JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2021, VOL. 30, NO. 2, 375–389
https://doi.org/10.1080/10618600.2020.1811103

Taylor & Francis
Taylor & Francis Group

Nonstationary Modeling With Sparsity for Spatial Data via the Basis Graphical Lasso

Mitchell Krock, William Kleiber, and Stephen Becker

Department of Applied Mathematics, University of Colorado–Boulder, Boulder, CO

ABSTRACT

Many modern spatial models express the stochastic variation component as a basis expansion with random coefficients. Low rank models, approximate spectral decompositions, multiresolution representations, stochastic partial differential equations, and empirical orthogonal functions all fall within this basic framework. Given a particular basis, stochastic dependence relies on flexible modeling of the coefficients. Under a Gaussianity assumption, we propose a graphical model family for the stochastic coefficients by parameterizing the precision matrix. Sparsity in the precision matrix is encouraged using a penalized likelihood framework—we term this approach the basis graphical lasso. Computations follow from a majorization-minimization (MM) approach, a byproduct of which is a connection to the standard graphical lasso. The result is a flexible nonstationary spatial model that is adaptable to very large datasets with multiple realizations. We apply the model to two large and heterogeneous spatial datasets in statistical climatology and recover physically sensible graphical structures. Moreover, the model performs competitively against the popular LatticeKrig model in predictive cross-validation but improves the Akaike information criterion score and a log score for the quality of the joint predictive distribution.

ARTICLE HISTORY
Received February 2019
Revised June 2020

KEYWORDS
Graphical lasso; Graphical model; Spatial basis functions

1. Introduction

Many modern spatial models express the stochastic variation component as a basis expansion with random coefficients. Low rank models, approximate spectral decompositions, multiresolution representations, stochastic partial differential equations, and empirical orthogonal functions all fall within this basic framework. The essential difference between these methods is the amount of modeling effort placed on the basis versus the coefficients.

We introduce a novel approach applicable to any model within this framework that allows for nonstationarity and easily adapts to large datasets using off-the-shelf popular basis models. The method allows for straightforward graphical interpretations of the conditional independence structure of the stochastic coefficients.

Most spatial statistical models for an observational process $Y(\mathbf{s})$ with $\mathbf{s} \in \mathbb{R}^d$ can be written

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + Z(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{1}$$

which decomposes the observations into a mean function μ, a spatially correlated random deviation Z, and a white noise process ε. Flexible models for the correlated deviation Z are necessary, and much research has been devoted to exploring classes of practical specifications.

In this work, we focus on a particularly popular framework where

$$Z(\mathbf{s}) = \sum_{j=1}^{\ell} c_j \phi_j(\mathbf{s}) \tag{2}$$

for ℓ fixed basis functions $\phi_1, \dots, \phi_\ell$ and stochastic coefficients $\mathbf{c} = (c_1, \dots, c_\ell)^{\mathrm{T}} \sim N(0, Q^{-1})$. Such a spatial basis expansion subsumes many popular approaches including discretized frequency domain models (Fuentes 2002; Bandyopadhyay and Lahiri 2009; Matsuda and Yajima 2009), empirical orthogonal functions (Cressie and Wikle 2011, chap. 5), low rank representations (Banerjee et al. 2008; Cressie and Johannesson 2008; Guhaniyogi and Banerjee 2018), multiresolution and wavelet representations (Nychka, Wikle, and Royle 2002; Nychka et al. 2015; Katzfuss 2017), and stochastic partial differential equation models (Bolin and Lindgren 2011; Lindgren, Rue, and Lindström 2011).

To set the stage, let us briefly describe some specific models of (2). Fixed rank kriging uses bisquare basis functions and sets ℓ to be relatively small compared to other approaches, with sample size as a reference, and Q^{-1} is not modeled but rather estimated by minimizing a squared Frobenius distance from a binned empirical covariance matrix (Cressie and Johannesson 2008). The more recent LatticeKrig model is a multiresolution model that places a large number of compactly supported basis functions with varying supports on a grid and specifies $\mathbf{c}$ as a Gaussian Markov random field (Nychka et al. 2015). Discretized frequency domain approaches and empirical orthogonal functions set the basis functions to be globally supported with independent coefficients (diagonal Q), mirroring the spectral representation theorem or the Karhunen–Loève expansion for stochastic processes. In this work, we attempt to relax the modeling assumptions on the structure of the stochastic coefficients $\mathbf{c}$ (equivalently Q) by assuming only that they arise from a Gaussian graphical model. We use the fact that the graph structure

CONTACT Mitchell Krock mitchell.krock@colorado.edu Department of Applied Mathematics, University of Colorado, Boulder, Boulder, CO 80309.

of a multivariate Gaussian is equivalent to the zero/nonzero pattern in the precision matrix Q (Rue and Held 2005). A major distinction of this work is that we do not specify the structure of the graph but instead try to infer its edges by estimating the entries of Q.

Our model setup is distinct from nonspatial methods that require direct realizations of $\mathbf{c}$ to estimate the structure of the undirected graphical model Q, such as the *graphical lasso*, which estimates Q by adding an ℓ_1 penalty to the negative log-likelihood to encourage sparsity, just as in lasso regression. In our problem, we observe realizations of $\sum_{j=1}^{\ell} c_j\phi_j(\mathbf{s}) + \varepsilon(\mathbf{s})$ and have the goal of fitting a graphical model to the (latent) vector $\mathbf{c}$. Thus, the graphical lasso and other methods for sparse inverse covariance estimation (Meinshausen and Bühlmann 2006; Cai, Liu, and Luo 2011) are not directly applicable in our setup. Latent variable graphical model selection has been studied in the context of directly observing a subset of the elements of $\mathbf{c}$ (Chandrasekaran, Parrilo, and Willsky 2012). To our knowledge, no one has considered the problem of estimating a latent graph given noisy realizations of $\Phi\mathbf{c}$ for an arbitrary basis matrix Φ. Penalized likelihood optimization for estimating a nonstationary covariance matrix appears in Nandy, Lim, and Maiti (2016), but their method considers a single realization of the process and regularizes the Cholesky factor of Q^{-1} rather than the precision matrix Q. Caveats of our approach are that it works best when multiple realizations of the observational process are available and that it can suffer from drawbacks of low rank models discussed in Stein (2014) depending on the choice of basis functions.

Extending the ℓ_1 penalization framework to the basis graphical lasso gives rise to a nonconvex optimization problem. We show it can be solved efficiently using an MM algorithm which amounts to iteratively solving the graphical lasso. We perform a relatively detailed simulation study to assess the algorithm and model's ability to recover unknown graphical structures, and also apply the method to two challenging large and heterogeneous datasets: the first a global reforecast dataset of surface temperature, and second a historical observational dataset of minimum temperatures over a portion of North America. The results from these real data applications suggest that our method can appropriately capture nonstationary spatial correlations with minimal modeling effort and produce better joint predictive distributions than a LatticeKrig competitor.

2. Methods

For ease of exposition, we suppose $\mu(\mathbf{s}) = 0$ in (1). Thus, the observational model is

$$Y(\mathbf{s}) = \sum_{j=1}^{\ell} c_j\phi_j(\mathbf{s}) + \varepsilon(\mathbf{s}). \quad (3)$$

We suppose we have m independent realizations $Y_1(\mathbf{s}), \ldots, Y_m(\mathbf{s})$ of the observational process at spatial locations $\mathbf{s} = \mathbf{s}_1, \ldots, \mathbf{s}_n$. Group a realization as $\mathbf{Y}_i = (Y_i(\mathbf{s}_1), \ldots, Y_i(\mathbf{s}_n))^{\mathrm{T}}$. A matrix representation of the model is

$$\mathbf{Y}_i = \Phi\mathbf{c}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, m, \quad (4)$$

where Φ is an $n \times \ell$ matrix with (i, j)th entry $\phi_j(\mathbf{s}_i)$, $\mathbf{c}_i = (c_{i1}, \ldots, c_{i\ell})^{\mathrm{T}}$ are m independent vectors of the stochastic coefficients, and $\boldsymbol{\varepsilon}_i = (\varepsilon_i(\mathbf{s}_1), \ldots, \varepsilon_i(\mathbf{s}_n))^{\mathrm{T}}$ are m independent realizations of the white noise process. The stochastic assumptions of our model are that ε_i is a mean zero white noise process with variance $\tau^2 > 0$, commonly referred to as the nugget effect in the geostatistical literature, and that $\mathbf{c}_i$ is a mean zero ℓ-variate Gaussian random vector with precision matrix Q. The zero structure of Q encodes the graphical model for $\mathbf{c}_i$.

The model (3) plays a crucial role in modern statistics: it is the framework for a variety of popular statistical techniques including factor analysis, principal component analysis, linear dynamical systems, hidden Markov models, and relevance vector machines (Roweis and Ghahramani 1999; Tipping 2001). In the spatial context, there are two main features that arise from using a model of the form (3): first, the resulting model of the spatial field is nonstationary, and second, common computations involving the covariance matrix can be sped up with particular choices for Φ and Q, in particular using compactly supported basis functions and a sparse precision matrix.

2.1. Basis Graphical Lasso

With m realizations $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ of (4), the negative log-likelihood can be written, up to multiplicative and additive constants, as

$$\log\det(\Phi Q^{-1}\Phi^{\mathrm{T}} + \tau^2 I_n) + \operatorname{tr}(S(\Phi Q^{-1}\Phi^{\mathrm{T}} + \tau^2 I_n)^{-1}), \quad (5)$$

where $S = \frac{1}{m}\sum_{i=1}^{m} \mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$ is an empirical covariance matrix.

Our goal is to estimate Q under the assumption that it follows a graphical structure. A few connections are worth noting. When $\Phi = I_\ell$ and $\tau^2 = 0$, (5) is, up to the regularization term, the graphical lasso problem studied in Friedman, Hastie, and Tibshirani (2008). In particular, this is equivalent to directly observing $\mathbf{c}_1, \ldots, \mathbf{c}_m$. The graphical lasso uses an ℓ_1 penalty to induce a graph structure on Q, from which we draw inspiration next. Our situation is substantially more complicated due to observational error and indirect observations of $\mathbf{c}_i$ that are modulated by Φ.

Our proposal is to estimate Q by minimizing a penalized version of (5). The estimator of Q, deemed the basis graphical lasso (BGL), is

$$\begin{aligned}\hat{Q} \in \underset{Q \succeq 0}{\arg\min}\; & \log\det(\Phi Q^{-1}\Phi^{\mathrm{T}} + \tau^2 I_n) \\ & + \operatorname{tr}(S(\Phi Q^{-1}\Phi^{\mathrm{T}} + \tau^2 I_n)^{-1}) + \|\Lambda \circ Q\|_1. \quad (6)\end{aligned}$$

The notation $Q \succeq 0$ indicates that Q must be positive semidefinite, and $\|\Lambda \circ Q\|_1 = \sum_{i,j} \Lambda_{ij}|Q_{ij}|$ is a penalty term that enforces sparsity on the elements of Q. Here, Λ_{ij} are nonnegative penalty parameters, with higher values in the matrix Λ encouraging more zeros in the estimate. In this article, we assume that the diagonal elements of Λ are zero, reflecting the fact that we are searching for sparsity in the off-diagonal elements of Q.

At first glance it is hard to determine whether the BGL (6) is convex or nonconvex. We address this issue in the next section. In either case, it will be difficult to work with the objective function presented in (6) due to the dependence on the spatial dimension n and the nested inverses surrounding Q. The following result's proof is in Appendix A.

