Electronic Journal of Statistics
Vol. 10 (2016) 2856–2893
ISSN: 1935-7524
DOI: 10.1214/16-EJS1125

REML estimation with intrinsic Matérn dependence in the spatial linear mixed model

Somak Dutta

Iowa State University, Ames, USA
e-mail: somakd@iastate.edu

and

Debashis Mondal

Oregon State University, Corvallis, USA
e-mail: debashis@stat.oregonstate.edu

Abstract: We present a new matrix-free residual maximum likelihood (REML) analysis for irregularly spaced spatial data, where observations usually represent average values over very small regions that are interpreted as points. The REML analysis is obtained after embedding the sampling locations in a fine scale rectangular lattice, treating unobserved sites as missing data. The spatial random fields considered here are based on fractional Laplacian differencing on the lattice and they are unique in approximating continuum intrinsic Matérn dependence. Here, using the h-likelihood method, we derive REML estimating equations that allow for singular precision matrices, estimation of covariate effects, prediction of unobserved spatial effects and REML estimation of precision parameters as a solution to an explicit gamma non-linear model. Furthermore, we devise a sophisticated computational algorithm that enables us to achieve scalable matrix-free statistical computations. In particular, these matrix-free computations include the use of (1) the two-dimensional discrete cosine transformation that arises in the spectral decomposition of the precision matrix of our spatial random fields and that allows fast matrix-free matrix-vector multiplication, (2) a matrix-free pre-conditioned Lanczos algorithm that solves non-sparse matrix equations with linear constraints, (3) a matrix-free Hutchinson's trace estimator that stochastically approximates the trace of a matrix, (4) a robust trust region method that always finds a local maximum of the non-concave residual log-likelihood function and (5) some preliminary computations of the log REML likelihood function based on Taylor series approximation. Using computer experiments, we provide further understanding on not just the number and values but also the basins of attraction of the local and global maxima of the REML function. This understanding significantly simplifies the problem of finding global maxima. We further demonstrate through computer experiments that our matrix-free REML estimators attain both efficiency and geostatistical inference, and surpass the widely used INLA methods in computational times. We provide an extensive application on mapping ground water arsenic concentration in Bangladesh, indicating numeric consistency of results and robustness of inference to changes of lattice spacing. The paper closes with some discussions that include computations in the stationary case, conditional simulations and matrix-free MCMC computations.

Keywords and phrases: Arsenic contamination, discrete cosine transform, fractional differencing, Hutchinson's trace approximations, incomplete Cholesky, Lanczos algorithm, long range dependence, power variogram, matrix-free computations, trust region method.

Received April 2015.

1. Introduction

In recent years, interest in connecting lattice-based Gaussian random fields with geostatistical models has increased significantly, as researchers begin to explore the extent to which lattice-based models offer an adequate fit or explanation to continuum spatial variations we see in real data. At the forefront of these discussions are the works of McCullagh (2002), Besag (2002), Besag and Mondal (2005), McCullagh and Clifford (2006), Lindgren et al. (2011), Dutta and Mondal (2015a), Mondal (2013), and many others. In particular, the work of McCullagh (2002) lays a theoretical foundation for the use of the de Wijs process or the logarithmic covariance model for spatial analysis. Besag and Mondal (2005) derive the connection between first-order intrinsic autoregression and the de Wijs process in which the latter arises as the scaling limit of the former. Using stochastic partial differential equation representations, Lindgren et al. (2011) (see also Gay and Heyde, 1990; Kelbert et al., 2005) provide useful Gaussian Markov random field approximations to a subclass of Matérn covariance models, and present Galerkin methods for statistical computations. Dutta and Mondal (2015a), on the other hand, derive fast matrix-free computations for spatial mixed models based on Gaussian Markov random fields with nugget effects, enhancing the computations developed in Lindgren et al. (2011). Encouraged by these developments, we undertake here the novel possibility of using the first-order Gaussian intrinsic autoregression on the regular lattice as a building block for constructing fractionally differenced random fields. The latter, as we shall see, is unique in approximating continuum intrinsic Matérn random fields that have an important place in the geostatistical literature. It is these lattice-based fractional random fields that give a fresh perspective to the conceptual framework of spatial Matérn dependence analysis, whose full potential has not been realized. We also develop novel, scalable, easy-to-implement and yet sophisticated matrix-free computations that allow us to argue for their wider relevance, and help us understand many of the strengths and weaknesses of intrinsic Matérn dependence models.

