Bayesian Analysis (2022) **17**, Number 1, pp. 291–351

Bayesian Nonstationary and Nonparametric Covariance Estimation for Large Spatial Data (with Discussion)[*]

Brian Kidd[†] and Matthias Katzfuss[‡]

Abstract. In spatial statistics, it is often assumed that the spatial field of interest is stationary and its covariance has a simple parametric form, but these assumptions are not appropriate in many applications. Given replicate observations of a Gaussian spatial field, we propose nonstationary and nonparametric Bayesian inference on the spatial dependence. Instead of estimating the quadratic (in the number of spatial locations) entries of the covariance matrix, the idea is to infer a near-linear number of nonzero entries in a sparse Cholesky factor of the precision matrix. Our prior assumptions are motivated by recent results on the exponential decay of the entries of this Cholesky factor for Matérn-type covariances under a specific ordering scheme. Our methods are highly scalable and parallelizable. We conduct numerical comparisons and apply our methodology to climate-model output, enabling statistical emulation of an expensive physical model.

Keywords: Bayesian linear regression, climate-model emulation, modified Cholesky factorization, ordered conditional independence, sparsity, Vecchia approximation.

1 Introduction

Modeling spatial data typically involves specification of spatial dependence in the form of a covariance function or matrix, under an implicit or explicit assumption of joint Gaussianity. This may involve many challenges, including small numbers of replicates, high-dimensional distributions, and complex, nonstationary dependence. Examples include gap-filling for satellite data (e.g., Cressie and Johannesson, 2008); forecast-covariance estimation in the ensemble Kalman filter (e.g., Furrer and Bengtsson, 2007; Katzfuss et al., 2016); conducting observing-system simulation experiments at the National Aeronautics and Space Administration (e.g., Zeng et al., 2021); and statistical climate-model emulation (e.g., Castruccio and Stein, 2013; Castruccio et al., 2014; Nychka et al., 2018; Wiens et al., 2020) based on an ensemble of spatial fields generated by an expensive computer model (Figure 1). Thus, there is a need for flexible and scalable methods for inferring high-dimensional spatial covariances.

Countless approximations have been proposed to address computational challenges in spatial statistics (see Heaton et al., 2019, for a recent review and comparison). In recent years, there has been increasing interest in the idea of Vecchia (1988), which

[*] MK's research was partially supported by National Science Foundation (NSF) Grants DMS–1654083, DMS–1953005, and CCF–1934904.

[†] Department of Statistics, Texas A&M University

[‡] Department of Statistics, Texas A&M University. Corresponding author, katzfuss@gmail.com

https://doi.org/10.1214/21-BA1273

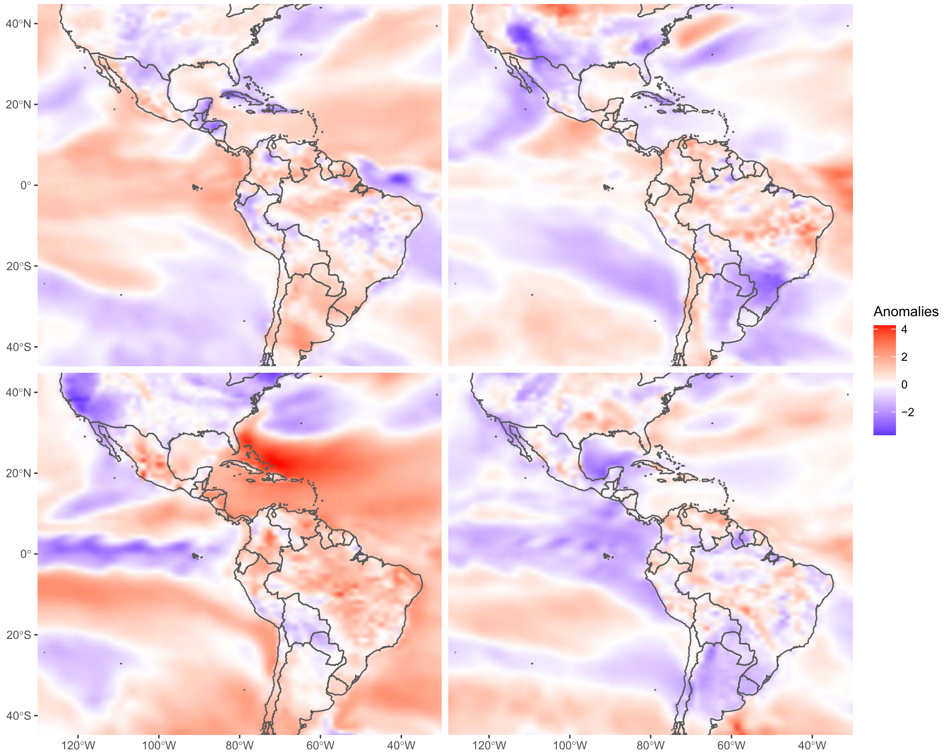


Figure 1: Four members of an ensemble of surface-temperature anomalies (in Kelvin) produced by a climate model, on a grid of size $n = 81 \times 96 = 7{,}776$ (see Section 4 for more details).

effectively approximates the Cholesky factor of the precision (i.e., inverse covariance) matrix as sparse. Under certain settings, the Vecchia approximation can provably provide ϵ-accurate approximations at near-linear computational complexity in the number of spatial locations (Schäfer et al., 2021a). A generalization of the Vecchia approach includes many popular spatial approximations as special cases (Katzfuss and Guinness, 2021). However, Vecchia approaches have mostly been used for approximating parametric and often isotropic covariance functions.

Isotropic, parametric covariance functions (e.g., Matérn) only depend on spatial distance and on a small number of unknown parameters. Despite being highly restrictive, this is the standard assumption in spatial statistics, especially in the absence of replicates. Approaches to relax these assumptions include parametric nonstationary covariances (e.g., as reviewed by Risser, 2016), stationary nonparametric covariances (e.g., Huang et al., 2011; Choi et al., 2013; Porcu et al., 2019), weighted sums of stationary covariances (e.g., Fuentes, 2002), and stationary covariances in transformed domains (e.g., Sampson and Guttorp, 1992; Damian et al., 2001; Qadir et al., 2019). In the context of local kriging, covariance functions are typically estimated locally from a parametric (e.g., Anderes and Stein, 2011) or nonparametric (e.g., Hsing et al., 2016) perspective; local estimation does not directly imply a valid joint model or positive-definite covariance matrix, but local parametric fits can be used to inform joint distributions (Nychka et al., 2018; Wiens et al., 2020).

Outside of spatial statistics, covariance estimation is often performed based on (modified) Cholesky decompositions of the precision matrix. This approach is attractive, because it automatically ensures positive-definiteness, because sparsity in the Cholesky factor directly corresponds to ordered conditional independence and hence to directed acyclic graphs, and because it allows covariance estimation to be reformulated as a series of regressions. Regularization can be achieved as in other regression settings, for example by enforcing sparsity using a Lasso-like penalty or a thresholding procedure (e.g., Huang et al., 2006; Levina et al., 2008) or via Bayesian prior distributions (e.g., Smith and Kohn, 2002). Motivated by a Gaussian Markov random field assumption for spatial data, Zhu and Liu (2009) estimate the Cholesky factor based on an ordering of the spatial locations intended to minimize the bandwidth, which amounts to coordinate ordering on a regular grid, and they regularize the entries of the Cholesky factor using a weighted Lasso penalty depending on spatial distance; this approach scales cubically in the number of spatial locations.

Here, we propose scalable nonparametric and nonstationary Bayesian inference on a high-dimensional spatial covariance matrix. The basic idea is to infer a near-linear number of nonzero entries in a sparse Cholesky factor of the inverse covariance matrix. Our model can be viewed as a nonparametric extension of the Vecchia approach, as regularized inference on a sparse Cholesky factor of the precision matrix, or as a series of Bayesian linear regression or spatial prediction problems. We specify prior distributions that are motivated by recent results (Schäfer et al., 2021b,a) on the exponential decay of the entries of the inverse Cholesky factor for Matérn-type covariances under a maximum-minimum-distance ordering of the spatial locations (Guinness, 2018; Schäfer et al., 2021b). Thus, we obtain a highly flexible method that enforces neither stationary nor parametric covariance structures, but instead regularizes the estimation and accounts for uncertainty via Bayesian priors. The resulting posterior contracts around the true covariance matrix as the number of replicates increases; an analysis of the climate data in Figure 1 indicates that this allows our method to outperform more restrictive approaches even for relatively small numbers of replicates. Our method scales well to very large datasets, as the number of nonzero entries in the Cholesky factor and the computational cost both scale near-linearly in the number of spatial locations, in effect inferring a near-linear number of parameters in the sparse inverse Cholesky factor instead of a square number of parameters in the dense covariance matrix. Further speedups are possible, as the main computational efforts are perfectly parallel. Our approach is applicable to a single realization of the spatial field, but the inference will be most useful and accurate if replicate observations are available.

The remainder of this document is organized as follows. Section 2 describes our methodology. Section 3 provides numerical comparisons using simulated data. In Section 4, our method is used for climate-model emulation. Section 5 concludes.

2 Methodology

2.1 Sparse inverse Cholesky approximation for spatial data

Consider a $N \times n$ matrix of spatial data,

$$\mathbf{Y} = \begin{pmatrix} y_1^{(1)} & \cdots & y_n^{(1)} \\ \vdots & \ddots & \vdots \\ y_1^{(N)} & \cdots & y_n^{(N)} \end{pmatrix} = \begin{pmatrix} — & \mathbf{y}^{(1)\prime} & — \\ & \vdots & \\ — & \mathbf{y}^{(N)\prime} & — \end{pmatrix} = \begin{pmatrix} | & & | \\ \mathbf{y}_1 & \cdots & \mathbf{y}_n \\ | & & | \end{pmatrix}, \tag{1}$$

where $y_i^{(\ell)}$ is the ℓth observation at spatial location $\mathbf{s}_i$. We assume that the locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, and hence the columns of $\mathbf{Y}$, are ordered according to a maximin ordering (Guinness, 2018; Schäfer et al., 2021b), which sequentially selects each location in the ordering to maximize the minimum distance from locations already selected (see Figure 2).

We model the rows $\mathbf{y}^{(\ell)} = (y_1^{(\ell)}, \ldots, y_n^{(\ell)})'$ of $\mathbf{Y}$ as independent n-variate Gaussians:

$$\mathbf{y}^{(\ell)} | \mathbf{\Sigma} \overset{iid}{\sim} \mathcal{N}_n(\mathbf{0}, \mathbf{\Sigma}), \qquad \ell = 1, \ldots, N. \tag{2}$$

We assume that the data are centered, either using an ad-hoc pre-processing step (e.g., by subtracting location-wise means) or using a more elaborate procedure (see Section 2.8).

Our goal is to make inference on the $n \times n$ spatial covariance matrix $\mathbf{\Sigma}$ based on the $N \times n$ observations $\mathbf{Y}$, in the case where n is large (at least in the thousands) and N is relatively small. Typically, a parametric, and often isotropic, covariance function is assumed such that $\mathbf{\Sigma}$ is a function of only a small number of parameters, which can then be estimated relatively easily. Here, we avoid explicit assumptions of isotropy or parametric structure.

We assume a form of ordered conditional independence,

$$p(y_i^{(\ell)} | \mathbf{y}_{1:i-1}^{(\ell)}, \mathbf{\Sigma}) = p(y_i^{(\ell)} | \mathbf{y}_{g_m(i)}^{(\ell)}, \mathbf{\Sigma}), \qquad i = 2, \ldots, n, \quad \ell = 1, \ldots, N, \tag{3}$$

where $g_m(i) \subset (1, \ldots, i-1)$ is an index vector consisting of the indices of the $\min(m, i-1)$ nearest neighbors to $\mathbf{s}_i$ among those ordered previously; that is, $\mathbf{s}_{(g_m(i))_j}$ is the jth nearest neighbor of $\mathbf{s}_i$ among $\mathbf{s}_1, \ldots, \mathbf{s}_{i-1}$ (see Figure 2). While (3) holds trivially for $m = n-1$, for many covariance structures it even holds (at least approximately) for $m \ll n$, as has been demonstrated numerically (e.g., Vecchia, 1988; Stein et al., 2004; Datta et al., 2016; Guinness, 2018; Katzfuss and Guinness, 2021; Katzfuss et al., 2020a,b) and theoretically (Schäfer et al., 2021a) in the context of Vecchia approximations of parametric covariance functions. Assume for now that m is known.

Consider the modified Cholesky decomposition of the precision matrix:

$$\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}', \tag{4}$$

where $\mathbf{D} = \text{diag}(d_1, \ldots, d_n)$ is a diagonal matrix with positive entries $d_i > 0$, and $\mathbf{U}$ is an upper triangular matrix with unit diagonal (i.e., $\mathbf{U}_{ii} = 1$). (To be precise, (4) is the

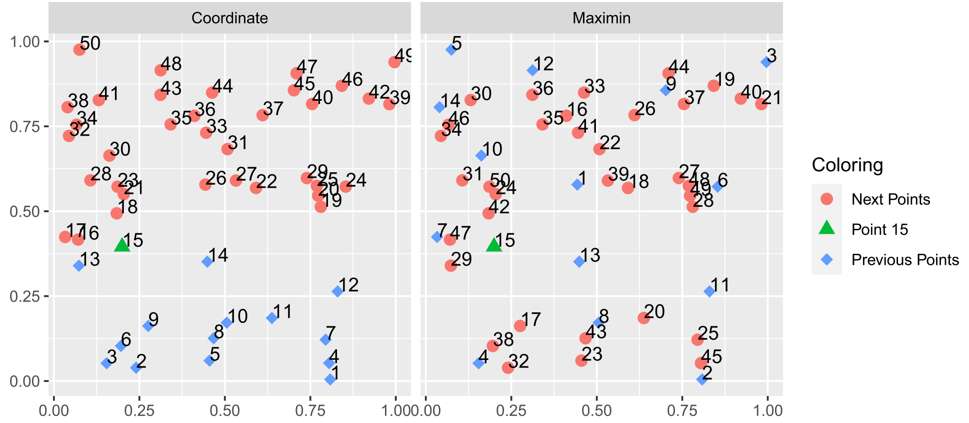


Figure 2: For $n = 50$ randomly sampled locations on the unit square, comparison of coordinate (bottom to top) and maximin ordering. For $i = 15$, previously ordered locations $\mathbf{s}_1, \ldots, \mathbf{s}_{n-1}$ are highlighted in blue to show their roughly equidistant spread over the domain for maximin. As an example, for $m = 4$, we would have conditioning sets $g_4(15) = (13, 9, 14, 6)$ for coordinate and $g_4(15) = (7, 13, 10, 1)$ for maximin.

reverse-ordered Cholesky factorization of the reverse-ordered $\mathbf{\Sigma}^{-1}$, which simplifies our notation later.) The ordered conditional independence assumed in (3) implies that $\mathbf{U}$ is sparse, with at most m nonzero off-diagonal elements per column (e.g., Katzfuss and Guinness, 2021, Proposition 3.1). We define $\mathbf{u}_i = \mathbf{U}_{g_m(i),i}$ as the nonzero off-diagonal entries in the ith column.

2.2 Covariance estimation via Bayesian regressions

From (4), we see that we can estimate the $\mathcal{O}(n^2)$ unknown entries of $\mathbf{\Sigma}$ by inferring the $\mathcal{O}(nm)$ variables $d_1, \ldots, d_n$ and $\mathbf{u}_1, \ldots, \mathbf{u}_n$. To do so, our data model (2) can be written as a series of n linear regression models (Huang et al., 2006):

$$p(\mathbf{Y}|\mathbf{\Sigma}) = \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{y}_{1:i-1}, \mathbf{\Sigma}) = \prod_{i=1}^{n} \mathcal{N}_N(\mathbf{y}_i|\mathbf{X}_i\mathbf{u}_i, d_i\mathbf{I}_N), \tag{5}$$

where the "response vector" $\mathbf{y}_i = (y_i^{(1)}, \ldots, y_i^{(N)})'$ is the ith column of $\mathbf{Y}$ in (1) consisting of the N observations at the ith spatial location, and the "design matrix" $\mathbf{X}_i$ consists of the observations at the m neighbor locations of $\mathbf{s}_i$, stored in the columns of $\mathbf{Y}$ with indices $g_m(i)$; specifically, $\mathbf{X}_i$ is an $N \times m$ matrix with ℓth row $-\mathbf{y}_{g_m(i)}^{(\ell)}{}'$.

The Bayesian regression models in (5) are completed by independent conjugate normal-inverse-gamma (NIG) priors:

$$\mathbf{u}_i|d_i, \boldsymbol{\theta} \overset{ind.}{\sim} \mathcal{N}(\mathbf{0}, d_i\mathbf{V}_i), \qquad d_i|\boldsymbol{\theta} \overset{ind.}{\sim} \mathcal{IG}(\alpha_i, \beta_i), \qquad i = 1, \ldots, n, \tag{6}$$

where $\boldsymbol{\theta}$ is a vector of hyperparameters determining m, $\mathbf{V}_i$, α_i, and β_i (see Section 2.3 below). Due to conjugacy, the posterior distributions (conditional on $\boldsymbol{\theta}$) are also NIG:

$$\begin{aligned} p(\mathbf{u}_1, \ldots, \mathbf{u}_n, d_1, \ldots, d_n | \mathbf{Y}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} p(\mathbf{u}_i, d_i | \mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{u}_i | d_i, \mathbf{Y}, \boldsymbol{\theta})\, p(d_i | \mathbf{Y}, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \mathcal{N}(\mathbf{u}_i | \hat{\mathbf{u}}_i, d_i \mathbf{G}_i)\, \mathcal{IG}(d_i | \widetilde{\alpha}_i, \widetilde{\beta}_i), \end{aligned} \tag{7}$$

where $\hat{\mathbf{u}}_i = \mathbf{G}_i \mathbf{X}_i' \mathbf{y}_i$, $\mathbf{G}_i = (\mathbf{X}_i'\mathbf{X}_i + \mathbf{V}_i^{-1})^{-1}$, $\widetilde{\alpha}_i = \alpha_i + N/2$, and $\widetilde{\beta}_i = \beta_i + (\mathbf{y}_i'\mathbf{y}_i - \hat{\mathbf{u}}_i' \mathbf{G}_i^{-1} \hat{\mathbf{u}}_i)/2 = \beta_i + (\mathbf{y}_i'(\mathbf{I}_N + \mathbf{X}_i \mathbf{V}_i \mathbf{X}_i')^{-1} \mathbf{y}_i)/2$.

Using (7), we can easily obtain samples or posterior summaries of the entries of $\mathbf{U}$ and $\mathbf{D}$ conditional on $\boldsymbol{\theta}$. However, in many applications, primary interest will be in computing posterior summaries of $\boldsymbol{\Sigma}$ and other quantities. If n is not too large ($n < 10^4$, say), we can simply compute $\boldsymbol{\Sigma}^{-1}$ (and hence $\boldsymbol{\Sigma}$) from $\mathbf{U}$ and $\mathbf{D}$. For large n, it is often not possible to even hold the entire dense matrix $\boldsymbol{\Sigma}$ in memory, but we can quickly compute useful summaries of it based on the sparse matrices $\mathbf{U}$ and $\mathbf{D}$ (e.g., Katzfuss et al., 2020a). For example, a selected inversion algorithm can compute the variances $\boldsymbol{\Sigma}_{ii}$ and all entries $\boldsymbol{\Sigma}_{ij}$ for which $i \in g_m(j)$ or $j \in g_m(i)$. We can also compute the covariance matrix for any set of linear combinations $\mathbf{H}\mathbf{y}^{(\ell)}$ as $\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' = \mathbf{A}'\mathbf{A}$, where $\mathbf{A} = \mathbf{D}^{1/2}\mathbf{U}^{-1}\mathbf{H}'$. In many applications, including climate-model emulation, it is of interest to sample new spatial fields from the model, which we can do by sampling $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and then setting $\mathbf{y}^\star = (\mathbf{U}')^{-1}\mathbf{D}^{1/2}\mathbf{z}$; if $\mathbf{U}$ and $\mathbf{D}$ are sampled from their posterior distribution given $\mathbf{Y}$, then we have obtained a sample from the posterior predictive distribution $p(\mathbf{y}^\star | \mathbf{Y})$.

2.3 Parameterization of the prior distributions

We now discuss parameterizing the NIG priors for $\mathbf{u}_i$ and d_i in (6) as a function of a small number of hyperparameters, $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$, inspired by the behavior of Matérn-type covariance functions. The parameter θ_1 is related to the marginal variance, while θ_2 and θ_3 are related to the range and smoothness. In general, our prior parameterizations are motivated by interpreting $\mathbf{u}_i$ and d_i as the kriging weights and variance, respectively, for the spatial prediction problem implied by (3), consisting of predicting $y_i^{(\ell)}$ from $\mathbf{y}_{g_m(i)}^{(\ell)}$; due to the maximin ordering, the locations of the variables in $\mathbf{y}_{g_m(i)}^{(\ell)}$ all have roughly similar distance to $\mathbf{s}_i$ (see Figure 2), and this distance decreases systematically with i.