Proposition 1. The minimizer of (6) is also the minimizer of

$$\log\det\left(Q+\tau^{-2}\Phi^{\mathrm{T}}\Phi\right)-\log\det Q \\ -\operatorname{tr}\left(\tau^{-4}\Phi^{\mathrm{T}}S\Phi\left(Q+\tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}\right)+\|\Lambda\circ Q\|_1. \quad (7)$$

After precomputing the matrix products $\Phi^{\mathrm{T}}\Phi$ and $\Phi^{\mathrm{T}}S\Phi$, evaluating (7) only involves $\ell\times\ell$ matrices. This is the essential computational strategy of fixed rank kriging and LatticeKrig which is harnessed in different ways: by making either ℓ noticeably smaller than n (fixed rank kriging), or by making ℓ large but ensuring the resulting matrices are sparse (LatticeKrig). It is important to observe that the $n\times n$ sample covariance matrix S does not need to be explicitly calculated in this procedure, so $\Phi^{\mathrm{T}}S\Phi$ can be computed more effectively than $\mathcal{O}(n^2\ell)$ naive matrix multiplication. Indeed, writing $S=\mathbf{Y}\mathbf{Y}^{\mathrm{T}}/m$ where $\mathbf{Y}$ has columns $\mathbf{Y}_1,\ldots,\mathbf{Y}_m$ leads to an $\mathcal{O}(n\ell m+\ell^2m)$ representation:

$$\Phi^{\mathrm{T}}S\Phi=\frac{1}{m}(\Phi^{\mathrm{T}}\mathbf{Y})(\mathbf{Y}^{\mathrm{T}}\Phi). \quad (8)$$

2.2. Optimization Approach

In the $\ell=1$ case, (7) is a univariate function that is twice differentiable on the positive real line. It is straightforward to select $\Phi^{\mathrm{T}}\Phi$, $\Phi^{\mathrm{T}}S\Phi$, and τ^2 so that the second derivative has a negative value at some point along the positive real line. Thus (7) is, in general, a nonconvex function on $Q\succeq 0$. We can, however, show that the four summands in (7) are concave, convex, concave, and convex, respectively, on $Q\succeq 0$; see Appendix A for details. Therefore, the objective function in (7) can be written as

$$\underset{Q\succeq 0}{\arg\min} f(Q)+g(Q)+\|\Lambda\circ Q\|_1, \quad (9)$$

where $f(Q)+\|\Lambda\circ Q\|_1$ is convex and $g(Q)$ is concave and differentiable. A natural approach for this nonconvex problem is a difference-of-convex (DC) program (Dinh Tao and Le Thi 1997) where we iteratively linearize the concave part $g(Q)$ at the previous guess Q_j and solve the resulting convex problem:

$$Q_{j+1}=\underset{Q\succeq 0}{\arg\min}\left(f(Q)+\operatorname{tr}(\nabla g(Q_j)Q)+\|\Lambda\circ Q\|_1\right). \quad (10)$$

Motivating the DC framework requires describing a more general scheme called an *MM algorithm* (Hunter and Lange 2004). We say that a function $h(\theta)$ is majorized by $m(\theta\mid\theta^*)$ at θ^* if $h(\theta)\le m(\theta\mid\theta^*)$ for all θ and $h(\theta^*)=m(\theta^*\mid\theta^*)$. Instead of directly minimizing $h(\theta)$, which can be very complicated, an MM algorithm solves a sequence of minimization problems where the majorizing function at the previous guess is minimized:

$$\theta_{j+1}=\underset{\theta}{\arg\min}\ m(\theta\mid\theta_j). \quad (11)$$

Combining (11) with the definition of a majorant yields the inequality

$$h(\theta_{j+1})\le m(\theta_{j+1}\mid\theta_j)\le m(\theta_j\mid\theta_j)=h(\theta_j)$$

and thus the algorithm is forced to a local minimum (or saddle point) of $h(\theta)$. The most famous instance of MM in statistics is the expectation-maximization (EM) algorithm, which under this framework uses Jensen's inequality to construct majorizing functions for the conditional expectation of log-likelihood equations.

DC programming, also called the concave-convex-procedure, is a subclass of MM where the supporting hyperplane inequality $g(\theta)\le g(\theta_j)+\langle\nabla g(\theta_j),\theta-\theta_j\rangle$ is used to construct a majorizing function when $h(\theta)$ is written as the sum of a concave differentiable function $g(\theta)$ and a convex function; that is, when $h(\theta)$ is a difference of convex functions. An added benefit under the DC framework is that the majorizing function is convex by construction and hence we solve a series of convex optimization problems in each step of (11).

In our likelihood function, the convex part is

$$f(Q)=-\log\det Q$$

and the concave part is

$$g(Q)=\log\det\left(Q+\tau^{-2}\Phi^{\mathrm{T}}\Phi\right) \\ -\operatorname{tr}\left(\tau^{-4}\Phi^{\mathrm{T}}S\Phi\left(Q+\tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}\right),$$

so the DC algorithm (10) becomes

$$Q_{j+1}=\underset{Q\succeq 0}{\arg\min}\left(-\log\det Q+\operatorname{tr}\left(\nabla g(Q_j)Q\right)+\|\Lambda\circ Q\|_1\right), \quad (12)$$

where $\nabla g(Q_j)=\left(Q_j+\tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}\left(Q_j+\tau^{-2}\Phi^{\mathrm{T}}\Phi+\tau^{-4}\Phi^{\mathrm{T}}S\Phi\right)\left(Q_j+\tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}$. The inner minimization problem in (12) is well-studied and known in statistics as the *graphical lasso* problem. Traditionally, the graphical lasso is used to estimate an undirected graphical model of a multivariate Gaussian vector $\mathbf{c}$ under the assumption that we observe $\mathbf{c}$ directly, without noise. The standard graphical lasso estimate is obtained from the penalized negative log-likelihood

$$\underset{Q\succeq 0}{\arg\min}\ -\log\det Q+\operatorname{tr}(S_{\mathbf{c}}Q)+\|\Lambda\circ Q\|_1, \quad (13)$$

where $S_{\mathbf{c}}$ is the sample covariance of $\mathbf{c}$. In summary, we have shown that the BGL (6) for estimating the graphical structure of Q given realizations from $\Phi\mathbf{c}+\boldsymbol{\varepsilon}$ can be discerned through a concave-convex-procedure where the inner solve is the graphical lasso (13) with "sample covariance" matrix depending upon the previous guess Q_j.

Another note is in order: the linearization step of (12) involves matrix solves using the matrix $Q_j+\tau^{-2}\Phi^{\mathrm{T}}\Phi$. If the basis functions are chosen to be compactly supported or orthogonal, then both Q_j and $\Phi^{\mathrm{T}}\Phi$ are sparse, and sparse matrix methods can potentially be used to speed up matrix solves (which is the computational technique exploited by LatticeKrig). For an arbitrary basis, common computations under our model have the same complexity, $\mathcal{O}(n\ell^2)$, as standard low rank or fixed rank methods (Cressie and Johannesson 2008).

A variety of numerical techniques have been proposed for the graphical lasso. Yuan and Lin (2007) and Banerjee, El Ghaoui, and d'Aspremont (2008) both use interior point methods, but the latter examine the *dual problem* of (13)

$$\underset{|U|\le\Lambda}{\arg\min}\ -\log\det(S_{\mathbf{c}}+U)-\ell, \quad (14)$$

where $|U| \leq \Lambda$ is understood elementwise, and solve (14) one column at a time via quadratic programming. Friedman, Hastie, and Tibshirani (2008) took an identical approach but write the dual problem of the columnwise minimization as a lasso regression, which they solve quickly using their own coordinate descent algorithm (Friedman et al. 2007). This implementation is available in the popular R package `glasso`.

Advances in solving (13) in recent years have stemmed from the use of second-order methods that incorporate Hessian information instead of simply the gradient. A current state of the art algorithm is `QUIC` (Hsieh et al. 2014) which also features an R package of the same name. Briefly, the `QUIC` algorithm uses coordinate descent to search for a Newton direction based on a quadratic expansion about the previous guess and then an Armijo rule to select the corresponding stepsize. During the coordinate descent update, only a set of free variables are updated, making the procedure particularly effective when Q is sparse. In the `BasisGraphicalLasso` package, available at `github.com/mlkrock/BasisGraphicalLasso`, we provide code to solve (12) using `QUIC`. A recent paper (Fattahi, Zhang, and Sojoudi 2019) shows that reformulating (13) as a maximum determinant matrix completion problem is a promising strategy.

2.2.1. Estimating the Nugget Variance

In practice, we must produce an estimate $\hat{\tau}^2$ which is fixed during the algorithm (12). For this purpose, we return to (5), now rewritten as

$$f(Q,\tau^2) = \log\det\left(Q+\tau^{-2}\Phi^{\mathrm{T}}\Phi\right) - \log\det Q$$
$$- \operatorname{tr}\left(\tau^{-4}\Phi^{\mathrm{T}}S\Phi\left(Q+\tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}\right) + n\log\tau^2 + \tau^{-2}\operatorname{tr}(S).$$

We minimize f jointly over τ^2 and α under the assumption that $Q = \alpha I_\ell$ with $\alpha > 0$. Our estimates $\hat{\tau}^2$ and $\hat{\alpha}$ are retrieved from an L-BFGS optimization routine via the `optim` function in R. In the simulation study below, this approach is seen to empirically work very well. Jointly estimating a full model of Q and τ^2 is complicated and unlikely to result in substantial empirical improvement (see Section 3.3.2).

2.2.2. Estimating the Penalty Weight

All that remains to specify is the penalty weight matrix Λ. One option follows Bien and Tibshirani (2011), in which a likelihood-based cross-validation approach is used to select Λ in the context of estimating a sparse covariance matrix. More formally, suppose we use k folds and consider t penalty matrices $(\Lambda_1,\dots,\Lambda_t)$. Let $\hat{Q}_\Lambda(S)$ be the estimate we get from applying our algorithm with empirical covariance $S = \frac{1}{m}\sum_{i=1}^m \mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$ and penalty Λ. For $A \subseteq \{1,\dots,m\}$, let $S_A = |A|^{-1}\sum_{i\in A}\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$. We seek Λ so that $\alpha(\Lambda) = \ell(\hat{Q}_\Lambda(S), S)$ is small, where

$$\ell(Q,S) = \log\det\left(Q+\tau^{-2}\Phi^{\mathrm{T}}\Phi\right) - \log\det Q$$
$$- \operatorname{tr}\left(\tau^{-4}\Phi^{\mathrm{T}}S\Phi\left(Q+\tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}\right) \qquad (15)$$

is the unpenalized version of (7). The cross-validation approach is to partition $\{1,\dots,m\}$ into disjoint sets $\{A_1,\dots,A_k\}$ and select $\hat{\Lambda} = \underset{\Lambda\in\{\Lambda_1,\dots,\Lambda_t\}}{\arg\min}\ \hat{\alpha}(\Lambda)$ where $\hat{\alpha}(\Lambda) = k^{-1}\sum_{i=1}^k \ell(\hat{Q}_\Lambda(S_{A_i^c}), S_{A_i})$.

Another option is through spatial cross-validation, where we use training data to estimate $\hat{Q}_{\Lambda_1}(S_{\text{train}}),\dots,\hat{Q}_{\Lambda_t}(S_{\text{train}})$ and then krig to the held out locations in the testing data. The penalty weight matrix producing the smallest RMSE is then used in conjunction with the full sample covariance S to obtain the final estimate $\hat{Q}$.

