Circulant embedding of approximate covariances for inference from Gaussian data on large lattices

Joseph Guinness and Montserrat Fuentes

July 25, 2014

Abstract

We introduce periodic covariance approximations designed for embedding covariance matrices for lattice-located observations inside of nested block circulant covariance matrices. The approximations are positive definite provided that the embedding lattice is at least as large as the observation lattice, in contrast with standard circulant embedding methods that require the embedding lattice to be at least twice the size of the observation lattice in each dimension. Recently proposed computationally efficient Markov chain Monte Carlo and Monte Carlo Expectation-Maximization (EM) methods for estimating covariance parameters rely on successive imputations of values on the larger embedding lattice. We demonstrate in simulations that the use of smaller embedding lattices, and thus smaller numbers of imputed values, leads to Markov chains with less autocorrelation and EM algorithms that converge more quickly, without sacrificing the accuracy of the parameter estimates. Our approximations are particularly advantageous in more than two dimensions. We also present numerical studies to guide the construction of the approximations. We conclude with an analysis and interpolation of photosynthetically available radiation data and show that our approximate procedures are faster to compute per iteration, in addition to the improved iterative performance.

1 Introduction

The Gaussian process model plays a central role in the analysis of spatially and spatial-temporally correlated data. It is used directly for modeling the data that can be assumed to be Gaussian and often used indirectly as a stage in a hierarchical process model when the data are not assumed to be Gaussian. Consider a stochastic process $Z(\boldsymbol{x}) \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^d$. The defining property of a Gaussian process is that for any $n \in \mathbb{N}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, the vector $\boldsymbol{Z} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))'$ has a multivariate normal distribution. A Gaussian process is characterized by its mean at every location $E(Z(\boldsymbol{x}))$ and the covariance between its observations at any two locations $\text{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{y}))$, and it is common to assume that both the means and the covariances are specified by parametric functions.

We write $\boldsymbol{\mu}_\beta = E(\boldsymbol{Z})$ and $K_\theta = E((\boldsymbol{Z} - \boldsymbol{\mu}_\beta)(\boldsymbol{Z} - \boldsymbol{\mu}_\beta)')$ to signify a mean vector with parameter β and covariance matrix with parameter θ for observations at a specific set of locations, understood from context. Statistical computation with Gaussian process models requires algebraic manipulations involving $\boldsymbol{\mu}_\beta$ and K_θ. For example, if we wish to simulate $\boldsymbol{Z}$, we draw a mean-zero vector with independent components $\boldsymbol{X} \sim N(0, I_n)$ and construct $\boldsymbol{Z} = \boldsymbol{\mu}_\beta + C\boldsymbol{X}$, where C is an $n \times n$ matrix for which $CC' = K_\theta$. The matrix C is generally not unique, but a popular choice is a triangular factorization, also known as a Cholesky factorization in this case where K_θ is symmetric and positive definite. The Cholesky factorization is also commonly used in computing the Gaussian loglikelihood function. If K_θ has no exploitable structure, the Cholesky factorization requires $O(n^3)$ floating point operations (flops) and $O(n^2)$ memory, so its computational burden begins to overwhelm modern standard computational facilities when n is greater than 10,000. The addition or subtraction of the mean vector requires only $O(n)$ flops and memory, so we focus our attention here on computations involving covariances and assume throughout that the mean is identically zero.

Some covariances and sets of observation locations induce structure in K_θ that allows for the exploitation of algorithms that sidestep these computational burdens. It is important to identify such scenarios so that, when they occur, we avoid unnecessary computational effort and streamline statistical analysis of spatial and spatial-temporal data. A well-known case is the pairing of regular lattice locations with stationary covariances. A process has stationary covariances if the covariance between observations at any two locations $\boldsymbol{x}$ and $\boldsymbol{y}$ depends only on the vector lag $\boldsymbol{x} - \boldsymbol{y}$. Then we write $\text{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{y})) = K_\theta(\boldsymbol{x} - \boldsymbol{y})$ and call $K_\theta(\cdot)$ the covariance function, which depends on parameter θ. We define $\delta\mathbb{Z}^d$ to be the d-dimensional regular lattice with spacing $\delta > 0$, that is, $\boldsymbol{x} = (x_1, \ldots, x_d) \in \delta\mathbb{Z}^d$ if $x_j/\delta \in \mathbb{Z}$ for each $j \in \{1, \ldots, d\}$. For any particular finite set of lattice locations $J = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \subset \delta\mathbb{Z}^d$, there exists $\boldsymbol{n} = (n_1, \ldots, n_d)$ such that $J \subset \delta\mathbb{J}_{\boldsymbol{n}}$, a finite rectangular lattice with dimension sizes given by the components of $\boldsymbol{n}$. We definite $n = n_1 \times \cdots \times n_d$ to be the number of locations in $\delta\mathbb{J}_{\boldsymbol{n}}$, and we call $\delta\mathbb{J}_{\boldsymbol{n}}$ the observation lattice.

Wood and Chan (1994) showed that if $J \subset \delta\mathbb{J}_{\boldsymbol{n}}$, and the covariances are stationary, then the resulting covariance matrix K_{11} for $\boldsymbol{Z}_1 = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_k))$ can always be embedded within a larger covariance matrix that is nested block circulant, meaning block circulant with each successive subblock being block circulant. We give a formal definition of nested block circulant in Appendix A. Letting $\lfloor y \rfloor$ denote the integer part of y, "circulant embedding" is achieved by defining $m_j := \lfloor \tau_j n_j \rfloor$ with $\tau_j \geq 1$ for each $j \in \{1, \ldots, d\}$, $\boldsymbol{m} = (m_1, \ldots, m_d)$, $m = m_1 \times \cdots \times m_d$, and by constructing a covariance function $\widetilde{K}_\theta(\cdot)$ that is periodic in each dimension with periods given by the components of $\boldsymbol{m}$, and for which $\widetilde{K}_\theta(\boldsymbol{x}-\boldsymbol{y}) = K_\theta(\boldsymbol{x}-\boldsymbol{y})$ for any $\boldsymbol{x}, \boldsymbol{y} \in \delta\mathbb{J}_{\boldsymbol{n}}$. The periodicity of $\widetilde{K}_\theta(\cdot)$ ensures that the covariance matrix $\widetilde{K}_\theta$ for the observations on $\delta\mathbb{J}_{\boldsymbol{m}}$ (ordered lexicographically) is nested block circulant. We call $\delta\mathbb{J}_{\boldsymbol{m}}$ the embedding lattice. Defining $\boldsymbol{Z}_2$ to be the set of observations on $\delta\mathbb{J}_{\boldsymbol{m}} \setminus J$, we write the covariance matrix for $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ as

$$P\widetilde{K}_\theta P^T = \begin{bmatrix} \widetilde{K}_{11} & \widetilde{K}_{12} \\ \widetilde{K}_{21} & \widetilde{K}_{22} \end{bmatrix},$$

where P is a permutation matrix for reordering the rows of $\widetilde{K}_\theta$ so that $\widetilde{K}_{ij} = E(\boldsymbol{Z}_i \boldsymbol{Z}_j')$ under $\widetilde{K}_\theta(\cdot)$. Since $\widetilde{K}_\theta(\boldsymbol{x} - \boldsymbol{y}) = K_\theta(\boldsymbol{x} - \boldsymbol{y})$ for any $\boldsymbol{x}, \boldsymbol{y} \in \delta\mathbb{J}_{\boldsymbol{n}}$, $\widetilde{K}_{11} = K_{11}$, which is the covariance matrix for $\boldsymbol{Z}_1$ under $K_\theta(\cdot)$, we say that K_{11} is embedded inside the nested block circulant matrix $\widetilde{K}_\theta$. Circulant embedding derives its usefulness from the fact that the d-dimensional discrete Fourier transform (DFT) diagonalizes nested block circulant matrices. Since fast Fourier transform (FFT) algorithms return the eigenvalues of $\widetilde{K}_\theta$ in $O(m \log m)$ flops and use only $O(m)$ memory, the authors proposed using circulant embedding to generate computationally efficient simulations of stationary Gaussian processes on lattices. This method proceeds by simulating a complete set of observations on the embedding lattice $\delta\mathbb{J}_{\boldsymbol{m}}$–using an FFT to factor $\widetilde{K}_\theta$–extracting the simulated values on J, and discarding the rest.

Circulant embedding has a rich literature and has been reinvented at least once independently for the purpose of Gaussian process simulation by Dietrich and Newsam (1997). Chan and Wood (1999) described extensions to the multivariate case, and circulant matrices are used extensively in statistical computations with Gaussian Markov random field models (Rue and Held, 2005). For given choices $\tau_j \geq 1$, it is not always possible to find a suitable function $\widetilde{K}_\theta(\cdot)$. Aside from covariance functions that are compactly supported, which is a somewhat limiting restriction, it is generally required that $\tau_j \geq 2$ for each j.[1] Even if we set $\tau_j \geq 2$ for each j, the existence of $\widetilde{K}_\theta(\cdot)$ for a particular $K_\theta(\cdot)$ may still not be guaranteed, and other authors, including Stein (2002) and Gneiting et al. (2006),

[1] It is sometimes possible to achieve positive definite circulant embedding with $m_j = 2n_j - 1$, but we ignore this possibility for simplicity of notation since it is not important asymptotically and negligible in practice for large lattices.

have provided methods for constructing valid $\widetilde{K}_\theta(\cdot)$ while minimizing the size of each τ_j. When there exists $\widetilde{K}_\theta(\cdot)$ with $\tau_j = 2$ for each j, this is referred to as minimal embedding.

Until recently, circulant embedding methods focused primarily on simulation. Stroud et al. (2014) proposed methodology for drawing inferences about covariance parameter θ from two-dimensional lattice data. The authors describe an MCMC algorithm for Bayesian inference that consists of alternating updates of the missing values $\boldsymbol{Z}_2$ and the possibly vector-valued parameter θ. The missing values are updated with conditional simulations given the observed values and the current parameter. The parameter is updated with a either a Gibbs or Metropolis-Hastings (MH) algorithm given the observed data and the current imputed missing values. Computing the MH acceptance probabilities is completed in $O(m \log m)$ flops since factoring $\widetilde{K}_\theta$ is the limiting computational task in evaluating the Gaussian loglikelihood for $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$. The computational effort required for the conditional simulations is dominated by solving two linear systems of the form $K_{11}y = z$. The linear systems are solved by applying a preconditioned conjugate gradient algorithm, which is efficient since the forward multiplication $K_{11}y$ is sped up by computing instead the embedded multiplication $\widetilde{K}_\theta[y, 0]'$, which requires $O(m \log m)$ flops with an FFT. Several efficiently computed preconditioners are proposed, with one based on the likelihood approximation in Stein et al. (2004) effective in several scenarios.

Frequentist inference employs a Monte Carlo Expectation-Maximization (EM) algorithm. At iteration k the expected loglikelihood for $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ given $\boldsymbol{Z}_1$ is approximated by averaging M loglikelihoods for $(\boldsymbol{Z}_1, \boldsymbol{Z}_2^{(j)})$, where $\boldsymbol{Z}_2^{(1)}, \cdots, \boldsymbol{Z}_2^{(M)}$ are mutually independent conditional simulations of $\boldsymbol{Z}_2$ given $\boldsymbol{Z}_1$ with parameter for iteration k, denoted by $\theta^{(k)}$. The M conditional simulations are computed efficiently using the same methods proposed for updating the missing values in the MCMC algorithm. Then $\theta^{(k+1)}$ is set to the value that maximizes the averaged loglikelihood. This process is repeated for N iterations, where N is selected by a convergence criterion, and the set of parameters $(\theta^{(1)}, \ldots, \theta^{(N)})$ is used to construct estimators for θ. We investigate in Section 4 the selection of M and N.