Thus, our purpose in studying Matérn models is distinct from classical works of Mardia and Marshall (1984), Handcock and Stein (1993), Stein (1999), Diggle et al. (2003), Guttorp and Gneiting (2006), Minasny and McBratney (2007), Pardo-Igúzquiza et al. (2009), Anitescu et al. (2012) and many others on conventional fitting of stationary Matérn covariances to spatial data. Furthermore, the purpose here is also distinct from studying Whittle's spectral approximations methods (see e.g., Guyon (1982), Dahlhaus and Künsch (1987), Kent and Mardia (1996), Fuentes (2007) and many others) for estimating parameters of

stationary covariances. Necessarily, we avoid direct modeling via covariances or spectral densities, and instead focus on the structure of the precision matrix (i.e., the canonical parameter) that arises through full conditional specifications and scaling limit connections. Furthermore, as geostatstics is largely about differencing and intrinsic random functions (Matheron, 1971, 1973; see also the review of Beran (1992) on long range dependence and the discussion of Diggle et al., 2010), we restrict our attention mostly to fractional differencing and approximations of intrinsic Matérn models. Corresponding methods for the stationary case, if needed, can be obtained easily and will be discussed later in the paper.

In geostatistical applications, observations usually represent average values over very small regions that are interpreted as points. Thus, for brevity of discussion, we assume that we can embed the locations in a fine scale rectangular lattice, treating unobserved sites as missing data. As we shall see in Sections 5 and 6, there is little loss in discretizing the space in the above way and in embedding irregularly spaced locations to a fine regular grid. We can then focus on estimation of parameters in an overall mixed effect formulation, which might include covariate information as well as a white noise component. In other words, we consider a mixed linear model of the form

$$\mathbf{y} = \mathbf{T}\boldsymbol{\tau} + \mathbf{F}\boldsymbol{\psi} + \boldsymbol{\epsilon}. \tag{1}$$

In the above $\mathbf{y}$ is the vector of observations around n sampling locations, $\mathbf{T}$ is an $n \times m$ matrix of some covariate values with $\boldsymbol{\tau}$ as the coefficients, $\boldsymbol{\psi}$ is the vector of latent spatial process coming from a fine $r \times c$ regular array on which the sampling locations are embedded, $\mathbf{F}$ is a known incidence matrix indicating whether an observation belongs to a particular array cell so that $\mathbf{F}\boldsymbol{\psi}$ gives back the vector of latent spatial variable values for the observed $\mathbf{y}$ and $\boldsymbol{\epsilon}$ indicates the vector of residual fluctuations that are left unexplained by the regularly varying spatial process or the covariate values. We assume that the covariate values are so adjusted that $\mathbf{T1} = \mathbf{F1} = \mathbf{1}$ or $\mathbf{F1} = \mathbf{1}$ where $\mathbf{1}$ is the vector of all ones of appropriate dimension and $\mathbf{1}$ belongs to the column space of $\mathbf{T}$. We interpret y_i, the ith entry of the $\mathbf{y}$ vector, as the average value on the array cell on which the corresponding sampling location falls. Similarly, we interpret $\psi_{u,v}$, the (u, v)th entry of $\boldsymbol{\psi}$, as the average value of the latent spatial random field on the (u, v)th array cell. Furthermore, we embrace the residual fluctuations as independent and identically distributed Gaussian random variables with an unknown precision parameter λ_y, and the random spatial effect $\boldsymbol{\psi}$ as an intrinsic Matérn Gaussian random field on an $r \times c$ array with singular precision matrix $\mathbf{R}$. In other words,

$$\boldsymbol{\epsilon} \sim (\lambda_y/(2\pi))^{\frac{1}{2}n} \exp\Big(-\tfrac{1}{2}\lambda_y \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\Big), \qquad \boldsymbol{\psi} \sim |\mathbf{R}/(2\pi)|_+^{\frac{1}{2}} \exp\Big(-\tfrac{1}{2}\boldsymbol{\psi}^T\mathbf{R}\boldsymbol{\psi}\Big).$$

where $|\mathbf{R}|_+$ denotes the product of nonzero eigenvalues of $\mathbf{R}$.

