial data have focused upon scalability of models and algorithms. There is still work to be done on evaluating the inferential performance of these models at such massive scales. How important is uncertainty quantification at such scales? How do GP based approaches compare with deep learning with neural networks in spatial analysis? Another area where the cross-validatory learning approaches for process hyperparameters will struggle is in multivariate contexts, where the number of hyperparameters is higher than here. These are some areas of research where we believe the statistical community still has much to offer.

References

[1] N. Cressie. *Statistics for Spatial Data*. Wiley-Interscience, revised edition, 1993.

[2] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, first edition, 1999.

[3] J. Moller and R. P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall, first edition, 2003.

[4] O. Schabenberger and C. A. Gotway. *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC Press, Boca Raton, FL, first edition, 2004.

[5] A.E. Gelfand, P.J. Diggle, M. Fuentes, and P Guttorp. *Handbook of Spatial Statistics*. Boca Raton, FL: CRC Press, 2010.

[6] Noel A. C. Cressie and Christopher K. Wikle. *Statistics for Spatio-temporal Data*. Wiley series in probability and statistics. Hoboken, N.J. Wiley, 2011.

[7] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC Press, Boca Raton, FL, 2014.

[8] T. Gneiting and P. Guttorp. Continuous-parameter spatio-temporal processes. In A.E. Gelfand, P.J. Diggle, M. Fuentes, and P Guttorp, editors, *Handbook of Spatial Statistics*, pages 427–436. CRC Press, Boca Raton, FL, 2010.

[9] Sudipto Banerjee. High-dimensional bayesian geostatistics. *Bayesian Analysis*, 12:583–614, 2017.

[10] M.J. Heaton, A. Datta, A.O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren, D. Nychka, F. Sun, and A. Zammit-Mangion. Methods for analyzing large spatial data: A review and comparison. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425, 2019.

[11] Matthias Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112:201–214, 2017.

[12] S. Abdulah, H. Ltaief, Y. Sun, M.G. Genton, and D.E. Keyes. Exageostat: A high performance unified software for geostatistics on manycore systems. *IEEE Transactions on Parallel and Distributed Systems*, 29:2771–2784, 2018.

[13] H. Huang and Y. Sun. Hierarchical low-rank approximation of likelihoods for large spatial datasets. *Journal of Computational and Graphical Statistics*, 27:110–118, 2018.

[14] Andrew O Finley, Abhirup Datta, Bruce C Cook, Douglas C Morton, Hans E Andersen, and Sudipto Banerjee. Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414, 2019.

[15] Lu Zhang, Abhirup Datta, and Sudipto Banerjee. Practical bayesian modeling and inference for massive spatial datasets on modest computing environments. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):197–209, 2019.

[16] C.K. Wikle and N. Cressie. A dimension reduced approach to space-time kalman filtering. *Biometrika*, 86:815–829, 1999.

[17] Christopher K. Wikle. Low-rank representations for spatial processes. *Handbook of Spatial Statistics*, pages 107–118, 2010. Gelfand, A. E., Diggle, P., Fuentes, M. and Guttorp, P., editors, Chapman and Hall/CRC, pp. 107-118.

[18] N. Cressie and G. Johannesson. Fixed rank kriging for very large data sets. *Journal of the Royal Statistical society, Series B*, 70:209–226, 2008.

[19] T. Shi and N. Cressie. Global statistical analysis of misr aerosol data: A massive data product from nasa's terra satellite. *Environmetrics*, 18:665–680, 2007.

[20] N. Cressie, T. Shi, and E. L. Kang. Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19:724–745, 2010.

[21] M Katzfuss and N Cressie. Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics*, 23:94–107, 2012.

[22] Matthias Katzfuss. Bayesian nonstationary modeling for very large spatial datasets. *Environmetrics*, 24:189–200, 2013.

[23] D. Nychka, C. Wikle, and J. A. Royle. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4):315–331, 2002.

[24] Douglas Nychka, Soutir Bandyopadhyay, Dorit Hammerling, Finn Lindgren, and Stephan Sain. A multiresolution gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015.