The construction of $\widetilde{K}_\theta(\cdot)$ adopted by Stroud et al. (2014) is a variation of circulant embedding called cutoff embedding that was originally proposed by Stein (2002). There are multiple forms of cutoff embedding; in this paper, we use the term "cutoff embedding" to refer to the specific form implemented in Stroud et al. (2014). They require $\tau_j \geq 2\sqrt{d}$ for square lattices and use $\tau_j = 3$ in two-dimensional simulations. The Bayesian and frequentist methods described above are exact in the sense that the covariance function that the likelihood implies for the actual observations is equal to the target covariance function, a desirable feature of their construction that stands in contrast with various approximate likelihood or composite likelihood methods (Whittle (1954), Vecchia (1988), Stein et al. (2004)). However, for inferential procedures, which are generally more computationally intensive than simulation, it can be beneficial to sacrifice exactness if computational gains are made by employing approximations, especially considering that all of the proposed estimators contain Monte Carlo error. Further, cutoff embedding is not guaranteed to define positive definite covariance functions and sometimes produces covariance matrices that are not positive definite.

In this paper, we provide methods for achieving positive definite circulant embedding with approximate covariances and demonstrate the computational gains that these approximations afford. In Section 2, we describe approximations to covariance functions based on their spectral densities, producing positive definite, periodic covariance functions on $\delta\mathbb{J}_{\boldsymbol{m}}$ for any choices $\tau_j \geq 1$. If each $\tau_j = 1$, the approximations we propose reduce to the approximation inherent to the Whittle likelihood (Whittle, 1954). We present numerical studies in Section 3 showing that, for the purposes of likelihood-based parameter estimation, the approximations improve as each τ_j increases, and the approximations can be extremely sharp even if each $\tau_j < 2$. Selecting smaller τ_j defines a smaller embedding lattice $\delta\mathbb{J}_{\boldsymbol{m}}$, which decreases the number of missing values to impute in the iterative inferential procedures. Section 4 includes the results of simulation studies demonstrating that using approximate covariances to reduce the number of imputed values has a dramatic effect on the performance of the iterative algorithms. Specifically, in two-dimensional simulations, our methods increased the number of effective samples per iteration by at least a factor of five over cutoff embedding methods, and in three-dimensional simulations, the increase in the number of effective samples was at least a factor of eight, without sacrificing the accuracy of parameter estimates. In Section 5 we apply the methodology to a spatial dataset consisting of satellite measurements of photosynthetically available radiation, and we show that the approximate procedures reduced the computing time required for each iteration by nearly a factor of four compared to cutoff embedding. Thus, our proposed methods provide computational benefits on two fronts in iterative estimation procedures: reducing the total number of iterations required and the computing time required for each iteration.

2 Circulant Embedding with Approximate Covariances

In this section we provide methods for constructing periodic covariance functions with period in dimension j given by m_j. The covariance functions are guaranteed to produce positive definite covariance matrices for observations on $\delta\mathbb{J}_{\boldsymbol{m}}$ as long as $m_j \geq n_j$ for each j, or equivalently, each expansion factor $\tau_j \geq 1$. This relaxes the requirement that each $m_j \geq 2n_j$ in standard circulant embedding, and thus our approximations allow for embedding lattices that are smaller than the smallest embedding lattices allowed in standard circulant embedding. Although not required in practice, we assume for notational brevity throughout the rest of the paper that the observation lattice is expanded by a common factor in each dimension and write τ for the common expansion factor. In Figure 1, we illustrate the sizes of the embedding lattices used in our approximate procedures with $\tau = 5/4$ and in cutoff embedding. We show in Sections 4 and 5 that the use of smaller embedding lattices offers substantial computational advantages. Our methods differ from standard circulant embedding methods in that the covariances for the observed values approximate the target covariances $K_\theta(\boldsymbol{h})$, with the accuracy of the approximations depending on the expansion factor, so some care must be taken to choose the expansion factor to ensure that approximations are sufficiently sharp. In Section 3, we provide numerical studies that guide our choice of the expansion factor.

A full understanding of our approximations requires some background on spectral representations of stationary covariance functions. Bochner's Theorem (Yaglom, 1987) states that every stationary covariance function $K(\cdot)$ corresponds to a unique spectral measure $F(\cdot)$, with the correspondence given by the continuous inverse Fourier transform

$$K(\boldsymbol{h}) = \int_{\mathbb{R}^d} \exp(i\boldsymbol{\omega}'\boldsymbol{h}) dF(\boldsymbol{\omega}). \tag{1}$$

If $F(\cdot)$ is differentiable, then $dF(\boldsymbol{\omega}) = f(\boldsymbol{\omega})d\boldsymbol{\omega}$, and we call $f(\boldsymbol{\omega})$ the spectral density. We assume throughout this article that the spectral density exists and is continuous. While covariance functions must be positive definite, which is often difficult to establish directly, the spectral density need only be positive almost everywhere and integrable, which is usually easy to check.

Spectral representations are useful for checking positive definiteness of covariance functions and for proving theoretical results about random fields, but they are also useful for statistical computations when data are observed on a regular lattice $\delta\mathbb{Z}^d$. In this case we need to consider only covariances $K(\boldsymbol{h})$ with $\boldsymbol{h} \in \delta\mathbb{Z}^d$, and thus the covariance function can be expressed as

$$\begin{aligned} K(\boldsymbol{h}) &= \sum_{\boldsymbol{j}\in\mathbb{Z}^d} \int_{[0,2\pi/\delta]^d} f(\boldsymbol{\omega} + 2\pi\boldsymbol{j}/\delta) \exp(i(\boldsymbol{\omega} + 2\pi\boldsymbol{j}/\delta)'\boldsymbol{h}) d\boldsymbol{\omega} \\ &= \int_{[0,2\pi/\delta]^d} f_\delta(\boldsymbol{\omega}) \exp(i\boldsymbol{\omega}'\boldsymbol{h}) d\boldsymbol{\omega}, \end{aligned} \tag{2}$$

with the second equality following by exchanging summation and integration and using the fact that $\exp(i\boldsymbol{\omega}'\boldsymbol{h})$ is indistinguishable from, or *aliased* with, $\exp(i(\boldsymbol{\omega} + 2\pi\boldsymbol{j}/\delta)'\boldsymbol{h})$ on $\boldsymbol{h} \in \delta\mathbb{Z}^d$. We call $f_\delta(\boldsymbol{\omega}) = \sum_{\boldsymbol{j}\in\mathbb{Z}^d} f(\boldsymbol{\omega} + 2\pi\boldsymbol{j}/\delta)$ the aliased spectral density.

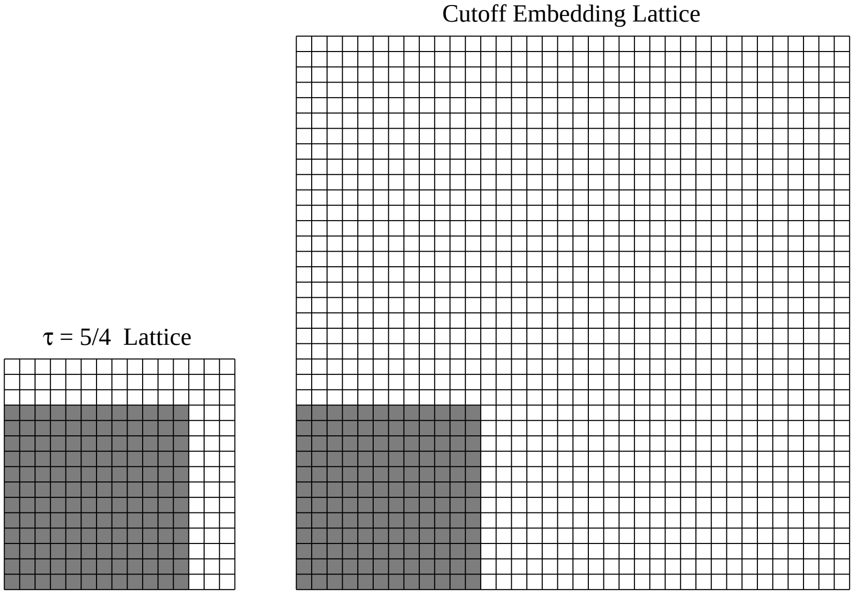

Figure 1: Illustration of the relative sizes of the embedding lattices ($\delta\mathbb{J}_{\boldsymbol{m}}$, gray+white) and the observation lattice ($\delta\mathbb{J}_{\boldsymbol{n}}$, gray) used in our approximate procedures with $\tau = 5/4$ and in cutoff embedding, which uses $\tau = 3$ in Stroud et al. (2014). Here, the observation lattice has size $\boldsymbol{n} = (12, 12)$.

The approximation for $K(\boldsymbol{h})$ that we propose for use in circulant embedding is a discretization of the integral in (2),

$$R_{\boldsymbol{m}}(\boldsymbol{h}) = \frac{(2\pi)^d}{m} \sum_{\boldsymbol{j} \in \mathbb{J}_{\boldsymbol{m}}} f_\delta(\boldsymbol{\omega_j}) \exp(i\boldsymbol{\omega}_{\boldsymbol{j}}'\boldsymbol{h}), \tag{3}$$

where $\boldsymbol{\omega_j} = (2\pi j_1/(\delta m_1), \dots, 2\pi j_d/(\delta m_d))$ are the d-dimensional Fourier frequencies on a grid of size $\boldsymbol{m}$. The expression in (3) has been called a discrete spectral approximation (Dietrich and Newsam, 1997). By varying the size of $\boldsymbol{m}$–or equivalently, τ–in $R_{\boldsymbol{m}}(\cdot)$, the practitioner has the ability to control its accuracy, and the approximations can be made arbitrarily accurate by choosing large τ. The approximation is powerful since $R_{\boldsymbol{m}}(\cdot)$ is automatically periodic with period m_j in each dimension, due to the periodicity of the complex exponentials $\exp(i\boldsymbol{\omega}_{\boldsymbol{j}}'\boldsymbol{h})$, which guarantees that the covariance matrix $R_{\boldsymbol{m}}$ for all observations on $\delta\mathbb{J}_{\boldsymbol{m}}$ (ordered lexicographically) is nested block circulant. This implies that the DFT diagonalizes $R_{\boldsymbol{m}}$, which has positive eigenvalues $f_\delta(\boldsymbol{\omega_j})$. Therefore $R_{\boldsymbol{m}}$ is automatically positive definite for any $\tau \geq 1$, and it is efficient to compute $R_{\boldsymbol{m}}(\boldsymbol{h})$ for all $\boldsymbol{h} \in \delta\mathbb{J}_{\boldsymbol{m}}$ with FFT algorithms when $f_\delta(\boldsymbol{\omega_j})$ is available. Rue and Held (2005) and Lindgren et al. (2011) discuss methods for constructing Markov random field models that are periodic on a domain that is slightly larger than the observation domain. Here, we do not assume that the model has a Markov or approximately Markov structure, only that it is stationary.