2.2.3. Initial Guess and Convergence

The nonconvex nature of this problem prohibits use of common convergence criterion available for convex optimization problems. Instead, we say that the DC scheme has “converged” when $\frac{\|Q_{j+1}-Q_j\|_F}{\|Q_j\|_F} < \epsilon$ for some loose tolerance $\epsilon = 0.01$. In this article, we use the initial guess $Q_0 = I_\ell$, and this choice can produce large diagonal entries $Q_{ii} \gg 0$ since the diagonal penalty weights of Λ are set to zero (see results in Section 4.1).

3. Simulated Data Studies

This section contains three perspectives of the proposed model. The first is a comparison study against two possible alternative estimation procedures, the second a timing study to illustrate the tradeoff between realizations and graph dimension, and the third a simulation study under different basis and graph structures.

3.1. Comparison Study

Let us consider alternatives to the BGL for estimating the precision (or covariance) matrix of the Gaussian vector $\mathbf{c}$ under the basis model. A naive but easy-to-compute estimator involves projecting the data onto the basis and using the projected data with standard shrinkage methods for estimating sparse precision matrices. Following notation in (4), the estimated regression coefficients of $\mathbf{Y}_i$ onto Φ comprise the columns of the least squares projected data matrix $\hat{\mathbf{c}}$. If Φ is not full rank, we cannot consider this least squares projection—a ridge regression may be more suitable, for example. Once $\hat{\mathbf{c}}$ is created, its sample covariance is substituted into the graphical lasso; we call this approach the *regression method*. For small samples, this approach is expected to be statistically inefficient as it does not explicitly use any likelihood information and moreover fails to incorporate the white noise process ε.

Another possible approach in the style of fixed rank kriging is to retrieve $K = Q^{-1}$ by minimizing the loss function

$$\underset{K\succeq 0}{\arg\min}\ \|\Phi K\Phi^{\mathrm{T}} + \tau^2 I_n - S\|_F \qquad (16)$$

as proposed by Cressie and Johannesson (2008). From equation (3.8) in Cressie and Johannesson (2008), the optimal parameter estimate is $\hat{K} = R^{-1}Q^{\mathrm{T}}(S-\tau^2 I_n)Q(R^{-1})^{\mathrm{T}}$ where $\Phi = QR$ is the QR decomposition of Φ. Again, if Φ is not full rank, we cannot consider this method. The R package `FRK` is designed for one realization and (16) is an analogous estimate for multiple realizations. By default, `FRK` parameterizes K as a block-exponential covariance matrix, but we use the “unstructured” option which is a more comparable nonparametric estimate. The connection to graphical modeling here is lost, as we cannot expect a sparse inverse of the estimated covariance matrix.

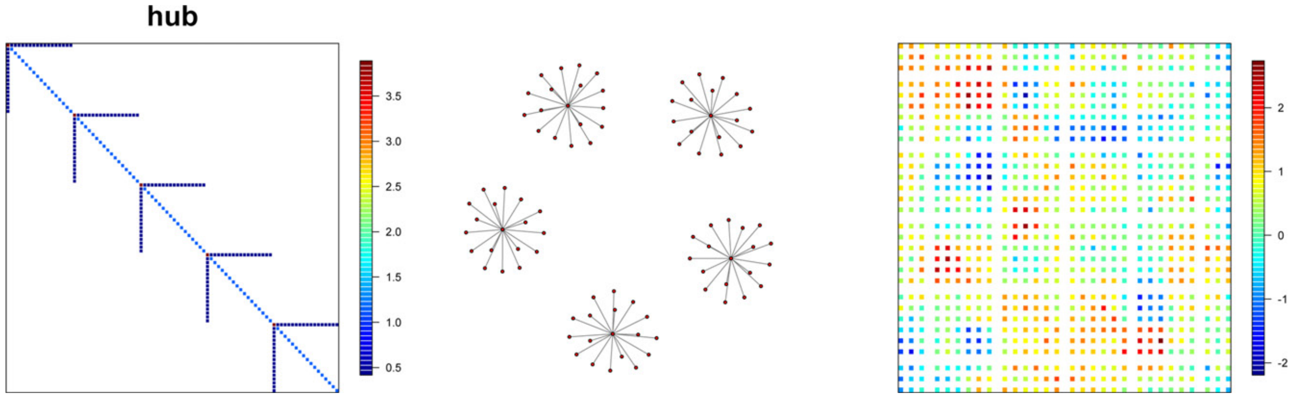

Figure 1. Displaying the hub graph structure for Q and a realization from the basis model.

Table 1. Comparing methods: loss is the fixed rank kriging approach for multiple realizations (FRK is used when $m = 1$), regression is the projection estimate, and BGL is the proposed basis graphical lasso.

Statistic	m	Loss/FRK	Regression	BGL
Frobenius error	1	1839.8	1.21	**1.10**
	5	97.7	0.78	**0.74**
	10	77.9	0.64	**0.63**
	20	53.6	0.59	0.59
	50	25.3	0.59	0.59
KL divergence	1	1,349,412	89.2	**78.2**
	5	–	48.3	**44.0**
	10	–	29.4	**27.8**
	20	–	17.4	**17.2**
	50	–	**12.3**	12.4

NOTE: Values are means based on 100 independent trials; the best score is indicated in bold. Missing values for Loss/FRK are due to numerically singular covariance matrix estimates.

The setup of the experiment is as follows: we have $n = 900$ observations on the two-dimensional grid $\{(i, j)\}_{i,j=1}^{30}$. Basis functions are a series of bisquare functions taken from the FRK package. Two resolutions are used in the basis for a total of $\ell = 90$ bisquare basis functions. We consider the graph Q to be a hub graph where the nodes are separated into groups and each member of that group is only neighbors with a central node. The hub graph is generated with default parameters from the R package huge (Zhao et al. 2015) for high-dimensional undirected graph estimation. The graphical structure is inherently tied to the spatial registration of the basis functions, but we are ignoring this and simply focusing on the ability of the methods to recover the precision matrix. Finally, the noise-to-signal ratio $\tau^2/(\mathrm{tr}(\Phi Q^{-1}\Phi^{\mathrm{T}})/n)$ is fixed at 0.1 and hence determines the true nugget variance $\tau^2 \approx 0.102$. See Figure 1 for an illustration of the graph structure and a realization of the process $\Phi\mathbf{c}$. The penalty matrix is populated with 0.25; we found that regularizing the diagonal helped when the number of realizations was small.

To compare these estimators, we conduct 100 trials and report summary statistics based on the Frobenius norm $\|\hat{Q} - Q\|_F/\|Q\|_F$ and the Kullback–Leibler (KL) divergence $\mathrm{tr}(\hat{Q}Q) - \log\det(\hat{Q}Q) - \ell$. Results are shown in Table 1. Generally, the estimates based on minimizing the loss function (16) were poorly behaved, and the versions which did not directly call the FRK package produced singular matrices. The BGL method is superior to the regression-based idea, especially for one or a handful of realizations. For data with more ensemble members the two are comparable. The next two sections consider the effect of multiple realizations in further detail.

3.2. Timing Study

Here, we consider timing results to solve the BGL (6) using our DC algorithm (10). The runtime is most influenced by the number of basis functions ℓ, the number of realizations m (i.e., the quality of the sample covariance S), the sparsity of Q, and the choice of penalty. Again we emphasize that the spatial dimension n is no longer relevant to the estimation of Q once $\Phi^{\mathrm{T}}\Phi$ and $\Phi^{\mathrm{T}}S\Phi$ are stored. The latter matrix $\Phi^{\mathrm{T}}S\Phi$ is also the only way that the data directly enters the BGL and should be computed using (8) to avoid storing the full sample covariance matrix S. With the QUIC algorithm, each inner solve of (10) is guaranteed to converge quadratically (Hsieh et al. 2014). It is difficult to say anything about the overall complexity of (6), however.

The spatial domain in consideration is the unit square $[0, 1] \times [0, 1]$, and basis functions are compactly supported Wendland bases from the LatticeKrig R package, described in detail in the next section. We observe data at $n = 2500$ uniformly randomly sampled spatial locations with a noise-to-signal ratio of 0.1. The fixed penalty matrix is populated with 0.2 and zeros on the diagonal—no cross-validation is performed. The initial guess is the identity matrix $Q_0 = I_\ell$, as in the rest of the article. We stop the BGL at a relative error of 0.01.

In Figure 2, we display the elapsed time for the BGL and the number of calls to QUIC in the MM scheme. In general, the computation time increases faster than linearly with respect to the number of basis functions, while the number of iterations scales linearly. Later MM iterations often spend increased amounts of time in QUIC without much reduction in relative error. Here, we see that having multiple realizations helps reduce computation time due to a more stable estimate of the sample spatial covariance matrix S.

3.3. Simulation Study

We close this section with a set of simulation studies to assess the ability of our proposed algorithm to recover unknown precision structures under the model (4). The section is broken into

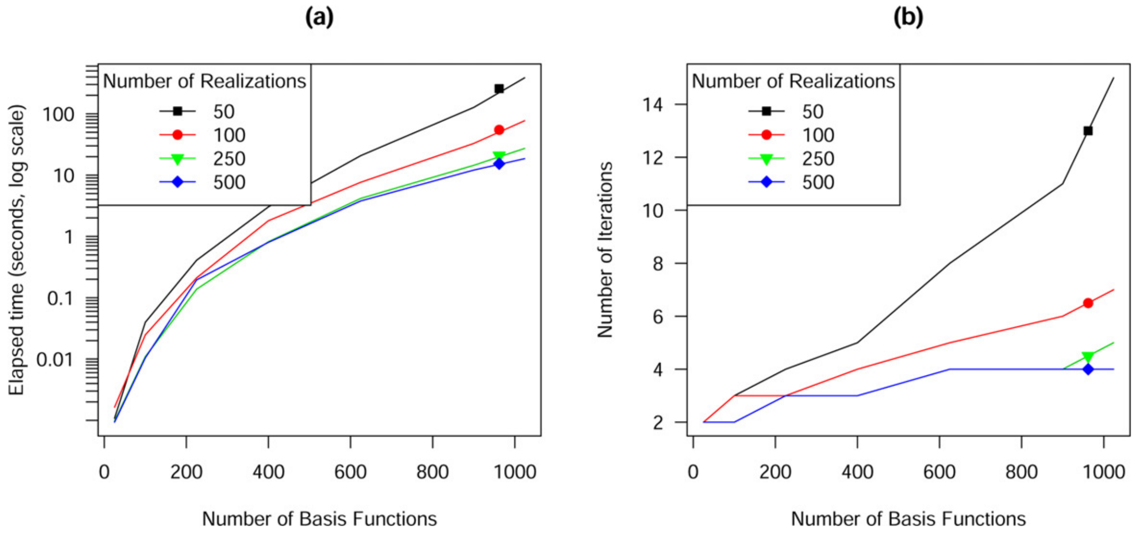

Figure 2. Illustrating the time to convergence for the algorithm (10). Plots show (a) the elapsed time and (b) the number of MM iterations.

two classes of basis functions: localized bases whose support is spatially compact, and global basis functions that are nonzero over the entire domain. For each class, we entertain multiple types of precision structures that are common in the graphical modeling literature.

For each choice of n, Φ, and Q, the noise-to-signal ratio $\tau^2/(\text{tr}(\Phi Q^{-1}\Phi^{\text{T}})/n)$ is fixed at 0.1 and hence determines the true nugget variance τ^2. Our estimated nugget variance $\hat{\tau}^2$ is retrieved from Section 2.2.1. Although in these simulations the population mean is zero, in practice it is unknown, so throughout we use the standard unbiased estimator S that includes an empirical demeaning which will reflect practical implications better than using the known mean.