Thus, what assumes importance is how we construct $\mathbf{R}$ and how we pursue subsequent statistical computations in order to draw meaningful statistical inference. In the case of first-order Gaussian intrinsic autoregression on the two

dimensional integer lattice, which is central to the Markov random field approach of spatial statistics, $\boldsymbol{\psi}$ is defined via the following conditional mean and variance formula

$$\mathrm{E}\,(\psi_{i,j} \mid ...) = \tfrac{1}{4}(\psi_{i,j-1} + \psi_{i,j+1} + \psi_{i-1,j} + \psi_{i+1,j}), \qquad \mathrm{var}\,(\psi_{i,j} \mid ...) = 1,$$

where the conditioning variables (namely ...) are all $\psi_{u,v}$ such that $(u, v) \neq (i, j)$. When restricted on a finite $r \times c$ array, this intrinsic autoregression actually follows

$$\boldsymbol{\psi} \sim |\mathbf{W}/(2\pi)|_{+}^{\frac{1}{2}} \ \exp\Big(-\tfrac{1}{2}\boldsymbol{\psi}^T\mathbf{W}\boldsymbol{\psi}\Big),$$

where the singular precision matrix $\mathbf{W}$ of $\boldsymbol{\psi}$ is the discrete (graph) Laplacian on the grid with $\mathrm{rank}(\mathbf{W}) = q - 1$ and can be written in terms of a quadratic form as

$$\boldsymbol{\psi}^T\mathbf{W}\boldsymbol{\psi} = \tfrac{1}{4}\sum\sum(\psi_{i,j} - \psi_{i-1,j})^2 + \tfrac{1}{4}\sum\sum(\psi_{i,j} - \psi_{i,j-1})^2.$$

In what follows, we shall draw upon the above construction of $\mathbf{W}$ to derive an explicit representation to $\mathbf{R}$. Specifically, we shall use

$$\mathbf{R} = \lambda_\psi \mathbf{W}^\nu, \qquad \lambda_\psi > 0, \quad \nu \geq 0, \tag{2}$$

with a precision parameter λ_ψ and a (long range) dependence parameter ν. The above may look rather simple, but as we shall see, it allows us to describe fractionally differenced random fields on regular arrays and, as lattice spacing diminishes, it allows us to approximate intrinsic Matérn models. Furthermore, it is the use of (2) that allows us to develop fast matrix-free REML computations for the mixed model in (1) and, in the process, allow us to gain new understanding on the statistical estimation and inference of intrinsic Matérn models.

It is worthwhile to point out that there is little literature on fitting exact intrinsic Matérn mixed linear model besides McCullagh and Clifford (2006). These exact REML estimation of McCullagh and Clifford (2006) require $O(n^3)$ computations and $O(n^2)$ storage when data are observed at n spatial locations. It is also worthwhile to point out that in Lindgren et al. (2011) the focus was on the values $\nu = 1, 2, \ldots$, which correspond to a subclass of conditional autoregressions with sparse precision matrices and which allow them to use certain sparse matrix computations that are of $O(n^{3/2})$. However, the sparse approximations and the fast computations of Lindgren et al. (2011) are not extensible for any arbitrary positive values of ν (e.g., for $\nu = 5/3$ that arises in the study of turbulence). In comparison, the effective order of matrix-free computations developed by Dutta and Mondal (2015a) for spatial mixed linear models based on a sparse precision matrix (i.e., $\nu = 1, 2, \ldots$) and nugget effects is just $O(n \log n)$ with $O(n)$ storage. In this paper, we not only extend the estimation to arbitrary $\nu > 0$, but also the computations that we develop here for any $\nu > 0$ effectively require only $O(n(\log n)^2)$ computations and $O(n)$ storage. Other than non-sparsity, the statistical problem considered here has many challenges. As,

for example, the intrinsic Matérn dependence leads to not just a non-sparse singular precision matrix but also an ill-conditioned one and finding an useful preconditioning method is a challenge. Similarly, non-integer values of ν give rise to non-convexity in the log REML function which poses an important optimization challenge. We offer here new and interesting solutions to these challenges without resorting to dimension reduction, tapering, or procedures such as block averaging. We also demonstrate that there is little loss of statistical efficiency (in terms of achieving Cramer-Rao lower bound) due to our matrix-free scalable computations and we can obtain answers that are as good as exact answers.