First, consider $d_i \sim \mathcal{IG}(\alpha_i, \beta_i)$ in (6). For an exponential covariance with variance θ_1 and range $2/\theta_2$, we have $\boldsymbol{\Sigma}_{i,j} = \theta_1 \exp(-\theta_2 \|\mathbf{s}_i - \mathbf{s}_j\|/2)$; assuming $m = 1$, we obtain

$$d_i = \operatorname{var}(y_i^{(\ell)} | \mathbf{y}_{g_m(i)}^{(\ell)}) = \theta_1 - \tfrac{(\theta_1 \exp(-\theta_2 \|\mathbf{s}_i - \mathbf{s}_g\|/2))^2}{\theta_1} = \theta_1(1 - e^{-\theta_2 \|\mathbf{s}_i - \mathbf{s}_g\|}), \tag{8}$$

where $g = g_1(i)$, and the distance $\|\mathbf{s}_i - \mathbf{s}_g\|$ between location $\mathbf{s}_i$ and its nearest previously ordered neighbor decreases roughly as $(i)^{-1/p}$ for a regular grid on a unit hypercube,

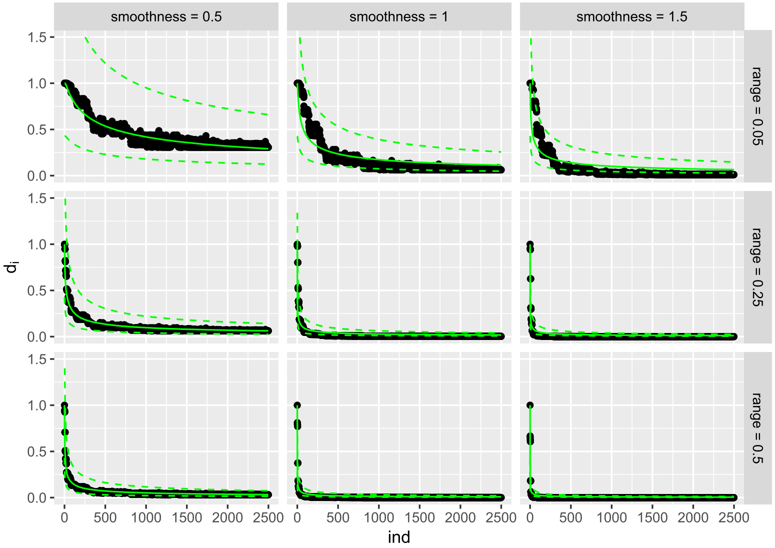


Figure 3: Illustration of the true entries $d_1, \ldots, d_n$ of $\mathbf{D}$ as a function of location index i for a Matérn covariance function on a regular $n = 50\times 50 = 2{,}500$ grid on the unit square. The columns correspond to different smoothness parameters, while the rows correspond to different range parameters. The dashed lines are approximate 95% pointwise intervals implied by our inverse-gamma prior, where θ_2 was chosen for illustration using a least-squares fitting procedure (`nls` in `R`) assuming known $\theta_1 = 1$.

$\mathbb{D} = [0,1]^p$. (Throughout, i is an index and not the imaginary number.) This motivates a prior for d_i that shrinks toward $d_i \approx \theta_1(1 - e^{-\theta_2 (i)^{-1/p}})$. While (8) only holds exactly for an exponential covariance with $m = 1$, Figure 3 illustrates that this functional form approximately holds for Matérn covariance functions in two dimensions with $m = n-1$ as well. Thus, we set the prior mean as $E(d_i|\boldsymbol{\theta}) = \beta_i/(\alpha_i - 1) = \theta_1 f_{\theta_2}(i)$, where $f_{\theta_2}(i) = 1 - e^{-\theta_2(i)^{-1/p}}$. In Figure 3, the empirically observed variance of the d_i elements around the fit line decreases with i as well, and so we set the prior standard deviation of d_i to be half of the mean. Solving for α_i and β_i, we obtain $\alpha_i = 6$ and $\beta_i = 5\theta_1 f_{\theta_2}(i)$, because $Var(d_i|\boldsymbol{\theta}) = \beta_i^2/((\alpha_i - 1)^2(\alpha_i - 2))$.

Recent results based on elliptic boundary-value problems (Schäfer et al., 2021b, Section 6.2) imply that the Cholesky entry $(\mathbf{u}_i)_j$, corresponding to the jth nearest neighbor, decays exponentially as a function of j, for Matérn covariance functions whose spectral densities are the reciprocal of a polynomial (ignoring edge effects). Thus, we assume $v_{ij} = \exp(-\theta_3 j)/(\theta_1 f_{\theta_2}(i))$ for $\mathbf{V}_i = \text{diag}(v_{i1}, \ldots, v_{im})$ in $\mathbf{u}_i | d_i, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, d_i \mathbf{V}_i)$ in (6). Note that we divide by $E(d_i|\boldsymbol{\theta})$ in v_{ij}, because the prior variance in $(\mathbf{u}_i)_j|\boldsymbol{\theta} \sim \mathcal{N}(0, d_i v_{ij})$ is multiplied by d_i. Figure 4 demonstrates this exponential decay as the neighbor number increases.

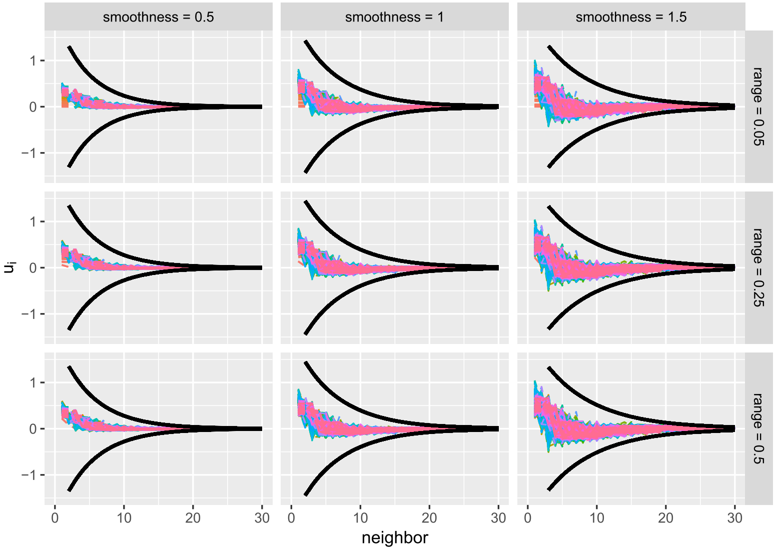


Figure 4: Illustration of the entries $(\mathbf{u}_i)_j$ of $\mathbf{U}$ as a function of neighbor number j for the same setting as in Figure 3. The dark lines correspond to approximate pointwise 95% prior intervals $(\pm 2\sqrt{\exp(-\theta_3 i)})$.

Finally, consider the choice of conditioning-set size m. Simply setting m to a fixed, reasonable value (e.g., $m \approx 10$, depending on computational constraints) works well in many settings, but the results can be highly inaccurate if m is chosen too small, and the computational cost is unnecessarily high if m is chosen too large. Hence, we prefer to allow the data to choose m by tying m to the prior decay of the elements of $\mathbf{U}$; for all of our numerical experiments, we set m as the largest j such that $\exp(-\theta_3 j) > 0.001$, where j denotes the neighbor number. This coincides with the amount of variation expected to be learnable from the data. Thus, entries of $\mathbf{U}$ with sufficiently small prior variance as implied by a specific θ_3 are set to zero, which ensures computational feasibility of our method.

2.4 Inference on the hyperparameters θ

The hyperparameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$ determine m, $\mathbf{V}_i$, α_i, and β_i as described in Section 2.3. We now discuss how $\boldsymbol{\theta}$ can be inferred based on the data $\mathbf{Y}$. All elements of $\boldsymbol{\theta}$ are assumed to be positive due to the decay previously discussed, and so we perform all inference on the logarithmic scale.

The crucial ingredient for inference on $\boldsymbol{\theta}$ is the marginal or integrated likelihood, which can be obtained by combining (5) and (6), moving the product over locations outside of the integral over the entries of $\mathbf{U}$ and $\mathbf{D}$, and simplifying using standard results for conjugate Gaussian models (e.g. Murphy, 2007):

$$
\begin{aligned}
p(\mathbf{Y}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} \int_{d_i} \int_{\mathbf{u}_i} \mathcal{N}_N(\mathbf{y}_i|\mathbf{X}_i\mathbf{u}_i, d_i\mathbf{I}_N)\mathcal{N}(\mathbf{0}, d_i\mathbf{V}_i)\mathcal{IG}(\alpha_i, \beta_i) d\mathbf{u}_i dd_i \\
&\propto \prod_{i=1}^{n} \Big(\frac{|\mathbf{G}_i|^{1/2}}{|\mathbf{V}_i|^{1/2}} \frac{\beta_i^{\alpha_i}}{\widetilde{\beta}_i^{\widetilde{\alpha}_i}} \frac{\Gamma(\widetilde{\alpha}_i)}{\Gamma(\alpha_i)} \Big),
\end{aligned} \tag{9}
$$

where Γ denotes the gamma function, the prior parameters $\alpha_i, \beta_i, \mathbf{V}_i$ are given in (6), and the posterior parameters $\widetilde{\alpha}_i, \widetilde{\beta}_i, \mathbf{G}_i$ are given in (7).

Based on this integrated likelihood, both empirical and fully Bayesian inference are straightforward. Empirical Bayesian inference is based on a point estimate of $\boldsymbol{\theta}$ obtained by numerically maximizing the log integrated likelihood. Fully Bayesian inference requires the specification of a hyperprior for $\boldsymbol{\theta}$, which we simply assume to be flat (on the log scale). As a result, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\theta})$ is proportional to the integrated likelihood in (9). While this distribution cannot be obtained analytically, we can sample from the posterior using the Metropolis-Hastings (MH) algorithm. To avoid slow mixing due to large negative correlation between θ_1 and θ_2, we employ an adaptive MH algorithm that jointly proposes $\boldsymbol{\theta}$ and learns its covariance matrix on-line; specifically, we use the implementation in R by Scheidegger (2012).

2.5 Computational complexity

The cost for inference, including computing the posteriors in (7), sampling $\mathbf{y}^\star$, or evaluating the integrated likelihood in (9), is dominated by computing the $m \times m$ matrix $\mathbf{G}_i$, which requires $\mathcal{O}(m^2 N)$ time, and decomposing $\mathbf{G}_i$, which requires $\mathcal{O}(m^3)$ time, for each $i = 1, \ldots, n$. Hence, the time complexity is $\mathcal{O}(n(m^2N + m^3))$ for each unique value of $\boldsymbol{\theta}$, where m is often very small (e.g., $m \approx 10$ in most of our numerical experiments). In addition, the most expensive computations can be carried out in parallel over $i = 1, \ldots, n$.

For very small numbers of replicates, with $N < m$, we can use alternative expressions (see below (7)) relying on computing and decomposing the $N \times N$ matrix $\mathbf{X}_i\mathbf{V}_i\mathbf{X}_i' + \mathbf{I}_N$ (instead of $\mathbf{G}_i$), which requires $\mathcal{O}(mN^2 + N^3) = \mathcal{O}(mN^2)$ time.

The maximin ordering and large nearest-neighbor conditioning sets (with $m_{\max} = 50$, say) can be computed in quasilinear time in n (Schäfer et al., 2021b,a). For any $m \leq m_{\max}$ implied by a specific $\boldsymbol{\theta}$, we can then simply select $g_m(i)$ as the first m entries of $g_{m_{\max}}(i)$.

2.6 Asymptotics

Assuming temporarily that (2) holds for some true $n \times n$ positive-definite covariance matrix $\boldsymbol{\Sigma}_0$, the data model with the true $\boldsymbol{\Sigma}_0$ can be written in the regression form (5) with $m = n - 1$. Holding n fixed and assuming $N \to \infty$, there are a fixed number (depending only on n, not on N) of variables in the regression models, and our prior distributions on the $\mathbf{u}_i$, d_i, and $\boldsymbol{\theta}$ place nonzero mass on the true model. Hence, using well-known asymptotic results based on the Bernstein–von Mises theorem (e.g., Van der Vaart, 2000), the posterior distributions will be asymptotically normal and our posterior of $\boldsymbol{\Sigma}$ will contract around the true covariance $\boldsymbol{\Sigma}_0$ as the number of independent replicates N approaches infinity. While we are most interested in the case $N \ll n$, our climate-data analysis in Section 4 will demonstrate that the posterior-contraction properties allow our method to outperform more restrictive approaches even for relatively small N.

2.7 Correlation-based ordering

For our methods, as discussed in Section 2.1, we recommend a maximin ordering of the variables $y_1, \ldots, y_n$, and then selecting the conditioning sets $g_m(i)$ based on the m nearest previously ordered variables, with m determined by $\boldsymbol{\theta}$ as described at the end of Section 2.3. So far, these tasks were assumed to be based on the Euclidean distance of the corresponding locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ (see Figure 2), which implies that our priors shrink toward isotropy (i.e., distributions for which dependence is only a function of distance). This shrinkage is not appropriate in some real-data applications. However, it is relatively straightforward to adapt our methods to processes (e.g., anisotropic or nonstationary) for which Euclidean distance is not meaningful. We merely require some prior guess of the correlation structure, based on expert knowledge, historical data, or (a regularized version of) the sample correlation of the data $\mathbf{Y}$; a simple choice used here is the element-wise product of the sample correlation and an (isotropic) exponential correlation with a large range parameter (e.g., half the maximum distance between any pair of locations in the dataset). Then, our procedures can be carried out as before, except that the ordering and nearest-neighbor conditioning is based on a correlation distance, defined as $(1 - |\text{correlation}|)^{1/2}$. This implicitly scales the space, so that the process is approximately isotropic in the transformed space. This approach can increase accuracy in the context of Vecchia approximations of parametric covariances (Kang and Katzfuss, in prep.); we propose it here for our nonparametric procedures. Schäfer et al. (2021a, Algorithm 7) allows us to compute the correlation-based ordering and conditioning sets in quasilinear time in n.

2.8 Noise or spatial trend

Our methodology described so far is most appropriate if the data are observed without any noise or nugget, meaning that realizations of the underlying spatial field are continuous over space; in this setting, approximations based on sparse inverse Cholesky factors of many popular covariance functions can be highly accurate (e.g., Katzfuss and Guinness, 2021; Schäfer et al., 2021a).

Now consider noisy observations $\mathbf{w}^{(\ell)} | \mathbf{y}^{(\ell)} \overset{iid}{\sim} \mathcal{N}_n(\mathbf{y}^{(\ell)}, \tau^2 \mathbf{I}_n)$, $\ell = 1, \ldots, N$, with $\mathbf{y}^{(\ell)}$ as in (2). One option is to simply apply our methodology directly to the data $\mathbf{w}^{(\ell)}$ as before; this will likely work well if the noise variance τ^2 is small, but the conditional-independence assumption in (3) is less appropriate if τ^2 is large (e.g., Katzfuss and Guinness, 2021), meaning that a much larger m might be necessary. A larger m results in higher computational cost and potentially less accuracy due to the higher number of Cholesky entries that must be estimated.

Hence, for large noise levels, we instead propose a Gibbs sampler that iterates between sampling $\mathbf{y}^{(\ell)}$ conditional on $\mathbf{w}^{(\ell)}$ and $\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}'$, and sampling $\boldsymbol{\theta}$ and the entries of $\mathbf{U}$ and $\mathbf{D}$ conditional on the $\mathbf{y}^{(\ell)}$ as in Sections 2.2 and 2.4. The former task can be accomplished without increasing the computational complexity for each Gibbs iteration, by exploiting the sparsity of the Cholesky factor $\mathbf{U}\mathbf{D}^{-1/2}$ of the prior precision, and approximating the Cholesky factor of the posterior precision using an incomplete Cholesky factorization to avoid fill-in as described in Schäfer et al. (2021a, Section 4.1). (If τ^2 is unknown, it is straightforward to sample from its full-conditional distribution as well.)

A similar Gibbs-sampling strategy can be employed to make inference on a spatial trend. For example, if the observations $\mathbf{w}^{(\ell)}$ are given by $\mathbf{y}^{(\ell)}$ plus a linear spatial trend with a Gaussian prior on the trend coefficients, the coefficients can be sampled in closed form conditional on $\mathbf{\Sigma}$, and all other unknown quantities can be sampled given the trend coefficients as before based on $\mathbf{y}^{(\ell)}$ obtained by subtracting the trend from $\mathbf{w}^{(\ell)}$.

2.9 Shrinkage toward a specific covariance

Our methodology can be modified to center the prior distributions at and thus shrink toward a specific covariance function C. For $i = 1, \ldots, n$, define

$$\begin{aligned} \mathbf{u}_i^{(m)} &= -C(\mathcal{S}_{g_m(i)}, \mathcal{S}_{g_m(i)})C(\mathcal{S}_{g_m(i)}, \mathbf{s}_i), \\ d_i^{(m)} &= C(\mathbf{s}_i, \mathbf{s}_i) + \mathbf{u}_i^{(m)\prime}C(\mathcal{S}_{g_m(i)}, \mathbf{s}_i), \end{aligned}$$

where $\mathbf{u}_1^{(m)} = 0$, and $\mathcal{S}_{g_m(i)}$ is the (ordered) set of locations corresponding to $\mathbf{y}_{g_m(i)}$. If we assume $\mathbf{\Sigma}_{ij} = C(\mathbf{s}_i, \mathbf{s}_j)$, then we can write $\mathbf{\Sigma}$ as in (4) with $\mathbf{u}_i = \mathbf{u}_i^{(i-1)}$ and $d_i = d_i^{(i-1)}$ (e.g., Katzfuss et al., 2020a, App. B). The Vecchia approximation essentially exploits that $(\mathbf{u}_i^{(i-1)})_j$ decays rapidly as a function of the neighbor number j for many covariance functions C; this is illustrated for various members of the Matérn family in Figure 4, and for generalized Cauchy covariances in Figure 5. Thus, we can set the prior mean of the $\mathbf{u}_i$ and d_i in (6) to the values implied by a Vecchia approximation of C with $m \ll n$.

Specifically, we first need to determine a sparsity level m; for example, we can choose m arbitrarily, or we can set it as the maximum integer j such that the average of $(\mathbf{u}_2^{(j)})_j^2, \ldots, (\mathbf{u}_n^{(j)})_j^2$ is above some small threshold. Given m, we then assume a NIG prior as in (6), except with $E(d_i) = d_i^{(m)}$ and with a nonzero mean for the normal distribution, $\mathbf{u}_i|d_i \sim \mathcal{N}(\boldsymbol{\mu}_i, d_i\mathbf{V}_i)$, where $\boldsymbol{\mu}_i = \mathbf{u}_i^{(m)}$. When determining α_i and β_i, we can again assume the prior standard deviation of d_i to be some multiple (e.g., half) of its mean; similarly, we can assume $v_{ij} = (a\mu_{ij})^2/E(d_i)$ (e.g., with $a = 1/2$).

Inference can then proceed as in Section 2.2, except that due to the nonzero prior mean of $\mathbf{u}_i$, we now have $\hat{\mathbf{u}}_i = \mathbf{G}_i(\mathbf{X}_i'\mathbf{y}_i + \mathbf{V}_i^{-1}\boldsymbol{\mu}_i)$ and $\widetilde{\beta}_i = \beta_i + (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\mu}_i)'(\mathbf{I}_N + \mathbf{X}_i\mathbf{V}_i\mathbf{X}_i')^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\mu}_i)/2$.

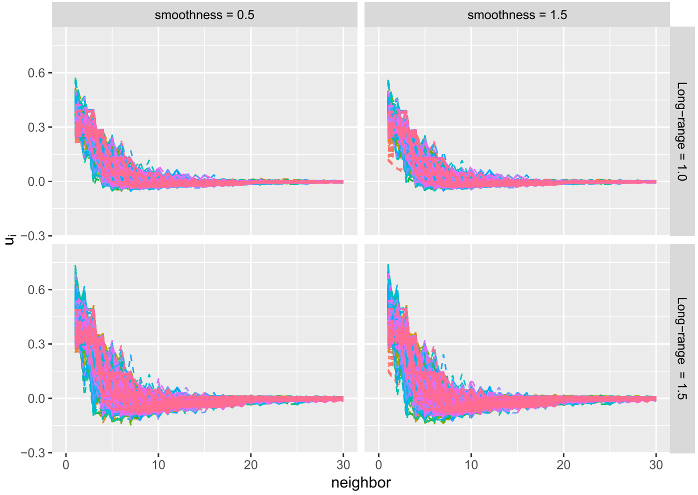


Figure 5: Illustration of $(\mathbf{u}_i^{(i-1)})_j$ as a function of neighbor number j as in Figure 4, except for a generalized Cauchy covariance, $C(\mathbf{s}_i, \mathbf{s}_j) = (1 + (\|\mathbf{s}_i - \mathbf{s}_j\|/\lambda)^\eta)^{(-\nu/\eta)}$, with range parameter $\lambda = 0.25$ (as in Figure 8), and different values of the smoothness $\nu \in \{0.5, 1.5\}$ and long-range dependence $\eta \in \{1.0, 1.5\}$.

In most cases, the goal will be to shrink toward a family of covariance functions, rather than a specific member of the family, and so we really have $C_{\boldsymbol{\theta}}$ that depends on an unknown parameter vector $\boldsymbol{\theta}$. In that case, we can make inference on $\boldsymbol{\theta}$ using the integrated likelihood as described in Section 2.4.

In practice, it may be unclear which covariance family to choose, and in many applications no standard family may be appropriate. Our nonparametric approach described in earlier sections avoids such arbitrary choices; it is also computationally cheaper than the parametric shrinkage here, which has time complexity $\mathcal{O}(nm^4)$ due to the search over m.

3 Simulation study

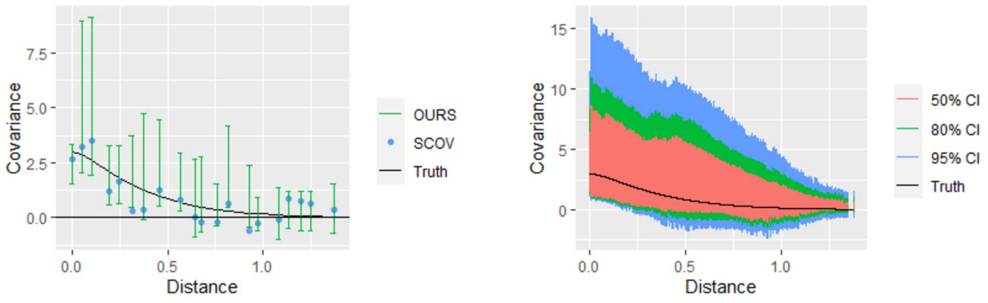
We compared the following methods:

SCOV: Basic sample covariance

OURS: Our method described in Sections 2.1–2.4

MLE: Covariance estimate based on the maximum likelihood estimates of $\mathbf{u}_i$ and d_i for the regressions in (5) (i.e., no prior shrinkage), with $m = \min(m_{\text{OURS}}, N-1)$, with m_{OURS} implied by OURS $\boldsymbol{\theta}$ estimate

LASSO: Least absolute shrinkage and selection operator for each regression in (5), with all previous points included as possible predictors (i.e., $m = n - 1$)



(a) (b)

Figure 6: Based on $N = 20$ draws from a Gaussian process with Matérn covariance at $n = 900$ locations (see Section 3.1): (a) Sample estimates (SCOV) and posterior 95% credible intervals using our fully Bayesian method (OURS) for 20 entries of the covariance matrix. (b) 50%, 80%, and 95% credible intervals using OURS for one randomly sampled entry of the covariance matrix corresponding to each unique distance.