3.3.1. Local Basis

First, we consider a localized problem where we use a basis of compactly supported functions on a grid using the LatticeKrig model setup (Nychka et al. 2015), which we briefly describe here: basis functions are compactly supported Wendland functions whose range of support is set so that each function overlaps with 2.5 other basis functions along axial directions. The model basis functions will correspond to either a single level or multiresolution model. In the single level setup, functions are placed on a regular grid. In the multiresolution setup, higher levels of resolution are achieved by increasing the number of basis functions and nodal points (e.g., the second level doubles the number of nodes in each axial direction). The precision matrix Q is set to a stationary spatial autoregressive structure, see Nychka et al. (2015) for details.

We specify the variance of the multiresolution levels to behave like an exponential covariance by setting parameter $\nu = 0.5$. In LatticeKrig, the precision matrix Q is constructed according to a spatial autoregression parameterized by the value α, which we fix at $\alpha = 4.05$. For simplicity, we employ no buffer region when constructing the Wendland bases; that is, there are no basis functions centered outside of the spatial domain. We use the R package `LatticeKrig` to set up the aforementioned basis and precision matrices. A total of $m = 500$ realizations from the process (4) under this model are generated, and we repeat this entire spatial data-generation process over 30 independent trials.

The spatial domain is $[0,1] \times [0,1]$, and n observation locations are chosen uniformly at random in this domain for different sample sizes $n \in \{100^2, 150^2, 200^2\}$. For the single level Wendland basis, we use $\ell \in \{100, 225, 400\}$ basis functions. Attempting to mirror these dimensions in the multiresolution basis, we use $\ell \in \{119, 234, 404\}$ which, respectively, corresponds to (1) four multiresolution levels, the coarsest containing two Wendland basis functions, (2) three multiresolution levels, the coarsest containing four Wendland basis functions, and (3) four multiresolution levels, the coarsest containing three Wendland basis functions.

We parameterize the penalty matrix Λ according to

$$\Lambda_{ij} = \begin{cases} \lambda, & i \neq j, \\ 0, & i = j, \end{cases} \tag{17}$$

allowing for free estimates of the marginal precision parameters. A 5-fold cross-validation as described in Section 2.2.2 is used to select a penalty matrix Λ from eight equally spaced values from 0.005 to 0.1. The optimal value is then used with the full set of simulated realizations to produce a best guess $\hat{Q}$. To validate our proposed estimation approach, we report several summary statistics, each averaged over the 30 trials: the Frobenius norm $\|\hat{Q} - Q\|_F/\|Q\|_F$, the KL divergence $\text{tr}(\hat{Q}Q) - \log\det(\hat{Q}Q) - \ell$, the percentage of zeros in Q that were missed by $\hat{Q}$, the percentage of nonzero elements in Q that were missed by $\hat{Q}$, the difference of the estimated nugget effect $\hat{\tau}^2$ and the true nugget effect τ^2, and the ratio of the estimated and true negative log-likelihoods $f(\hat{Q}, \hat{\tau}^2)$ and $f(Q, \tau^2)$, where f is defined in Section 2.2.1.

3.3.2. Comments

Tables 2 and 3 contain results from this simulation study. We see that estimating the nugget effect τ^2 by treating the process as stationary is quite accurate; here and in the rest of the article, the unreported standard deviations of the averaged differences $\hat{\tau}^2 - \tau^2$ are many orders of magnitude smaller than the magnitude of

Table 2. Simulation study results for the single level case.

n	ℓ	Frob	KL	%MZ	%MNZ	$\hat{\tau}^2 - \tau^2$	$f(\hat{Q}, \hat{\tau}^2)/f(Q, \tau^2)$
10,000	100	0.41	6.5	21	5	0.000008	1.0001
	225	0.49	26	20	2	−0.000042	1.00081
	400	0.72	110	4.7	0.8	−0.00016	0.998709
22,500	100	0.38	5.6	22	6	0.0000093	1.00005
	225	0.43	21	21	2	−0.0000041	1.00037
	400	0.65	82	4.8	1	−0.000033	0.999366
40,000	100	0.37	5.3	23	6	0.0000084	1.00002
	225	0.41	19	22	3	0.0000038	1.00021
	400	0.61	70	4.9	1	−0.0000079	0.999617

NOTE: Scores are averaged over 30 independent trials. Each column represents the number of observation samples, number of basis functions, Frobenius norm, KL divergence, percent of true zeros missed, percent of true nonzeros missed, estimated nugget minus true nugget, and the estimated negative log-likelihood divided by the true negative log-likelihood.

Table 3. Simulation study results for the multiple level case.

n	ℓ	Frob	KL	%MZ	%MNZ	$\hat{\tau}^2 - \tau^2$	$f(\hat{Q}, \hat{\tau}^2)/f(Q, \tau^2)$
10000	119	0.82	645	5.6	7	−0.000035	1.00002
	234	0.89	2020	6.4	4	−0.00013	1.00022
	404	0.85	3790	0.07	3	−0.0002	0.999782
22,500	119	0.77	640	7.1	6	−0.0000078	1
	234	0.85	2050	10	3	−0.000043	1.00017
	404	0.93	4280	0.085	2	−0.0001	0.99983
40,000	119	0.79	635	7.8	6	0.0000012	0.999998
	234	0.84	1940	11	3	−0.000015	1.00011
	404	0.94	4440	0.087	2	−0.000046	0.99986

NOTE: Scores are averaged over 30 independent trials. Each column represents the number of observation samples, number of basis functions, Frobenius norm, KL divergence, percent of true zeros missed, percent of true nonzeros missed, estimated nugget minus true nugget, and the estimated negative log-likelihood divided by the true negative log-likelihood.

the true nugget effect. Estimates under the multiresolution basis are clearly lackluster when compared to the single resolution counterpart. The Frobenius norm and KL divergence tend to increase with the size of ℓ, but this is to be expected as the dimensions of the target precision matrix Q grow in ℓ. The percentage of zeros in Q that are missed (i.e., nonzero) in $\hat{Q}$ drops sharply as ℓ increases to 400, but this is the consequence of a harsher penalty weight matrix selected in the cross-validation scheme.

3.3.3. *Global Basis*

Next, we consider a spatial basis defined globally, that is, without compact support. In particular, we set up a harmonic basis via the model $\phi_i(\mathbf{s}) = \cos(2\pi \boldsymbol{\omega}_i^{\mathrm{T}} \mathbf{s})$ where the frequencies $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_\ell \in \mathbb{R}^2$ are all pairwise combinations of the form $(\frac{k}{\sqrt{n}}, \frac{j}{\sqrt{n}})$ for $k, j = 1, \ldots, \sqrt{\ell} - 1$. Given n samples, the corresponding $n \times \ell$ basis matrix Φ has (i, j)th entry $\cos(2\pi \boldsymbol{\omega}_j^{\mathrm{T}} \mathbf{s}_i)$. Such a model can be interpreted as a discretized approximation to the representation theorem for stationary processes (see, e.g., Stein 1999).

Whereas compactly supported basis functions laid on a grid suggest natural nearest-neighbor structures for Q, in the case of global basis functions it is not as clear what natural models might be. We consider four traditional undirected graphical models from the literature, described briefly below and depicted in Figure 3.

1. Random graph: Edges of the graph are randomly selected.
2. Cluster graph: The diagonal of Q consists of block matrices with random graph edges in each block.
3. Scale-free graph: Generated from the algorithm of Barabási and Albert (1999).
4. Band graph: Q is tridiagonal.

Our experiment is similar to the local basis experiment: the n spatial observation locations are randomly uniformly sampled from the square $[0, \sqrt{n}] \times [0, \sqrt{n}]$ where we entertain $n \in \{100^2, 150^2, 200^2\}$ and $\ell \in \{100, 225, 400\}$. The four graphs above are each generated with the R package `huge` (Zhao et al. 2015) using default parameters. The same test statistics as reported in the previous section, again averaged over 30 trials, are recorded in Tables A.1–A.4 in Appendix A. We display an abbreviated version in Table 4 with $n = 10{,}000$ as we saw little variability in the summary statistics upon increasing n.

3.3.4. *Comments*

Tables A.1–A.4 contain the results of the global basis simulation study and are shown in Appendix A. As in the compactly supported basis study, we note the same behavior in $\hat{\tau}^2$ in that the independent identical coefficient assumption (constant diagonal Q) yields robust estimates of τ^2. The cluster graph appears highest in Frobenius norm and KL divergence, but this should not be surprising given the fact that there are more nonzero elements in the cluster graph than its counterparts and we are searching for a sparse estimate. The percentage of missed zeros and missed nonzeros in Q also behave similarly to the local basis study with respect to the dimension ℓ. Overall, the proposed method seems to supply reasonable estimates of the

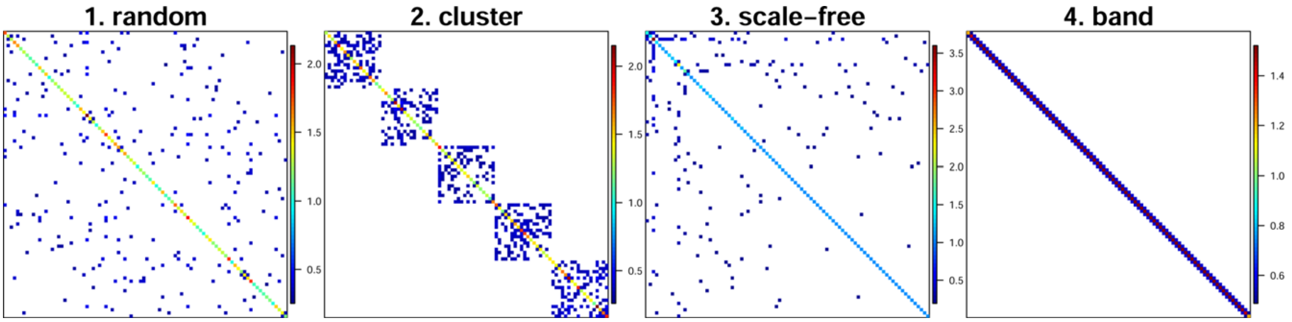

Figure 3. Illustration of the various graph structures of the precision matrix used in our simulation study.

Table 4. Condensed simulation study results for the various graphical models.

Graph	ℓ	Frob	KL	%MZ	%MNZ	$\hat{\tau}^2 - \tau^2$	$f(\hat{Q}, \hat{\tau}^2)/f(Q, \tau^2)$
Random	100	0.19	1.7	9.6	0	0.00066	0.999961
	225	0.2	4.8	5.2	0	−0.00042	0.999931
	400	0.22	9.8	2.6	0.0002	0.0019	0.99995
Cluster	100	0.26	3.1	16	0.026	0.0008	0.999948
	225	0.3	8.9	8.9	0.036	−0.00086	0.999913
	400	0.32	18	5	0.049	0.0026	0.999922
Scale-free	100	0.21	1.4	6.1	0.01	0.00061	0.999978
	225	0.21	3.5	2.8	0.05	−0.00081	0.999972
	400	0.21	6.3	2.6	0.06	0.0032	0.999883
Band	100	0.17	1.5	8.5	0	0.0008	0.999967
	225	0.19	4.2	4.6	0	0.0015	0.999944
	400	0.21	8.7	2.3	0	0.0017	0.999963

NOTE: Here, n is fixed at 10,000. Full results with $n = 22{,}500$ and 40,000 are shown in Appendix A. Scores are averaged over 30 independent trials. Each column represents the graph type, number of basis functions, Frobenius norm, KL divergence, percent of true zeros missed, percent of true nonzeros missed, estimated nugget minus true nugget, and the estimated negative log-likelihood divided by the true negative log-likelihood.

zero structure of the precision matrix as well as its nonzero values.