The rest of the paper is laid out as follows. In Section 2, we provide interpretation of (2) based on the fractional Laplacian difference equations and derive some justifications for its validity using both the frequency domain and the spatial domain results. Specifically, we show that 1) as the lattice spacing diminishes the eigenvalues of the singular precision matrix from (2) trace out the eigenvalues of an intrinsic Matérn process, and 2) the differences between the variograms of the fractionally Laplacian differenced process at a moderately fine lattice and the continuum Matérn process are just a small constant. Thus, at a moderately fine lattice scale, our approximations work remarkably well by the routine use of a random error term along with the spatial component. In Sections 3 and 4, we develop REML estimation procedure of the mixed model in (1). Specifically, we derive of matrix-free statistical computations of best linear unbiased estimators (BLUEs) of the contrasts of fixed effect $\boldsymbol{\tau}$, best linear unbiased predictors (BLUPs) of the contrasts of random effect $\boldsymbol{\psi}$, and REML estimation of $\boldsymbol{\theta} = (\lambda_y, \lambda_\psi, \nu)^T$. We also provide explicit characterization of the non-convex nature of REML estimation. Our computational steps consist of (1) matrix vector multiplications using the discrete cosine transformations (2) linear equations solving with a non-sparse Lanczos algorithm, (3) derivation of effective matrix-free preconditioner using an incomplete Cholesky factor decomposition, (4) calculating traces of matrices using Hutchinson's trace estimators and (5) optimization using a robust matrix-free trust region method. We also include some preliminary computations of the log REML likelihood function based on Taylor series approximation. In Section 5, we present through simulations the performance of the REML estimators. These simulation runs suggest that despite non-convex nature of REML optimization, the trust region method can pick the global estimates, at least when $n = O(rc)$ and n is large. We also run simulations to demonstrate that there is little loss in embedding irregularly spaced locations to a fine regular grid. In particular, we show numeric consistency of results and highlight robustness of inference to changes of lattice spacing. We further demonstrate that our matrix-free REML estimators attain efficiency properties that are very close to those of the exact REML estimators. In terms of practical gain, we also report actual computational times that are better than those from Lindgren et al. (2011). In Section 6, we provide an extensive application to mapping ground water arsenic contamination in Bangladesh and again highlight robustness of statistical inference to the changes of scales. Finally, in Section 7, we close the paper with some discussions on computations of the stationary case, conditional simulations, and on potential challenges ahead

in accommodating spatial anisotropy and heterogeneity within the framework developed here.

2. Fractionally differenced random fields and intrinsic Matérn processes

The objective of this section is to provide some details of the fractional Laplacian differenced random fields on finer and finer regular arrays as approximations to continuum intrinsic Matérn random fields. Here, we shall consider some frequency domain results, exact variogram calculations and precise algebraic eigenvalue expressions that lead to (2). We also discuss certain meaning and significance to its use.

First, as discussed by Mondal (2011) with regard to the paper of Lindgren et al. (2011), we consider a sequence of Gaussian random fields $\{Z^{(m)}(u,v)\}$ on regular two dimensional sub-lattices $\mathcal{Z}_m^2$ with spacing $1/m$, $m = 1, 2, \ldots$, which have individual spectral densities of the form

$$f_m(\omega,\eta) = \frac{\sigma_m^2}{m^2\Big[\sin^2(\frac{1}{2m}\omega) + \sin^2(\frac{1}{2m}\eta)\Big]^{\nu}}, \tag{3}$$

with $\omega, \eta \in (-\pi m, \pi m]$, $\sigma_m > 0$ and $\nu > 0$. In the above, $\nu = 1$ corresponds to the first-order intrinsic autoregression, $\nu = 2$ suggests a Whittle's simultaneous intrinsic autoregression and so on. The integer values of ν, namely, $\nu = 1, 2, 3, \ldots$ were the focus of Lindgren et al. (2011) and they correspond to various lattice-based intrinsic autoregressions with sparse precision matrices. In contrast, the non-integer values of ν, which is the primary focus of this paper, lead to fractionally differenced random fields. Typically, they correspond to random fields with non-sparse precision matrices and offer greater flexibility in modeling long range spatial dependencies. Specifically, following Duffin (1953), let D_m be the Laplace difference operator on the sub-lattice $\mathcal{Z}_m^2$, i.e.,