SLASSO: Spatial LASSO with penalty scaled by the spatial distance to favor inclusion of nearer points as predictors, intended to be similar to Zhu and Liu (2009)

autoFRK: Resolution-adaptive automatic fixed rank kriging (Tzeng and Huang, 2018; Tzeng et al., 2021) with approximately $\sqrt{n}$ basis functions, resulting in a similar number of parameters as OURS

The spatial domain for all comparisons was the unit square.

3.1 Uncertainty quantification

First, we fit a fully Bayesian version of OURS to simulated data, to demonstrate the uncertainty quantification in the covariance estimation. Specifically, we considered $N =$ 20 realizations of a Gaussian process with Matérn covariance function with variance 3, smoothness 1, and range parameter 0.25, at $n = 900$ randomly sampled locations. We obtained 50,000 posterior samples of $\boldsymbol{\theta}$. The trace plots showed good mixing and convergence, and the individual effective sample sizes for the three parameters were all larger than 1,000. After conservatively discarding the first half of the samples for burn-in and thinning by a factor of 50, a covariance matrix was calculated from a sample from (7) for each $\boldsymbol{\theta}$ draw.

Figure 6a shows the resulting 95% posterior credible intervals (CIs) along with the SCOV estimates for 20 randomly sampled matrix entries $\boldsymbol{\Sigma}_{ij}$ as a function of $\|\mathbf{s}_i - \mathbf{s}_j\|$, the distance between the corresponding spatial locations. Most of the OURS CIs contained the true value and tracked the decay of the covariance as a function of distance. This is also the general trend for CIs at all distances shown in Figure 6b.

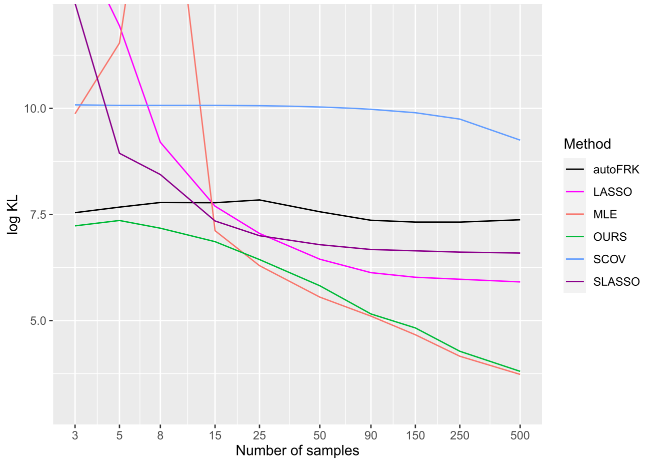


Figure 7: For the comparison in Section 3.2, KL divergence (on a log scale) for different covariance estimation methods for varying numbers N of samples from a Matérn covariance at $n = 900$ locations.

3.2 Comparison to LASSO for small n

We compared estimation accuracy using the Kullback-Leibler (KL) divergence between the estimated distribution $\mathcal{N}_n(\mathbf{0}, \hat{\mathbf{\Sigma}})$ and the true distribution $\mathcal{N}_n(\mathbf{0}, \mathbf{\Sigma})$:

$$\mathrm{KL}(\hat{\mathbf{\Sigma}} \| \mathbf{\Sigma}) = \mathrm{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}) - \log|\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}| - n,$$

where $\mathrm{tr}(\cdot)$ denotes the trace and $|\cdot|$ denotes the determinant. This exclusive KL divergence does not require inverting the estimate $\hat{\mathbf{\Sigma}}$ and thus avoids issues with SCOV for $N < n$. For ease of computation and comparison to the non-Bayesian methods, we computed the KL divergence for OURS based on a point estimate $\hat{\mathbf{\Sigma}} = (\hat{\mathbf{U}}^{-1})'\hat{\mathbf{D}}\hat{\mathbf{U}}^{-1}$, where $\hat{\mathbf{U}}$ and $\hat{\mathbf{D}}$ were the maximum a posteriori (MAP) estimates from (7), using the value of $\boldsymbol{\theta}$ that maximized the integrated likelihood (9).

Figure 7 shows the results, using the same set-up with $n = 900$ as in Section 3.1, for various numbers of replicates N. autoFRK performed nearly as well as our method for small N, but it did not meaningfully improve with larger N. MLE was similarly accurate as OURS for large N, as expected, but it performed worse for small N due to the lack of prior shrinkage. Similarly, the inclusion of spatial information in SLASSO resulted in higher accuracy than LASSO for small N. LASSO and SLASSO were not competitive with OURS and MLE, despite increased flexibility in selecting predictors (i.e., conditioning sets) in the regressions (5), and despite much higher computational cost due to calculations involving all $O(n)$ possible predictors.

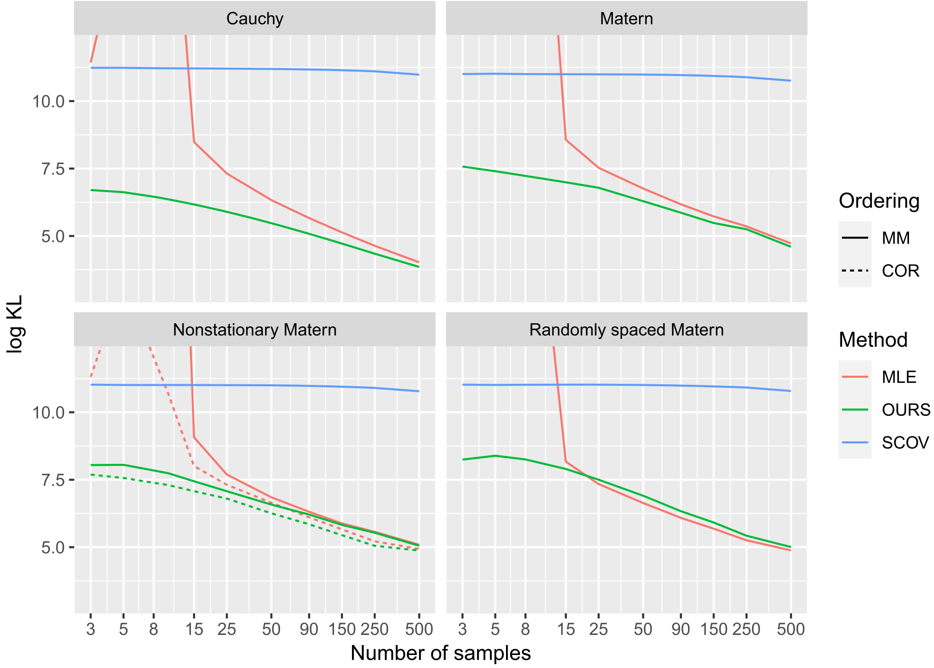


Figure 8: Comparison of KL divergence (on a log scale) for four different settings with n = 2,500 described in Section 3.3. MM: maximin ordering. COR: Correlation-based ordering (only used in the nonstationary setting).

3.3 Comparison for larger n

Figure 8 shows further comparisons with n = 2,500 spatial locations using the KL divergence (again based on the MAP estimate for OURS) in four different settings (counter-clockwise from top right), all with a marginal variance of 5: Matérn with smoothness 1 and range parameter 0.5 on a regular 50×50 spatial grid (corresponding to the middle panel in the bottom row of Figures 3 and 4); a Cauchy covariance with range 0.25 and memory parameters 1 and 0.5 on a regular 50×50 grid; Matérn covariance with varying anisotropy (Paciorek and Schervish, 2006), for which the range parameter is constant at 0.05 in the x direction but varies as $0.05 + 0.45\, s_y$ (as a function of the y-coordinate s_y) in the y direction, on a regular 50×50 grid; Matérn with smoothness 1 and range 0.25 at n = 2,500 randomly spaced locations sampled uniformly.

For all scenarios, MLE was roughly as accurate as OURS for very large N, but performed poorly for small N, indicating that the added shrinkage from our prior improved the accuracy. OURS strongly outperformed SCOV in all settings. For the nonstationary covariance, we also considered the correlation-based ordering described in Section 2.7. While we used the true correlation for the comparison here, the element-wise product of the sample covariance and an exponential correlation proposed in Section 2.7 resulted in comparable accuracy (not shown). As expected, correlation-based ordering performed better than maximin-ordering in this nonstationary setting. We also conducted some experiments (not shown) using a natural ordering by one of the spatial coordinates, which performed comparably to maximin ordering for isotropic covariances on a regular grid, but was much less accurate for randomly sampled locations. We did not consider (S)LASSO or autoFRK here, because they were not competitive in the similar setting of Section 3.2.

Overall, our method performed well across all simulations, even though our prior distributions were motivated by isotropic Matérn-like covariances. In addition, the computational burden for OURS was relatively low, with the estimated m often around ten and always below 30. While we only considered moderate n here in order to be able to carry out many comparisons using the KL divergence, it is also possible to run our method on much larger datasets. For example, using a C++ implementation, evaluating the integrated likelihood (9) only took about 6 seconds on a 4-core laptop (Intel i7-7560U) for $n = 250{,}000$, $m = 10$, $N = 50$.

4 Climate-model emulation

We analyzed climate-model output from the Community Earth System Model (CESM) Large Ensemble Project (Kay et al., 2015). Specifically, we considered daily mean surface temperature (in Kelvin) on July 1 in 98 consecutive years starting in the year 402, on a roughly 1° longitude-latitude grid of size $n = 81 \times 96 = 7{,}776$ containing much of the Americas (see Figure 1). The chosen region features ocean, land, islands, and mountain ranges, leading to a complicated, nonstationary dependence structure. The data $\mathbf{Y}$ were defined as the temperature anomalies obtained by standardizing the climate-model output at each grid point to unit mean and variance. We found no evidence of temporal correlation in the data, and so the assumption of independent replicates in (2) was at least approximately satisfied.

First, we compared several covariance estimates: an exponential covariance with a range parameter estimated from the data (EXP); a tapered sample covariance given by the element-wise product of the sample covariance and an exponential correlation with a range of 6,000 km, with a small added nugget with variance 10^{-5} for numerical stability (SCOVT); the MAP estimate (as in Section 3.2) using our method with correlation ordering (Section 2.7) based on the SCOVT matrix (OURS); autoFRK, described at the beginning of Section 3; and a nonstationary, locally parametric method specifically developed for gridded climate data (Wiens et al., 2020), which locally fits anisotropic Matérn covariances in small windows around every grid point and then synthesizes these local fits into a global model (LOCAL). Of the 98 replicates (i.e., years), we randomly selected and withheld 18 as test data, and fit the models on subsets of various sizes N between 6 and 80. As the true distribution was unknown, it was not possible to compute the KL divergence. Instead, we used the strictly proper log score (e.g. Gneiting and Katzfuss, 2014) given by the average negative log posterior predictive density of the test data based on (2), with $\mathbf{\Sigma}$ replaced by each of the methods' estimates.

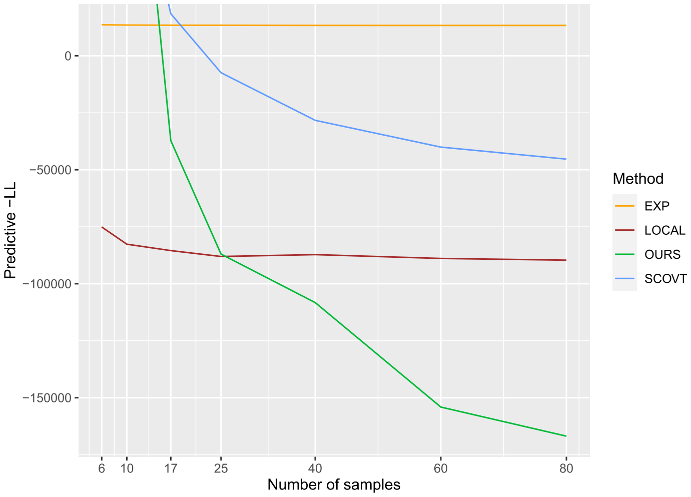


Figure 9: Comparison using the log score (lower is better) of methods fitted on climate-model temperature anomalies with varying numbers of replicates N (see Section 4).

Figure 9 shows the resulting scores, averaged over three random training/test splits. OURS was more accurate than SCOVT for all values of N, and its posterior-contraction properties (Section 2.6) enabled it to be more accurate than EXP for all $N \geq 10$. We also tried OURS with Euclidean (instead of correlation-based) ordering, which resulted in similar scores for large N but required almost twice the $N = 17$ replicates to surpass EXP (not shown). To our surprise, autoFRK performed worse than EXP and was thus not included in Figure 9. LOCAL performed best for small $N < 25$, but did not meaningfully improve and was thus less accurate than OURS for larger N; this indicates that OURS was able to capture some non-Matérn behavior in the climate data that LOCAL was not, due to its local Matérn assumption. Computing each covariance-matrix estimate on a 4-core laptop (Intel i7-7560U) without parallelization took over 17 hours using LOCAL but at most a few minutes using OURS, although both of these computing times could be reduced via parallelization.

We created a stochastic simulator emulating the climate model, by fitting a fully Bayesian version of OURS to the full dataset with $N = 98$ and sampling from the posterior predictive distribution $p(\mathbf{y}^\star|\mathbf{Y})$ as described at the end of Section 2.2. Four such samples are shown in Figure 10; they look qualitatively similar to the actual samples from the climate model in Figure 1, including reproducing features corresponding to land/ocean effects despite using no explicit information on land boundaries. These results were based on 50,000 Metropolis-Hastings (MH) samples of $\boldsymbol{\theta}$ (after a burn-in of 50,000) with trace plots showing good mixing and effective sample sizes all larger than 1,000; the samples were then thinned by a factor of 50. On the laptop, it took about 200 minutes to train the emulator (i.e., for 100,000 MH iterations), and it took 2.5 seconds to obtain a sample $\mathbf{y}^\star$ for a given value of $\boldsymbol{\theta}$.

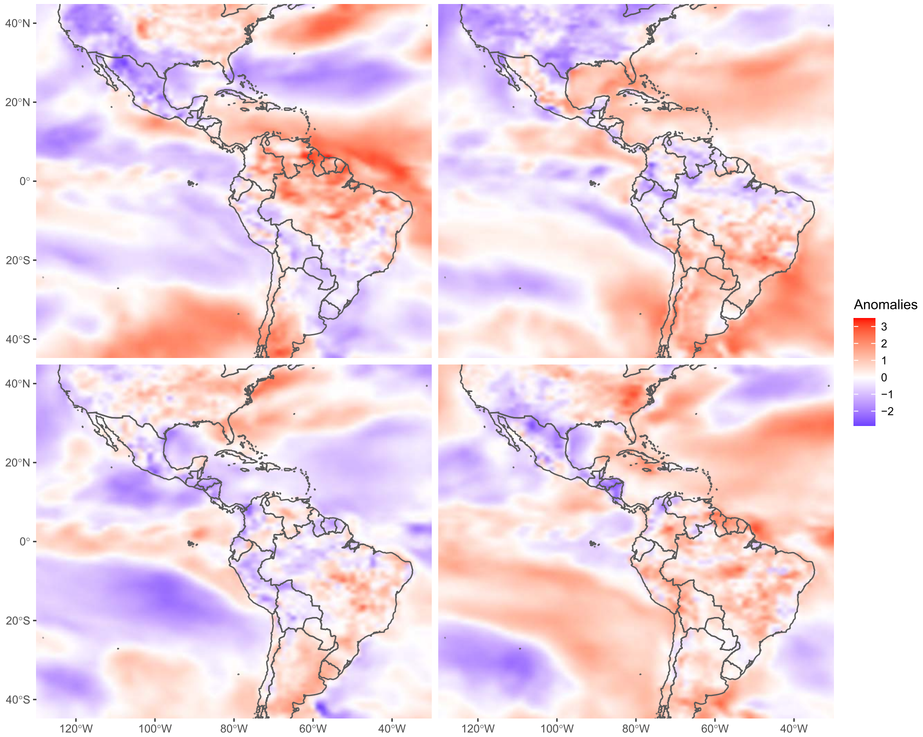


Figure 10: Four temperature-anomaly fields (in Kelvin) sampled from the posterior predictive distribution using our fully Bayesian method, computed as described in Section 4 based on climate-model output as in Figure 1.

5 Conclusions

We have developed a scalable, flexible Bayesian model for spatial covariance estimation and emulation. We regularize our method by taking advantage of a form of ordered conditional independence often assumed for spatial data. This motivates the assumption of sparsity in the Cholesky of the precision matrix, which greatly improves scalability and reduces the number of unknown parameters from quadratic to near-linear in the number of spatial locations. We describe three hyperparameters related to the marginal variance and the decay of Cholesky entries; these hyperparameters can be quickly optimized or sampled, resulting in an automatic data-based selection of the sparsity structure. Hence, our method requires no manual tuning or cross-validation. While our approach was motivated by the behavior of isotropic covariances on regular grids, our numerical comparisons demonstrated its generality with more complex covariances and irregularly spaced locations. We also applied the method to climate-model emulation, where it was able to capture the nonstationary and nonparametric behavior better than existing methods. `R` code implementing our method can be found at `https://github.com/katzfuss-group/NPVecchia`.

There are several interesting extensions for our spatial covariance estimation procedure. Our method can be extended to handle missing values by imputation using a Gibbs sampler similar to the samplers described in Section 2.8. However, if the number of observations at a particular location is very small or even zero, the posterior distribution at that location will be very vague and thus generally not particularly useful, unless some additional assumptions about the covariance between the unobserved and observed locations are made; for example, our prior can be modified to shrink toward a specific parametric covariance (see Section 2.9). Another potential extension is to estimate the covariance as a function of external variables by including them as additional covariates in the regressions in (5); for instance, for climate-model emulation, the covariance could depend on season, year, elevation, or land versus ocean. Finally, our approach can be extended to data assimilation, by using it to infer the forecast covariance matrices in an ensemble Kalman filter (Boyles and Katzfuss, 2021).

References

Anderes, E. B. and Stein, M. L. (2011). "Local likelihood estimation for nonstationary random fields." *Journal of Multivariate Analysis*, 102(3): 506–520. MR2755012. doi: https://doi.org/10.1016/j.jmva.2010.10.010. 292

Boyles, W. and Katzfuss, M. (2021). "Ensemble Kalman filter updates based on regularized sparse inverse Cholesky factors." *Monthly Weather Review*. doi: https://doi.org/10.1175/MWR-D-20-0299.1. 309

Castruccio, S., McInerney, D. J., Stein, M. L., Crouch, F. L., Jacob, R. L., and Moyer, E. J. (2014). "Statistical emulation of climate model projections based on precomputed GCM runs." *Journal of Climate*, 27(5): 1829–1844. 291

Castruccio, S. and Stein, M. L. (2013). "Global space-time models for climate ensembles." *Annals of Applied Statistics*, 7(3): 1593–1611. MR3127960. doi: https://doi.org/10.1214/13-AOAS656. 291

Choi, I. K., Li, B., and Wang, X. (2013). "Nonparametric estimation of spatial and space-time covariance function." *Journal of Agricultural, Biological, and Environmental Statistics*, 18(4): 611–630. MR3142603. doi: https://doi.org/10.1007/s13253-013-0152-z. 292

Cressie, N. and Johannesson, G. (2008). "Fixed rank kriging for very large spatial data sets." *Journal of the Royal Statistical Society, Series B*, 70(1): 209–226. MR2412639. doi: https://doi.org/10.1111/j.1467-9868.2007.00633.x. 291

Damian, D., Sampson, P. D., and Guttorp, P. (2001). "Bayesian estimation of semiparametric non-stationary spatial covariance structures." *Environmetrics*, 12(2): 161–178. 292

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets." *Journal of the American Statistical Association*, 111(514): 800–812. MR3538706. doi: https://doi.org/10.1080/01621459.2015.1044091. 294

Fuentes, M. (2002). “Spectral methods for nonstationary spatial processes.” *Biometrika*, 89(1): 197–210. MR1888368. doi: https://doi.org/10.1093/biomet/89.1.197. 292

Furrer, R. and Bengtsson, T. (2007). “Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants.” *Journal of Multivariate Analysis*, 98(2): 227–255. MR2301751. doi: https://doi.org/10.1016/j.jmva.2006.08.003. 291

Gneiting, T. and Katzfuss, M. (2014). “Probabilistic forecasting.” *Annual Review of Statistics and Its Application*, 1(1): 125–151. doi: https://doi.org/10.1146/annurev-statistics-062713-085831. 306

Guinness, J. (2018). “Permutation and grouping methods for sharpening Gaussian process approximations.” *Technometrics*, 60(4): 415–429. MR3878098. doi: https://doi.org/10.1080/00401706.2018.1437476. 293, 294

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F., and Zammit-Mangion, A. (2019). “A case study competition among methods for analyzing large spatial data.” *Journal of Agricultural, Biological, and Environmental Statistics*, 24(3): 398–425. 291

Hsing, T., Brown, T., and Thelen, B. (2016). “Local intrinsic stationarity and its inference.” *Annals of Statistics*, 44(5): 2058–2088. MR3546443. doi: https://doi.org/10.1214/15-AOS1402. 293

Huang, C., Hsing, T., and Cressie, N. (2011). “Nonparametric estimation of the variogram and its spectrum.” *Biometrika*, 98(4): 775–789. MR2860323. doi: https://doi.org/10.1093/biomet/asr056. 292

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). “Covariance matrix selection and estimation via penalised normal likelihood.” *Biometrika*, 93(1): 85–98. MR2277742. doi: https://doi.org/10.1093/biomet/93.1.85. 293, 295

Katzfuss, M. and Guinness, J. (2021). “A general framework for Vecchia approximations of Gaussian processes.” *Statistical Science*, 36(1): 124–141. MR4194207. doi: https://doi.org/10.1214/19-STS755. 292, 294, 295, 300