3.3.5. Changing the Number of Realizations and the Noise-to-Signal Ratio

We consider a brief study to assess the effect realizations on the algorithm's ability to recover Q. In the prior section, we fixed the number of realizations at $m = 500$. Here we fix $\ell = 100$ to ease computation times but vary the number of realizations according to $m \in \{100, 200, 500, 1000\}$ and number of spatial locations according to $n \in \{100^2, 150^2, 200^2, 250^2, 300^2\}$. The same Wendland basis and precision matrix Q as in the local basis study (Section 3.3.1) are used to generate m realizations of the additive model (4) where we have observations at n uniformly randomly sampled locations in $[0, 1] \times [0, 1]$. We record the KL divergence $\mathrm{tr}(\hat{Q}Q) - \log\det(\hat{Q}Q) - \ell$ and the percentage of zeros in Q that $\hat{Q}$ fails to capture. The penalty parameter is fixed at $\lambda = 0.005$, the value which was favored in our previous simulations when $(\ell, m, n) = (100, 500, 10{,}000)$. Figure 4 shows results from this study. The plots suggest that the number of realizations m has a prominent effect on KL divergence and missed zeros, with larger values of m increasing the quality of the estimate, but increasing the sample size n has much less of an effect. We also noticed that increasing n noticeably lowered the Frobenius norm $\|\hat{Q} - Q\|_F/\|Q\|_F$ (not displayed in the figure), but more unexpectedly, increasing m did not necessarily decrease the Frobenius error, an oddity which we attribute to instability of this norm (Tropp 2015).

Up to this point, the noise-to-signal ratio has been fixed at 0.1. We conducted a brief experiment where, for fixed Φ and Q, we increased the noise-to-signal ratio to 0.25 and 0.5. The penalty matrix was also kept constant across these various noise levels. Increasing the ratio from 0.1 to 0.25 slightly worsened the Frobenius norm and KL divergence, but the jump from 0.25 to 0.5 demonstrated a substantial decrease in ability to accurately capture the true precision matrix. At a noise-to-signal ratio of 0.5, however, the resulting model is extremely noisy and expecting accurate estimates is not entirely reasonable.

4. Data Analysis

4.1. Reforecast Data

The Global Ensemble Forecast System from the National Center for Environmental Prediction provides an 11 member daily reforecast of climate variables available from December 1984 to present day. We take all readings from January of each year through 2018, giving a total of $m = 1054$ global fields of the two-meter temperature variable. Measurements were recorded at each integer valued longitude and latitude combination, totaling $n = 65{,}160$ spatial locations. We use a set of Wendland basis functions spread over the globe in a way that ensures that the $\ell = 2531$ centers are equispaced with respect to the great circle distance. See Figure 5 for an illustration of the data and basis functions. We opt for 2531 as it provides a reasonably dense network of basis functions that still allows for computational tractability.

Throughout this section, we work with temperature anomalies, that is, residuals after removing a pixelwise mean over realizations. Although these residuals are not strictly independent, they do decorrelate over a few days and we use the assumption of

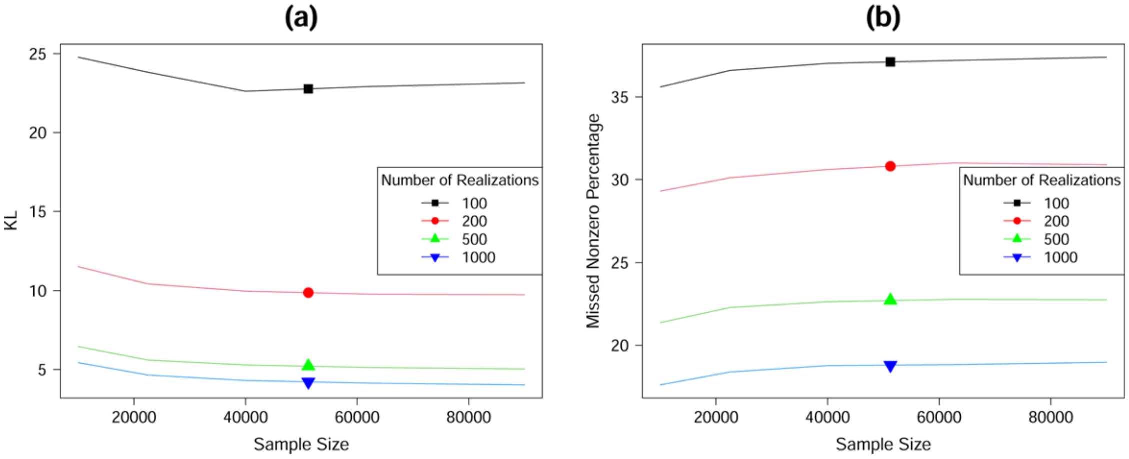

Figure 4. Results of a small simulation study where we fix the basis size $\ell = 100$ but vary the number of realizations m and sample size n. Plots (a) and (b) show our estimate's error in the form of KL divergence and the missed nonzero percentage.

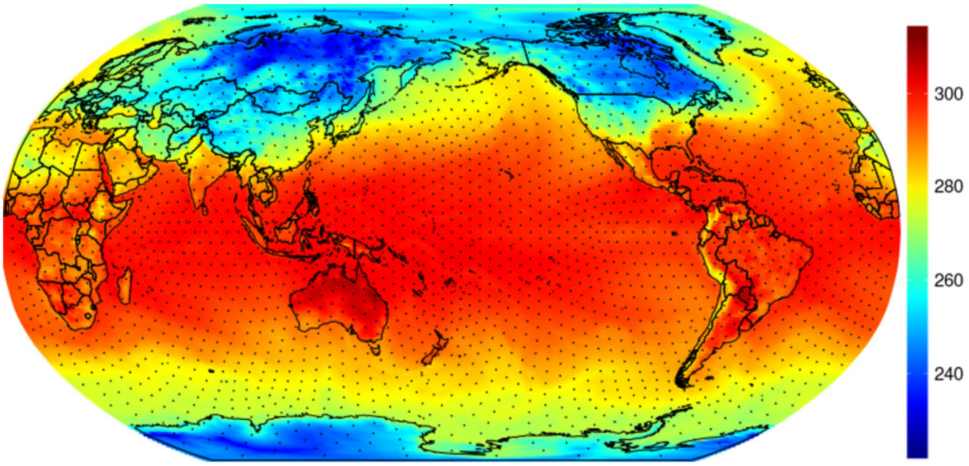

Figure 5. Two meter temperature (Kelvin) on January 1, 1984. Dots indicate nodal points.

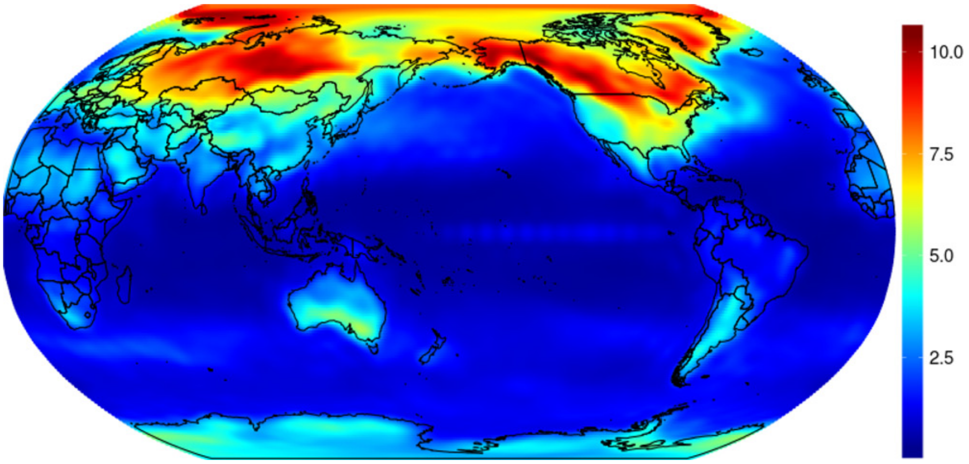

Figure 6. Estimated pointwise standard deviations under the model (3) using Q estimated from the DC algorithm. Units are Kelvins.

independence for convenience. Moreover, considering data over an approximately 30 year period is standard for climatological studies, even if there are slowly changing climate signals within the period. The technique introduced in Section 2.2.1 is used to estimate the nugget effect $\hat{\tau}^2 = 1.74$.

An interesting idea when using localized basis functions with a notion of distance between them is to adjust the penalty matrix so that neighbors are encouraged to remain in close proximity to the center point. In particular, we make Λ proportional to the distance matrix of the centers of the basis functions. This idea was pursued in Davanloo Tajbakhsh, Serhat Aybat, and Del Castillo (2014) but in the context of direct spatial observations with the graphical lasso rather than working through basis functions with our DC algorithm.

We consider $\Lambda = \lambda D$ where D is the pairwise distance matrix of the nodal points registering the Wendland basis functions. To select the penalty parameter λ, we use 2-fold likelihood-based cross-validation as described in Section 2.2.2 for values $\lambda \in \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$. Smaller values than 0.0001 were examined but failed to converge after a reasonable runtime. Despite the fact that it sits on the boundary of our parameter set, the value $\lambda = 0.0001$ was chosen as yielded a significant drop in negative log-likelihood when compared to larger penalties.

Figure 6 contains a global plot of the implied estimated local standard deviations. Note similar behavior in Figure 4 of Legates and Willmott (1990), which depicts standard deviations for mean air surface temperature over the globe. Both plots illustrate a clear land-ocean difference and increased variability in higher latitudes where the overall land area is greater.

Figure 7 shows a plot of estimated spatial correlation functions centered at three different geographical locations in the southern tip of South America, the Middle East, and central North America. There is clear evidence of nonstationarity in all three cases as well as negative correlation at medium distances. An interesting feature of the estimated covariance structure is the negative correlation between Alaska and the central United States, indicative of medium-range teleconnections, and may be a result of Rossby waves that occur during winter in the northern hemisphere.

An interesting byproduct of the BGL is that we can examine the GMRF neighborhood structure of the estimated precision matrix. Recall that each random coefficient c_i in our model is registered to a nodal grid shown in Figure 5, and thus we can identify estimated spatial neighborhood patterns according to this grid. We illustrate some of these neighborhood structures, colored by their respective Q values, in Figure 8. For nodal points over the ocean, the tendency is large,

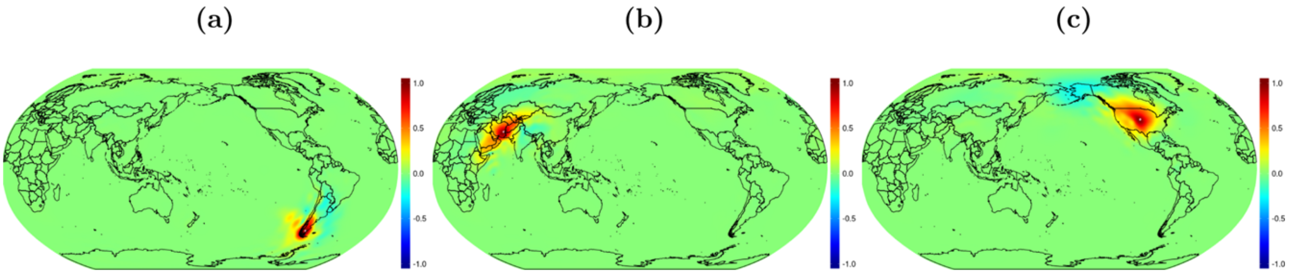

Figure 7. Estimated spatial correlation functions centered at a point in southern South America (a), central Middle East (b), and central North America (c).