The approximation in (3) has the additional attractive feature that the covariance function $R_{\boldsymbol{n}}(\cdot)$, which is obtained by setting $\tau = 1$, is identical to the covariance function implied by the Whittle likelihood (Whittle, 1954). To see this, write $R_{\boldsymbol{n}} = F_{\boldsymbol{n}}^\dagger D F_{\boldsymbol{n}}$, where † is conjugate transpose, $F_{\boldsymbol{n}}$ is the matrix that performs the DFT, and D is a diagonal matrix containing appropriately ordered eigenvalues $f_\delta(\boldsymbol{\omega_j})$. Then the Gaussian loglikelihood for $\boldsymbol{Z}_1$ (assuming no missingness on $\delta\mathbb{J}_{\boldsymbol{n}}$) with covariance matrix $R_{\boldsymbol{n}}$ is

$$-\frac{1}{2}\log\det F_{\boldsymbol{n}}^{\dagger}DF_{\boldsymbol{n}} - \frac{1}{2}\boldsymbol{Z}_1'F_{\boldsymbol{n}}^{\dagger}D^{-1}F_{\boldsymbol{n}}\boldsymbol{Z}_1 = -\frac{1}{2}\sum_{\boldsymbol{j}\in\mathbb{J}_{\boldsymbol{n}}} f_\delta(\boldsymbol{\omega_j}) - \frac{1}{2}\sum_{\boldsymbol{j}\in\mathbb{J}_{\boldsymbol{n}}}\frac{|I(\boldsymbol{\omega_j})|^2}{f_\delta(\boldsymbol{\omega_j})}, \tag{4}$$

where $I(\boldsymbol{\omega_j})$ is the DFT of $\boldsymbol{Z}_1$ ($|I(\boldsymbol{\omega_j})|^2$ is called the periodogram). The expression on the right of (4) is a common form of the Whittle likelihood. We now see that the Whittle likelihood implies a covariance model that is periodic on the observation domain $\delta\mathbb{J}_{\boldsymbol{n}}$, so observations on the edges of the lattice are assumed to be correlated with each other. It is for this reason that the performance of the Whittle likelihood degrades with increasing the number of dimensions d, since in higher dimensions, greater proportions of observations are near the edges (Guyon, 1982). Dahlhaus and Künsch (1987) showed that multiplying the boundary observations by a tapering function to downweight their influence can improve the asymptotic behavior of parameter estimates when an adjusted Whittle likelihood is used, but this comes at the expense of altering the data.

The covariance function in (3) is periodic on the embedding lattice $\delta\mathbb{J}_{\boldsymbol{m}}$, so the resulting covariance matrix $R_{\boldsymbol{m}}$ for the set of all values on $\delta\mathbb{J}_{\boldsymbol{m}}$ (ordered lexicographically) has a nested block circulant structure, and after permutation of rows and columns, can be written in block form as

$$PR_{\boldsymbol{m}}P' = \left[\begin{array}{cc} R_{11} & R_{12} \\ R_{21} & R_{22} \end{array}\right],$$

where P is a permutation matrix for reordering the rows of $R_{\boldsymbol{m}}$ so that $R_{ij} = E(\boldsymbol{Z}_i\boldsymbol{Z}_j')$ under $R_{\boldsymbol{m}}(\cdot)$. We propose to use the Bayesian and frequentist methods in Stroud et al. (2014) and discussed in the Introduction. The difference here is that we replace $\widetilde{K}_\theta$ with $R_{\boldsymbol{m}}$ (which depends on θ as well). We show in the Sections 4 and 5 that this replacement has dramatic impacts on the computational efficiency of the methods without sacrificing the accuracy of the parameter estimates.

The drawback of our approach is that R_{11} is not generally equal to K_{11}, but the advantage is that since the approximations can often be made very sharp with $\tau < 2$, the number of imputed values required for the Monte Carlo methods is not large compared to the number of observed values, allowing the observed values to drive the parameter updates. We demonstrate the accuracy of this approximation for small τ in Section 3. The simulations in Section 4 show that the computational advantage of using fewer imputed values far outweighs the drawback of the approximate nature of the method, and in Section 5, we show that the overall computational effort is smallest when $f_\delta(\cdot)$ is modeled directly with flexible elementary parametric functions.

2.1 Computation of $R_{\boldsymbol{m}}$

In this subsection we discuss three methods for computing and approximating the covariances $R_{\boldsymbol{m}}(\boldsymbol{h})$. We show in Appendix C that

$$\begin{aligned} R_{\boldsymbol{m}}(\boldsymbol{h}) &= \sum_{\boldsymbol{j}\in\mathbb{Z}^d} K(\boldsymbol{h} + \delta\boldsymbol{j}\circ\boldsymbol{m}) \\ &= \lim_{N\to\infty}\sum_{j_1=-N}^{N-1}\cdots\sum_{j_d=-N}^{N-1} K((h_1+\delta j_1 m_1, \ldots, h_d + \delta j_d m_d)), \end{aligned} \tag{5}$$

so when $K(\boldsymbol{h})$ decays quickly with $\|\boldsymbol{h}\|$, where $\|\cdot\|$ denotes Euclidean distance, as in the exponential covariance function, $R_{\boldsymbol{m}}(\boldsymbol{h})$ is well-approximated by a truncation of (5). This truncation does not guarantee positive definiteness, but the commonly used covariance functions often decay very quickly with $\|\boldsymbol{h}\|$, so the resulting covariance matrices rarely fail to be positive definite with $N = 2$ or 3, especially when the components of $\boldsymbol{n}$ are large.

If the spectral density $f(\boldsymbol{\omega})$ is available in closed-form, and it decays quickly with $\|\boldsymbol{\omega}\|$, one may approximate $f_\delta(\boldsymbol{\omega_j})$ by truncating $\sum_{\boldsymbol{k}\in\mathbb{Z}^d} f(\boldsymbol{\omega_j}+2\pi\boldsymbol{k}/\delta)$. This approximation is guaranteed to generate a positive definite function. Most of the commonly-used covariance models in spatial statistics possess the property that either the spectral density or the covariance function decays quickly. The Matérn covariance function (Matérn, 1960), for example, decays faster than any polynomial, as do the spectral densities for the Gaussian and Cauchy covariance functions.

In Section 5 we show that our methods are particularly efficient when the aliased spectral density $f_\delta(\cdot)$ is modeled directly in closed form. Then the array of eigenvalues $f_\delta(\boldsymbol{\omega_j})$, $\boldsymbol{j}\in\mathbb{J}_{\boldsymbol{m}}$, can be formed directly without any truncations, and $R_{\boldsymbol{m}}$ is thus guaranteed to be positive definite. We recommend this modeling approach due to its computational advantages, and we discuss in Section 5 a parametric model for $f_\delta(\cdot)$ that mimics the flexibility of the Matérn covariance function.

3 Numerical Studies

In the previous section, we presented a periodic covariance approximation whose accuracy depends on an expansion factor τ. The numerical studies in this section concern the resulting approximate covariance matrix R_{11} and the effect that the number of observations, the strength of spatial correlation, and the choice of τ have on how well R_{11} approximates K_{11}. Since both the Bayesian and frequentist methods discussed in this paper are likelihood-based, we study the approximations with respect to Kullback-Leibler (KL) divergences. Let $\boldsymbol{Z}_1$ be the set of all observations on $\delta\mathbb{J}_{\boldsymbol{n}}$. Defining $L_\tau(\theta;\, \boldsymbol{Z}_1)$ to be the Gaussian loglikelihood function for $\boldsymbol{Z}_1$ under covariance function $R_{\boldsymbol{m}}(\cdot)$ with $\boldsymbol{m}=\tau\boldsymbol{n}$, and defining $L(\theta;\, \boldsymbol{Z}_1)$ to be the Gaussian loglikelihood function for $\boldsymbol{Z}_1$ under the target model, which has covariance function $K_\theta(\cdot)$, the KL divergence of our approximate model from the target model is

$$E_0(L(\theta_0;\, \boldsymbol{Z}_1) - L_\tau(\theta;\, \boldsymbol{Z}_1)), \tag{6}$$

where the expectation is taken with respect to the target model with parameter θ_0. The θ that maximizes $L_\tau(\theta;\, \boldsymbol{Z}_1)$ is consistent for θ^τ, the minimizer of (6), under replication of $\boldsymbol{Z}_1$ (Varin et al., 2011). Since $E_0(L(\theta_0;\, \boldsymbol{Z}_1))$ does not depend on θ, θ^τ is the minimizer of

$$E_0(-L_\tau(\theta;\, \boldsymbol{Z}_1)) = \frac{1}{2}\log\det R_{11} + \frac{1}{2}\mathrm{tr}(R_{11}^{-1}K_{11}),$$

where R_{11} implicitly depends on θ and τ, and K_{11} is formed using the covariance function $K_{\theta_0}(\cdot)$. In this section, we compute θ^τ for various choices of τ in two different asymptotic scenarios and for various values of true parameter θ_0. While we do not propose using maximizers of $L_\tau(\theta;\, \boldsymbol{Z}_1)$ as estimators in practice, the results of these computations are nonetheless useful for understanding how the quality of the approximate covariance functions depends on τ, the size of the observation lattice, and the strength of spatial correlation.

The numerical studies use the isotropic exponential covariance function $K_\lambda(\boldsymbol{h})=\exp(-\|\boldsymbol{h}\|/\lambda)$, where we refer to the parameter λ as the range parameter; increasing λ decreases the rate at which $K_\lambda(\boldsymbol{h})$ decays with $\|\boldsymbol{h}\|$. All of the studies assume a square two-dimensional lattice. The lattice in the first set of calculations has spacing $\delta=(32\sqrt{2})^{-1}$ and $\lambda_0=0.15$, and we specify lattice sizes of $n\in\{32^2,48^2,64^2,80^2\}$. Hence the spacing is fixed, and the number of locations increases; this is sometimes called increasing domain asymptotics. The covariance approximations are calculated with values of $\tau\in\{1,17/16,9/8,5/4,3/2,5\}$, and we recall that setting $\tau=1$ corresponds to the approximation implied by the Whittle likelihood. Choosing $\tau=5$ is intended to show how λ^τ behaves when R_{11} is a very good approximation to K_{11}. In Table 1, we present the results of the first numerical study. When $\tau=1$, $\lambda^\tau<\lambda_0$, although λ^τ increases with n. This is not surprising because we expect the Whittle likelihood to underestimate range parameters since the Whittle likelihood assumes periodic correlation on $\delta\mathbb{J}_{\boldsymbol{n}}$ when the true covariance function is not periodic at all. For

	n			
τ	32^2	48^2	64^2	80^2
1	0.1234	0.1310	0.1353	0.1380
17/16	0.1457	0.1484	0.1493	0.1496
9/8	0.1485	0.1495	0.1498	0.1499
5/4	0.1496	0.1499	0.1500	0.1500
3/2	0.1499	0.1500	0.1500	0.1500
5	0.1500	0.1500	0.1500	0.1500

Table 1: Numerical values of λ^τ for various choices of τ and n with constant lattice spacing $(32\sqrt{2})^{-1}$ and $\lambda_0 = 0.15$.