Katzfuss, M., Guinness, J., Gong, W., and Zilber, D. (2020a). “Vecchia approximations of Gaussian-process predictions.” *Journal of Agricultural, Biological, and Environmental Statistics*, 25(3): 383–414. MR4139037. doi: https://doi.org/10.1007/s13253-020-00401-7. 294, 296, 301

Katzfuss, M., Guinness, J., and Lawrence, E. (2020b). “Scaled Vecchia approximation for fast computer-model emulation.” URL http://arxiv.org/abs/2005.00386. 294

Katzfuss, M., Stroud, J. R., and Wikle, C. K. (2016). “Understanding the ensemble Kalman filter.” *The American Statistician*, 70(4): 350–357. MR3574787. doi: https://doi.org/10.1080/00031305.2016.1141709. 291

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J. F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M. (2015). "The Community Earth System Model (CESM) Large Ensemble Project: A community resource for studying climate change in the presence of internal climate variability." *Bulletin of the American Meteorological Society*, 96(8): 1333–1349. 306

Levina, E., Rothman, A., and Zhu, J. (2008). "Sparse estimation of large covariance matrices via a nested Lasso penalty." *Annals of Applied Statistics*, 2(1): 245–263. MR2415602. doi: https://doi.org/10.1214/07-AOAS139. 293

Murphy, K. P. (2007). "Conjugate Bayesian analysis of the Gaussian distribution." *Technical Report*. 299

Nychka, D. W., Hammerling, D., Krock, M., and Wiens, A. (2018). "Modeling and emulation of nonstationary Gaussian fields." *Spatial Statistics*, 28: 21–38. MR3887154. doi: https://doi.org/10.1016/j.spasta.2018.08.006. 291, 293

Paciorek, C. and Schervish, M. (2006). "Spatial modelling using a new class of nonstationary covariance functions." *Environmetrics*, 17(5): 483–506. MR2240939. doi: https://doi.org/10.1002/env.785. 305

Porcu, E., Bissiri, P. G., Tagle, F., and Quintana, F. (2019). "Nonparametric Bayesian modeling and estimation of spatial correlation functions for global data." *Technical Report*. 292

Qadir, G. A., Sun, Y., and Kurtek, S. (2019). "Estimation of spatial deformation for nonstationary processes via variogram alignment." URL http://arxiv.org/abs/1911.02249. 292

Risser, M. D. (2016). "Review: Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches." URL http://arxiv.org/abs/1610.02447. 292

Sampson, P. D. and Guttorp, P. (1992). "Nonparametric estimation of nonstationary spatial covariance structure." *Journal of the American Statistical Association*, 87(417): 108–119. 292

Schäfer, F., Katzfuss, M., and Owhadi, H. (2021a). "Sparse Cholesky factorization by Kullback-Leibler minimization." *SIAM Journal on Scientific Computing*, 43(3): A2019–A2046. doi: https://doi.org/10.1137/20M1336254. 292, 293, 294, 299, 300, 301

Schäfer, F., Sullivan, T. J., and Owhadi, H. (2021b). "Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity." *Multiscale Modeling & Simulation*, 19(2): 688–730. MR4243658. doi: https://doi.org/10.1137/19M129526X. 293, 294, 297, 299

Scheidegger, A. (2012). *adaptMCMC: Implementation of a generic adaptive Monte Carlo Markov Chain sampler*. R package version 1.0.3. 299

Smith, M. and Kohn, R. (2002). "Parsimonious covariance matrix estimation for longitudinal data." *Journal of the American Statistical Association*, 97(460): 1141–1153. MR1951266. doi: https://doi.org/10.1198/016214502388618942. 293

Stein, M. L., Chi, Z., and Welty, L. (2004). "Approximating likelihoods for large spatial data sets." *Journal of the Royal Statistical Society: Series B*, 66(2): 275–296. MR2062376. doi: https://doi.org/10.1046/j.1369-7412.2003.05512.x. 294

Tzeng, S. and Huang, H.-C. (2018). "Resolution adaptive fixed rank kriging." *Technometrics*, 60(2): 198–208. MR3804248. doi: https://doi.org/10.1080/00401706.2017.1345701. 303

Tzeng, S., Huang, H.-C., Wang, W.-T., Nychka, D., and Gillespie, C. (2021). *autoFRK: Automatic Fixed Rank Kriging*. R package version 1.4.3. URL https://CRAN.R-project.org/package=autoFRK. 303

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press. MR1652247. doi: https://doi.org/10.1017/CBO9780511802256. 300

Vecchia, A. (1988). "Estimation and model identification for continuous spatial processes." *Journal of the Royal Statistical Society, Series B*, 50(2): 297–312. MR0964183. 291, 294

Wiens, A., Nychka, D. W., and Kleiber, W. (2020). "Modeling spatial data using local likelihood estimation and a Matérn to spatial autoregressive translation." *Environmetrics*, 31(6): 1–15. MR4151871. doi: https://doi.org/10.1002/env.2652. 291, 293, 306

Zeng, X., Atlas, R., Birk, R. J., Carr, F. H., Carrier, M. J., Cucurull, L., Hooke, W. H., Kalnay, E., Murtugudde, R., Posselt, D. J., Russell, J. L., Tyndall, D. P., Weller, R. A., and Zhang, F. (2021). "Use of observing system simulation experiments in the United States." *Bulletin of the American Meteorological Society*, 101(8): E1427–E1438. 291

Zhu, Z. and Liu, Y. (2009). "Estimating spatial covariance using penalised likelihood with weighted L1 penalty." *Journal of Nonparametric Statistics*, 21(7): 925–942. MR2572592. doi: https://doi.org/10.1080/10485250903023632. 293, 303

Acknowledgments

We would like to thank Mohsen Pourahmadi, Florian Schäfer, Will Boyles, and Joe Guinness for helpful comments and suggestions.

Invited Discussion[*]

Bo Li[†] and Lyndsay Shand[‡,§]

1 Overview

We admire the authors for developing this computationally efficient Bayesian method to estimate the nonstationary correlation structure in large spatial data, without relying on a restrictive parametric model. The method is presented as a nonparametric extension of the Vecchia approach (Vecchia, 1988) and is based on the ordered conditional independence assumption that holds or approximately holds for many data sets arising from a Gaussian random process. The conditional independence leads to a sparse precision matrix and consequently a sparse Cholesky factorization. It has been shown that an n-variate Gaussian model can be expressed as a series of linear regression models with the nonzero elements in the Cholesky factor matrix as the regression coefficients (Huang et al., 2006). This enables us to estimate the Cholesky factorization and thus the precision matrix through Bayesian regression. The authors carefully studied the properties of unknown parameters and selected independent conjugate normal-inverse-gamma (NIG) priors that lead to closed-form posteriors and thus further improve computational efficiency.

There have been various approaches to nonstationary or nonparametric covariance modeling for spatial or spatiotemporal data. Kidd and Katzfuss (2022) (referred to KK22 thereafter) has provided an excellent review of previous literature. In addition to all methods reviewed for nonstationary covariance modeling in the Introduction, another semiparametric approach for nonstationary covariance modeling is through dimension expansion (Bornn et al., 2012; Shand and Li, 2017). Bornn et al. (2012) also requires replicates to estimate the spatial model.

KK22 thoroughly discussed many aspects of their method, including the theory, computation complexity, solutions for the presence of noise or trend, data ordering, conditioning-set size, and how to adapt priors to allow posterior converging to a covariance structure other than Matérn. We find several aspects very interesting and worth additional discussions. In the following, we adopt all notations from KK22.

[*] Li's research is partially supported by NSF-DMS-1830312 and -2124576. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SAND2022-0656 J.

[†] Department of Statistics, University of Illinois Urbana-Champaign, libo@illinois.edu

[‡] Department of Statistics, University of Illinois Urbana-Champaign

[§] Sandia National Laboratories, Albuquerque, New Mexico, lshand@sandia.gov

2 Further considerations

2.1 Requirement for replicates

The KK22 method was developed for spatial data with independent replicates and requires an ordering of spatial locations. KK22 used a maximin ordering that makes locations in $g_m(i)$ all have a roughly similar distance to the location $\mathbf{s}_i$ and this distance decreases systematically as i increases. Observing how the mean and variance of the d_i in $\mathbf{D}$ decrease exponentially with i given an underlying isotropic Matérn covariance structure (Sháfer et al., 2021a,b), the authors developed an inverse gamma prior for d_i with an appropriate form for α_i and β_i. Via the same scheme, they chose the normal prior for $\mathbf{u}_i$ with a tailored form for the correlation matrix. These carefully chosen conjugate priors ensure fast posterior sampling of unknown parameters, making the proposed method computationally efficient. The posteriors hold the nice property that the estimated covariance matrix contracts around the true covariance matrix as the number of replicates N increases. The numerical results showed that the Kullback-Leibler (KL) divergence of the KK22 method is lower than other methods in comparison, even for a very small N relative to the number of spatial locations n. In particular, the method works more efficiently than the maximum likelihood (ML) method for small N due to the inclusion of prior information, and then performs similarly to ML without a surprise when N increases. This conclusion holds for large n as well.

Given that this method demonstrates reasonable performance when N is as low as 3, do we really need replicates, i.e., $N > 1$, for the proposed method to be valid? This seems a rather constraining requirement. It is very common to observe spatial data without independent replicates. For example, suppose we are interested in an annual data of last year that can be either temperature over North America, or the county level Midwest crop yield or zip code level human immunodeficiency virus (HIV) new diagnoses in Philadelphia; all these spatial data likely have only one observation at each location. KK22 proposed their method based on Vecchia (1988), who developed a procedure with a spatial process not necessarily with replicates. We conjecture the requirement for replicates is mainly to attain the nice posterior-contraction property and reduce uncertainty in the parameter estimation. Is it possible to find a way to relax the replicates requirement (i.e., $N = 1$) but still approximately obtain the posterior-contraction? The authors investigated how the parameters decay with i, but is there any other spatial structure not in the order of i but on the distance between different i's that can be exploited to reduce the dependence on replicates? More specifically, can d_i and $\mathbf{u}_i$ be also modeled as spatially dependent processes in addition to their dependence on i? Of course, modeling additional spatially dependent processes can increase computation, so some special techniques such as those modeling dependence only on the nearest neighbors (Datta et al., 2016) may be considered. The maximin ordering makes the spatial dependence between i's unclear, but it may be worth a deliberation.

2.2 Choice of m

The choice of m is a trade-off between estimation accuracy and computation. KK22 suggested to set m as the largest j such that $\exp(-\theta_3 j) > 0.001$, where j denotes the neighbor number. We wonder whether it is better to set m as adaptive for different i. With the maximin ordering, the distance between the conditioning-set to $\mathbf{s}_i$ decreases as i increases, but the size of the pool, $\mathbf{y}_{1:i-1}^{(l)}$, is always increasing. So it seems reasonable to have m as an increasing function of i to better approximate the full conditional distribution with the conditional distribution given a few neighbors.

2.3 Spatiotemporal data

Spatial data often have temporally correlated "replicates", i.e., spatiotemporal data. The temperatures of North America, the Midwest crop yield, and the disease data in our early examples can all become spatiotemporal if we now collect the annual data for the last 10 years. For spatiotemporal data, the temporal correlation and its interaction with spatial correlation need to be considered. This brings an additional challenge as the observations in the temporal dimension rapidly inflate the size of the covariance matrix $\mathbf{\Sigma}$ unless some simplified assumption such as space-time separability is assumed. If the KK22 method can be extended to this wealth of data, it would certainly expand its applicability. There is more impetus to relax the replication requirement in this case though, as independent replicates for spatiotemporal data are rarely available.

Similar to KK22, the nonstationary covariance modeling method in Bornn et al. (2012) also requires independent replicates of spatial data. Shand and Li (2017) extends Bornn et al.'s idea to model nonstationary covariance in both space and time for spatiotemporal data, and discusses the scalability of the method by taking random samples for latent dimension estimation. In Shand and Li (2017), observations in space and time are treated somewhat as independent replicates when estimating temporal and spatial correlation, respectively. Those intermediate results are then taken as inputs when dependency in all different forms is considered holistically. This strategy helps us estimate both space-time separable and nonseparable covariance structures while eliminating the dependence on replicates.

In the context of KK22, the ordering of space-time observations can be a challenge because both spatial and temporal distances are involved and the maximin ordering cannot be directly applicable. However, KK22 also mentioned other ordering strategies such as the ordering based on correlation distance which would more naturally extend to space-time data. Once the space-time observations are ordered, the KK22 method can readily apply to such data. Regarding the inflated size of $\mathbf{\Sigma}$, Section 2.2 in KK22 already discussed how to deal with a very large covariance matrix. If we can assume space-time separability, the precision matrix will be a Kronecker product of the precision matrices in space and time. In that case, we wonder if the ordering can be calculated for each dimension separately, and then $\mathbf{u}_i$ in space and time can be estimated separately as well. On the other hand, if we assume a simple temporal correlation structure such as an autoregressive model of order 1 (AR(1)), we wonder whether the estimates of $\mathbf{u}_i$ in the spatial dimension can approximately attain the posterior-contraction property, because the observations in time may act as dependent replicates for spatial correlation estimation. All these discussions for spatiotemporal data can be generalized to multivariate spatial data modeling.

3 Miscellaneous discussion

The smoothness of random fields is difficult to capture for nonparametric modeling. Im et al. (2007) proposes a semiparametric method that models the spectral density as a linear combination of B-splines up to a certain frequency threshold ω_0 and then an algebraic power function with a smoothness parameter similar to Matérn model for high frequencies beyond ω_0. However, many other nonparametric models, including Choi et al. (2013) that constructs the spatial or space-time covariance function using completely monotone functions do not directly consider the smoothness of covariance models. It is inspiring that KK22 has a smoothness parameter in their priors. We are curious how θ_3 explicitly relates to the smoothness of random fields.

There are different measures to evaluate covariance structure estimation. For example, mean squared prediction error is common for comparing different covariance estimates as prediction is a typical task for spatial data analysis. We think it would be informative if the authors could briefly comment on whether KL divergence relates to the prediction performance measure in general.

Spatial random effects and multi-resolution models (e.g. Nychka et al., 2014) are very popular for capturing either stationary or nonstationary spatial structures of massive datasets. KK22 also included the resolution adaptive fixed rank kriging approach by Tzeng and Huang (2018) (autoFRK), as one of the competitive methods. The author Katzfuss published a very high-impact paper on multi-resolution approximation (Katzfuss, 2017). We would like to learn how the authors view the connection, the discrepancies and the comparison between the KK22 and multi-resolution models.

4 Summary

We congratulate the authors on developing a very useful method for large spatial data. This method would find many applications for modeling the complex dependency structure of global climate data, e.g., the teleconnection of climate variables (Choi et al., 2015). Studies for spatial extremes often approximate the block maxima as independent replicates (e.g. Cao and Li, 2018), so the KK22 method can be naturally used to model dependence in spatial extremes by combining with the copula technique that takes care of the marginal extreme value distribution, among many other exciting applications.

References

Bornn, L., Shaddick, G., and Zidek, J. V. (2012). "Modeling nonstationary processes through dimension expansion." *Journal of the American Statistical Association*, 107(497): 281–289. MR2949359. doi: https://doi.org/10.1080/01621459.2011.646919. 313, 315

Cao, Y. and Li, B. (2018). "Assessing models for spatial extremes and methods for uncertainty quantification on return level estimation." *Environmetrics*. MR3919899. doi: https://doi.org/10.1002/env.2508. 316

Choi, I., Li, B., and Wang, X. (2013). "Nonparametric estimation of spatial and space-time covariance function." *Journal of Agricultural, Biological, and Environmental Statistics*, 18(4): 611–630. MR3142603. doi: https://doi.org/10.1007/s13253-013-0152-z. 316

Choi, I., Li, B., Zhang, H., and Li, Y. (2015). "Modelling space–time varying ENSO teleconnections to droughts in North America." *Stat*, 4(1): 140–156. MR3405397. doi: https://doi.org/10.1002/sta4.85. 316

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets." *Journal of the American Statistical Association*, 111(514): 800–812. PMID: 29720777. MR3538706. doi: https://doi.org/10.1080/01621459.2015.1044091. 314

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). "Covariance matrix selection and estimation via penalised normal likelihood." *Biometrika*, 93(1): 85–98. MR2277742. doi: https://doi.org/10.1093/biomet/93.1.85. 313

Im, H. K., Stein, M. L., and Zhu, Z. (2007). "Semiparametric estimation of spectral density with irregular observations." *Journal of the American Statistical Association*, 102(478): 726–735. MR2381049. doi: https://doi.org/10.1198/016214507000000220. 316

Katzfuss, M. (2017). "A multi-resolution approximation for massive spatial datasets." *Journal of the American Statistical Association*, 112(517): 201–214. MR3646566. doi: https://doi.org/10.1080/01621459.2015.1123632. 316

Kidd, B. and Katzfuss, M. (2022). "Bayesian nonstationary and nonparametric covariance estimation for large spatial data." *Bayesian Analysis*, 1(1): 1–22. 313

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2014). "A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets." *Journal of Computational and Graphical Statistics*, 24. MR3357396. doi: https://doi.org/10.1080/10618600.2014.914946. 316

Sháfer, F., Katzfuss, M., and Owhadi, H. (2021a). "Sparse Cholesky Factorization by Kullback-Leibler Minimization." *SIAM Journal on Scientific Computing*, 43(3): A2019–A2046. MR4267493. doi: https://doi.org/10.1137/20M1336254. 314

Sháfer, F., Sullivan, T. J., and Owhadi, H. (2021b). "Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity." *Multiscale modeling & Simulation*, 19(2): 688–730. MR4243658. doi: https://doi.org/10.1137/19M129526X. 314

Shand, L. and Li, B. (2017). "Modeling nonstationarity in space and time." *Biometrics*, 73(3): 759–768. MR3713110. doi: https://doi.org/10.1111/biom.12656. 313, 315

Tzeng, S. and Huang, H. (2018). "Resolution Adaptive Fixed Rank Kriging." *Technometrics*, 60: 1–11. doi: https://doi.org/10.1080/00401706.2017.1345701. 316

Vecchia, A. V. (1988). "Estimation and model identification for continuous spatial processes." *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2): 297–312. MR0964183. 313, 314

Invited Discussion

Sudipto Banerjee[*] and Michele Peruzzi[†]

We congratulate the authors for an interesting article on a very relevant topic in spatial statistics. Nonstationary spatial modeling and inference holds significant value in spatial statistics and has attracted significant attention over the years to produce a substantial body of original contributions; a fairly comprehensive account is provided by Sampson (2010). Nonstationarity in spatial models can refer to nonstationarity in the mean or in the spatial covariance. Nonstationarity in the mean, or trends, is customarily addressed by introducing the effects of known explanatory processes in the mean either using spatially-varying regression models (see, e.g., Gelfand et al., 2003).

Nonstationarity in spatial covariance deals with characterizing and constructing spatial processes that will describe nonstationarity. The rich spectral theory available for stationary processes is far less accessible for nonstationary processes. Valid nonstationary processes, therefore, have largely emerged from tractable operations on stationary processes. Examples include *deformations* of stationary processes (Sampson and Guttorp, 1992; Damian et al., 2001; Schmidt and O'Hagan, 2000); kernel convolutions of stationary processes (Higdon et al., 1999; Fuentes, 2001, 2002a,b; Calder, 2008); and spatially-varying covariance kernels (Gelfand et al., 2004; Paciorek and Schervish, 2006; Risser and Calder, 2015).

Much of the aforementioned literature focuses upon the construction of a nonstationary spatial covariance function $\mathrm{Cov}(y(\mathbf{s}), y(\mathbf{s}')) = C(\mathbf{s}, \mathbf{s}')$ for pairs of spatial locations $\mathbf{s}$ and $\mathbf{s}'$, which will legitimately define a spatial stochastic process over an uncountable set of locations $\{y(\mathbf{s}) : \mathbf{s}' \in \mathcal{D} \subseteq \mathbb{R}^d\}$. Kidd and Katzfuss (2022) appear to broadly classify the above approaches as "parametric", avoid the construction of a nonstationary process and, therefore, describe their approach as "nonparametric" with regard to modeling spatial covariances. The use of the term "nonparametric" may appear as somewhat misplaced since the distribution on the realizations of the process are still parametric (Gaussian) with an unspecified covariance matrix. We remark that nonparametric spatial models with no parametric specifications on the distribution of process realizations can be constructed using spatial Dirichlet processes (Gelfand et al., 2005; Duan et al., 2007).

A distinctive feature of Kidd and Katzfuss (2022) is that they do not attempt to construct a nonstationary spatial process, instead working with the probability law of replicated finite-dimensional realizations of the process to ease the computational burden of Bayesian inference for nonstationary processes for large spatial data sets. Here, we should remark that process-based dimension reduction, which has become pervasive in analyzing large spatial data, often leads to classes of nonstationary models in a natural way. Examples include low rank spatial processes derived from empirical orthogonal functions (Holland et al., 1999), multiresolution basis functions (Nychka et al., 2002),

[*] Department of Biostatistics, University of California, Los Angeles, sudipto@ucla.edu

[†] Department of Statistical Science, Duke University, michele.peruzzi@duke.edu

fixed-rank kriging (Cressie and Johannesson, 2008), and predictive processes (Banerjee et al., 2008; Finley et al., 2009; Katzfuss, 2017). However, low rank processes with truncated basis functions are limited in their capabilities to emulate the underlying process because of a natural tendency to oversmooth (Banerjee, 2017), which can be very pronounced in very large or massive data sets where the number of basis functions is a small fraction of the number of data points.

Faced with increasingly massive spatial and spatial-temporal data, spatial processes built from directed acyclic graphs (DAGs), following Gaussian process likelihood approximations outlined in Vecchia (1988), have attracted much attention (Datta et al., 2016a,b; Guinness, 2018; Katzfuss and Guinness, 2021; Peruzzi et al., 2020; Jin et al., 2021). The authors exploit this idea to develop classes of intuitive, simple, yet effective hierarchical models for analyzing large spatial data. Their elegant solutions connect the emerging literature on so-called Vecchia approximations of Gaussian processes (GPs) to the existing literature on sparse covariance estimation. Of particular interest is that parametric (stationary) covariance models motivate the prior distribution specification for entries in the (modified) Cholesky factor of the Gaussian precision matrix. The idea of allowing nonstationarity while shrinking towards a simple (perhaps stationary) model is intuitive and very appealing. Additionally, the proposed strategies enable massive parallelization of all operations, leading to scalability to large scale data.