$$D_m f(u,v) = f(u,v) - \tfrac{1}{4}\{f(u+\tfrac{1}{m},v) + f(u-\tfrac{1}{m},v) + f(u,v+\tfrac{1}{m}) + f(u,v-\tfrac{1}{m})\},$$

where f is any real valued function defined at the lattice points of $\mathcal{Z}_m^2$. Then, by extending the results of Hosking (1981), we can represent $\{Z^{(m)}(u,v)\}$ as

$$D_m^{\nu/2}\, Z_{u,v}^{(m)} = \xi_{u,v}, \quad \nu > 0,$$

where $\xi_{u,v}$ is a Gaussian white noise random field on the sub-lattice $\mathcal{Z}_m^2$ with var $\xi_{u,v} = \sigma_m^2/m^2$. Thus, $\{Z^{(m)}(u,v)\}$ can be interpreted as a fractional Laplacian differenced random field on the sub-lattice $\mathcal{Z}_m^2$ and they enjoy properties similar to those of fractionally differenced time series.

Now, just like the fractionally differenced time series, the fractionally differenced random fields $\{Z^{(m)}(u,v)\}$ can be thought of as discrete space analogue of certain continuum fractional random fields. In fact, these continuum random fields, say, $\{Z(u,v)\}$ can be obtained by taking scaling limits of $\{Z^{(m)}(u,v)\}$.

To elaborate on this, we assume that, as $m \to \infty$, $m^{\nu-1}\sigma_m \to \sigma/2^\nu$. Then, it is not difficult to see that $f_m(\omega, \eta)$ converges to

$$f(\omega, \eta) = \frac{\sigma^2}{(\omega^2 + \eta^2)^\nu}, \quad \sigma > 0, \tag{4}$$

The above gives the spectral density formula of a continuum Gaussian intrinsic Matérn random field and thus, from the above convergence result, it follows that continuum Gaussian intrinsic Matérn random fields are scaling limits of fractionally differenced random fields. Furthermore, the above convergence result explicitly describes the rescaling of parameters needed, particularly when we want to choose a suitable sub-lattice (e.g., to embed irregular sampling locations into a grid or to approximate irregular regions by unions of grid cells when observations themselves are aggregates over such regions). As we shall see, this will become important to show numeric consistency and robustness of inference to changes of lattice spacing.

The key question is, as approximations to the continuum intrinsic Matérn process, how good are the above lattice-based model $\{Z^{(m)}(u, v)\}$ for small values of m. For brevity of discussion, we focus on the case $1 < \nu < 2$. The intrinsic case with $\nu = 1$ is discussed in details in Mondal (2005), Besag and Mondal (2005) and Dutta and Mondal (2015a). The intrinsic case with $\nu \geq 2$ requires use of higher-order contrasts and some calculations are outlined in Appendix and in the supplement to the paper. The intrinsic case with $0 < \nu < 1$ is similar to the case $\nu = 1$ in the sense it also requires calculations using (contrasts of) averages over nun-null regions, and some calculations are given in the supplement to the paper. If $1 < \nu < 2$, following Matheron (1973), it is well known that the continuum intrinsic Matérn random field $\{Z(u, v)\}$ has the following exact power variogram formula

$$\gamma(s, t) = \sigma^2 \int_{\mathbb{R}^2} \frac{1 - \cos(s\omega)\cos(t\eta)}{4\pi^2(\omega^2 + \eta^2)^\nu} d\omega d\eta = -\frac{\sigma^2 \pi^{3/2} \Gamma(\nu - \frac{1}{2})}{4\pi^2 \Gamma(\nu)\Gamma(2\nu - 1)\sin(\nu\pi)}\ h^{2\nu-2},$$

where $h = \sqrt{s^2 + t^2}$. On the other hand, the variogram of the fractional Laplacian differenced random field on $\mathcal{Z}_m^2$ at lag (s, t) is given by the double integral formula