[25] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, first edition, 2005.

[26] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society, Series B*, 70:825–848, 2008.

[27] S. Banerjee, A. O. Finley, P. Waldmann, and T. Ericcson. Hierarchical spatial process models for multiple traits in large genetic trials. *Journal of the American Statistical Association*, 105:506–521, 2010.

[28] H. Sang, M. Jun, and J.Z. Huang. Covariance approximation for large multivariate spatial datasets with an application to multiple climate model errors. *Annals of Applied Statistics*, 4:2519–2548, 2011.

[29] H. Sang and J. Z. Huang. A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical society, Series B*, 74:111–132, 2012.

[30] Andrew O. Finley, Sudipto Banerjee, and Alan E. Gelfand. spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63(13):1–28, 2015.

[31] R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15:503–523, 2006.

[32] C. G. Kaufman, M. J. Scheverish, and D. W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103:1545–1555, 2008.

[33] J. Du, H. Zhang, and V. S. Mandrekar. Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Annals of Statistics*, 37:3330–3361, 2009.

[34] Havard Rua and Leonard Held. *Gaussian Markov Random Fields : Theory and Applications*. Monographs on statistics and applied probability. Chapman and Hall/CRC Press, Boca Raton, FL, 2005.

[35] Havard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

[36] Finn Lindgren, Havard Rue, and Johan Lindstrom. 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, 2011.

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

[38] A. Datta, S. Banerjee, A. O. Finley, N. A. S. Hamm, and M. Schaap. Non-separable dynamic nearest-neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Annals of Applied Statistics*, 10:1286–1316, 2016b.

[39] A. V. Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical society, Series B*, 50:297–312, 1988.

[40] M. L. Stein, Z. Chi, and L. J. Welty. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical society, Series B*, 66:275–296, 2004.

[41] J.R. Stroud, M. L. Stein, and S. Lysen. Bayesian and maximum likelihood estimation for gaussian processes on an incomplete lattice. *Journal of Computational and Graphical Statistics*, 26:108–120, 2017.

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

[43] Matthias Katzfuss and Joseph Guinness. A general framework for vecchia approximations of gaussian processes. *arXiv preprint arXiv:1708.06302*, 2017.

[44] Matthias Katzfuss, Joseph Guinness, Wenlong Gong, and Daniel Zilber. Vecchia approximations of gaussian-process predictions. *arXiv preprint arXiv:1805.03309*, 2018.

[45] Hongxiang Shi, Emily L. Kang, Bledar A. Konomi, Kumar Vemaganti, and Sandeep Madireddy. Uncertainty quantification using the nearest neighbor gaussian process. In Ding-Geng Chen, Zhezhen Jin, Gang Li, Yi Li, Aiyi Liu, and Yichuan Zhao, editors, *New Advances in Statistics and Data Science*, pages 89–107. Springer International Publishing, Cham, Switzerland, 2017.

[46] Pulong Ma and Emily L. Kang. Fused gaussian process for very large spatial data. *arXiv:1702.08797v3*, 2017.

[47] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, 3rd Edition*. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC, 2013.

[48] Hao Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.

[49] Ozgür Yeniay and Atill Goktas. A comparison of partial least squares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics*, 31(99):99–101, 2002.

[50] Paulo J. Ribeiro Jr and Peter J. Diggle. *geoR: a package for geostatistical analysis*, June 2012. R package version 1.7-4.

[51] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

[52] J.M. Flegal and G.L. Jones. Implementing markov chain monte carlo: Estimating with confidence. In S. Brooks, A. Gelman, G.L. Jones, and X. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 175–197. Chapman and Hall/CRC Press, Boca Raton, FL, 2011.

[53] S Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21:2308–2335, 2015.

[54] Rajarshi Guhaniyogi and Sudipto Banerjee. Meta-kriging: Scalable bayesian modeling and inference for massive spatial datasets. *Technometrics*, 60(4):430–444, 2018.

[55] Rajarshi Guhaniyogi and Sudipto Banerjee. Multivariate spatial meta-kriging. *Statistics and Probability Letters*, 144:3–8, 2019.