We offer some remarks with respect to predictive inference at new arbitrary spatial locations $\mathbf{s}_i^*$'s, which seems to have been glossed over by Kidd and Katzfuss (2022). A key objective of extending finite-dimensional approximations to well-defined spatial processes, such as the Nearest-Neighbor GP (NNGP; Datta et al. 2016a) or the meshed-GP (MGP; Peruzzi et al. 2020), is to facilitate spatial or spatial-temporal interpolation or prediction at arbitrary points. Unlike process-based approaches such as the NNGP or the MGP, where the modified Cholesky decomposition is parametrized using a parent covariance function, the $\mathbf{U}$ and $\mathbf{D}$ in (4) of Kidd and Katzfuss (2022) are free parameters that do not depend on covariance function parameters. This assumption together with the availability of replicates is cleverly exploited by the authors to estimate $\mathbf{U}$ using essentially the familiar distribution theory from conjugate Bayesian linear regression models. Indeed, conjugate Bayesian spatial models, often criticized for their lack of flexibility, deliver substantial computational benefits in handling massive datasets on modest computing architectures (Banerjee, 2020).

It is worth comparing predictive inference from process-based frameworks with that proposed by the authors. Assume that the $N \times n$ matrix of spatial data, $\mathbf{Y}$, as defined in (1) of Kidd and Katzfuss (2022) consists of observed measurements and let $\mathbf{s}_{n+1}$ be a new location where we wish to predict the value of the process, say $Y(\mathbf{s}_{n+1})$. The posterior predictive distribution $p(Y(\mathbf{s}_{n+1}) \,|\, \mathbf{Y})$ is then obtained by sampling from $Y(\mathbf{s}_{n+1}) \sim N(\mathbf{x}_{n+1}^\top \mathbf{u}_{n+1}, d_{n+1})$ for each $\{\mathbf{u}_{n+1}, d_{n+1}\}$ sampled from their posterior distribution. However, it is unclear how effectively $\{\mathbf{u}_{n+1}, d_{n+1}\}$ is learned from the spatial data matrix $\mathbf{Y}$. In process-based frameworks, such as for the NNGP and the MGP, $\mathbf{u}_{n+1}$ and d_{n+1} explicitly depend on the covariance function parameters, which act as a bridge between the observed data and the predictive distribution. Here, we do not see such parameters and, hence, we do not see how the higher values of spatial covariance between neighboring points will affect spatial predictions. The effectiveness of predictive inference for the nonparametric framework of Kidd and Katzfuss (2022) should, we opine, be further investigated and compared with process-based counterparts based on parametric spatial covariance functions.

We also note that the authors focus on models of a single, continuous, noise-free, Gaussian spatially referenced outcome. This Gaussian response model is often contrasted with a latent model (Finley et al., 2019; Katzfuss and Guinness, 2021), in which case the spatial GP enters the Bayesian hierarchy as a random effect. In latent models, dependence must flow through the random effects: assumptions of conditional independence of the outcomes given the latent effects enable great flexibility in modeling, e.g., multivariate non-Gaussian data via latent GPs (Peruzzi and Dunson, 2022a). In attempting to extend their proposal to a latent plus noise model, the authors suggest a posterior sampling strategy based on a block Gibbs sampler. One iterates between (1) sampling the latent process $\mathbf{y}^{(\ell)}$ given the data $\mathbf{w}^{(\ell)}$ and the other parameters $\mathbf{\Theta} = (\boldsymbol{\theta}, \mathbf{U}, \mathbf{D})$, and (2) sampling $\mathbf{\Theta}$ given $\mathbf{y}^{(\ell)}$ and $\mathbf{w}^{(\ell)}$. This Gibbs sampler ultimately produces correlated samples from the joint posterior $p(\mathbf{y}^{(\ell)}, \mathbf{\Theta} \mid \mathbf{w}^{(\ell)}) \propto p(\mathbf{w}^{(\ell)} \mid \mathbf{y}^{(\ell)}, \mathbf{\Theta}) p(\mathbf{y}^{(\ell)} \mid \mathbf{\Theta}) p(\mathbf{\Theta})$. There are difficulties in performing (1) because the conditional independence assumptions encoded by $\mathbf{U}$ are not immediately useful in calculating the modified Cholesky factor of the posterior precision matrix $\mathbf{\Sigma}^{-1} + \frac{1}{\tau^2}\mathbf{I}_n$. In fact, $\mathbf{\Sigma}^{-1} + \frac{1}{\tau^2}\mathbf{I}_n = \mathbf{V}\mathbf{R}^{-1}\mathbf{V}^\top$ where $\mathbf{V}$ and $\mathbf{U}$ do not in general share the same sparsity structure. Since $\mathbf{V}$ may be considerably denser than $\mathbf{U}$, using $\mathbf{V}$ may result in bottlenecks when sampling the latent process from its full conditional distribution.

The authors propose to replace $\mathbf{V}$ with $\mathbf{V}^*$ obtained via an incomplete Cholesky factorization which forces it to share the same sparsity structure of $\mathbf{U}$. However, using $\mathbf{V}^*$ leads to sampling $\mathbf{y}^{(\ell)}$ from an approximate full conditional distribution. This approximation breaks the coherence between steps (1) and (2) and leads to a sampling algorithm which is not a Gibbs sampler and is not guaranteed to converge (what distribution would it converge to?). In order to maintain coherence between the two steps, one can either use the exact Cholesky factor $\mathbf{V}$, or accept/reject the approximate sample $\widetilde{\mathbf{y}}^{(\ell)}$ based on the Hastings ratio. The performance of both strategies will deteriorate when the nugget is large: in the former case, computing exact Cholesky of $\mathbf{V}$ will be slow due to substantial fill-in relative to $\mathbf{U}$, whereas in the latter, incomplete Cholesky decompositions will lead to a poor approximation of the full conditional distribution, resulting in a low acceptance probability. Alternatives to sample from the *exact* full posterior of the latent process involve visiting the nodes of the spatial DAG used to build $\mathbf{U}$; while sequential visits may lead to inefficiencies (Finley et al., 2019), blocking via domain partitioning, graph coloring, parameter expansions, and gradient-based sampling methods have all been proposed to improve algorithmic performance in these settings (Peruzzi et al., 2020, 2021; Peruzzi and Dunson, 2022a). Alternatively, one can use models that lead to no fill-in by construction (Katzfuss, 2017; Peruzzi and Dunson, 2022b).

In summary, we enjoyed reading this interesting and stimulating article and, in particular, are in agreement that conjugate Bayesian modeling frameworks have much to offer for the data science community when it comes to scalable inference. We raise some potential concerns with respect to predictive inference and offer some additional thoughts that, we hope, will motivate further developments of Bayesian models for nonstationary data and related scalable computing algorithms.

References

Banerjee, S. (2017). "High-dimensional Bayesian geostatistics." *Bayesian Analysis*, 12(2): 583–614. MR3654826. doi: https://doi.org/10.1214/17-BA1056R. 319

Banerjee, S. (2020). "Modeling massive spatial datasets using a conjugate Bayesian linear modeling framework." *Spatial Statistics*, 37: 100417. MR4109600. doi: https://doi.org/10.1016/j.spasta.2020.100417. 319

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. MR2523906. doi: https://doi.org/10.1111/j.1467-9868.2008.00663.x. 319

Calder, C. A. (2008). "A dynamic process convolution approach to modeling ambient particulate matter concentrations." *Environmetrics*, 19(1): 39–48. MR2416543. doi: https://doi.org/10.1002/env.852. 318

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. MR2412639. doi: https://doi.org/10.1111/j.1467-9868.2007.00633.x. 319

Damian, D., Sampson, P. D., and Guttorp, P. (2001). "Bayesian estimation of semi-parametric non-stationary spatial covariance structures." *Environmetrics*, 12(2): 161–178. doi: https://doi.org/10.1002/1099-095X(200103)12:2<161::AID-ENV452>3.0.CO;2-G. 318

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets." *Journal of the American Statistical Association*, 111: 800–812. doi: https://doi.org/10.1080/01621459.2015.1044091. 319

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A. S., and Schaap, M. (2016b). "Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis." *The Annals of Applied Statistics*, 10: 1286–1316. MR3553225. doi: https://doi.org/10.1214/16-AOAS931. 319

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). "Generalized spatial Dirichlet process models." *Biometrika*, 94(4): 809–825. MR2416794. doi: https://doi.org/10.1093/biomet/asm071. 318

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). "Efficient algorithms for Bayesian nearest neighbor Gaussian processes." *Journal of Computational and Graphical Statistics*, 28: 401–414. MR3974889. doi: https://doi.org/10.1080/10618600.2018.1537924. 320

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). "Improving the performance of predictive process modeling for large datasets." *Computational Statistics and Data Analysis*, 53: 2873–2884. MR2667597. doi: https://doi.org/10.1016/j.csda.2008.09.008. 319

Fuentes, M. (2001). "A high frequency kriging approach for non-stationary environmental processes." *Environmetrics*, 12(5): 469–483. doi: https://doi.org/10.1002/env.473. 318

Fuentes, M. (2002a). "Interpolation of nonstationary air pollution processes: a spatial spectral approach." *Statistical Modelling*, 2(4): 281–298. MR1951586. doi: https://doi.org/10.1191/1471082x02st034oa. 318

Fuentes, M. (2002b). "Spectral methods for nonstationary spatial processes." *Biometrika*, 89(1): 197–210. MR1888368. doi: https://doi.org/10.1093/biomet/89.1.197. 318

Gelfand, A., Schmidt, A., Banerjee, S., and Sirmans, C. F. (2004). "Nonstationary multivariate process modeling through spatially varying coregionalization." *Test*, 13(2): 263–312. doi: https://doi.org/10.1007/BF02595775. 318

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). "Spatial modeling with spatially varying coefficient processes." *Journal of the American Statistical Association*, 98(462): 387–396. MR1995715. doi: https://doi.org/10.1198/016214503000170. 318

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). "Bayesian nonparametric spatial modeling with Dirichlet process mixing." *Journal of the American Statistical Association*, 100(471): 1021–1035. doi: https://doi.org/10.1198/016214504000002078. 318

Guinness, J. (2018). "Permutation and grouping methods for sharpening Gaussian process approximations." *Technometrics*, 60(4): 415–429. doi: https://doi.org/10.1080/00401706.2018.1437476. 319

Higdon, D., Swall, J., and Kern, J. (1999). "Non-stationary spatial modeling." In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (eds.), *Bayesian Statistics 6*, 761–768. Oxford: Oxford University Press. MR1723490. 318

Holland, D. M., Saltzman, N., Cox, L. H., and Nychka, D. (1999). "Spatial prediction of sulfur dioxide in the eastern United States." In *geoENV II — Geostatistics for Environmental Applications*, 65–76. Springer Netherlands. 318

Jin, B., Peruzzi, M., and Dunson, D. B. (2021). "Bag of DAGs: Flexible & Scalable Modeling of Spatiotemporal Dependence." arXiv:2112.11870. 319

Katzfuss, M. (2017). "A multi-resolution approximation for massive spatial datasets." *Journal of the American Statistical Association*, 112: 201–214. MR3646566. doi: https://doi.org/10.1080/01621459.2015.1123632. 319, 320

Katzfuss, M. and Guinness, J. (2021). "A general framework for Vecchia approximations of Gaussian processes." *Statistical Science*, 36(1): 124–141. MR4194207. doi: https://doi.org/10.1214/19-STS755. 319, 320

Kidd, B. and Katzfuss, M. (2022). "Bayesian nonstationary and nonparametric covariance estimation for large spatial data." *Bayesian Analysis*, 1–34. doi: https://doi.org/10.1214/21-BA1273. 318, 319, 320

Nychka, D., Wikle, C., and Royle, J. A. (2002). "Multiresolution models for nonstationary spatial covariance functions." *Statistical Modelling*, 2(4): 315–331. MR1951588. doi: https://doi.org/10.1191/1471082x02st037oa. 318

Paciorek, C. J. and Schervish, M. J. (2006). "Spatial modelling using a new class of nonstationary covariance functions." *Environmetrics*, 483–506. MR2240939. doi: https://doi.org/10.1002/env.785. 318

Peruzzi, M., Banerjee, S., Dunson, D. B., and Finley, A. O. (2021). "Grid-Parametrize-Split (GriPS) for Improved Scalable Inference in Spatial Big Data Analysis." arXiv:2101.03579. 320

Peruzzi, M., Banerjee, S., and Finley, A. O. (2020). "Highly scalable Bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains." *Journal of the American Statistical Association*. In press. doi: https://doi.org/10.1080/01621459.2020.1833889. 319, 320

Peruzzi, M. and Dunson, D. B. (2022a). "Spatial meshing for general Bayesian multivariate models." arXiv:2201.10080. 320

Peruzzi, M. and Dunson, D. B. (2022b). "Spatial multivariate trees for big data Bayesian regression." *Journal of Machine Learning Research*, 23(17): 1–40. http://jmlr.org/papers/v23/20-1361.html. 320

Risser, M. D. and Calder, C. A. (2015). "Regression-based covariance functions for nonstationary spatial modeling." *Environmetrics*, 26(4): 284–297. MR3340964. doi: https://doi.org/10.1002/env.2336. 318

Sampson, P. D. (2010). "Constructions for nonstationary spatial processes." In Gelfand, A. E., Diggle, P., Fuentes, M., and Guttorp, P. (eds.), *Handbook of Spatial Statistics*, chapter 9. CRC Press. doi: https://doi.org/10.1201/9781420072884. 318

Sampson, P. D. and Guttorp, P. (1992). "Nonparametric estimation of nonstationary spatial covariance structure." *Journal of the American Statistical Association*, 87(417): 108–119. doi: https://doi.org/10.1080/01621459.1992.10475181. 318

Schmidt, A. M. and O'Hagan, A. (2000). "Bayesian inference for nonstationary spatial covariance structure via spatial deformations." *Journal of the Royal Statistical Society, Series B*, 65: 745–758. MR1998632. doi: https://doi.org/10.1111/1467-9868.00413. 318

Vecchia, A. V. (1988). "Estimation and model identification for continuous spatial processes." *Journal of the Royal Statistical Society, Series B*, 50: 297–312. doi: https://doi.org/10.1111/j.2517-6161.1988.tb01729.x. 319

Contributed Discussion

Stefano Peluso[*]

Congratulations to the Authors for the interesting work. Starting from an assumed ordering of the variable nodes $i \in \{1, \dots, n\}$, in the model formulation the set of parents of i is fixed to be $g_m(i) \subseteq \{1, \dots, i-1\}$, with $|g_m(i)| \leq m$. This is equivalent to assume the knowledge of a *base graph* $\mathcal{G}_0$, where the parents of node i are m at most. A posteriori the Authors provide an estimator $\hat{\mathcal{G}}$ of the graph $\mathcal{G}$ of dependences which is not necessarily equal to $\mathcal{G}_0$, but dependent on the latter: with the introduction of a threshold, say a small $\tau > 0$, for which v_{ij} will be fixed to zero whenever $\exp\{-\theta_3(i-j)\} < \tau$, $g_m(i)$ is replaced by

$$g_{m,\tau}(i) := \{j \in g_m(i) : j \geq i + \log \tau / \theta_3\},$$

estimated substituting θ_3 with $\hat{\theta}_3$. The graph $\hat{\mathcal{G}}$ is derived accordingly, by removing from $\mathcal{G}_0$ all those edges $j \to i$ for which $j \in g_m(i)$ but $j \notin \hat{g}_{m,\tau}(i)$.

Through a slight change in the priors proposed by Kidd and Katzfuss (2021), it is of interest to see the relationship with the model of Ben-David et al. (2015). When $\mathcal{G}$ is restricted to be a Directed Acyclic Graph (DAG), the proposed priors are equivalent to assign $(\boldsymbol{D}, \boldsymbol{U})$ a DAG-Wishart prior with hyperparameter $\boldsymbol{L}$ (a $q \times q$ positive definite matrix, diagonal for simplicity) and shape hyperparameter $a^{\mathcal{G}} = (a_1^{\mathcal{G}}, \dots, a_q^{\mathcal{G}})^\top$; see also Cao et al. (2019) and Castelletti and Peluso (2022) for further analyses of the model with, respectively, observational and interventional data. Hierarchically, this means to further characterize the Inverse Gamma (IG) prior $d_i|\theta \sim IG(\alpha_i, \beta_i)$ as

$$IG\left((\alpha - n + |g_m(i)| + 1)/2, \boldsymbol{L}_{i|g_m(i)}/2\right),$$

where $\boldsymbol{L}_{i|g_m(i)}$ is the Schur complement of $\boldsymbol{L}_{g_m(i),g_m(i)}$ in $\boldsymbol{L}_{i \cup g_m(i), i \cup g_m(i)}$. Also, the prior $\boldsymbol{u}_i | d_i, \theta \sim N(0, d_i V_i)$ becomes

$$N\left(0, d_i (\boldsymbol{L}_{g_m(i),g_m(i)})^{-1}\right).$$

This choice of hyperparameters guarantees *compatibility* among DAGs, so that different DAGs implying the same conditional independences will have equal integrated likelihoods; see Peluso and Consonni (2020). Then, following the Authors,

$$E\left(d_i|\theta\right) = \boldsymbol{L}_{i|g_m(i)} / \left(\alpha - n + |g_m(i)| - 1\right) \overset{fixed}{=} \theta_1 f_{\theta_2}(i)$$

and note that $\boldsymbol{L}_{i|g_m(i)} = \boldsymbol{L}_{ii}^{-1}$. Then v_{ij} is fixed as

$$v_{ij} = (\boldsymbol{L}_{g_m(i),g_m(i)})^{-1}_{jj} = \boldsymbol{L}^{-1}_{jj} = \theta_1 f_{\theta_2}(j) \left(\alpha - n + |g_m(j)| - 1\right)$$

and we still see an explicit dependence on j, also in the expression

$$Var\left(u_{ij}|\theta\right) = \theta_1^2 f_{\theta_2}(i) f_{\theta_2}(j) \left(\alpha - n + |g_m(i)| - 1\right) \left(\alpha - n + |g_m(j)| - 1\right).$$

[*] Università degli Studi di Milano-Bicocca, Milan, Italy, stefano.peluso@unimib.it

Finally, with this prior choice, we have

$$Var\left(d_i|\theta\right) = \left(4\theta_1^2 f_{\theta_2}(i)^2\right) / \left(\alpha - n + |g_m(i)| - 3\right),$$

so that the hyper-parameter α is free to choose, but under the constraint $\alpha > n + 3$.

The reformulation in terms of DAG-Wishart prior also suggests a direct way to extend to a model where the whole space of graphs is investigated, without the strong restrictions to the known ordering of the variables and to graphs which are *absolutely continuous* to the base $\mathcal{G}_0$. The Metropolis-Hastings algorithm could move jointly, or conditional to one another, on both the graph space and the space of θ, by using the algorithm suggested by Kidd and Katzfuss (2021) as a step on θ together with, for instance, the sampler in Consonni et al. (2017) as a step on DAGs.

References

Ben-David, E., Li, T., Massam, H., and Rajaratnam, B. (2015). “High dimensional Bayesian inference for Gaussian directed acyclic graph models.” *arXiv pre-print*. URL https://arxiv.org/abs/1109.4371 324

Cao, X., Khare, K., and Ghosh, M. (2019). “Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models.” *The Annals of Statistics*, 47(1): 319–348. MR3909935. doi: https://doi.org/10.1214/18-AOS1689. 324

Castelletti, F. and Peluso, S. (2022). “Network structure learning under uncertain interventions.” *Journal of the American Statistical Association*, forthcoming. doi: https://doi.org/10.1080/01621459.2022.2037430. 324

Consonni, G., La Rocca, L., and Peluso, S. (2017). “Objective Bayes covariate-adjusted sparse graphical model selection.” *Scandinavian Journal of Statistics*, 44(3): 741–764. MR3687971. doi: https://doi.org/10.1111/sjos.12273. 325

Kidd, B. and Katzfuss, M. (2021). “Bayesian nonstationary and nonparametric covariance estimation for large spatial data.” *Bayesian Analysis*, 1(1): 1–22. 324, 325

Peluso, S. and Consonni, G. (2020). “Compatible priors for model selection of high-dimensional Gaussian DAGs.” *Electronic Journal of Statistics*, 14(2): 4110–4132. MR4170698. doi: https://doi.org/10.1214/20-EJS1768. 324

Contributed Discussion

Diego Andrés Pérez Ruiz[*] and Tom Leonard[†]

Maybe the authors should refer to the Bayesian Econometric approach pioneered by LeSage and Kelley Pace (2007). Le Sage and Kelley proposed the matrix exponential spatial specification (MESS) as a way of simplification of the log-likelihood allowing a closed form solution to the problem of maximum likelihood estimation and simplification of the Bayesian estimation of the model. MESS can produce estimates and inferences similar to those from conventional spatial auto-regressive (AR) models, but has analytical, computational, and interpretive advantages.

Inference and estimation of traditional spatial autoregressive (SAR) models requires non-linear optimization for estimation and inference. The conventional spatial autoregressive approach introduces additional theoretical complexity relative to non-spatial autoregressive models and is difficult to implement in large samples.

MESS replaces the conventional geometric decay of influence over space with an exponential pattern of decay. It results in theoretical simplicity as well as improved numerical performance relative to the conventional spatial autoregression. MESS models the dependence of the covariances on explanatory variables by observing that for any real symmetric matrix **A** the matrix exponential transformation **C** is a positive definite matrix.

Le Sage and Kelley utilise an approach proposed by Chiu et al. (1996). Chiu et al. develop a generalized linear model for covariance matrices together with a linear model for the means, using the matrix logarithmic transformation

$$\mathbf{A} = \log \mathbf{C}.$$

This provides a very general paradigm for modeling a multitude of spatial processes, particularly when random effects are included with fixed effects. Why is another approach needed?