	λ_0				
τ	0.05	0.10	0.15	0.20	0.25
1	0.0482	0.0932	0.1353	0.1748	0.2120
17/16	0.0500	0.0997	0.1493	0.1986	0.2477
9/8	0.0500	0.0999	0.1498	0.1996	0.2494
5/4	0.0500	0.1000	0.1500	0.1999	0.2499
3/2	0.0500	0.1000	0.1500	0.2000	0.2500
5	0.0500	0.1000	0.1500	0.2000	0.2500

Table 2: Numerical values of λ^τ for various choices of τ and λ_0 for a lattice with 64^2 points and spacing $(32\sqrt{2})^{-1}$. The third column of this table is equivalent to the third column of Table 1

every n, λ^τ approaches λ_0 as τ increases, with λ^τ converging more quickly for larger n. For every $\tau > 1$, λ^τ approaches λ_0 as n increases. This last remark is an important one because it suggests that we obtain very accurate approximations with small τ when the number of observations is large, which is desirable because the methods are designed for analyzing very large datasets.

The second numerical study considers the behavior of λ^τ for various choices of λ_0 to understand how the performance of the approximation depends on the strength of the spatial correlation. We fix the lattice spacing at $\delta = (32\sqrt{2})^{-1}$ and the number of lattice locations at 64^2. We vary $\lambda_0 \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$. The results presented in Table 2 show that λ^τ approaches λ_0 as τ increases. We also observe that λ^τ converges faster to λ_0 when λ_0 is small, that is, when the spatial correlation is weak. In every case, $\tau = 3/2$ is large enough to ensure that λ^τ and λ_0 agree to four decimal places. In Figure 2, we plot the target covariance functions and several of the approximations with $\lambda = 0.25$. It is important to note that approximations need not be accurate at all spatial lags in order for λ^τ to be very close to λ_0. To say this more concretely with a specific example, when $\lambda_0 = 0.25$ and $\tau = 5/4$, $\lambda^\tau = 0.2499$ even though $R_{\boldsymbol{m}}(\delta(63,0)) = 0.2283$ is not close to the target covariance $K_{\lambda_0}(\delta(63,0)) = 0.0038$. This suggests that it is not necessary for $R_{\boldsymbol{m}}(\boldsymbol{h})$ to well approximate $K(\boldsymbol{h})$ at large lags in order for R_{11} to produce a likelihood function that returns accurate parameter estimates.

The third numerical study addresses how the approximations perform when the size of the spatial domain is fixed, and the lattice spacing decreases. This increasing resolution scenario is sometimes called fixed domain asymptotics. We use lattice spacing $(\sqrt{n}\sqrt{2})^{-1}$ with $n \in \{32^2, 48^2, 64^2, 80^2\}$, and we set $\lambda_0 = 0.15$. The results are reported in Table 3. When $\tau = 1$, λ^τ is again less than λ_0, but in this fixed domain scenario, λ^1 does not improve much as n increases; it changes from $\lambda^1 = .1234$ when $n = 32^2$ to $\lambda_1 = .1237$ when $n = 80^2$, as opposed to $\lambda^1 = .1380$ when $n = 80^2$ in the increasing domain scenario. This is perhaps not surprising because increasing the resolution does not necessarily give much more information about the range of spatial dependence. When $\tau > 1$, however, λ^τ does appear to be approaching λ_0 as n increases, even when τ is as small as 17/16. In that case $\lambda^{17/16} = 0.1457$ when $n = 32^2$ versus $\lambda^{17/16} = .1488$ when $n = 80^2$. The error $\lambda^{17/16} - \lambda_0$

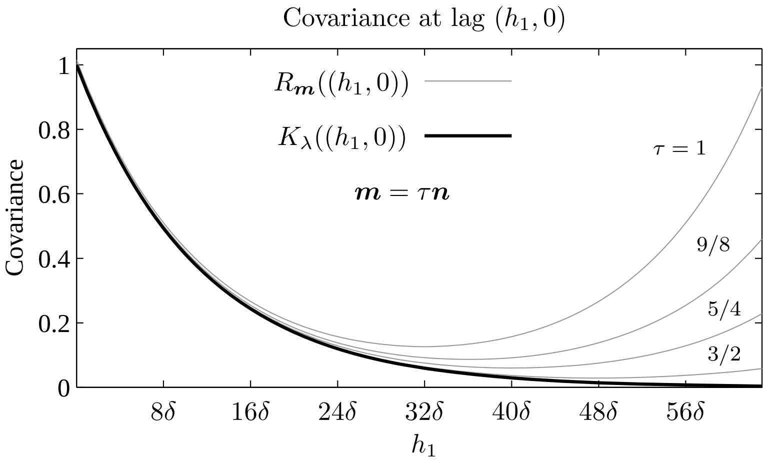

Figure 2: Target covariance function and approximations with several values of τ for $\lambda = 0.25$.

	n			
τ	32^2	48^2	64^2	80^2
1	0.1234	0.1235	0.1236	0.1237
17/16	0.1457	0.1474	0.1483	0.1488
9/8	0.1485	0.1492	0.1495	0.1497
5/4	0.1496	0.1498	0.1499	0.1499
3/2	0.1499	0.1500	0.1500	0.1500
5	0.1500	0.1500	0.1500	0.1500

Table 3: Numerical values of λ^τ for various choices of τ and n when the resolution of the lattice increases on a fixed domain. The lattice spacing is $(\sqrt{n}\sqrt{2})^{-1}$, and $\lambda_0 = 0.15$. The first column of this table is the same as the first column of Table 1.

decreases by 72% as we increase n from 32^2 to 80^2 versus a 1.3% decrease in the error $\lambda^1 - \lambda_0$ for the same increase in n. As before, λ^τ approaches λ_0 as τ increases in every case.

4 Simulations

The previous section showed that τ need not be large in order for the parameter minimizing the KL divergence to be close to the true parameter, especially when the number of observations is large. This section presents simulations that provide further support that we can obtain very accurate parameter estimates with small values of τ. Further, choosing a small value of τ defines a relatively small embedding lattice, which we show has a dramatic effect on the speed of convergence of the Monte Carlo inferential procedures. With small embedding lattices, the number of imputed values is not large relative to the number of observations, so the observations hold greater authority in driving the parameter updates in the iterative algorithms. Using a large embedding lattice leads to highly correlated Markov chains and slowly converging Monte Carlo EM algorithms. Standard circulant embedding requires an embedding lattice that is at least two times larger–and in practice often three or more times larger–than the observation lattice in each dimension. We show that using approximate covariances allows us to obtain accurate parameter estimates with τ as small as 1.25.

To demonstrate these points, we focus our simulation studies on the estimation of a single parameter in the powered exponential covariance function. In Section 5, we pursue the estimation of multiple parameters in the powered exponential covariance function, as well as in the Matérn covariance function for the photosynthetically available radiation data. The isotropic powered exponential covariance is a flexible and commonly used covariance function defined by

$$K(\boldsymbol{h}) = \sigma^2 \exp\big(- (\|\boldsymbol{h}\|/\lambda)^\alpha \big) + \gamma \mathbf{1}(\boldsymbol{h} = 0),$$

where $\sigma^2, \lambda, \gamma > 0$ and $\alpha \in (0, 2]$, with λ interpreted as a range parameter, γ as a nugget parameter, and α controlling the local behavior of $K(\cdot)$.

4.1 Two-dimensional simulations

We simulate 100 spatial data sets on a lattice of size $\boldsymbol{n} = (32, 32)$ with spacing $(32\sqrt{2})^{-1}$ from a mean-zero Gaussian process model with powered exponential covariance function with $\sigma^2 = 4$, $\lambda = 0.1$, $\alpha = 1$ (exponential model), and $\gamma = 0.01$. We focus on the estimation of α and specify its prior to be uniform over $(0, 2]$. The lattice dimensions, covariance model and parameters, and prior are the same as those used in Stroud et al. (2014).

The MCMC procedure consists of Metropolis-Hastings updates of α, where the acceptance probabilities are tuned to 0.50 during a burn-in period of 1000 iterations. The lattice contains $n = 1024$ observation locations, so MCMC estimation using the exact Gaussian likelihood is feasible. When using the exact likelihood, no imputations are necessary. Cutoff embedding is achieved with the same procedures outlined in Stroud et al. (2014), which give an embedding lattice of size $\boldsymbol{m} = (96, 96)$. Circulant embedding with approximate covariances is carried out using the methods in Section 2 with various choices of expansion parameter $\tau \in \{17/16, 9/8, 5/4, 3/2, 2, 3\}$, giving embedding lattices of size $\boldsymbol{m} = (32\tau, 32\tau)$. The choice of $\tau = 3$ matches the amount of imputation used in the cutoff embedding procedure. We compute $R_{\boldsymbol{m}}(\boldsymbol{h})$ using a truncation of the wrapping of $K(\cdot)$ with $N = 3$, which always produced positive definite covariance matrices.

We write $\alpha_E(k, j)$ to denote the sample mean after k post-burn-in iterations in the exact likelihood chain for the jth simulated data set, $\alpha_C(k, j)$ to denote the corresponding mean using cutoff embedding, and $\alpha_A(k, j, \tau)$ to denote the corresponding mean using embedding of approximate covariances with expansion factor τ. To evaluate the various procedures, we compare the root mean squared deviations from the exact likelihood estimate at 10,000 iterations,

$$\left(\frac{1}{100} \sum_{j=1}^{100} (\alpha_C(k, j) - \alpha_E(10000, j))^2 \right)^{1/2} \quad \text{and} \quad \left(\frac{1}{100} \sum_{j=1}^{100} (\alpha_A(k, j, \tau) - \alpha_E(10000, j))^2 \right)^{1/2},$$

for various choices of k and τ.

Table 4 includes results of this comparison. For small numbers of iterations, $k < 2000$, embedding with approximate covariances outperformed cutoff embedding with respect to this metric, even when τ is as small as $17/16 = 1.0625$. At 6000 iterations, the parameter deviation using cutoff embedding was no better than that of the approximate procedures with $\tau \in \{9/8, 5/4\}$ at 1000 iterations, an iteration speedup of 6 times at this tolerance. Even at 10,000 iterations, cutoff embedding still underperformed compared to approximate procedures, which were better at just 3000 iterations with $\tau \in \{9/8, 5/4\}$. An interesting result is that the deviations using cutoff embedding and embedding of approximate covariances with $\tau = 3$ were roughly equal at every number of iterations. This is evidence that the slow convergence of cutoff embedding procedures can be attributed to the increased number of imputed values, since setting $\tau = 3$ produces an embedding lattice that is the same size as that used in cutoff embedding.