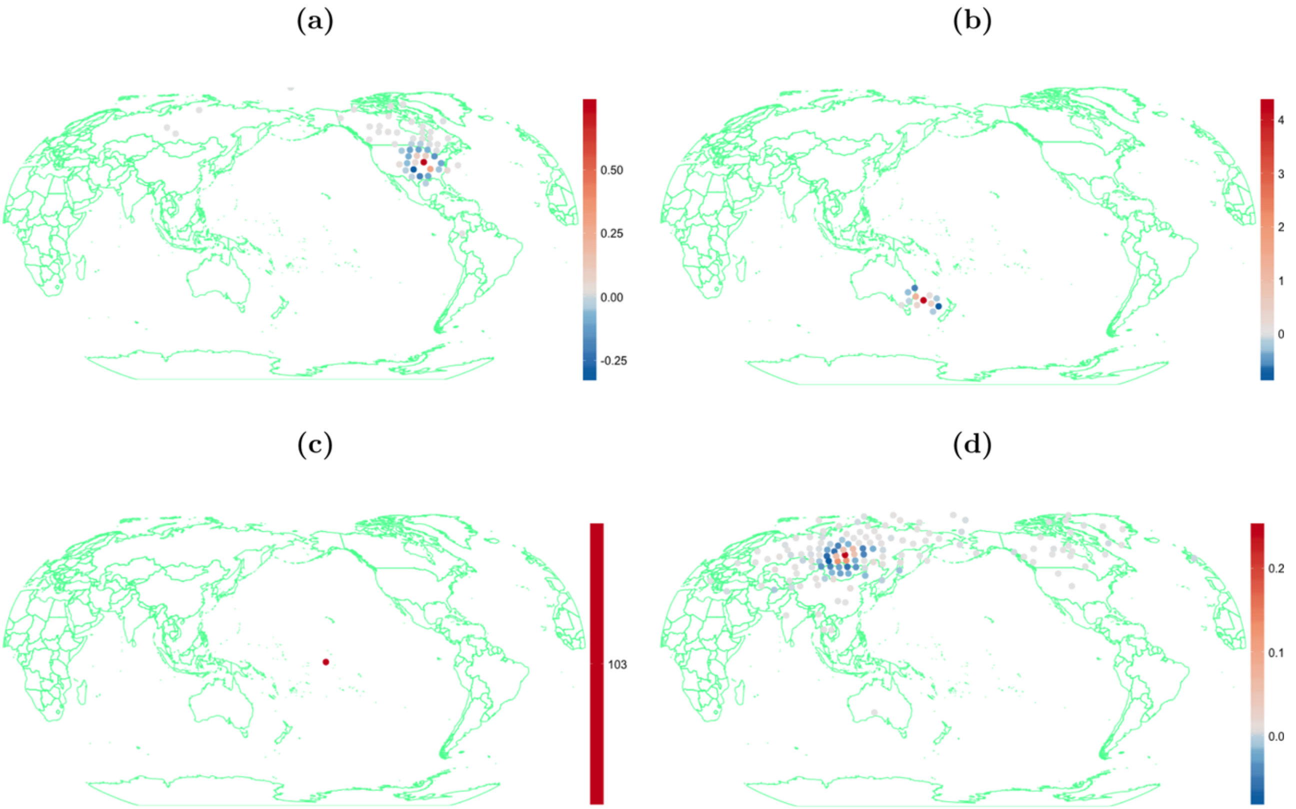

Figure 8. The estimated neighborhood structure of Q registered at the nodal points of the basis functions. Colors correspond to neighborhood entries in Q with a center point that is clockwise starting top left: over United States (a); in Pacific Ocean near Australia (b); over the Pacific Ocean near the equator (c); over Russia (d).

positive precision values with few neighbors; we note that this does not necessarily correspond to low spatial correlation of temperature over the ocean, in fact the opposite is true—the fact that there are smooth, overlapping basis functions over the ocean imply spatially correlated temperatures as we would expect. For nodal points over land, we observe that the most significantly nonzero neighbor elements are geographically near the center node, implying more complicated spatial covariance structure.

The utility of examining neighborhood structures as in Figure 8 allows for model interpretation (with respect to the chosen basis), but also model interrogation. In particular, a model like LatticeKrig or the SPDE approach of Lindgren, Rue, and Lindström (2011) uses compactly supported bases distributed throughout the domain, but both models use simple stationary spatial autoregressive processes for the random amplitudes **c**. The substantial variation in neighborhoods seen in Figure 8 suggests that these stationary models are not appropriate for highly nonstationary processes, but our approach accommodates such nonstationarity readily and without much extra modeling effort.

Finally, we display simulations from the continuous stochastic process model (i.e., without white noise) in Figure 9. The bottom row of this figure illustrates a limitation of our low rank model—we cannot adequately study the small scale structure of the process (Stein 2014). A multiresolution design in the style of LatticeKrig (or, more simply, using more basis functions) could help capture more small scale variability, but this would require innovations for computing with larger graph structures.

4.2. Observational Data

The Topography Weather dataset (Oyler et al. 2015) contains observed air temperatures from a set of observation networks over the continental United States. We consider daily minimum temperatures during the month of June from 2010 to 2014, giving a total of $m = 150$ realizations. Network locations are chosen to have no missing values, yielding $n = 4577$ spatial

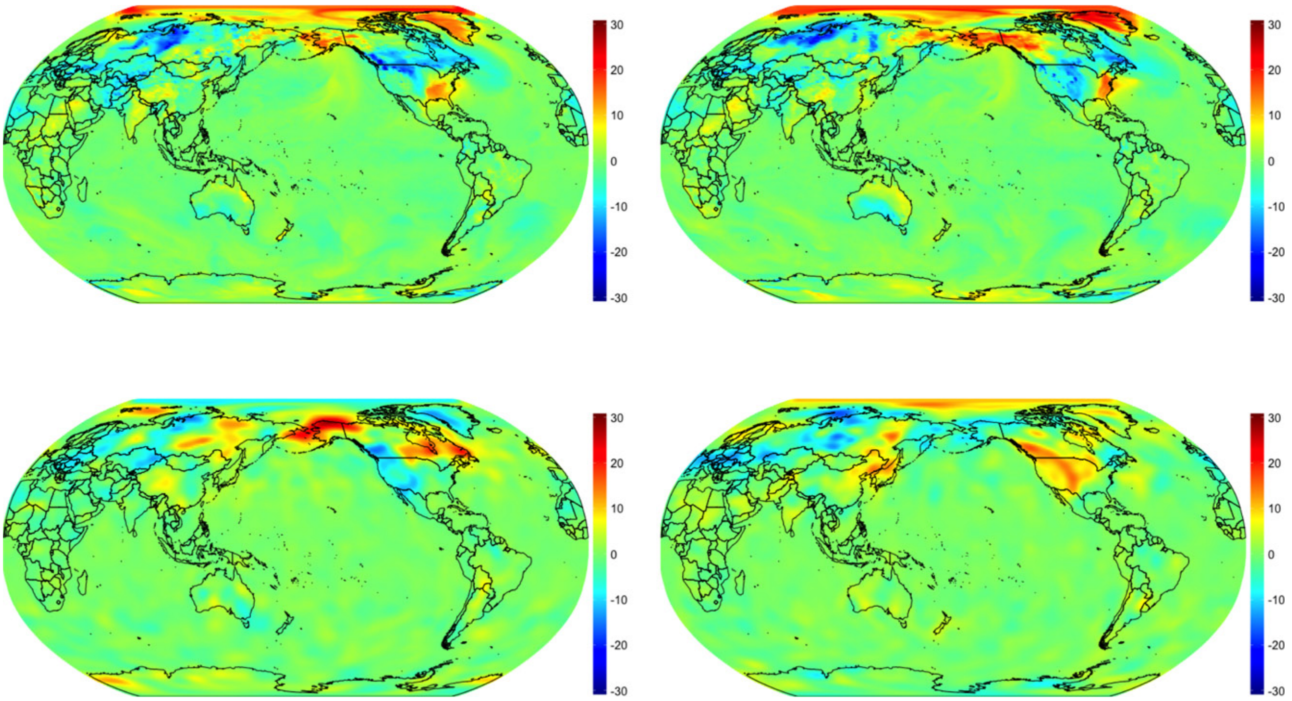

Figure 9. Example residual fields (top row) and simulations (bottom row).

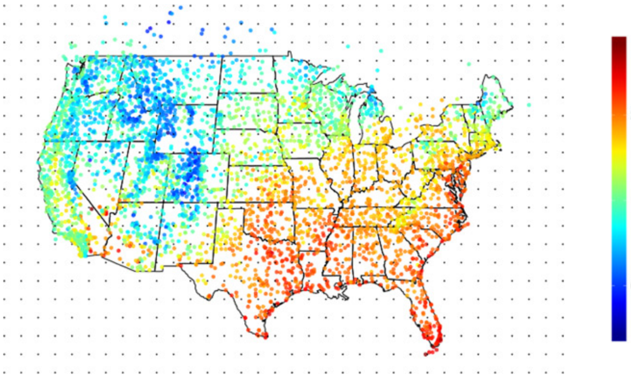

Figure 10. Minimum temperature (Celsius) on June 1, 2010, overlaid with a grid of basis function nodes.

locations. Figure 10 shows an example day of data on June 1, 2010.

We work with minimum temperature residuals after removing a pixelwise mean over realizations and also transform the raw spatial coordinates with a sinusoidal projection. Our statistical model for the temperature residuals uses Wendland basis functions centered at nodes displayed in Figure 10. We opt for $\ell = 1160$ functions using a single level of resolution. The nodal grid and Wendland functions are chosen to match up with a LatticeKrig model specification but with relaxed assumptions on the precision matrix governing the random coefficients.

The nugget estimate $\hat{\tau}^2 = 2.18$ is retrieved following Section 2.2.1. As with the previous dataset, the penalty matrix Λ is parameterized according to $\Lambda = \lambda D$, where D is the distance matrix of node points. We selected a scaling parameter of 7 from $\lambda \in \{1, 2, \dots, 30\}$ using the likelihood-based cross-validation scheme in Section 2.2.2. Code to reproduce the estimated precision matrix with this data is available at `github.com/mlkrock/BasisGraphicalLasso`.

Figures 11 and 12 show graphical model neighborhoods and estimated correlation functions centered at locations in Utah and Kansas. Clear anisotropy and nonstationarity is present in the estimated correlation functions with greater north-south directionality of correlation, while the neighborhood structure for the Utah nodal point (a) displays greater complexity than the relatively nearby neighbors of the nodal point in the midwest (b). In particular, the estimated graphs in Figure 11 suggest that the stationary spatial autoregressive assumption underlying both LatticeKrig and the standard SPDE approach are inappropriate for these data.

Due to lack of data availability over the ocean there is an identifiability problem with our method. Since several of the Wendland basis functions lie over the ocean where there is no observed data, we cannot expect the algorithm to give reasonable estimates for the diagonal elements of Q corresponding to those nodes. Moreover, the diagonal of the penalty matrix Λ is identically zero, and thus the corresponding diagonal elements of Q remained unchanged no matter the initial guess Q_0, which we fixed at $Q_0 = I_\ell$.