Perhaps the authors should also consider the large literature for Bayesian inference for a covariance matrix **C** that refers to the matrix transformation $\mathbf{A} = \log \mathbf{C}$. Key papers include Leonard and Hsu (1992) and Hsu et al. (2012), who assume a matrix normal prior distribution for the upper triangular elements of **A**. In particular, Deng and Tsui (2013) address the estimation of large sparse covariance matrices. Most recently, Magnus et al. (2021) derive an explicit expression for the Jacobian of the matrix exponential transformation, with even further applications in Econometrics.

[*] School of Social Statistics, University of Manchester, diego.perezruiz@manchester.ac.uk

[†] Retired – 4/3 Hopetoun Crescent, Edinburgh EH7 4AY, leonardthomas70@googlemail.com

References

Chiu, T. Y., Leonard, T., and Tsui, K.-W. (1996). "The matrix-logarithmic covariance model." *Journal of the American Statistical Association*, 91(433): 198–210. MR1394074. doi: https://doi.org/10.2307/2291396. 326

Deng, X. and Tsui, K.-W. (2013). "Penalized covariance matrix estimation using a matrix-logarithm transformation." *Journal of Computational and Graphical Statistics*, 22(2): 494–512. MR3173726. doi: https://doi.org/10.1080/10618600.2012.715556. 326

Hsu, C.-W., Sinay, M. S., and Hsu, J. S. (2012). "Bayesian estimation of a covariance matrix with flexible prior specification." *Annals of the Institute of Statistical Mathematics*, 64(2): 319–342. MR2878908. doi: https://doi.org/10.1007/s10463-010-0314-5. 326

Leonard, T. and Hsu, J. S. (1992). "Bayesian inference for a covariance matrix." *The Annals of Statistics*, 20(4): 1669–1696. MR1193308. doi: https://doi.org/10.1214/aos/1176348885. 326

LeSage, J. P. and Kelley Pace, R. (2007). "A matrix exponential spatial specification." *Journal of Econometrics*, 140(1): 190–214. Analysis of spatially dependent data. MR2395921. doi: https://doi.org/10.1016/j.jeconom.2006.09.007. 326

Magnus, J. R., Pijls, H. G., and Sentana, E. (2021). "The Jacobian of the exponential function." *Journal of Economic Dynamics and Control*, 127: 104122. MR4256919. doi: https://doi.org/10.1016/j.jedc.2021.104122. 326

Contributed Discussion

Matthew J. Heaton[*]

I thank Kidd and Katzfuss (2021) for a fascinating article. The core idea to flexibly model spatial data in a computationally feasible way through a series of regressions is novel. The structure of their method is exciting for the spatial statistics community because it opens the door to allowing the data to infer the type of underlying covariance function rather than requiring the user to specify one. Further, their careful attention to the choice in priors to allow shrinking towards a given covariance function is particularly appealing for situations in which the data may not be able to fully inform the covariance function. Overall, their methods provide a fascinating approach to modeling spatial data.

While the approach of Kidd and Katzfuss (2021) provides a promising way forward for spatial analysis, there are various aspects of their approach that I worry may limit the applicability of their methods and, I believe, are not well discussed in the original article. Hence, I submit this public discussion in hopes of further illuminating these shortcomings and to create a forum for potentially addressing them.

1. **Lack of Repetitions.** Following the notation of Kidd and Katzfuss (2021), let $\boldsymbol{y}_i = (y_i^{(1)}, \ldots, y_i^{(N)})'$ be the vector of N repeated measures of a response at location $\boldsymbol{s}_i$ for $i = 1, \ldots, n$. Assuming independence between repeated measures and a Gaussian process within a measure, the Vecchia approximation gives the likelihood as

   $$\prod_{i=1}^{n} \mathcal{N}\left(\boldsymbol{y}_i \mid \boldsymbol{X}_i \boldsymbol{u}_i, d_i \boldsymbol{I}_N\right), \tag{0.1}$$

   where $\boldsymbol{X}_i$ is an $N \times m$ matrix of $\{y_i^{(j)} : i \in \mathcal{S}_i, j \in \{1, \ldots, N\}\}$ where $\mathcal{S}_i$ is the set of m neighbors of location i, $\boldsymbol{u}_i$ are m unknown regression coefficients and d_i is a common variance (see Kidd and Katzfuss 2021 Equation (5) for more details on the derivation). Inference for the above model is carried out via conjugate normal-inverse gamma priors for $\boldsymbol{u}_i$ and d_i.

   Under this regression framework, in order for $\boldsymbol{u}_i$ and d_i to be identifiable, the number of repetitions N would have to exceed the number of neighbors m. To see this in the extreme sense, notice that if $N = 1$ then $\boldsymbol{y}_i = y_i$ is a scalar used to infer m coefficients in $\boldsymbol{u}_i$ (a situation where the number of parameters is m times the number of observations). Admirably, the authors recognize this and specify highly informative priors that shrink towards a certain covariance function. However, I am not convinced that the derived shrinkage priors are sufficient to perform inference. This is evidenced by the fact that the authors only consider situations where as few as $N = 3$ repetitions are available and never consider the extreme case where $N = 1$. In my experience, repeated measures in spatial data

[*] 2196 WVB, Brigham Young University, Provo, UT 84602, mheaton@stat.byu.edu

are actually quite rare (see the recent spatial data competitions in Heaton et al., 2019; Huang et al., 2021, for examples) with the more common situation being a spatio-temporal field rather than actual repeated measures.

A potential solution to the lack of identifiability would be to parameterize $\{\boldsymbol{u}_i, d_i\}$ as a function of a set of parameters rather than parameterizing the priors as such. This was done in a similar setup by Messick et al. (2017) who used basis functions to parameterize the coefficients (albeit in a conditional setup for a multivariate spatial response rather than a univariate response as is the case here). However, such parameterization would sacrifice conjugacy of the model (and, hence, computational speed) to obtain identifiability.

2. **Kriging.** Under the usual Vecchia process setup, prediction is possible because the conditional distributions are given by a small set of covariance parameters (see Datta et al., 2016a). However, under the regression setup in (0.1) above, when considering prediction to a new location $\boldsymbol{s}^\star$, the associated coefficients, say $\boldsymbol{u}^\star$, are unknown and, hence, how to use the model of Kidd and Katzfuss (2021) for prediction is unclear. This is demonstrated by the fact that the authors focus primarily on covariance estimation (a worthy endeavor in its own right) rather than prediction. It would seem that $\{\boldsymbol{u}_i, d_i\}$ would have to, again, be parameterized using a small set of parameters to facilitate prediction.

3. **Extensions.** While the authors mention several interesting extensions to their work in their concluding section, I can see several additional extensions that were not mentioned but, I feel, should be highlighted to illustrate the flexibility and novelty of their approach. First, the approach of Kidd and Katzfuss (2021) can be intertwined with the methods of Cressie and Zammit-Mangion (2016) or Messick et al. (2017) to capture non-stationary multivariate spatial processes. Second, a similar conditional approach could be used for space-time processes potentially mimicking a dynamic linear model (Petris et al., 2009) wherein regression coefficients are estimated not only for spatial neighbors but space-time neighbors (Datta et al., 2016b).

In conclusion, I again thank Kidd and Katzfuss (2021) for what will surely be a highly read and influential paper. I believe their methods will be a go-to-method in spatial statistician's toolboxes for years to come.

References

Cressie, N. and Zammit-Mangion, A. (2016). "Multivariate spatial covariance models: a conditional approach." *Biometrika*, 103(4): 915–935. MR3620448. doi: https://doi.org/10.1093/biomet/asw045. 329

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets." *Journal of the American Statistical Association*, 111(514): 800–812. MR3538706. doi: https://doi.org/10.1080/01621459.2015.1044091. 329

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A., and Schaap, M. (2016b). "Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis." *The annals of applied statistics*, 10(3): 1286. MR3553225. doi: https://doi.org/10.1214/16-AOAS931. 329

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., et al. (2019). "A case study competition among methods for analyzing large spatial data." *Journal of Agricultural, Biological and Environmental Statistics*, 24(3): 398–425. MR3996451. doi: https://doi.org/10.1007/s13253-018-00348-w. 329

Huang, H., Abdulah, S., Sun, Y., Ltaief, H., Keyes, D. E., and Genton, M. G. (2021). "Competition on spatial statistics for large datasets." *Journal of Agricultural, Biological and Environmental Statistics*, 26(4): 580–595. MR4330397. doi: https://doi.org/10.1007/s13253-021-00457-z. 329

Kidd, B. and Katzfuss, M. (2021). "Bayesian Nonstationary and Nonparametric Covariance Estimation for Large Spatial Data." *Bayesian Analysis*, 1 – 22. doi: https://doi.org/10.1214/21-BA1273. 328, 329

Messick, R. M., Heaton, M. J., and Hansen, N. (2017). "Multivariate spatial mapping of soil water holding capacity with spatially varying cross-correlations." *The Annals of Applied Statistics*, 11(1): 69–92. MR3634315. doi: https://doi.org/10.1214/16-AOAS991. 329

Petris, G., Petrone, S., and Campagnoli, P. (2009). "Dynamic linear models." In *Dynamic Linear Models with R*, 31–84. Springer. MR2730074. doi: https://doi.org/10.1007/b135794. 329

Contributed Discussion

Lamiae Azizi[*], Sara Wade[†], and Weichang Yu[‡]

1 Introduction

First, we would like to congratulate the authors for producing interesting and important contributions to advance Bayesian inference of spatial data. In the present paper, the authors investigated the important issue of nonparametric and nonstationary Bayesian inference of a high dimensional covariance matrix. The novel framework relies on an extension of the Vecchia approximation, which effectively approximates the Cholesky factor of the precision matrix with a sparse matrix; a common approach for covariance estimation. The extension relaxes the usual restrictive assumptions of isotropy and the need for a specific parametric form of the covariance function, an unrealistic one in many real world applications. The authors exploited recent results on the exponential decay of the entries of the inverse Cholesky factor for covariance matrices under a specific ordering scheme of the spatial locations in order to specify suitable prior distributions. The resulting framework is a scalable and flexible approach that enforces neither stationary nor parametric structures while accounting for uncertainty and allowing for regularization through the choice of priors. We note that the proposed inference approach achieves a remarkable computational performance with a complexity that is linear in the number of locations n, i.e., $O(n(m^2N+m^3))$. This is a significant reduction from the original computational complexity of inferring a general covariance matrix $O(n^3)$. The flexible specification of the hyperparameters allows for the incorporation of prior information about the underlying covariance structure when available. The authors provided a comprehensive repository for the accompanying codes and tutorial, which allows their results to be reproduced and facilitates their model to be adopted in other applications (**we noted a small typo in the Github tutorial:** `find_nn_dist(fields::rdist(dataa), n_locs)` **should be** `find_nn_dist(fields::rdist(locs), n_locs)`).

We start our discussion by commenting on the authors' choice of priors. In particular, both the d_i's and $\boldsymbol{u}_i$'s are assigned conditionally independent priors, across $i = 1, \ldots, n$. This prior independence is motivated by conjugacy, resulting in closed form expressions, and to reduce the computation burden. Nonetheless, if one were to simulate from the prior, the behavior of the d_i's and $\boldsymbol{u}_i$'s could be quite erratic across i due to this independence assumption. This may result in a strange covariance structure, and we invite the authors to simulate from the prior and comment on this aspect.

We further note that while independence of $u_{i\,j}$ for $j = 1, \ldots, m$ is assumed, i.e. diagonal $\boldsymbol{V}_i$, a general form of $\boldsymbol{V}_i$ could be used, allowing for dependence, without increasing the computational cost. Moreover, the authors specified the prior hyperparameters for d_i to match the exponential decay in the true values of d_i of an exponential covariance

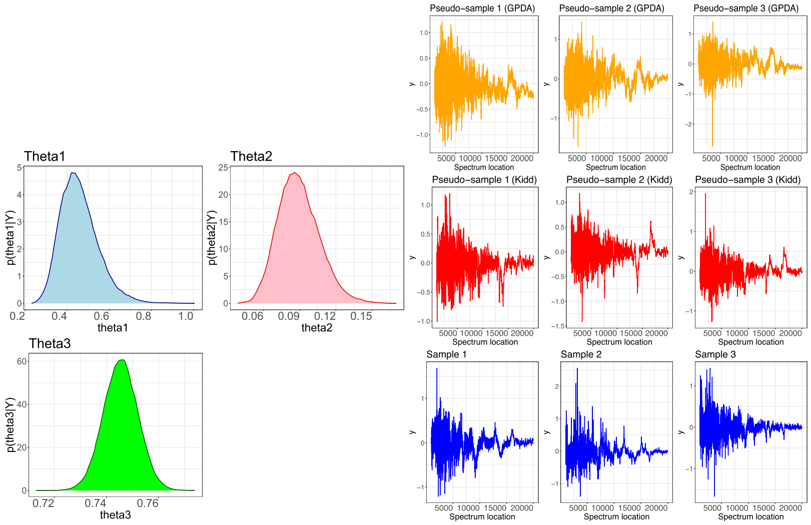
[*] NABLAS AI - Australia, lamiae.azizi@gmail.com
[†] School of Mathematics, University of Edinburgh, sara.wade@ed.ac.uk
[‡] Melbourne Centre for Data Science, University of Melbourne, weichang.yu@unimelb.edu.au

function. However, the prior credible intervals in Figure 3 of their paper suggest that the true values in some of the panels are not included and point towards a not-well chosen hyperparameter. The choice of hyperparameters α_i and β_i is guided by the rule that the prior SD of d_i is half of the prior mean. While it looks reasonable to us, to some extent, to assume that the prior SD decays exponentially, the authors did not provide convincing justification for the choice of the multiplicative factor 0.5; indeed this factor may need to be larger for better prior coverage of the true values. We follow by commenting on the independence assumption in the data model. We wonder how restrictive is this in cases where the $\boldsymbol{y}^{(l)}$'s are correlated. The potential heterogeneous behavior across the replicates is not accounted for by the approach, and it is unclear to us how difficult the extension to account for this would be or how robust the results would be when applied to such cases in its current form? Lastly, we ask the author's to clarify the data-driven choice of m. When using MCMC, would m change at every iteration based on the sampled θ_3? and how do you combine MCMC draws across varying m?

2 Functional Data Analysis

While the author's work was motivated by spatial data, we believe that this piece of work has the potential to become an important building block of more general models for *functional data* as well. Similar to spatial data, the issues of nonstationarity and nonparametric structures of the covariance matrices are prevalent and require special design when developing an appropriate model for the application of interest. This observation stems from our recent work in Yu et al. (2021) for high dimensional nonstationary functional data, where we developed a Bayesian approach that relies on a non-stationary exponential covariance structure and a novel variational scheme that exploits advances in the use of sparse precision matrices to achieve scalability. To illustrate briefly our suggested extension and its usefulness in this context, we use the author's approach to model the breast cancer data analyzed respectively in Shi et al. (2006) and Yu et al. (2021) and check its ability to recover the non-stationarity structure present in the data. The dataset consists of $N = 64$ spectra of breast cancer patients observed at $n = 10451$ locations. To assess briefly how the author's proposed approach can be used for covariance estimation in functional data, we sampled three draws from the posterior predictive distribution $p(\boldsymbol{y}^\star|\boldsymbol{Y})$, and we compare these to the observed data as well as to samples from the posterior predictive distribution using the model in Yu et al. (2021). Figure 1 suggests that both approaches behave similarly in recovering the nonstationary behavior and range of functions, indicating that the proposed approach has potential to be useful beyond the context it was initially developed for. Moreover, the experimental results in the authors' paper suggest that a sufficient number of replicates N (at least 20) is required for competitive covariance estimation with their method, and this is typically the case in functional data, with N often greater than 50.

We should however emphasize that in its current form some extensions are necessary in order for it to be applied accurately to functional data, notably in classification tasks (e.g. discriminating breast cancer patients from healthy individuals based mass-spectrometry data). In particular, we require the specification of a non-zero mean function, which must be inferred and typically also exhibits non-stationarity. Moreover,



(a)

(b)

Figure 1: (a) Kernel density plots based on approximately 1×10^6 MCMC samples of $\boldsymbol{\theta}$ with $m = 3$. (b) Three draws from the predictive distributions based on: model in Yu et al. (2021) (**Top row**), the present paper's method (**Middle row**), randomly selected observed spectra in the dataset (**Bottom row**).

heterogeneity may exist across the observed functional replicates; and it would be interesting to understand if one could build a semi-parametric extension that allows for some parametric heterogeneity across replicates.

References

Shi, Q., Harris, L., Lu, X., Li, X., Hwang, J., Gentleman, R., Iglehart, J., and Miron, A. (2006). "Declining plasma fibrinogen alpha fragment identifies HER2-positive breast cancer patients and reverts to normal levels after surgery." *Journal of Proteome Research*, 5(11): 2947–2955. 332

Yu, W., Wade, S., Bondell, H., and Azizi, L. (2021). "Non-stationary Gaussian process discriminant analysis with variable selection for high-dimensional functional data." *arXiv*. MR4085868. doi: `https://doi.org/10.1080/10618600.2019.1637748`. 332, 333

Contributed Discussion

Suman Guha[*]

We would like to start by congratulating the authors for their brilliant effort in putting together the nonparametric Bayesian method and Vecchia approximation to create a scalable Bayesian nonstationary and nonparametric model for spatial data with/without replication.

To the best of our knowledge, there are only a few articles that consider Bayesian nonstationary and nonparametric modeling of the covariance function underlying a spatial process and two early notable works in this direction are Damian et al. (2001) and Schmidt and O'Hagan (2003). Unfortunately, both of these models require implementing the MCMC algorithm on high dimensional parameter space, which is computationally demanding. On the other hand, fitting the model proposed by Kidd and Katzfuss (2021) requires merely $\mathcal{O}(n(m^2N + m^3))$ time and hence can be applied to massive spatial data. However, note that Kidd and Katzfuss (2021) assume that the spatial locations are already ordered following a maximin ordering, which iteratively selects the $(i + 1)$-th location to be the furthest one from the already selected i locations. In reality, quite opposite happens and the spatial locations are often ordered according to clusters formed by geographical and political boundaries, for example, terrains, countries, zip codes, etc. In that case, the true computational cost for running an MCMC up to L iterations would be the sum of computational cost for finding a maximin ordering, computational cost for finding the nearest-neighbor conditioning sets, and the computational cost exclusively attributed to the MCMC iterations. At this point, we differ with Kidd and Katzfuss (2021) to state that finding a maximin ordering of n spatial locations requires $\mathcal{O}(n^3)$ time (see Guinness, 2018) and hence the total cost would be $\mathcal{O}(n^3) + \mathcal{O}(n \log n) + \mathcal{O}(Ln(m^2N + m^3)) = \mathcal{O}(n^3)$. Moreover, the algorithm for finding a maximin ordering is not parallelizable. However, there are approximate maximin orderings that preserve the salient features of a maximin ordering and can be found in $\mathcal{O}(n \log n)$ time (see Guinness, 2018). Using such an approximate maximin ordering the model proposed by Kidd and Katzfuss (2021) can be fitted in $\mathcal{O}(n \log n) + \mathcal{O}(n \log n) + \mathcal{O}(Ln(m^2N + m^3)) = \mathcal{O}(n \log n)$ time albeit providing similar performance.

Another critical aspect is the selection of the neighbor sets that determines the sparsity of the model which is an approximation to the full and dense spatial Gaussian process model. Kidd and Katzfuss (2021) use m nearest locations to form the neighbor set for any particular location and suggest a data-driven dynamic choice of m which is neither too small nor too large. However, we anticipate that a judiciously chosen fixed m would produce an equally good result as the dynamic one and sensitivity analysis over a range of values of m would confirm it. In a purely parametric setting as specified in Datta et al. (2016) different reasonable values of m which are neither too small nor too large

[*] Department of Statistics, Presidency University, 86/1 College Street Kolkata 700073, India, suman.stat@presiuniv.ac.in

show an insignificant difference in model performance. Similar behavior is anticipated also for this model. Alternatively, one may elicit a prior on the natural numbers and select m in a fully Bayesian manner from the associated posterior. A different line of thought is exploring the effect of considering mostly the nearest neighbors with a few far neighbors. What if the neighbor sets include a few far neighbors besides the nearest ones as sometimes that leads to dramatic improvements in the efficiencies of the resulting estimators (see Stein et al., 2004).

Selection of prior distribution and prior hyperparameters is an integral part of any Bayesian modeling and more so in the Bayesian spatial modeling. The form of the prior distribution and the exact values of the hyperparameters are to be set so as to ensure that a wide range of spatial structures can be accommodated within the proposed Bayesian model. Kidd and Katzfuss (2021) propose the general pragmatic solution of selecting proper, but weakly informative independent normal-inverse-gamma priors for each of $(\boldsymbol{u}_i, d_i)$ for $i = 1, \cdots, n$. The normal-inverse-gamma priors lead to normal-inverse-gamma posteriors for $(\boldsymbol{u}_i, d_i)$, from which it is easy to simulate. However, we wonder whether one can elicit a non-informative prior for the model in Kidd and Katzfuss (2021) as in Berger et al. (2001), who carried out objective Bayesian analysis for spatial data using reference and Jeffreys priors for variance-covariance parameters in a Gaussian random field without nugget effect.

Finally, it is important to assess how weak is the prior (for $\boldsymbol{\Sigma}$) induced on the cone of positive definite matrices. Kidd and Katzfuss (2021) point out that the prior does not center around any specific covariance matrix or any specific family of covariance matrices. Indeed, in section 2.9 they describe how little modification of their approach lands with a prior that centers at and hence shrinks towards a specific covariance matrix/family of covariance matrices. That said, we still believe that the prior is mostly supported on covariance matrices induced by Matérn kernels, by its very construction. Won't it conflict with spatial data originating from kernels which are very different from Matérn i.e. Gaussian kernels or periodic kernels? One way to answer this question is to simulate $\boldsymbol{\Sigma}$ from the proposed prior and investigate the histogram of $tr((\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0)'(\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0))$, $\lambda_{max}((\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0)'(\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0))$ or $KL(\boldsymbol{\Sigma} \mid\mid \boldsymbol{\Sigma}_0)$ where $\boldsymbol{\Sigma}_0$ is associated with the kernel of interest.