The results in Table 4 can be explained by considering the autocorrelation in each chain and the resulting effective sample sizes of the correlated Markov chains. To begin the exploration of this issue, we include in Figure 3 trace plots of a Markov chain that used cutoff embedding and a Markov chain that used circulant embedding with approximate covariances ($\tau = 5/4$) for one of the

		Approximate					
	Cutoff	τ					
k	Embed	17/16	9/8	5/4	3/2	2	3
1000	60	42	25	24	26	42	59
2000	41	38	20	20	22	30	39
3000	34	37	16	16	18	25	32
4000	29	37	17	15	19	21	29
5000	27	36	16	15	16	19	26
6000	25	36	16	14	15	17	24
7000	23	35	16	13	13	16	22
8000	21	35	16	12	13	16	21
9000	20	35	15	12	12	15	20
10000	18	35	15	11	12	14	19

Table 4: Root mean squared deviations from exact likelihood estimates after k MCMC iterations with $d = 2$. We report the deviations for the various methods multiplied by 10,000.

simulated data sets. Both chains started at an initial value of $\alpha = 1.2$. In this example, we do not see an appreciable difference in mixing time–the number of iterations until the chain began to oscillate around its mean–although we sometimes did see faster mixing in the chains that used approximate covariances. On the other hand, we do see a noticeable difference in the two chains' autocorrelations; the chain that used approximate covariances with $\tau = 5/4$ appears to be less correlated with itself than does the chain that used cutoff embedding

The qualitative behavior we see in Figure 3 can be made more formal by analyzing the empirical autocorrelations among the chains for the various methods and the 100 simulated datasets. To show that the example in Figure 3 is not an isolated one, we plot in Figure 4 histograms of empirical lag 1 autocorrelations among the 100 chains and various embedding approaches. The empirical lag 1 correlations were computed using 10,000 paramter iterates after a burn-in period of 1000 iterations. The exact likelihood Markov chains, which used no imputation, offered the smallest lag 1 correlations. The lag 1 correlation increased as we increased τ in the approximate procedures. Cutoff embedding and embedding with our approximate procedure with $\tau = 3$ produce essentially the same lag 1 correlations. This is strong evidence that the amount of autocorrelation in the Markov chains is well explained by the number of imputed observations

The strength of correlation in a Markov chain is directly related to the computational effort required for MCMC parameter estimation. Effective sample size is a useful measure intended to summarize how much information about a parameter is contained in a correlated Markov chain. The variance of the sample average of parameter iterates in a Markov chain of length k is given by $(\mathbf{1}'\Sigma\mathbf{1})/k^2$, where Σ is the covariance matrix of the parameter iterates, and $\mathbf{1}$ is a vector of ones. As k increases, the variance of the sample average tends to

$$\frac{\text{Var}\left(\alpha^{(i)}\right)}{k} \sum_{j=\infty}^{\infty} \text{Corr}\left(\alpha^{(i)}, \alpha^{(i+j)}\right),$$

which does not depend on i for a stationary chain. The effective sample size is defined as

$$ESS(k) = \frac{k}{\sum_{j=-\infty}^{\infty} \text{Corr}\left(\alpha^{(i)}, \alpha^{(i+j)}\right)},$$

which is equal to the actual sample size k if the chain is uncorrelated. Therefore, effective sample size is a measure of how much information the sample average of a correlated Markov chain contains relative to a sample average of uncorrelated draws from the posterior distribution.

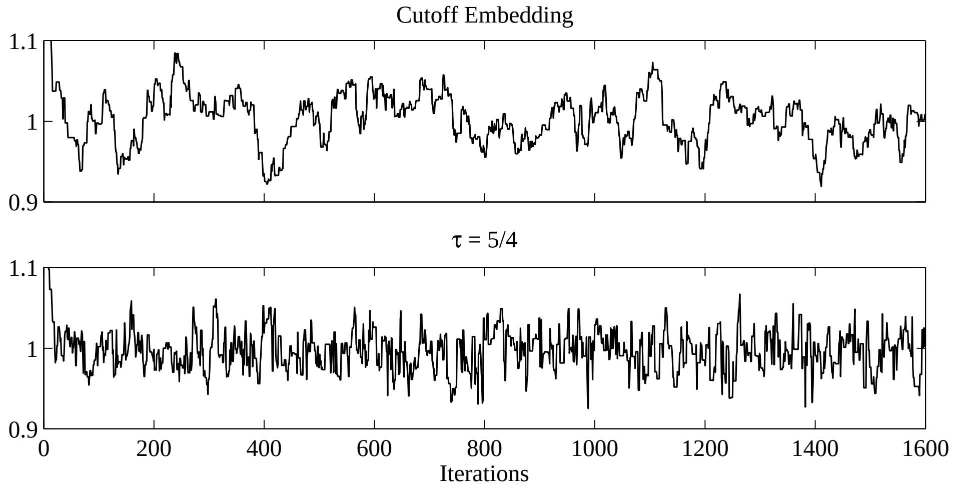

Figure 3: Trace plots of Markov chains for α with cutoff embedding and with our approximate methods for $\tau = 5/4$.

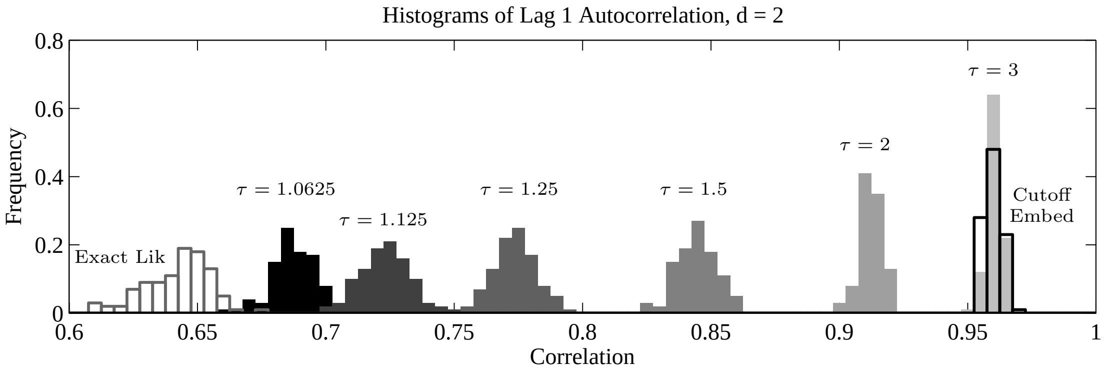

Figure 4: Histograms of lag 1 correlation in the Markov chains among 100 simulations and various methods for $d = 2$ simulations.

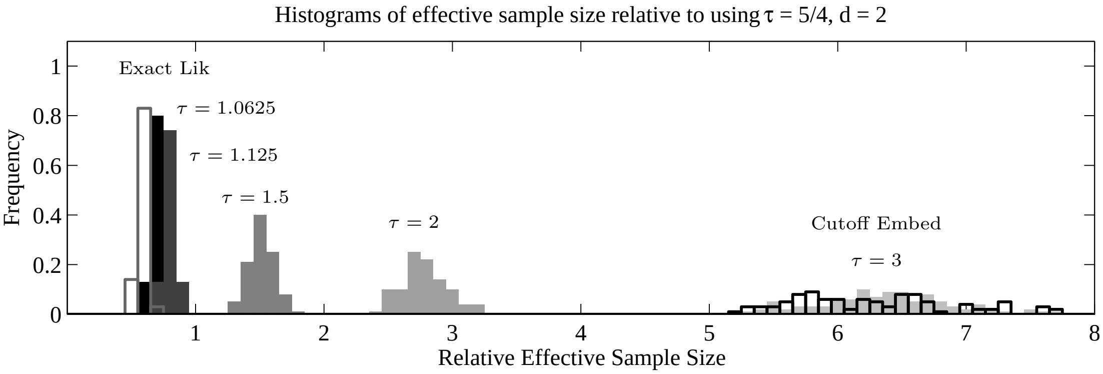

Figure 5: Histograms of effective sample size in the Markov chains relative to approximate methods with $\tau = 5/4$ among 100 simulations and various methods for the $d = 2$ simulations.

If we assume that the correlation between successive parameter iterates decays exponentially, as is the case for an autoregressive model of order 1, the effective sample size is $k(1-\rho)/(1+\rho)$, where ρ is the lag 1 correlation. We define the relative effective sample size of two autocorrelated chains as $ESS_1(k)/ESS_2(k)$, which does not depend on k. This is a measure of the relative computational cost of reaching a desired number of effective samples. In Figure 5, we plot histograms consisting of the relative effective sample sizes of the various methods among the 100 simulated datasets, where we take as a reference the effective sample size of the approximate procedures with $\tau = 1.25$. The histograms in Figure 5 represent the proportional changes (relative to the approximate procedure with $\tau = 1.25$) in the number of MCMC samples required to reach a desired number of effective samples. As we see from Figure 5, compared to our approximate procedures, cutoff embedding procedures required between 5 and 8 times the number of iterations to reach a desired number of effective samples, a range that roughly agrees with the analysis reported in Table 4.

Using the same 100 simulated datasets, we implement the Monte Carlo EM algorithm proposed in Stroud et al. (2014). The Monte Carlo EM algorithm specifies M, the number of conditional simulations over which the loglikelihood is averaged in each iteration of the algorithm. To see the effect of the choice of M, we use $M = 20$ and $M = 100$ to analyze each simulated data set. Since this is a Monte Carlo EM algorithm, the parameter iterates do not converge to any particular value. For this reason, we suggest estimating parameters by averaging the parameter iterates after a "burn-in" period has concluded. In this simulation study, since Cholesky decompositions of the exact covariance matrix can be stored in memory, it is possible to obtain maximum likelihood estimates with standard procedures. Thus we can compare root mean squared differences between the maximum likelihood estimates and the Monte Carlo EM estimates found using various choices of burn-in iterations and averaging iterations. The results for cutoff embedding and our approximate procedures with $\tau = 1.5$ are plotted in Figure 6. We see that the Monte Carlo EM algorithm with cutoff embedding requires more burn-in iterations for the estimates to stabilize. Even if we set the burn-in time to 100 iterations, the estimates found using approximate covariances converge faster to the maximum likelihood estimates; when $M = 20$, the approximations need only 50 iterations in order for the root mean squared differences to fall below 10^{-3}, whereas cutoff embedding needs 200 iterations to reach this tolerance. Thus, our approximate procedures provide remarkable reductions in the number of iterations required for both burn-in and averaging in the Monte Carlo EM algorithm.

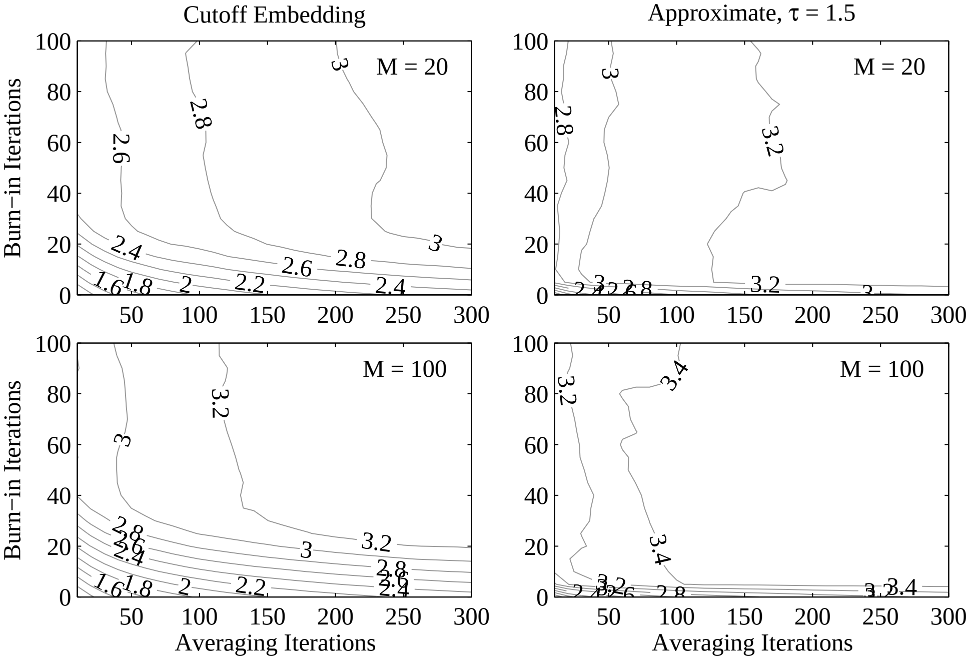

Figure 6: Contour plots of $-\log_{10}$ of root mean squared deviations between Monte Carlo EM estimates and maximum likelihood estimates of α, where the mean squared differences are taken over the 100 simulated datasets. The approximate covariances use $\tau = 1.5$.