We compare our BGL model against the analogous LatticeKrig model using the same nodal grid and same Wendland basis functions but with the spatial autoregressive precision matrix of LatticeKrig. We estimate LatticeKrig parameters by maximum likelihood within the `LatticeKrig` package in R. In particular, the central `a.wght` parameter is estimated at 4.495, and the nugget variance $\hat{\tau}^2 = 2.23$ is close to our estimate. The smoothness parameter ν in the LatticeKrig setup is set at 0.5, which is a typical assumption for observational temperature data. However, using this few basis functions downplays the capabilities of LatticeKrig, so we include a multiresolution LatticeKrig model created with 3 levels, the coarsest of which

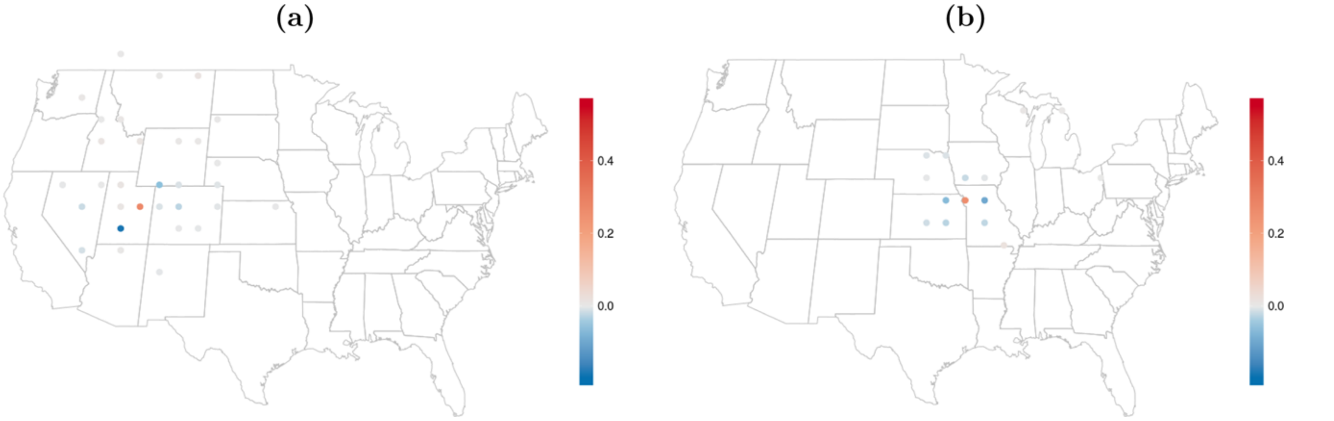

Figure 11. Displaying the neighborhood structure of *Q*, where the two center points are located in central Utah (a) and near Kansas City (b). Neighbors are colored according to the corresponding nonzero elements in *Q*.

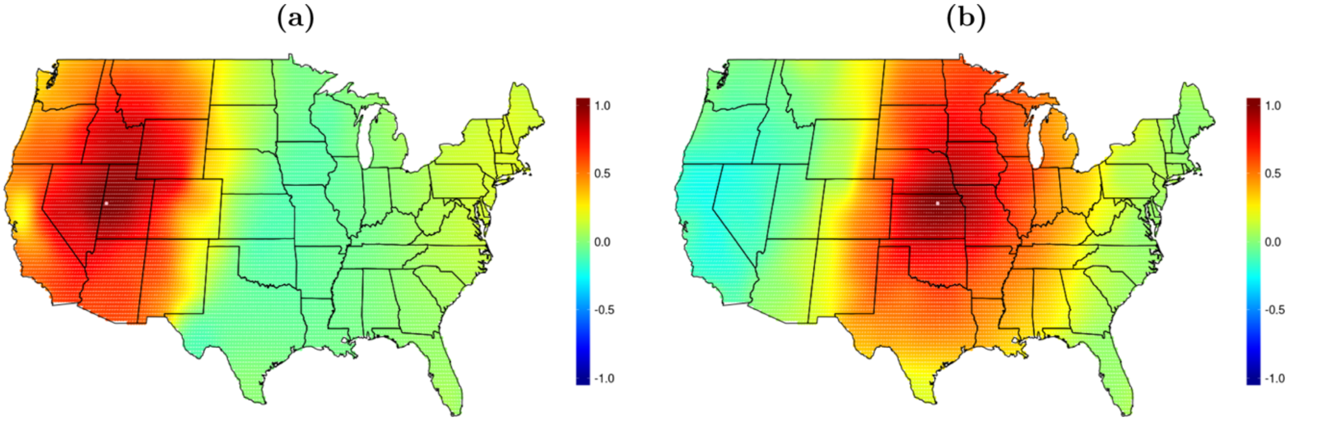

Figure 12. Estimated spatial correlation functions centered at the pink-colored locations in Utah (a) and Kansas (b).

matches the nodepoints from the smaller example. In total, there are 15,394 basis functions over the 3 levels, and the estimated `a.wght` and nugget variance parameters are 4.161 and 2.05, respectively.

The three models are compared based on cross-validation prediction accuracy and standard Akaike information criterion. We randomly hold out 400 locations and calculate the kriging predictive distribution at these locations for each day of data. We compare point predictions using the average root mean squared error (RMSE). To compare the predictive distributions, we use two proper scoring rules: the continuous ranked probability score (CRPS) and the (negative) log score (Gneiting and Raftery 2007). The former quantifies the quality of the marginal predictive distributions at each location separately, while the latter is a measure of the quality of the *joint* predictive distribution over all validation locations simultaneously. To calculate the AIC, we note that the number of degrees of freedom of a spatial model can be identified with the trace of the spatial smoothing hat matrix (Nychka 2000).

Table 5 contains the averaged scores over all days. It is perhaps surprising that both LatticeKrig models give a marginally better CRPS despite being designed for a single spatial field $m = 1$, but these scores only measure the marginal behavior of the predictive distributions. The AIC and log scores quantify the quality of the *joint* distributions and suggest that the BGL more accurately represents such joint distributions of the process than the LatticeKrig models.

Table 5. Cross-validation results comparing the proposed basis graphical lasso (BGL) to two versions of LatticeKrig on the TopoWX data.

	RMSE	CRPS	Negative log score	AIC
BGL	1.47	2.15	714.4	9018.5
Single-level LatticeKrig	1.48	2.13	1084.6	9063.2
Multi-level LatticeKrig	1.46	2.14	997.6	9341.6

5. Conclusion

In this work, we introduce a novel approach for estimating the precision matrix of the random coefficients of a basis representation model that is pervasive in the spatial statistical literature. The only assumption we make about the precision matrix is that it is sparse. In the case that the basis functions are registered to a grid, the precision entries can be interpreted as a spatial Gaussian Markov random field, while graphical model interpretations are still viable with global bases.

The proposed BGL estimator minimizes an ℓ_1 penalized negative log-likelihood equation. We show that the optimization problem is equivalent to one involving a sum of a convex and concave functions, suggesting a DC algorithm in which we iteratively linearize the concave part at the previous guess and solve the resulting convex problem. The linearization in our case gives rise to a graphical lasso problem with its “sample covariance” depending upon the previous guess. The graphical lasso problem is well-studied and a number of user-friendly

R packages exist, headed by the second-order method QUIC. Our method has important practical applications in spatial data analysis, since we obtain a nonparametric, penalized maximum likelihood estimate of Q which can subsequently be used in kriging or simulation with computational complexity $\mathcal{O}(n\ell^2)$ under the basis model.

In our data examples, we see that the proposed method performs competitively with existing alternatives such as LatticeKrig on marginal predictions but substantially improves the quality of joint predictive distributions. Moreover, our model results in interpretable fields, allowing for checking of graphical neighborhood structures or implied nonstationary covariance functions. Future work may be directed toward other penalties, increasing the number of basis functions to better accommodate multiple levels of resolution, or pushing these notions to space-time modeling.

Appendix A

We start the appendix with the proof of Proposition 1.

Proof of Proposition 1. For matrices A, U, C, and V of appropriate size, the Sherman–Morrison–Woodbury formula is

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1} VA^{-1}$$

and the matrix determinant lemma is

$$\det(A + UCV) = \det(C^{-1} + VA^{-1}U)\det(C)\det(A).$$

In our case, these two equations read

$$\left(\Phi Q^{-1}\Phi^{\mathrm{T}} + \tau^2 I_n\right)^{-1} = \tau^{-2}I_n - \tau^{-4}\Phi\left(Q + \tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}\Phi^{\mathrm{T}} \tag{A.1}$$

and

$$\det(\Phi Q^{-1}\Phi^{\mathrm{T}} + \tau^2 I_n) = \det\left(Q + \tau^{-2}\Phi^{\mathrm{T}}\Phi\right)\det(Q^{-1})\det(\tau^2 I_n). \tag{A.2}$$

Combining (A.1) with linearity and the cyclic property of trace gives

$$\begin{aligned}\mathrm{tr}(S(\Phi Q^{-1}\Phi^{\mathrm{T}} + \tau^2 I_n)^{-1}) &= \tau^{-2}\mathrm{tr}(S)\\ &\quad - \mathrm{tr}\left(\tau^{-4}\Phi^{\mathrm{T}}S\Phi\left(Q + \tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}\right),\end{aligned}$$

and taking the logarithm of (A.2) immediately yields

$$\begin{aligned}&\log\det(\Phi Q^{-1}\Phi^{\mathrm{T}} + \tau^2 I_n)\\ &\quad = \log\det\left(Q + \tau^{-2}\Phi^{\mathrm{T}}\Phi\right) - \log\det Q + n\log\tau^2.\end{aligned}$$

□

A.1. Convexity, Gradients, and Hessians

The penalized negative log-likelihood in Proposition 1 reads

$$\begin{aligned}&\log\det\left(Q + \tau^{-2}\Phi^{\mathrm{T}}\Phi\right) - \log\det(Q)\\ &\quad - \mathrm{tr}\left(\tau^{-4}\Phi^{\mathrm{T}}S\Phi\left(Q + \tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}\right) + \|\Lambda \circ Q\|_1.\end{aligned}$$

Let us explain the classifications of convexity and concavity stated in the opening paragraph of Section 2.2. The penalty function $Q \mapsto \|\Lambda \circ Q\|_1$ is trivially convex on $Q \succeq 0$. A proof of concavity for $Q \mapsto \log\det Q$ on $Q \succeq 0$ is given in Boyd and Vandenberghe (2004, p. 74). The convexity of $\mathrm{tr}(AQ^{-1})$ on $Q \succeq 0$ for an arbitrary positive semidefinite matrix A can be shown in a similar fashion, as the authors suggest in their Exercise 3.18(a). Composition with an affine mapping preserves both convexity and concavity, so $Q \mapsto \log\det\left(Q + \tau^{-2}\Phi^{\mathrm{T}}\Phi\right)$ is concave on $Q \succeq 0$ and $Q \mapsto \mathrm{tr}\left(A\left(Q + \tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}\right)$ is convex on $Q \succeq 0$.

Below we report the gradient and Hessian matrices of the first three terms in the negative log-likelihood, with $\otimes$ indicating the Kronecker product of two matrices. Let $W = Q^{-1}$ and $M = \left(Q + \tau^{-2}\Phi^{\mathrm{T}}\Phi\right)^{-1}$ for shorthand.

	Gradient	Hessian
$\log\det\left(Q + \tau^{-2}\Phi^{\mathrm{T}}\Phi\right)$	M	$-(M \otimes M)$
$-\log\det(Q)$	$-W$	$W \otimes W$
$-\mathrm{tr}\left(\tau^{-4}\Phi^{\mathrm{T}}S\Phi M\right)$	$\tau^{-4}M\Phi^{\mathrm{T}}S\Phi M$	$-\tau^{-4}\left(M\Phi^{\mathrm{T}}S\Phi M \otimes M\right)$ $-\tau^{-4}\left(M \otimes M\Phi^{\mathrm{T}}S\Phi M\right)$

Our claims of convexity and concavity in the above paragraphs can be further verified with the following fact: if $\{\lambda_i\}$ and $\{\mu_i\}$ are the eigenvalues of A and B, then $A \otimes B$ has eigenvalues $\{\lambda_i\mu_j\}$ (Seber 2007, 11.5).