References

Berger, J. O., De Oliveira, V., and Sansó, B. (2001). "Objective Bayesian analysis of spatially correlated data." *Journal of the American Statistical Association*, 96(456): 1361–1374. MR1946582. doi: https://doi.org/10.1198/016214501753382282. 335

Damian, D., Sampson, P. D., and Guttorp, P. (2001). "Bayesian estimation of semi-parametric non-stationary spatial covariance structures." *Environmetrics: The Official Journal of the International Environmetrics Society*, 12(2): 161–178. 334

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets." *Journal of the*

American Statistical Association, 111(514): 800–812. MR3538706. doi: https://doi.org/10.1080/01621459.2015.1044091. 334

Guinness, J. (2018). "Permutation and grouping methods for sharpening Gaussian process approximations." *Technometrics*, 60(4): 415–429. MR3878098. doi: https://doi.org/10.1080/00401706.2018.1437476. 334

Kidd, B. and Katzfuss, M. (2021). "Bayesian nonstationary and nonparametric covariance estimation for large spatial data." *Bayesian Analysis*, 1(1): 1–22. 334, 335

Schmidt, A. M. and O'Hagan, A. (2003). "Bayesian inference for non-stationary spatial covariance structure via spatial deformations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3): 743–758. MR1998632. doi: https://doi.org/10.1111/1467-9868.00413. 334

Stein, M. L., Chi, Z., and Welty, L. J. (2004). "Approximating likelihoods for large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2): 275–296. MR2062376. doi: https://doi.org/10.1046/j.1369-7412.2003.05512.x. 335

Contributed Discussion

Andrea Sottosanti[*], Davide Risso[†], and Cristian Castiglione[‡]

We first want to congratulate the authors for the impressive work, which consists in a non-parametric method for estimating the spatial dependence structure of both stationary and non-stationary fields. For convenience, we refer to their method as `NPVecchia`. We are convinced that their work represents a notable advance in spatial statistics and brings a powerful and flexible analysis tool into many real-data problems. Nevertheless, to better understand which domains of application could benefit of such innovation, some open issues should be further discussed.

Due to the absence of an explicit model for the underlying continuous spatial field, we are concerned about the possibility of using `NPVecchia` for performing common operations in spatial data analysis, such as spatial interpolation and prediction. This limitation would restrict the range of applications to contexts of in-sample analysis, where prediction over unobserved sites is not the main goal. Moreover, geo-spatial data are frequently observed upon a collection of sites distributed extremely irregularly over the space and so the distances between different points can vary considerably. On the contrary, Kidd and Katzfuss (2021) implicitly assume an almost uniform distribution of the observed locations.

It therefore appears that `NPVecchia` is appropriate for applications whose data are characterized by roughly equally-spaced sites, with many observations per site, and whose main goal is not the prediction over unobserved locations. Based on these considerations, we believe that `NPVecchia` would be ideal for modelling the data processed by a new, groundbreaking class of technologies for DNA sequencing, called *spatial transcriptomics* (*s.t.*). For the substantial contributions that *s.t.* is carrying into the study of biological organisms, it was named *method of the year 2020* (Marx, 2021). The 10X Visium sequencing platform (Rao et al., 2020), one among several *s.t.* technologies, collects the cells of a tissue sample through a grid of equally spaced spots on the surface of a chip. The transcriptome is sequenced within each spot, where a few neighbour cells are collected. The output of the procedure is the expression of thousands of genes within each spot, together with the coordinates of the spots. Figure 1 is an example of a human prostate cancer tissue sample processed with 10X Visium.[1]

The growing popularity of *s.t.* has allowed researchers to identify the so-called *spatially expressed* (*s.e.*) genes, i.e., genes that show spatial variation patterns across the tissue. Discovering and comprehending the functions of *s.e.* genes is of great scientific interest and might lead to new insights and discoveries of specific biological processes.

[*] Department of Statistical Sciences, University of Padova, via C. Battisti 241, Padova, Italy, andrea.sottosanti@unipd.it

[†] Department of Statistical Sciences, University of Padova, via C. Battisti 241, Padova, Italy, davide.risso@unipd.it

[‡] Department of Statistical Sciences, University of Padova, via C. Battisti 241, Padova, Italy, cristian.castiglione@phd.unipd.it

[1] `https://www.10xgenomics.com/resources/datasets`.

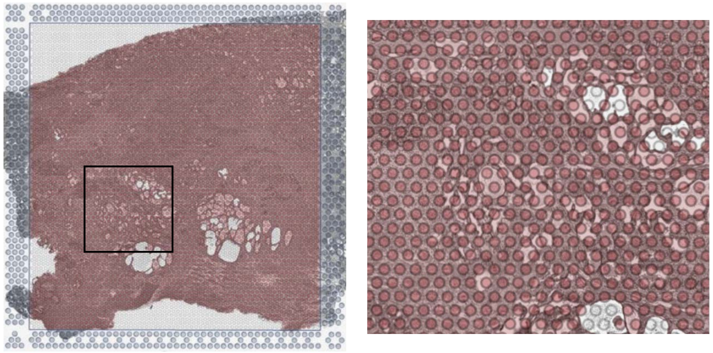


Figure 1: Left: human prostate cancer sample analysed with the 10X Visium platform. The tissue covers a total of 4,371 spots. Right: detail of the left figure corresponding to the black square. The spots on the chip are visible as circles over the whole surface.

Svensson et al. (2018) and Sun et al. (2020) tackle this research question as a statistical hypothesis testing problem where, for each gene, the presence of spatial variation patterns is tested. Both these methods assume a stationary field and express the dependency across the spots using parametric spatial correlation functions. To overcome these limitations and perform an accurate inferential process, we discuss a possible use of the idea of Kidd and Katzfuss (2021) into the analysis of *s.t.* experiments, with the aim of improving the discovery of *s.e.* genes.

Let $\mathbf{Y} = (\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(N)})^T$ be an $N \times n$ experiment matrix, where $\mathbf{y}^{(\ell)}$ is the expression of gene ℓ over the n observational sites (spots) with spatial coordinates $\mathbf{s}_1, \dots, \mathbf{s}_n$. We assume that the data have been centered and pre-processed in such a way that $y_i^{(\ell)} \in \mathbb{R}$ and the histogram of each $\mathbf{y}^{(\ell)}$ is approximately symmetric. Then, we assume the following model:

$$\mathbf{y}^{(\ell)} | \mathbf{f}^{(\ell)}, \lambda_\ell^2, \sigma_\varepsilon^2 \sim \mathcal{N}_n(\mathbf{f}^{(\ell)}, \lambda_\ell^2 \mathbf{I}_n + \sigma_\varepsilon^2 \mathbf{I}_n), \quad \mathbf{f}^{(\ell)} | \tau_\ell^2, \mathbf{\Sigma} \sim \mathcal{N}_n(\mathbf{0}_n, \tau_\ell^2 \mathbf{\Sigma}), \tag{1}$$

where $\mathbf{f}^{(\ell)}$ is the gene-specific spatial field with marginal variance τ_ℓ^2 and common covariance matrix $\mathbf{\Sigma}$, while λ_ℓ^2 and σ_ε^2 are the variances of idiosyncratic error terms. We assume the prior structure on the precision matrix $\mathbf{\Sigma}^{-1}$ proposed by Kidd and Katzfuss (2021), and non-informative priors for the variance parameters λ_ℓ^2 and σ_ε^2 as suggested by Gelman (2006). Last, taking inspiration from the recent literature on shrinkage priors and on the extraction of sparse signals (Bhadra et al., 2019), we propose to consider a prior model for τ_ℓ^2 that performs an aggressive shrinkage toward 0 if no spatial patterns arise, while leaving a high level of flexibility when the genes show a significant amount of spatial correlation. Within this framework, an interesting choice with optimal theoretical properties is the Horseshoe prior (Carvalho et al., 2010), corresponding to a hierarchical Half-Cauchy distribution on the standard deviation parameter τ_ℓ.

Formula (1) can be seen as a generalized, Bayesian version of the SpatialDE model proposed by Svensson et al. (2018), where all the unknown parameters, including the spatial covariance matrix $\mathbf{\Sigma}$, are inferred directly from the data using, for example, a Gibbs sampling algorithm as described in Section 2.8 of Kidd and Katzfuss (2021). Thanks to the shrinkage imposed on τ_ℓ^2 through its a priori setup, the *s.e.* genes can be determined by evaluating the posterior distribution of $\delta_\ell = \tau_\ell^2/(\tau_\ell^2 + \lambda_\ell^2)$, that is the percentage of spatial variability specific of gene ℓ. For example, one may define an operating rule based on some threshold conditions, classifying as *s.e.* only those genes which have $\mathbb{P}(\delta_\ell \geq t|\mathbf{Y}) \geq p$ for t close to 0 and p close to 1.

Although we see a lot of promise in applying the work of Kidd and Katzfuss (2021) to the problem of identifying *s.e.* genes, it remains an open question whether irregularities on the edges and within the surface of tissues, as the one that appears in Figure 1 (left), could somehow affect the estimate of $\mathbf{\Sigma}$.

Several generalizations of the model in Formula (1) could be explored. First, it is often of clinical interest to evaluate biological processes common to a cohort of patients. Hence, the model could be extended to identify *s.e.* genes by simultaneously evaluating multiple tissue samples. Second, since *s.t.* raw data are highly variable, possibly zero-inflated counts, a Poisson or Negative Binomial extension could be considered, similarly to what has been done by Sun et al. (2020).

References

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). "Lasso meets horseshoe: a survey." *Statistical Science*, 34(3): 405–427. MR4017521. doi: https://doi.org/10.1214/19-STS700. 338

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 338

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–533. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 338

Kidd, B. and Katzfuss, M. (2021). "Bayesian nonstationary and nonparametric covariance estimation for large spatial data." *Bayesian Analysis*. 337, 338, 339

Marx, V. (2021). "Method of the Year 2020: spatially resolved transcriptomics." *Nature Methods*, 18: 9–14. 337

Rao, N., Clark, S., and Habern, O. (2020). "Bridging genomics and tissue pathology." *Genetic Engineering & Biotechnology News*, 40(2): 50–51. 337

Sun, S., Zhu, J., and Zhou, X. (2020). "Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies." *Nature Methods*, 17(2): 193–200. 338, 339

Svensson, V., Teichmann, S. A., and Stegle, O. (2018). "SpatialDE: identification of spatially variable genes." *Nature Methods*, 15(5): 343–346. 338

Contributed Discussion

Isa Marques[*], Thomas Kneib[†], and Finn Lindgren[‡]

Vecchia approximations can, by construction, be seen as a special case of general Gaussian Markov random field (GMRF) computational Cholesky techniques. We are therefore surprised by the general lack of references to and discussion of other non-stationary GMRF methods and their continuous domain counterparts, stochastic partial differential equations (SPDE), in the introduction as well as in the simulation study and discussion. The SPDE representations from Lindgren et al. (2011) provide direct methods for representing continuous non-stationary Gaussian random field (GRF) precision operators as GMRFs on the coefficients of locally supported or nested basis expansions. Crucially, the SPDE models have spatially coherent interpretability via the differential operators, regardless of the discretised node ordering, which also generalises to non-Markovian models (Lindgren et al., 2022). Some of these computationally efficient models, such as Ingebrigtsen et al. (2015) or Fuglstad et al. (2015), would have been relevant approaches worth considering, at least in the simulation study. In particular, the former model considers replicates and accommodates covariates.

As shown by Guinness (2018), the order in which the observations are included has an impact on the quality of Vecchia approximations. The method presented (OURS) involves a choice of discretisation ordering and resulting directed acyclic graph (Katzfuss and Guinness, 2021). The resulting coefficients are strongly tied to this graph, and cannot easily be interpreted in a spatially coherent manner. In contrast, the continuous domain SPDE-approach provides both closed form expressions for the precision matrix elements, and spatially coherent interpretability on the continuous space, as opposed to only on the discrete subset of locations that have observations (in OURS). However, we note that the maximin ordering bears a qualitative resemblance to hierarchical wavelet basis methods (Bolin and Lindgren, 2013), and adapting the computational methods to such basis expansions could potentially improve the interpretability and generalisability, especially if combined with a more spatially coherent shrinkage prior.

The authors suggest to parameterise their prior distribution such that it is centered around the prior expectation of a Matérn-like covariance. While the prior mean may be interpreted as the centre of the prior in some sense, a more principled approach for constructing priors that are appropriately centered around a base model has been developed in Simpson et al. (2017). They quantify the deviance between a prior and the base model with a distance measure and then construct an exponential decay prior on the distance scale. This approach has the advantage of centering around the complete base model, rather than working with the prior expectations, and is also equivariant under parameter transformations. We are curious whether such an approach could be transferred to the construction of the prior distribution for the model suggested here.

[*] Georg-August-University Göttingen, Germany, imarques@uni-goettingen.de

[†] Georg-August-University Göttingen, Germany

[‡] The University of Edinburgh, Scotland

On a related note, the numerical results in the paper highlight that a sufficiently large number of replicates N is important to get useful and accurate inferences. In the simulation study, the Bayesian model (OURS) seems quite robust, even for a small N, while the maximum likelihood estimated (MLE) version performs poorly for $N < 10$. As a potential explanation, the authors state that the added shrinkage from their prior improves accuracy when compared to MLE. While shrinkage would certainly play an important role, we wonder about the concrete specification used – if it follows Section 2.9 of the paper and shrinks towards the Vecchia approximation of the true covariance, the results might not translate well to the application, where the true covariance is unknown. Ultimately, this could explain the behavior of OURS in Figure 9 for $N < 10$.

In summary, while there are clear computational speed advantages to the proposed method, it is unclear how to adjust it to achieve spatially coherent models.

References

Bolin, D. and Lindgren, F. (2013). "A comparison between Markov approximations and other methods for large spatial data sets." *Computational Statistics & Data Analysis*, 61: 7–21. MR3062997. doi: https://doi.org/10.1016/j.csda.2012.11.011. 340

Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. (2015). "Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy." *Statistica Sinica*, 115–133. MR3328806. 340

Guinness, J. (2018). "Permutation and grouping methods for sharpening Gaussian process approximations." *Technometrics*, 60(4): 415–429. MR3878098. doi: https://doi.org/10.1080/00401706.2018.1437476. 340

Ingebrigtsen, R., Lindgren, F., Steinsland, I., and Martino, S. (2015). "Estimation of a non-stationary model for annual precipitation in southern Norway using replicates of the spatial field." *Spatial Statistics*, 14: 338–364. MR3431045. doi: https://doi.org/10.1016/j.spasta.2015.07.003. 340

Katzfuss, M. and Guinness, J. (2021). "A general framework for Vecchia approximations of Gaussian processes." *Statistical Science*, 36(1): 124–141. MR4194207. doi: https://doi.org/10.1214/19-STS755. 340

Lindgren, F., Bolin, D., and Rue, H. (2022). "The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running." *Spatial Statistics*, 100599. 340

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." *Journal of the Royal Statistical Society: Series B*, 73(4): 423–498. MR2853727. doi: https://doi.org/10.1111/j.1467-9868.2011.00777.x. 340

Simpson, D., Rue, T. G., H. Martins, Riebler, A., and Sørbye, S. H. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32(1): 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 340

Rejoinder*

Brian Kidd[†] and Matthias Katzfuss[‡]

We would like to thank all discussants for their stimulating comments. In this rejoinder, we first discuss a few bigger issues and common threads, and then give some shorter responses to individual comments. We use the same notation as in the main paper (Kidd and Katzfuss, 2022). All section, figure, and equation numbers in this rejoinder are prefixed by the letter R, to distinguish them from elements in the main paper.

R1 Shrinkage toward a parametric covariance

Our methodology can be modified to center the prior distributions at and thus shrink toward a parametric covariance function C, as briefly described in Section 2.9. This modification directly and indirectly addresses a large number of comments by several of the discussants, and so we elaborate on it here. The idea is to set the prior mean of the $\mathbf{u}_i$ and d_i to the values $\mathbf{u}_i^{(m)}$ and $d_i^{(m)}$, respectively, implied by a Vecchia approximation of C with $m \ll n$. Specifically, we assume $\mathbf{u}_i | d_i \sim \mathcal{N}(\mathbf{u}_i^{(m)}, d_i \mathbf{V}_i)$ and $d_i \sim \mathcal{IG}(\alpha_i, \beta_i)$ with $E(d_i) = \beta_i/(\alpha_i - 1) = d_i^{(m)}$. The prior standard deviations, which determine the degree of shrinkage toward C, can be set to be roughly fractions c_u and c_d of the respective prior means, such that $\mathbf{V}_i = \text{diag}(v_{i1}, \ldots, v_{im})$ with $v_{ij} = (c_u \mathbf{u}_{i,j}^{(m)})^2 / E(d_i)$ and $sd(d_i) = \beta_i/((\alpha_i - 1)(\alpha_i - 2)^{1/2}) = c_d E(d_i)$.

This modification of our methodology is especially useful for small numbers N of training replicates, and it can even be used when $N = 1$, a situation of interest to several discussants. We implemented this parametric-shrinkage modification based on a fixed Matérn covariance C, with fixed $c_u = 1/2$ and $c_d = 1$. Figure R1 shows a comparison in this setting, illustrating that both our original method and the parametric-shrinkage modification can outperform a Gaussian process (GP) with a slightly misspecified parametric covariance. We also added the parametric-shrinkage approach to the climate-data comparison in Figure 9; as shown in Figure R2, the modification was more accurate than our original approach for small N (roughly, $N < 20$), and less accurate for large N.

The parametric-shrinkage modification provides a solution to Banerjee and Peruzzi's concern that our approach is not parameterized according to a parent covariance. Further, Azizi was concerned about erratic behavior of the d_i's and $\mathbf{u}_i$'s when sampling from our original prior; this is alleviated using the parametric-shrinkage modification, especially for small c_u and c_d.

Our modification could be further improved and extended. Most importantly, instead of fixing C, we can consider a covariance family $C_{\boldsymbol{\theta}}$ that depends on an unknown

*MK's research was partially supported by National Science Foundation (NSF) Grants DMS–1654083, DMS–1953005, and CCF–1934904.

[†]Department of Statistics, Texas A&M University

[‡]Department of Statistics, Texas A&M University, katzfuss@gmail.com

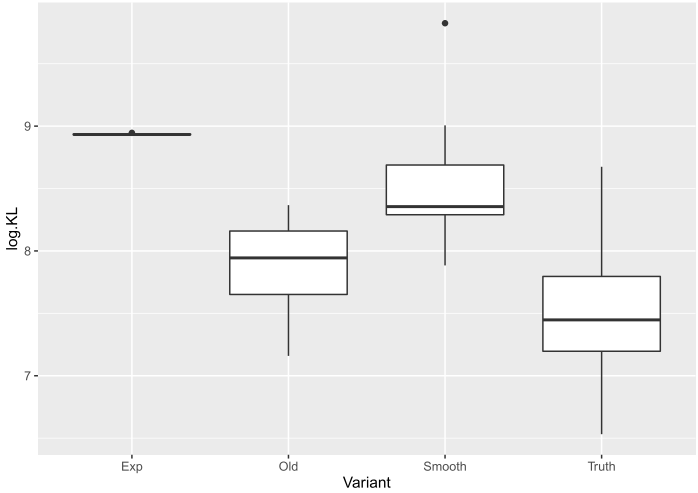


Figure R1: Comparison of Kullback-Leibler (KL) divergences for different priors based on a single replicate ($N = 1$) in the setting of Figure 7, which is based on simulating data from a Matérn with smoothness 1 and range 0.25 at 900 locations. The boxplots show results over 20 repetitions of the simulation. Exp: GP with exponential covariance. Old: Our original method from the main paper. Smooth and Truth are variants of the parametric-shrinkage modification in Section R1 with $m = 10$ neighbors, shrinking toward Matérn covariances with range 0.25 and smoothness 1.5 and 1, respectively. Old often performed better than shrinking toward an incorrect covariance model (Smooth). Shrinking toward the true covariance performed best.

parameter vector $\boldsymbol{\theta}$, which can be inferred using the integrated likelihood as described in Section 2.4. It will also likely be useful to add c_u and c_d to $\boldsymbol{\theta}$, to let the data decide how strongly the covariance should be shrunk toward C. These extensions may substantially improve the results of the parametric-shrinkage methods in Figures R1–R2. Further, it would even be possible to let $\boldsymbol{\theta}$ (and hence the implied covariance) vary as a function of covariates (e.g., spatial location).

R2 Prediction at unobserved locations

Several discussants asked about using or extending our methodology to make predictions at unobserved locations. We want to emphasize that our methodology was intended for data on a (regular or irregular) grid, where data may be missing at each grid location for some but not all replicates (see Section 5). However, it is in principle possible to obtain predictions at entirely unobserved locations using our method, although these predictions will be most useful under the parametric-shrinkage modification in Section R1.