4.2 Three-dimensional Simulations

When $d = 3$, the number of imputed values required in the imputation-based procedures is generally larger than when $d = 2$, since $m = n\tau^d$, where τ is the factor by which the lattice is expanded in each dimension. For this reason, our methods for reducing the size of τ are especially important for computations in three and higher dimensions. We include here results from a simulation study where $\boldsymbol{n} = (10, 10, 10)$, giving $n = 1000$ lattice locations with spacing $\delta = (10\sqrt{3})^{-1}$. We simulate from a powered exponential covariance model with $\sigma^2 = 4$, $\lambda = 0.1$, $\alpha = 1$, and $\gamma = 0.01$, and use the MCMC to estimate α assuming that all of the other parameters are known. A straightforward way to apply cutoff embedding in three dimensions is to define

$$\widetilde{K}_\theta(\boldsymbol{h}) = \begin{cases} K_\theta(\boldsymbol{h}) & \|\boldsymbol{h}\| \le 1 \\ b(\|\boldsymbol{h}\|) & 1 < \|\boldsymbol{h}\| < 1 + \varepsilon \\ b(1+\varepsilon) & \|\boldsymbol{h}\| \ge 1 + \varepsilon \end{cases} .$$

We use $\widetilde{K}_\theta(\boldsymbol{h})$ to define covariances for observations within $\delta\mathbb{J}_{(1+\varepsilon)\boldsymbol{n}}$ and then perform three-dimensional minimum embedding of those covariances. We set $\varepsilon = 1.5/\sqrt{2}$, which is the same value that was used in two-dimensional cutoff embedding, giving $\boldsymbol{m} = (35, 35, 35)$, which corresponds to setting $\tau = 3.5$ in our approximate procedures. This defines an embedding lattice that contains an absurdly large 41,875 locations at which to impute the data at each iteration compared to 1,000 actual observations, so we expect cutoff embedding to produce highly correlated Markov chains and slowly converging Monte Carlo EM algorithms in three dimensions. We compare cutoff embedding to our approximate procedures with $\tau \in \{1, 1.2, 1.4, 1.6, 2, 3.5\}$ in MCMC. In contrast to cutoff embedding, setting $\tau = 1.6$ in our approximate procedures requires just 3,096 imputed values at each iteration. In Table 5 we report the root mean squared deviations between the various estimates and

		Approximate				
	Cutoff	τ				
k	Embed	1.2	1.4	1.6	2	3.5
1000	231	150	97	83	94	240
2000	169	149	82	62	76	192
3000	155	147	77	57	63	157
4000	139	149	74	53	55	134
5000	114	149	72	50	49	114
6000	112	150	72	51	45	110
7000	103	150	71	49	42	101
8000	95	150	72	48	42	91
9000	94	150	71	47	40	90
10000	90	150	71	47	37	82

Table 5: Root mean squared deviations from exact likelihood estimates after k MCMC iterations with $d = 3$. We report the deviations for the various methods multiplied by 10,000.

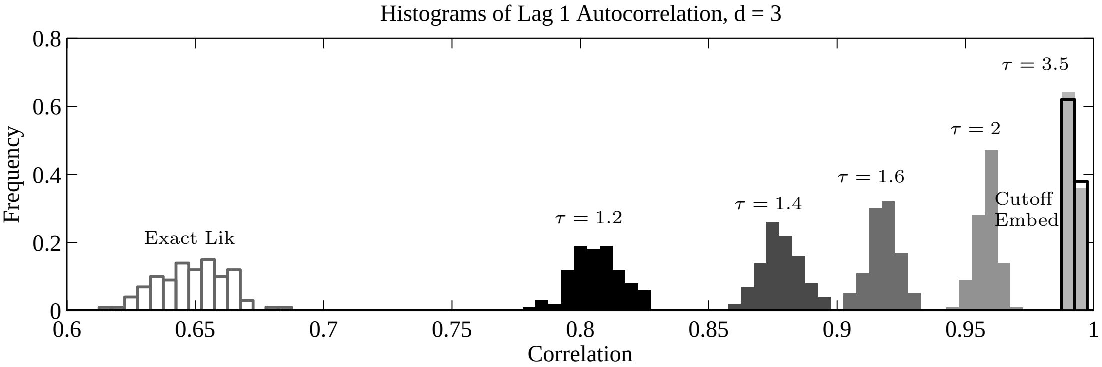

Figure 7: Histograms of lag 1 correlation in the Markov chains among 100 simulations and various methods for $d = 3$ simulations.

the estimate found using MCMC with the exact likelihood (no imputations) after 10,000 post burn-in iterations. For $\tau \in \{1.4, 1.6, 2\}$, circulant embedding with approximate covariances outperforms exact embedding at every number of iterations, and setting $\tau = 1.6$ produces smaller root mean squared deviations in 1,000 iterations than does cutoff embedding after 10,000 iterations. Setting $\tau = 3.5$ roughly recovers the deviations obtained with exact circulant embedding.

In Figure 7, we plot the empirical lag 1 correlations for the various embedding procedures, and in Figure 8, we plot the effective sample sizes relative to using approximate covariances with $\tau = 1.6$ based on an AR(1) model for the parameter iterates. All of the lag 1 correlations–aside the exact likelihood MCMC–were larger than they were in the $d = 2$ case, and cutoff embedding required between 8 and 14 times the number of iterations that the approximate procedure with $\tau = 1.6$ required to reach a desired number of effective samples, which is consistent with our observation that cutoff embedding required 10,000 iterations to achieve the accuracy that our approximate procedures achieved in 1,000 iterations.

In summary, the simulations in this section showed that circulant embedding with approximate covariances can provide substantial computational gains over cutoff embedding without sacrificing the accuracy of parameter estimates. Compared to cutoff embedding in two dimensions, our approximate procedures increased the number of effective samples per MCMC iteration by at least a

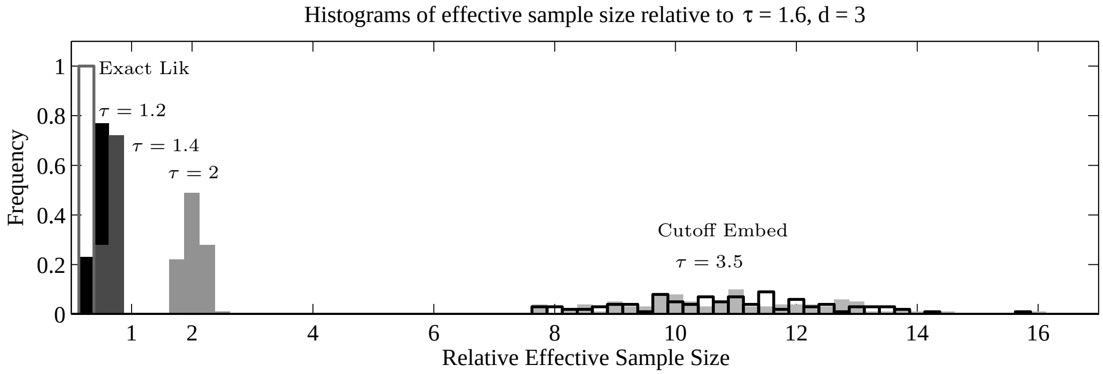

Figure 8: Histograms of effective sample sizes in the Markov chains relative to using $\tau = 1.6$ among 100 simulations and various methods for $d = 3$ simulations.

factor of five when $\tau = 5/4$, and in three dimensions, the increase was at least a factor of eight when $\tau = 1.6$. The Monte Carlo EM algorithms that used approximate covariances also burned in and converged in fewer iterations. In the following section, we apply the various embedding techniques to satellite data and provide timing results to show that the approximate procedures also serve to reduce the computing time required for each iteration.

5 Photosynthetically Available Radiation Data

Aqua is NASA satellite mission whose central aim is to collect information about Earth's water cycle. As is typical of most polar-orbiting satellites, Aqua's measurements do not attain complete global coverage on short time scales; a typical daily map of Aqua data contains large swaths of missing values at locations over which Aqua did not orbit. Our goal in this section is to provide complete spatial maps of a quantity called photosynthetically available radiation (PAR) over a region for which there are a substantial numer of missing values. PAR, which is detected by the Moderate Resolution Imaging Spectrometer (MODIS), quantifies the abundance of light at wavelengths between 400 and 700 nm, the spectral range of radiation that organisms use in photosynthesis, and thus is an important quantity affecting biological systems. In Figure 9, we plot a map of a daily gridded data product of PAR values located west of Mexico's Baja California peninsula. The data can be downloaded from http://oceancolor.gsfc.nasa.gov, and this particular dataset is from December 1, 2013. PAR values derived from Aqua's measurements are reported only over the oceans. There is a triangular region of missing observations, as well as a few missing along the coasts. We aim to interpolate the missing observations with values that match the statistical properties of the observed process to obtain physically plausible reconstructions of PAR. To accomplish this, we use the conditional simulations of the missing values that are required as part of the computationally efficient estimation methods presented in this paper, and we report an ensemble of the conditional simulations to provide accurate indications of the uncertainty in the interpolations.

The PAR lattice presented in Figure 9 contains 120 evenly-spaced longitude values and 100 evenly-spaced latitude values at a resolution of 1/12° in both latitude and longitude, for a total of 12,000 total lattice locations. There are 2,412 lattice locations for which PAR is missing, due either to the pixel being a land pixel or the value being genuinely missing, giving 9,588 observed PAR values. The data do not possess any obvious deviations from the isotropic Gaussian assumption, nor are there any discernible trends in the data. After subtracting the empirical mean of the observations, we consider three covariance models for PAR anomalies: (1) mean-zero isotropic powered exponential covariance with zero nugget and unknown $(\sigma^2, \lambda, \alpha)$, (2) mean-zero isotropic Matérn covariance with

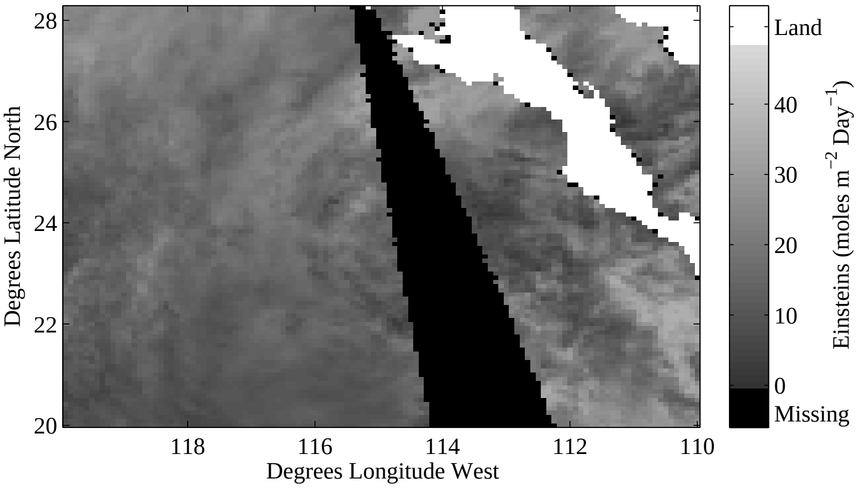

Figure 9: Grayscale map of photosynthetically available radiation from December 1, 2013. Missing values are black, and land pixels are white; the diagonal strip of land is Mexico's Baja California peninsula. The resolution is 1/12° in both latitude and longitude.