A.2. Additional Tables

Table A.1. Simulation study results for the random graphical model.

n	ℓ	Frob	KL	%MZ	%MNZ	$\hat{\tau}^2 - \tau^2$	$f(\hat{Q}, \hat{\tau}^2)/f(Q, \tau^2)$
10,000	100	0.19	1.7	9.6	0	0.00066	0.999961
	225	0.2	4.8	5.2	0	−0.00042	0.999931
	400	0.22	9.8	2.6	0.0002	0.0019	0.99995
22,500	100	0.19	1.8	9.7	0.0007	0.00033	0.999983
	225	0.2	4.8	5.2	0	−0.00048	0.999968
	400	0.22	9.7	2.6	0.0003	0.001	0.999976
40,000	100	0.19	1.7	9.7	0	0.00018	0.99999
	225	0.2	4.7	5.2	0	−0.00063	0.999981
	400	0.22	9.7	2.6	0.0003	0.00019	0.999986

NOTE: Scores are averaged over 30 independent trials. Each column represents the number of observation samples, number of basis functions, Frobenius norm, KL divergence, percent of true zeros missed, percent of true nonzeros missed, estimated nugget minus true nugget, and the estimated negative log-likelihood divided by the true negative log-likelihood.

Table A.2. Simulation study results for the cluster graphical model.

n	ℓ	Frob	KL	%MZ	%MNZ	$\hat{\tau}^2 - \tau^2$	$f(\hat{Q}, \hat{\tau}^2)/f(Q, \tau^2)$
10,000	100	0.26	3.1	16	0.026	0.0008	0.999948
	225	0.3	8.9	8.9	0.036	−0.00086	0.999913
	400	0.32	18	5	0.049	0.0026	0.999922
22,500	100	0.26	3.1	17	0.021	0.00035	0.999973
	225	0.29	8.9	9	0.042	−0.00047	0.999957
	400	0.31	18	5	0.05	0.0007	0.999961
40,000	100	0.26	3.1	18	0.02	0.0002	0.999984
	225	0.29	8.7	9.1	0.041	−0.00053	0.999976
	400	0.31	18	5.1	0.05	0.00041	0.999977

NOTE: Scores are averaged over 30 independent trials. Each column represents the number of observation samples, number of basis functions, Frobenius norm, KL divergence, percent of true zeros missed, percent of true nonzeros missed, estimated nugget minus true nugget, and the estimated negative log-likelihood divided by the true negative log-likelihood.

Table A.3. Simulation study results for the scale-free graphical model.

n	ℓ	Frob	KL	%MZ	%MNZ	$\hat{\tau}^2 - \tau^2$	$f(\hat{Q}, \hat{\tau}^2)/f(Q, \tau^2)$
10,000	100	0.21	1.4	6.1	0.01	0.00061	0.999978
	225	0.21	3.5	2.8	0.05	−0.00081	0.999972
	400	0.21	6.3	2.6	0.06	0.0032	0.999883
22,500	100	0.2	1.4	5.9	0.009	0.00029	0.999991
	225	0.21	3.5	2.8	0.04	−0.00054	0.999988
	400	0.21	6.3	2.7	0.06	0.00071	0.999946
40,000	100	0.2	1.3	6.1	0.01	0.0002	0.999994
	225	0.21	3.5	2.8	0.05	−0.00061	0.999993
	400	0.21	6.2	2.7	0.06	0.00029	0.999969

NOTE: Scores are averaged over 30 independent trials. Each column represents the number of observation samples, number of basis functions, Frobenius norm, KL divergence, percent of true zeros missed, percent of true nonzeros missed, estimated nugget minus true nugget, and the estimated negative log-likelihood divided by the true negative log-likelihood.

Table A.4. Simulation study results for the band graphical model.

n	ℓ	Frob	KL	%MZ	%MNZ	$\hat{\tau}^2 - \tau^2$	$f(\hat{Q}, \hat{\tau}^2)/f(Q, \tau^2)$
10,000	100	0.17	1.5	8.5	0	0.0008	0.999967
	225	0.19	4.2	4.6	0	0.0015	0.999944
	400	0.21	8.7	2.3	0	0.0017	0.999963
22,500	100	0.17	1.5	8.5	0	0.00031	0.999985
	225	0.19	4.2	4.6	0	0.00053	0.999973
	400	0.2	8.6	2.4	0	0.00058	0.999982
40,000	100	0.17	1.5	8.6	0	0.00017	0.999992
	225	0.19	4.2	4.7	0	0.00034	0.999985
	400	0.2	8.6	2.4	0	0.00011	0.99999

NOTE: Scores are averaged over 30 independent trials. Each column represents the number of observation samples, number of basis functions, Frobenius norm, KL divergence, percent of true zeros missed, percent of true nonzeros missed, estimated nugget minus true nugget, and the estimated negative log-likelihood divided by the true negative log-likelihood.

Acknowledgments

The authors thank the reviewers and associate editor for constructive comments that greatly improved the article.

Funding

This research was supported by NSF DMS-1811294 and DMS-1923062.

References

Bandyopadhyay, S., and Lahiri, S. N. (2009), "Asymptotic Properties of Discrete Fourier Transforms for Spatial Data," *Sankhyā*, 71, 221–259. [375]

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008), "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data," *Journal of Machine Learning Research*, 9, 485–516. [377]

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 825–848, DOI: 10.1111/j.1467-9868.2008.00663.x. [375]

Barabási, A.-L., and Albert, R. (1999), "Emergence of Scaling in Random Networks," *Science*, 286, 509–512, DOI: 10.1126/science.286.5439.509. [381]

Bien, J., and Tibshirani, R. J. (2011), "Sparse Estimation of a Covariance Matrix," *Biometrika*, 98, 807–820, DOI: 10.1093/biomet/asr054. [378]

Bolin, D., and Lindgren, F. (2011), "Spatial Models Generated by Nested Stochastic Partial Differential Equations, With an Application to Global Ozone Mapping," *The Annals of Applied Statistics*, 5, 523–550, DOI: 10.1214/10-AOAS383. [375]

Boyd, S., and Vandenberghe, L. (2004), *Convex Optimization*, New York: Cambridge University Press. [387]

Cai, T., Liu, W., and Luo, X. (2011), "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 106, 594–607, DOI: 10.1198/jasa.2011.tm10155. [376]

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012), "Rejoinder: Latent Variable Graphical Model Selection via Convex Optimization," *The Annals of Statistics*, 40, 2005–2013, DOI: 10.1214/11-AOS949. [376]

Cressie, N., and Johannesson, G. (2008), "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 209–226, DOI: 10.1111/j.1467-9868.2007.00633.x. [375,377,378]

Cressie, N., and Wikle, C. (2011), *Statistics for Spatio-Temporal Data*, CourseSmart Series, Hoboken, NJ: Wiley. [375]

Davanloo Tajbakhsh, S., Serhat Aybat, N., and Del Castillo, E. (2014), "Sparse Precision Matrix Selection for Fitting Gaussian Random Field Models to Large Data Sets," arXiv no. 1405.5576. [383]

Dinh Tao, P., and Le Thi, H. A. (1997), "Convex Analysis Approach to DC Programming: Theory, Algorithm and Applications," *Acta Mathematica Vietnamica*, 22, 289–355. [377]

Fattahi, S., Zhang, R. Y., and Sojoudi, S. (2019), "Linear-Time Algorithm for Learning Large-Scale Sparse Graphical Models," *IEEE Access*, 7, 12658–12672. [378]

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1, 302–332, DOI: 10.1214/07-AOAS131. [378]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [376,378]

Fuentes, M. (2002), "Spectral Methods for Nonstationary Spatial Processes," *Biometrika*, 89, 197–210, DOI: 10.1093/biomet/89.1.197. [375]

Gneiting, T., and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378, DOI: 10.1198/016214506000001437. [386]

Guhaniyogi, R., and Banerjee, S. (2018), "Meta-Kriging: Scalable Bayesian Modeling and Inference for Massive Spatial Datasets," *Technometrics*, 60, 430–444, DOI: 10.1080/00401706.2018.1437474. [375]

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. (2014), "QUIC: Quadratic Approximation for Sparse Inverse Covariance Estimation," *Journal of Machine Learning Research*, 15, 2911–2947. [378,379]

Hunter, D. R., and Lange, K. (2004), "A Tutorial on MM Algorithms," *The American Statistician*, 58, 30–37. [377]

Katzfuss, M. (2017), "A Multi-Resolution Approximation for Massive Spatial Datasets," *Journal of the American Statistical Association*, 112, 201–214, DOI: 10.1080/01621459.2015.1123632. [375]

Legates, D. R., and Willmott, C. J. (1990), "Mean Seasonal and Spatial Variability in Global Surface Air Temperature," *Theoretical and Applied Climatology*, 41, 11–21, DOI: 10.1007/BF00866198. [383]

Lindgren, F., Rue, H., and Lindström, J. (2011), "An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach" (with discussion and a reply by the authors), *Journal of the Royal Statistical Society*, Series B, 73, 423–498, DOI: 10.1111/j.1467-9868.2011.00777.x. [375,384]

Matsuda, Y., and Yajima, Y. (2009), "Fourier Analysis of Irregularly Spaced Data on $\mathbb{R}^d$," *Journal of the Royal Statistical Society*, Series B, 71, 191–217, DOI: 10.1111/j.1467-9868.2008.00685.x. [375]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462, DOI: 10.1214/009053606000000281. [376]

Nandy, S., Lim, H. Y., and Maiti, T. (2016), "Estimating Non-Stationary Spatial Covariance Matrix Using Multi-Resolution Knots," Technical Report. [376]

Nychka, D. (2000), "Spatial-Process Estimates as Smoothers," in *Smoothing and Regression: Approaches, Computation and Application*, ed. M. G. Schimek, New York: Wiley, pp. 393–424. [386]

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015), "A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 24, 579–599, DOI: 10.1080/10618600.2014.914946. [375,380]

Nychka, D., Wikle, C., and Royle, J. A. (2002), "Multiresolution Models for Nonstationary Spatial Covariance Functions," *Statistical Modelling*, 2, 315–331, DOI: 10.1191/1471082x02st037oa. [375]

Oyler, J. W., Ballantyne, A., Jencso, K., Sweet, M., and Running, S. W. (2015), "Creating a Topoclimatic Daily Air Temperature Dataset for the Conterminous United States Using Homogenized Station Data and Remotely Sensed Land Skin Temperature," *International Journal of Climatology*, 35, 2258–2279. [384]

Roweis, S., and Ghahramani, Z. (1999), "A Unifying Review of Linear Gaussian Models," *Neural Computation*, 11, 305–45. [376]

Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Monographs on Statistics and Applied Probability, Boca Raton, FL: Chapman & Hall/CRC. [376]

Seber, G. A. F. (2007), *A Matrix Handbook for Statisticians* (1st ed.), New York: Wiley-Interscience. [387]

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer-Verlag. [381]

——— (2014), "Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data," *Spatial Statistics*, 8, 1–19, DOI: 10.1016/j.spasta.2013.06.003. [376,384]

Tipping, M. E. (2001), "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, 1, 211–244, DOI: 10.1162/15324430152748236. [376]

Tropp, J. A. (2015), "An Introduction to Matrix Concentration Inequalities," *Foundations and Trends in Machine Learning*, 8, 1–230, DOI: 10.1561/2200000048. [382]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35, DOI: 10.1093/biomet/asm018. [377]

Zhao, T., Li, X., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2015), "huge: High-Dimensional Undirected Graph Estimation," R Package Version 1.2.7, available at *https://CRAN.R-project.org/package=huge*. [379,381]