Assume we would like to predict $y_{n+1}^{(\ell)}$ at location $\mathbf{s}_{n+1}$ based on $\mathbf{Y}$. The most straightforward way to do this is to order $\mathbf{s}_{n+1}$ (and hence $y_{n+1}^{(\ell)}$) after the n observed

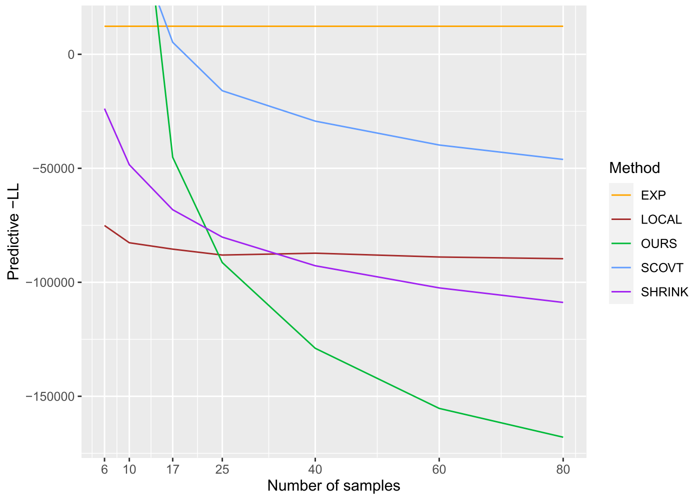


Figure R2: The climate-data comparison of Figure 9 with an added method SHRINK that shrinks toward the fitted exponential (EXP) with $m = 10$ neighbors (see Section R1). SHRINK performed better than OURS for small N, and provided similar results as EXP for $N = 1$.

locations. In that case, we can simply train the model as before based on $\mathbf{Y}$, and then consider prediction as a separate step. Specifically, we consider here the empirical Bayes approach of Section 2.4, in which we obtain a hyperparameter estimate $\hat{\boldsymbol{\theta}}$ based on the training data. Then, the predictive distribution is

$$p(y^{(\ell)}_{n+1}|\mathbf{Y}) = \int \mathcal{N}(y^{(\ell)}_{n+1}| - \mathbf{y}^{(\ell)}_{g_m(n+1)}{}'\mathbf{u}_{n+1}, d_{n+1})p(\mathbf{u}_{n+1}|d_{n+1}, \hat{\boldsymbol{\theta}})p(d_{n+1}|\hat{\boldsymbol{\theta}})d\mathbf{u}_{n+1}dd_{n+1}$$
$$= t_{2\alpha_{n+1}}\big(y^{(\ell)}_{n+1}| - \mathbf{y}^{(\ell)}_{g_m(n+1)}{}'\boldsymbol{\mu}_{n+1}, \tfrac{\beta_{n+1}}{\alpha_{n+1}}(1 + \mathbf{y}^{(\ell)}_{g_m(n+1)}{}'\mathbf{V}_{n+1}\mathbf{y}^{(\ell)}_{g_m(n+1)})\big),$$

where $\mathbf{u}_{n+1}$ and d_{n+1} (and m) are informed by the training data $\mathbf{Y}$ only through $\hat{\boldsymbol{\theta}}$ (i.e., we simply consider their prior distribution for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$), and so $\boldsymbol{\mu}_{n+1}$ is the prior mean of $\mathbf{u}_{n+1}$. In our original model, we had $\boldsymbol{\mu}_{n+1} = \mathbf{0}$, meaning that the prediction is centered at zero, no matter how close $\mathbf{s}_{n+1}$ is to an observed location. Much more useful predictions can be obtained using the parametric-shrinkage modification in Section R1, with $\boldsymbol{\mu}_{n+1} = \mathbf{u}^{(m)}_{n+1}$ as implied by a Vecchia approximation of a parametric covariance $C_{\hat{\boldsymbol{\theta}}}$. If the prior variability converges to zero (i.e., $c_u, c_d \to 0$), then $y^{(\ell)}_{n+1}|\mathbf{Y} \sim \mathcal{N}(-\mathbf{y}^{(\ell)}_{g_m(n+1)}{}'\mathbf{u}^{(m)}_i, d^{(m)}_i)$, which is equivalent to a (parametric) Vecchia GP prediction based on $C_{\hat{\boldsymbol{\theta}}}$. For $c_u, c_d > 0$, the predictive distribution is centered at the same value but has larger variance and heavier tails, due to the uncertainty in $\mathbf{u}_{n+1}$ and d_{n+1}. This is illustrated numerically in Figure R3.

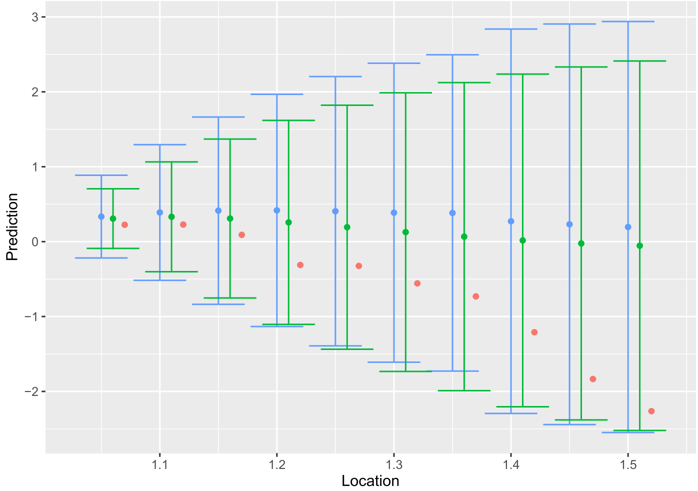


Figure R3: Predictive distributions (means and 80% intervals) using our method from Sections R1–R2 with $m = 9$ neighbors (blue) versus standard GP prediction (green) using the true covariance (Matérn with variance 5, smoothness 1.5, and range 0.25), along with the true values (red). Training data consisted of $N = 1$ replicate of a GP with the true covariance at $n = 250$ randomly sampled locations on the unit square $[0, 1]^2$; predictions are made at coordinates $(x, 0.5)$ for various locations x on the x-axis of the plot, with predictions offset slightly for better visibility.

R3 Gibbs sampler for noisy data

In Section 2.8, we considered noisy observations $\mathbf{w}^{(\ell)}|\mathbf{y}^{(\ell)} \overset{iid}{\sim} \mathcal{N}_n(\mathbf{y}^{(\ell)}, \tau^2\mathbf{I}_n)$, $\ell = 1, \ldots, N$, with the latent fields $\mathbf{y}^{(\ell)}$ modeled using our Bayesian nonstationary approach. We proposed a Gibbs sampler, which requires sampling $\mathbf{y}^{(\ell)}$ conditional on $\mathbf{w}^{(\ell)}$ and $\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}'$. When n is at most moderately large (say, $n < 10^5$), this can be done exactly via Cholesky factorization of the sparse posterior precision matrix $\mathbf{A} := \mathbf{\Sigma}^{-1} + \tau^{-2}\mathbf{I}_n$ after applying a fill-reducing ordering (e.g., approximate minimum degree).

For very large n, however, the fill-in and computational cost may become too high, and so we suggested approximating the Cholesky factor of $\mathbf{A}$ using an incomplete Cholesky factorization as described in Schäfer et al. (2021a, Sect. 4.1). We agree with Banerjee and Peruzzi, who point out that this approximation breaks the coherence of the Gibbs sampler. Our hope is that the resulting sampler may still produce useful results, because the incomplete-Cholesky error is often small.

Alternatively, it is possible to carry out exact sampling of $\mathbf{y}^{(\ell)}$ conditional on $\mathbf{w}^{(\ell)}$ and $\mathbf{\Sigma}$ without substantially increasing the computational complexity. If we sample $\mathbf{z}^{(\ell)} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ and $\widetilde{\mathbf{w}}^{(\ell)} \sim \mathcal{N}_n(\mathbf{w}^{(\ell)}, \tau^2\mathbf{I}_n)$, then

$$\mathbf{y}^{(\ell)} = \mathbf{A}^{-1}(\mathbf{U}\mathbf{D}^{-1/2}\mathbf{z}^{(\ell)} + \tau^{-2}\widetilde{\mathbf{w}}^{(\ell)}) \tag{R1}$$

is a sample from the desired distribution, as has been exploited, for example, in the ensemble Kalman filter (e.g., Hunt et al., 2007; Boyles and Katzfuss, 2021). In (R1), we need to solve a linear system in $\mathbf{A}$, which can be done to any desired accuracy via the conjugate gradient algorithm, using its incomplete Cholesky factor as a preconditioner (Schäfer et al., 2021a, Sect. 4.1 and Fig. 9).

R4 Extensions for spatio-temporal fields

As pointed out by several discussants, extending our method to spatio-temporal fields is important for many environmental applications.

Ordering and nearest-neighbor selection for space-time coordinates could be obtained using the correlation distance in Section 2.7. In the simplest case, this would be equivalent to a scaled space-time distance, which would require estimating (either as a pre-processing step or as part of $\boldsymbol{\theta}$) the ratio of the temporal correlation range relative to the spatial range.

Ensembles (i.e., $N > 1$ replicates) of spatio-temporal fields are available in some applications, such as climate models. In this setting, our methods can be applied directly. Shrinkage toward parametric (see Section R1) spatio-temporal covariance functions, including separable ones (as mentioned by Li and Shand), is also straightforward.

However, in many applications only $N = 1$ spatio-temporal field is available, as pointed out by Li and Shand. Our method could be extended to this setting under more restrictive assumptions. For example, we could order the spatio-temporal coordinates first by time and then by maximin ordering in space within a given time point. Inference would then be possible under the assumption that the process value at a particular spatial location depends on its spatial neighbors at the current and previous time point(s) in the same way at each time point, in which case time can act as a pseudo-replicate.

R5 Further applications

We were excited to see that several of the discussants shared ideas for interesting applications or extensions of our methodology. Li and Shand suggested combining our method with copula techniques to model dependence in spatial extremes. Azizi et al. made a case for using our method for functional data, with some potential extensions. And Sottosanti et al. suggested applying our method to spatial transcriptomics; we believe that our method would be able to handle irregularities on the edges and within the surface of tissues relatively well given sufficiently large numbers of training replicates. We are hoping that our approach will be used successfully in these and other applications.

R6 Brief responses to miscellaneous comments

- Li and Shand:
  - Letting m_i depend on i is an interesting idea, which is explored in Figure R4. However, we observed little indication that m_i should increase with i — on

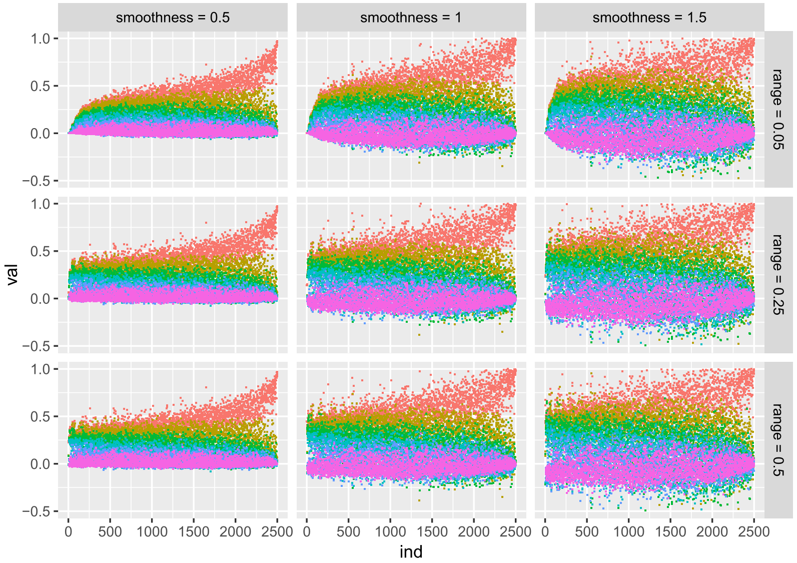
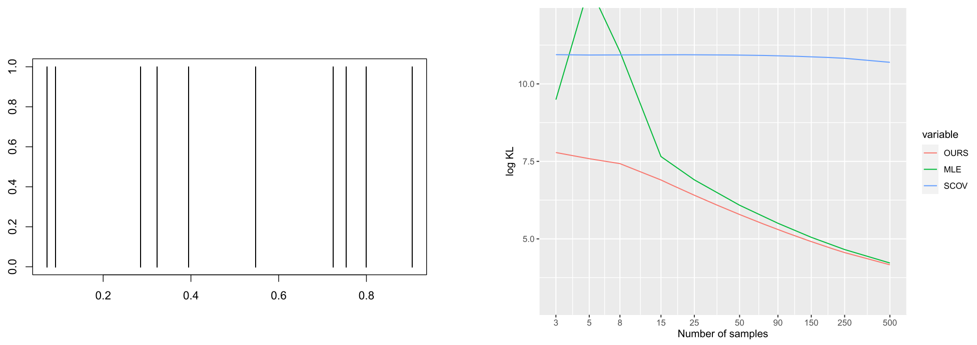


Figure R4: Illustration of $(\mathbf{u}_i^{(i-1)})_j$ as a function of maximin index i (as in Figure 4) for several values of range and smoothness in a Matérn covariance, with points colored by neighbor number (larger values correspond to closer neighbors).

the contrary, the importance of the first nearest neighbor (NN) relative to, say, the fifth NN increased with i, and so m_i could potentially decrease with i.

- Regarding relating θ_3 to the smoothness of random fields, some empirical results can be found in Figure 4. When centering the prior on a Matérn covariance as in Section R1, the smoothness of this covariance can directly be regarded as one of the hyperparameters of the resulting model, making the link more explicit.
- The multi-resolution approximation (Katzfuss, 2017; Katzfuss and Gong, 2020) can be viewed as a variant of the Vecchia approximation (Katzfuss and Guinness, 2021), and so it could be extended nonparametrically in a similar way to our model here. The main difficulty would be to come up with suitable prior distributions, although shrinkage to a parametric covariance (Section R1) is one possibility.

- Banerjee and Peruzzi:
  - Our model indeed only infers the covariance structure nonparametrically, under the strong parametric assumption of a Gaussian joint distribution. Katzfuss and Schäfer (2021) propose an extension of our approach to non-Gaussian distributions using Bayesian transport maps, including a further nonparametric extension based on Dirichlet process mixtures for flexible marginal distributions.



(a) "Satellite" sampling pattern

(b) log KL divergence

Figure R5: Comparison of KL divergence in the setting of Figure 8, except for simulated data at highly irregular locations on ten strips of 250 points each, mimicking satellite orbits that are a common source of spatial data. The results are similar to those for gridded or randomly sampled locations in Figure 8.

- Sottosanti et al.:
  - Our method does not necessarily require the observation locations to be on a regular grid. We also obtained good results for uniformly sampled locations (Figure 8). Figure R5 shows that similarly good results can be obtained for extremely irregular locations mimicking the sampling patterns of polar-orbiting satellites.
- Azizi et al.:
  - We appreciate pointing out the typo in our GitHub tutorial. It has been corrected.
  - It is possible to account for correlation between replicates. The expression $p(\mathbf{Y}|\mathbf{\Sigma}) = \prod_{i=1}^{n} \mathcal{N}_N(\mathbf{y}_i|\mathbf{X}_i\mathbf{u}_i, d_i\mathbf{I}_N)$ in (5) could be generalized to

    $$p(\mathbf{Y}|\mathbf{\Sigma}) = \prod_{i=1}^{n} \mathcal{N}_N(\mathbf{y}_i|\mathbf{X}_i\mathbf{u}_i, d_i\mathbf{C}_i)$$

    for some nondiagonal correlation matrix $\mathbf{C}_i$, which may not depend on i and may depend on parameters that can be included in $\boldsymbol{\theta}$. We conjecture that even without this extension, our model may produce useful results based on correlated replicates, although the resulting posterior uncertainty may then be an underestimate of the true uncertainty.
  - When using Markov chain Monte Carlo (MCMC), m may change at every iteration based on the sampled θ_3. However, the resulting MCMC draws of $\mathbf{u}_i$ can always be viewed (at least conceptually) as vectors of length $i-1$ with all but m entries equal to zero, and so combining MCMC draws is straightforward.

- Marques et al.:
  - We agree that the stochastic partial differential equation (SPDE) approach is very useful in many applications and provides spatially coherent interpretability, which our approach does not. To our knowledge, the form or structure of the nonstationarity in SPDE approaches is typically specified manually as a function of hyperparameters.
  - Shrinkage toward a parametric covariance was not used for any numerical results in the main paper.
  - We agree that it would be interesting to investigate whether our method could be combined with the principled approach of centering priors around a base model proposed in Simpson et al. (2017). However, this appears difficult to us, especially when trying to avoid increasing the computational burden.
- Guha:
  - An exact maximin ordering of n spatial locations can indeed be computed in quasilinear time in n using the algorithms provided in Schäfer et al. (2021b,a).
  - It would certainly be possible to include a few far-away locations in the neighbor sets, as originally suggested by Stein et al. (2004) for parametric Vecchia approximations. However, it is not obvious how these locations should be selected and what the corresponding priors should be. Guinness (2018) did not observe an improvement due to including far-away points (as opposed to using only nearest neighbors) for parametric Vecchia approximations under maximin ordering.
- Pérez Ruiz and Leonard:
  - While our literature review was focused on the existing approaches most closely related to our specific methodology based on modified Cholesky decomposition (MCD), we agree that there is a vast literature on covariance estimation. Instead of MCD, positive-definiteness constraints can also be avoided by estimating the matrix log of the covariance (e.g., Leonard and Hsu, 1992; Chiu et al., 1996; Hsu et al., 2012), but this may only be feasible for large n (and small N) under additional sparsity assumptions (e.g., Deng and Tsui, 2013). If we consider extending our approach to a spatial autoregressive (SAR) model, there seem to be potential insights and overlap with the matrix exponential spatial specification (LeSage and Pace, 2007), which fixes the sparsity pattern of $\mathbf{U}$ based on nearest neighbors as in our approach, but it also fixes the values up to a scaling constant, making it less flexible. However, their use of a matrix exponential to approximate $\mathbf{U}^{-1}$ has major computational advantages for SAR models. Similarly, Mukherjee et al. (2011) use a Bayesian SAR model with the nonzeros in $\mathbf{U}$ being a function of distance, with further spatially varying extensions in Mukherjee et al. (2014).
- Peluso:
  - Reformulating our prior as a directed acyclic graph (DAG)–Wishart prior provides insight into the hyperparameter choices of α_i, β_i and connects the

idea to a diagonal scale matrix parameter of the Wishart. (Rather than α_i being fixed, at the start of the ordering it could decrease by 0.5 per neighbor until the number of neighbors is constant.) Then, as shown in Peluso and Consonni (2020), the prior is compatible under DAGs, so our method could be considered jointly with graph estimation (i.e., not a-priori fixing the graph and ordering), though this may not be computationally feasible for large n.

References

Boyles, W. and Katzfuss, M. (2021). "Ensemble Kalman filter updates based on regularized sparse inverse Cholesky factors." *Monthly Weather Review*, 149(7): 2231–2238. doi: https://doi.org/10.1175/MWR-D-20-0299.1. 346

Chiu, T. Y., Leonard, T., and Tsui, K.-W. (1996). "The matrix-logarithmic covariance model." *Journal of the American Statistical Association*, 91(433): 198–210. MR1394074. doi: https://doi.org/10.2307/2291396. 349

Deng, X. and Tsui, K.-W. (2013). "Penalized covariance matrix estimation using a matrix-logarithm transformation." *Journal of Computational and Graphical Statistics*, 22(2): 494–512. MR3173726. doi: https://doi.org/10.1080/10618600.2012.715556. 349

Guinness, J. (2018). "Permutation and grouping methods for sharpening Gaussian process approximations." *Technometrics*, 60(4): 415–429. MR3878098. doi: https://doi.org/10.1080/00401706.2018.1437476. 349

Hsu, C.-W., Sinay, M. S., and Hsu, J. S. (2012). "Bayesian estimation of a covariance matrix with flexible prior specification." *Annals of the Institute of Statistical Mathematics*, 64(2): 319–342. MR2878908. doi: https://doi.org/10.1007/s10463-010-0314-5. 349

Hunt, B. R., Kostelich, E. J., and Szunyogh, I. (2007). "Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter." *Physica D: Nonlinear Phenomena*, 230(1-2): 112–126. MR2345207. doi: https://doi.org/10.1016/j.physd.2006.11.008. 346

Katzfuss, M. (2017). "A multi-resolution approximation for massive spatial datasets." *Journal of the American Statistical Association*, 112(517): 201–214. MR3646566. doi: https://doi.org/10.1080/01621459.2015.1123632. 347

Katzfuss, M. and Gong, W. (2020). "A class of multi-resolution approximations for large spatial datasets." *Statistica Sinica*, 30(4): 2203–2226. 347

Katzfuss, M. and Guinness, J. (2021). "A general framework for Vecchia approximations of Gaussian processes." *Statistical Science*, 36(1): 124–141. MR4194207. doi: https://doi.org/10.1214/19-STS755. 347

Katzfuss, M. and Schäfer, F. (2021). "Scalable Bayesian transport maps for high-dimensional non-Gaussian spatial fields." arXiv:2108.04211. 347

Kidd, B. and Katzfuss, M. (2022). "Bayesian nonstationary and nonparametric covariance estimation for large spatial data." *Bayesian Analysis*, accepted. doi: https://doi.org/10.1214/21-BA1273. 342

Leonard, T. and Hsu, J. S. (1992). "Bayesian inference for a covariance matrix." *The Annals of Statistics*, 20(4): 1669–1696. MR1193308. doi: https://doi.org/10.1214/aos/1176348885. 349

LeSage, J. P. and Pace, R. K. (2007). "A matrix exponential spatial specification." *Journal of Econometrics*, 140(1): 190–214. MR2395921. doi: https://doi.org/10.1016/j.jeconom.2006.09.007. 349

Mukherjee, C., Kasibhatla, P., and West, M. (2011). "Bayesian statistical modeling of spatially correlated error structure in atmospheric tracer inverse analysis." *Atmospheric Chemistry and Physics*, 11(11): 5365–5382. 349

Mukherjee, C., Kasibhatla, P., and West, M. (2014). "Spatially-varying SAR models and Bayesian inference for high-resolution lattice data." *Annals of the Institute of Statistical Mathematics*, 66: 473–494. MR3211871. doi: https://doi.org/10.1007/s10463-013-0426-9. 349

Peluso, S. and Consonni, G. (2020). "Compatible priors for model selection of high-dimensional Gaussian DAGs." *Electronic Journal of Statistics*, 14(2): 4110–4132. MR4170698. doi: https://doi.org/10.1214/20-EJS1768. 350

Schäfer, F., Katzfuss, M., and Owhadi, H. (2021a). "Sparse Cholesky factorization by Kullback-Leibler minimization." *SIAM Journal on Scientific Computing*, 43(3): A2019–A2046. MR4267493. doi: https://doi.org/10.1137/20M1336254. 345, 346, 349

Schäfer, F., Sullivan, T. J., and Owhadi, H. (2021b). "Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity." *Multiscale Modeling & Simulation*, 19(2): 688–730. MR4243658. doi: https://doi.org/10.1137/19M129526X. 349

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32(1): 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 349

Stein, M. L., Chi, Z., and Welty, L. (2004). "Approximating likelihoods for large spatial data sets." *Journal of the Royal Statistical Society: Series B*, 66(2): 275–296. MR2062376. doi: https://doi.org/10.1046/j.1369-7412.2003.05512.x. 349