zero nugget,

$$K_\theta(\boldsymbol{h}) = \sigma^2 \frac{(\|\boldsymbol{h}\|/\lambda)^\nu \mathcal{K}_\nu(\|\boldsymbol{h}\|/\lambda)}{2^{\nu-1}\Gamma(\nu)},$$

and unknown (σ^2, λ, ν), and (3) a mean-zero model with aliased spectral density

$$f_\delta(\boldsymbol{\omega}) = \sigma^2 c_{\lambda,\nu} \left[\left(1 + \left(\frac{\lambda}{\delta}\right)^2 \left(\sin^2\left(\frac{\delta\omega_1}{2}\right) + \sin^2\left(\frac{\delta\omega_2}{2}\right)\right)\right)^{-\nu-1} + \gamma \right]. \tag{7}$$

We refer to the model in (7) as the quasi Matérn model due to its similarity to the Matérn spectral density and give an asymptotic justification for this name in Appendix B, along with a more general specification. The parameter σ^2 controls the variance of the process, λ can be interpreted as a range parameter, ν interpreted as a smoothness parameter, and γ as a nugget, which we set to zero in this analysis. The coefficient $c_{\lambda,\nu}$ is a normalizing constant, computed numerically. The quasi Matérn is defined in terms of its aliased spectral density, and thus no wrapping of covariances or spectral densities is required for the computations described in this paper; we simply evaluate $f_\delta(\boldsymbol{\omega})$ at the Fourier frequencies associated with $\boldsymbol{m}$ and transform the resulting array of spectral density values with an inverse DFT to obtain the associated covariances. We assume (σ^2, λ, ν) are unknown.

We implement the Bayesian MCMC methods to estimate the parameters in all models. To simplify notation across the models, we define $\boldsymbol{\theta} = (\lambda, \alpha)$ if the model is the powered exponential or $\boldsymbol{\theta} = (\lambda, \nu)$ if the model is either the Matérn or quasi Matérn. We specify prior $\pi(\sigma^2, \boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta})/\sigma^2$, where $\pi(\boldsymbol{\theta}) = 1/2(1 + \lambda/2)^{-2}$ in the powered exponential model, which places a uniform prior on α over $(0, 2)$, and $\pi(\boldsymbol{\theta}) = 1/4(1 + \lambda/2)^{-2}(1 + \nu/2)^{-2}$ in the Matérn and quasi Matérn models. We update $\boldsymbol{\theta}$ with a MH algorithm with a bivariate normal proposal distribution on the log scale. The posterior $\pi(\sigma^2|\boldsymbol{\theta}, \boldsymbol{Z})$ is inverse gamma $IG((m-1)/2, S^2(\boldsymbol{\theta})/2)$, where $S^2(\boldsymbol{\theta}) = \boldsymbol{Z}'C(\boldsymbol{\theta})^{-1}\boldsymbol{Z}$, and $C(\boldsymbol{\theta})$ is the correlation matrix corresponding to parameter vector $\boldsymbol{\theta}$. This is the standard conjugate family for variance parameter σ^2. The bivariate lognormal proposal distribution for λ and ν is tuned to have acceptance probability of 0.5 during 5,000 burn-in iterations. In cutoff embedding,

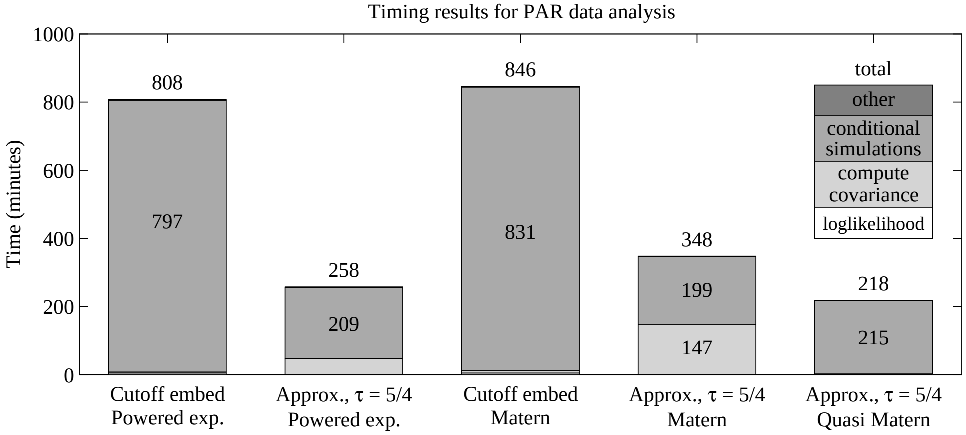

Figure 10: CPU time required to reach 50,000 MCMC iterations with various models and embedding methods, separated by three main computational tasks.

we use an embedding lattice of size $\boldsymbol{m} = (336, 336)$, which is highly composite and has 112,876 lattice locations. Our approximate methods are implemented with $\tau = 1.25$ for both the isotropic Matérn model and the quasi Matérn model, which gives embedding lattices of size $\boldsymbol{m} = (125, 150)$ and $m = 18,750$ total lattice locations. We wrap the isotropic Matérn covariance with $N = 3$, which always resulted in positive definite covariance matrices, and no wrapping is required for the quasi Matérn covariance.

The Markov chains are each run for 50,000 iterations, and with the help of Matlab's profiling capabilities, we record the cpu time attributed to the various computational tasks required for the MCMC. We report those results in Figure 10. All computations are completed with a machine running Matlab R2013a on an Intel Core-i7 2600 processor at 3.4GHz. The three most time-consuming tasks are the conditional simulations, the loglikelihoods required for evaluating the MH acceptance probabilities, and the construction of the covariance arrays. The approximate methods are faster overall than cutoff embedding, which required 808 minutes to reach 50,000 iterations with the powered exponential covariance and 846 minutes with the Matérn covariance, whereas the approximate methods required 258 minutes and 348 minutes for the powered exponential and the Matérn models. The approximate methods devoted a significant amount of time–147 minutes for the Matérn–to constructing the periodic covariance arrays, a consequence of the wrapping of the covariances. Cutoff embedding, on the other hand, required a relatively negligible amount of time–8 minutes for the Matérn–to constructing the covariance arrays. The approximate methods for the quasi Matérn also devote a negligible amount of time to constructing the covariance arrays and is the fastest overall of the three methods, taking 218 minutes to reach 50,000 iterations, nearly four times faster than the Matérn with cutoff embedding. All methods use a preconditioner corresponding to the submatrix of the complete data precision matrix and require roughly 50 iterations for the preconditioned conjugate gradient algorithms to converge on average. We experimented with several forms of a preconditioner based on the Stein et al. (2004) likelihood approximation, which can be made to converge in a smaller number of iterations but was slower overall in this instance.

We now investigate the quality of the fitted models in terms of exact loglikelihood. There are several ways to construct fitted models from the Markov chains, one of which is to compute $K_{\widehat{\theta}}(\boldsymbol{h})$, where $\widehat{\theta}$ as an average of the parameter iterates in the chain. We refer to this as the *averaged parameter* fitted model. For this estimate, we thin the chain, taking only every tenth iterate of the 45,000 post-burn-in iterations for the average. We also construct separate fitted models by averaging

		Loglikelihood	
Model	Method	Averaged Covariance	Averaged Parameter
Powered exp.	Cutoff Embedding	-51.75	-52.14
Powered exp.	Approximate, $\tau = 5/4$	-51.26	-52.26
Matérn	Cutoff Embedding	-0.34	-0.58
Matérn	Approximate, $\tau = 5/4$	0	-0.60
quasi Matérn	Approximate, $\tau = 5/4$	-1.28	-2.20

Table 6: Table of loglikelihoods for the three models and two methods. We compute the exact Gaussian loglikelihoods for the model estimate constructed by averaging the covariances and by averaging the parameters. Loglikelihood differences from that of the Matérn model fit with the approximate method are reported.

the covariances associated with the parameter iterates. Specifically, we compute

$$\widehat{K}(\boldsymbol{h}) = \frac{1}{N}\sum_i K_{\theta^i}(\boldsymbol{h}), \tag{8}$$

where the sum is over a thinned version of the Markov chain that has N iterations. We refer to the estimate in (8) as the *averaged covariance* fitted model. In cases where the loglikelihood has irregularly shaped contours, the averaged parameter model could differ substantially from the averaged covariances model. When using the quasi Matérn model we approximate $K_\theta(\boldsymbol{h})$ by discretizing the integral in (2) over a very fine grid with $\boldsymbol{m} = 4\boldsymbol{n}$, which is still efficient to compute with FFT algorithms. In Table 6, we include the exact loglikelihood values of the various model estimates. The Matérn and quasi matern models provide better fits than the powered exponential model in terms of loglikelihood. All of the fitted Matérn and quasi Matérn models agree to within a few loglikelihood units, which is negligible for a dataset of this size. It is actually quite remarkable that even though the quasi Matérn model is not equivalent to the Matérn model, in the sense that the aliased spectral density of the Matérn is not equal to the spectral density in (7), both fitted models give nearly the same loglikelihoods, an indication of the flexibility of the two models. Averaging covariances provided slightly better fits than averaging parameters for every model. The models obtained by approximate methods are roughly equal in terms of loglikelihood to the corresponding model estimate obtained by using cutoff embedding, so the approximations provided computational benefits without any sacrifice in the quality of the fitted models they produced.

Finally, in Figure 11, we plot conditional simulations of the PAR process over the ocean pixels of the observation region. The three conditional simulations use three different sets of parameters taken from the quasi Matérn Markov chain at iterations 10,000, 20,000, and 30,000, so the conditional simulations incorporate the uncertainty of the parameters. The conditional simulations produce PAR values over the land pixels as well, but we do not plot those since they are not reported in Aqua MODIS datasets. The PAR values in the three conditional simulations are exactly the same except for a few pixels along the coasts and the pixels in the triangular swath indicated by the thin black lines, which merge neatly with the observed values on the borders of the swath. The interpolated values also match the statistical properties of the rest of the dataset because they are simulated from a covariance model that is fit to the observed data. In scientific applications where it is necessary to have a complete map of PAR values as an input into larger model, the three (or possibly more) conditional simulations could be used to propagate the uncertainty associated with the interpolations and the fitted spatial model through the analysis. The methods discussed in this paper offer a way to produce an ensemble of complete interpolated maps in an computationally efficient manner.

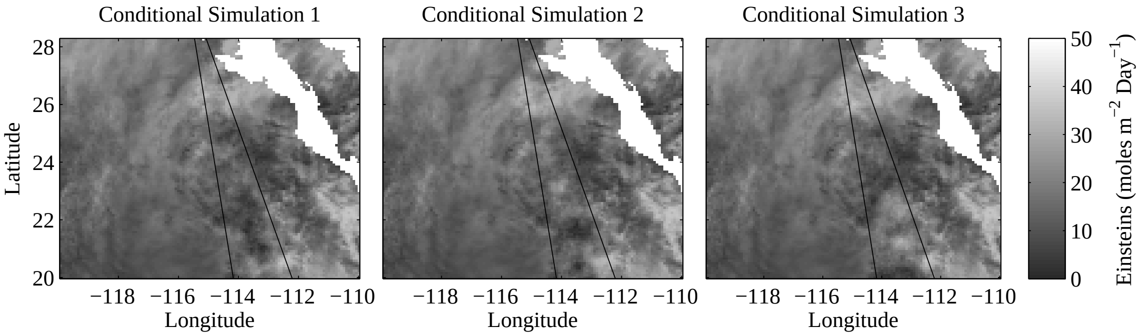

Figure 11: Three conditional simulations with Quasi Matérn covariance parameters taken from MCMC chain at iterations 10,000, 20,000, and 30,000.

6 Discussion

Numerical methods based on circulant embedding of covariance matrices are powerful tools for statistical computations involving lattice data and stationary covariance models. Not only do they avoid the $O(n^3)$ flops required for Cholesky decompositions of covariance matrices, they circumvent the need to store the $O(n^2)$ covariance matrices. The use of circulant embedding for simulation of stationary Gaussian processes on regular lattices has a mature history, and the recent work of Stroud et al. (2014) is an important step forward in uncovering ways to exploit circulant embedding for making inference from lattice data.

We demonstrate that, for inferential purposes, it is often advantageous to turn to approximations that reduce the size of the embedding lattice compared to what is required in standard circulant embedding. We present such an approximation that arises naturally from a discretization of the spectral density of a stationary covariance function and that reduces to the approximation implied by the Whittle likelihood in a special case. We show with numerical studies that the approximations can be made very sharp–especially for large lattices–with embeddings that are only a factor of 1.25 or 1.5 times the size of the original lattice in each dimension. Our simulation studies show that reducing the size of the embedding lattice decreases the amount of autocorrelation in the Markov chains used for Bayesian inference and speeds the convergence of the Monte Carlo EM algorithm without sacrificing the accuracy of the parameter estimates. In two-dimensional simulations, cutoff embedding required between 5 and 8 times more iterations to reach a given number of effective samples compared to the approximate methods with $\tau = 5/4$. In three dimensions, the speedup was more substantial; our approximate methods with $\tau = 1.6$ reached a desired number of effective samples 8 to 14 times more quickly than methods based on cutoff embedding. We demonstrate the usefulness of these methods for interpolating gridded satellite observations of photosynthetically available radiation. In a timing study, our approximate methods were also faster per iteration, with nearly a factor of four decrease in cpu time when the model is written in terms of its aliased spectral density.

Based on our numerical studies and simulations, we recommend setting $\tau = 5/4$ (or set each m_j to the smallest $m_j \geq 5/4n_j$ giving highly composite m_j) if the range of spatial correlation is less than half the size of the observation lattice. If the spatial correlation is stronger, we recommend setting $\tau = 3/2$. Extremely strong spatial correlation may require even larger embedding lattices to ensure that the approximations are sufficiently sharp, which could make cutoff embedding seem attractive, but we expect that in situations where the spatial correlation is very strong, cutoff embedding with $\boldsymbol{m} = 3\boldsymbol{n}$ will likely not produce positive definite covariance arrays. We expect that the embedding lattice required for cutoff embedding to produce positive definite covariance arrays will generally always be larger than the lattice required for the approximate methods to produce extremely sharp approximations. Making the previous conjecture more precise is an avenue for future work.

Several aspects of the material presented here can be easily generalized. The powered exponential and Matérn covariance functions are isotropic, but we do not require isotropy or even geometric anisotropy, only stationarity. The quasi Matérn model has a generalization to d dimensions, which we present in Appendix B. We also assumed that the lattice had equal spacing in every dimension, but the models and methods are easily generalized to situations with different spacing δ_j in each dimension j, although we still require regular spacing within each dimension. In this case, the aliased spectral densities are defined on $\prod_{j=1}^{d}[0, 2\pi/\delta_j]$. The analyses that we presented used a common expansion factor for each dimension, but this is not required, and one can see how it may be computationally advantageous to use a smaller expansion factor in a dimension that is very large or has weak correlation along that dimension. The methods are applicable for multivariate spatial data as well. Guinness et al. (2014) provide a framework for defining multivariate spatial lattice models in the spectral domain.

Acknowledgements

The authors are supported by the National Science Foundation's Research Network for Statistical Methods for Atmospheric and Oceanic Sciences, award number 1107046.

A Nested Block Circulant Matrices

A matrix is block circulant if it can be written in block form as

$$K = \begin{bmatrix} K_0 & K_1 & K_2 & \cdots & K_{n-1} \\ K_{n-1} & K_0 & K_1 & \cdots & K_{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ K_1 & K_2 & K_3 & \cdots & K_0 \end{bmatrix},$$

where the blocks are all the same size and may be of size 1, in which case we also say that the matrix is circulant. Nested block circulant matrices are defined recursively: a matrix is nested block circulant if it is block circulant, and each subblock K_j is also nested block circulant. Covariance matrices have the additional properties that they are symmetric and positive definite. Symmetry implies that K_0 is symmetric, and $K_{n-j} = K_j'$.

B Quasi Matérn covariance

The quasi Matérn model presented in Section 5 has a generalization to d dimensions. In this case, the spectral densities are defined on $[-\pi/\delta, \pi/\delta]^d$, and the model is

$$f_\delta(\boldsymbol{\omega}) = \sigma^2 \left(1 + \left(\frac{\alpha}{\delta}\right)^2 \left(\sum_{j=1}^{d} \sin^2\left(\frac{\delta\omega_j}{2}\right) \right) \right)^{-\nu-d/2}.$$

This expression converges pointwise as $\delta \to 0$ to

$$\sigma^2 \left(1 + \left(\frac{\alpha}{2}\right)^2 \|\boldsymbol{\omega}\|^2 \right)^{-\nu-d/2},$$

which is one parametric form for the spectral density of the isotropic Matérn covariance function, justifying the name quasi Matérn.

C Proofs

The following lemma is of particular use for approximating the covariance functions introduced in Section 2.

Lemma C.1. *If f is the continuous spectral density for K, then $R_{\boldsymbol{m}}(\boldsymbol{h}) = \sum_{\boldsymbol{j}\in\mathbb{Z}^d} K(\boldsymbol{h} + \delta \boldsymbol{j} \circ \boldsymbol{m})$.*

To simplify the notation, we set $\delta = 1$ in the proof. It is no more difficult to prove with arbitrary δ, only more cumbersome notationally.

Proof. We write:

$$\begin{aligned}\sum_{j_1=-N}^{N} \cdots \sum_{j_d=-N}^{N} & K((h_1 + j_1 m_1, \ldots, h_d + j_d m_d)) \\ &= \sum_{j_1=-N}^{N} \cdots \sum_{j_d=-N}^{N} \int_{[0,2\pi]^d} f(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}'\boldsymbol{h}} e^{i\omega_1 j_1 m_1} \cdots e^{i\omega_d j_d m_d} d\boldsymbol{\omega} \\ &= \int_{[0,2\pi]^d} f(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}'\boldsymbol{h}} \left(\sum_{j_1=-N}^{N} e^{i\omega_1 j_1 m_1} \right) \cdots \left(\sum_{j_d=-N}^{N} e^{i\omega_d j_d m_d} \right) d\boldsymbol{\omega} \qquad (9)\end{aligned}$$

Since the integrand in (9) is periodic in each dimension, integrating over $[0, 2\pi]^d$ is equivalent to integrating over

$$\prod_{k=1}^{d} \left[-\frac{\pi}{m_k}, 2\pi - \frac{\pi}{m_k} \right] = \bigcup_{\boldsymbol{\ell}\in\mathbb{J}_{\boldsymbol{m}}} \prod_{k=1}^{d} \left[\frac{2\pi}{m_k}(\ell_k - 1/2), \frac{2\pi}{m_k}(\ell_k + 1/2) \right] := \bigcup_{\boldsymbol{\ell}\in\mathbb{J}_{\boldsymbol{m}}} A_{\boldsymbol{\ell}},$$

so that the integral in (9) can be written as

$$\sum_{\boldsymbol{\ell}\in\mathbb{J}_{\boldsymbol{m}}} \int_{A_{\boldsymbol{\ell}}} f(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}'\boldsymbol{h}} \left(\sum_{j_1=-N}^{N} e^{i\omega_1 j_1 m_1} \right) \cdots \left(\sum_{j_d=-N}^{N} e^{i\omega_d j_d m_d} \right) d\boldsymbol{\omega}.$$

The quantities in parentheses converge to periodic delta functions with period $2\pi/m_k$ in ω_k (DLMF, Section 1.17(iii)). Since f is continuous,

$$\begin{aligned}&\lim_{N\to\infty} \sum_{\boldsymbol{\ell}\in\mathbb{J}_{\boldsymbol{m}}} \int_{A_{\boldsymbol{\ell}}} f(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}'\boldsymbol{h}} \left(\sum_{j_1=-N}^{N} e^{i\omega_1 j_1 m_1} \right) \cdots \left(\sum_{j_d=-N}^{N} e^{i\omega_d j_d m_d} \right) d\boldsymbol{\omega} \\ &= \frac{(2\pi)^d}{m} \sum_{\boldsymbol{\ell}\in\mathbb{J}_{\boldsymbol{m}}} f(\boldsymbol{\omega}_{\boldsymbol{\ell}}) e^{i\boldsymbol{\omega}_{\boldsymbol{\ell}}\boldsymbol{h}} = R_{\boldsymbol{m}}(\boldsymbol{h})\end{aligned}$$

where $\boldsymbol{\omega}_{\boldsymbol{\ell}} = (2\pi\ell_1/m_1, \ldots, 2\pi\ell_d/m_d)$. □

References

Grace Chan and Andrew TA Wood. Simulation of stationary Gaussian vector fields. *Statistics and Computing*, 9(4):265–268, 1999.

R Dahlhaus and H Künsch. Edge effects and efficient parameter estimation for stationary random fields. *Biometrika*, 74(4):877–882, 1987.

CR Dietrich and Garry Neil Newsam. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107, 1997.

DLMF. NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.8 of 2014-04-25, 2014. URL `http://dlmf.nist.gov/`.

Tilmann Gneiting, Hana Ševčíková, Donald B Percival, Martin Schlather, and Yindeng Jiang. Fast and exact simulation of large Gaussian lattice systems in r_2: Exploring the limits. *Journal of Computational and Graphical Statistics*, 15(3), 2006.

Joseph Guinness, Montserrat Fuentes, Dean Hesterberg, and Matthew Polizzotto. Multivariate spatial modeling of conditional dependence in microscale soil elemental composition data. *Spatial Statistics*, 9:93–108, 2014.

Xavier Guyon. Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69(1):95–105, 1982.

Finn Lindgren, Håvard Rue, and Johan Lindström. 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.

Bertil Matérn. *Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations.* Meddelanden fran statens Skogsforskningsinstitut, 1960.

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications.* CRC Press, 2005.

Michael L Stein. Fast and exact simulation of fractional Brownian surfaces. *Journal of Computational and Graphical Statistics*, 11(3):587–599, 2002.

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

Jonathan R Stroud, Michael L Stein, and Shaun Lysen. Bayesian and maximum likelihood estimation for Gaussian processes on an incomplete lattice. *arXiv preprint arXiv:1402.4281*, 2014.

Cristiano Varin, Nancy Margaret Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

Aldo V Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 297–312, 1988.

Peter Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

Andrew T. A. Wood and Grace Chan. Simulation of stationary Gaussian processes in $[0, 1]^d$. *Journal of Computational and Graphical Statistics*, 3(4):409–432, 1994.

A. M. Yaglom. *Correlation theory of stationary and related random functions.* Springer-Verlag, 1987.