$$\gamma_m(s, t) = \sigma_m^2 \int_{-m\pi}^{m\pi} \int_{-m\pi}^{m\pi} \frac{1 - \cos(s\omega)\cos(t\eta)}{4\pi^2 m^2 \left(\sin^2 \frac{\omega}{2m} + \sin^2 \frac{\eta}{2m}\right)^\nu} d\omega d\eta,$$

which is computed numerically with great accuracy by a two-dimensional quadrature method discussed in Dutta and Mondal (2015b). For large m, $\gamma_m(s, t)$ will be very close to $\gamma(s, t)$. However, we want to know whether one can work with $\{Z^{(m)}(u, v)\}$ for moderately small values of m instead of working with the continuum process $\{Z(u, v)\}$ and essentially get the same result. To address this question, Figure 1 plots the difference $\{\gamma(s, t) - \gamma_m(s, t)\}$ against the lag distance $\sqrt{s^2 + t^2}$ for $\nu = 1.25$ and $\nu = 1.5$ and $\sigma^2 = 1$. When $\nu = 1.25$, we see that the difference decreases as m increases and this difference becomes almost

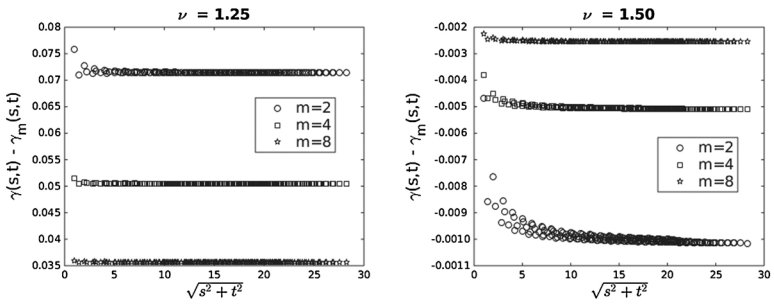

FIG 1. *Plots of the differences in semivariograms of the intrinsic Matérn process and the fractionally differenced process for* $\nu = 1.25$ *(left) and* 1.5 *(right). We took* $\sigma^2 = 1$ *and* $\sigma_m^2 = 4^{-\nu} m^{2-2\nu}$. *Each difference is plotted against the lag distance* $h = \sqrt{s^2 + t^2}$ *for lag* (s, t). *The values of* $\gamma(1, 0)$ *for* $\nu = 1.25$ *and* $\nu = 1.5$ *are* 0.304 *and* 0.159 *respectively.*

a positive constant at $m = 8$. Thus, in this case it is to our advantage to add a small nugget effect to the process $\{Z^{(m)}(u, v)\}$ so that statistical analyses based on the fractional Laplacian differenced random field on sub-lattices $\mathcal{Z}_m^2$ are essentially equivalent to the same based on continuum power variogram model on integer lattice $\mathcal{Z}^2$. On the other hand, when $\nu = 1.5$, the difference of the variograms is negative, increases with respect to m and essentially becomes a small negative constant when $m = 8$. In fact, notice that this difference is only 1.5% of $\gamma(0, 1)$ when $m = 8$. Thus, the lattice-based approximation would be good for moderate values of m even if a little nugget effect is present in the data.

In general, we have observed similar phenomena for other values of ν. In particular, for other reasonable values of $1 < \nu < 2$, we found that the difference $\gamma(s, t) - \gamma_m(s, t)$ is essentially a small constant (positive or negative) for moderately small values of m. Thus, it not unreasonable to anticipate that lattice based approximations are enhanced by routine use of a random error term along with the spatial component. In fact, these results are typical to the ones found in Mondal (2005), Besag and Mondal (2005) and Dutta and Mondal (2015a) and they can be further justified by writing down the asymptotic expansions of $\gamma_m(s, t)$ using results from Duffin and Shaffer (1960), as done in Mondal (2005), Besag and Mondal (2005) and Dutta and Mondal (2015a) for the case of $\nu = 1$. Overall, a fractional Laplacian differenced random field plus a white noise at a reasonable sub-lattice is a very good approximation of the intrinsic Matérn plus white noise model on the actual lattice.

Next we explore how the finite grid representation in (2) is connected with the infinite lattice representation in (3). To this end, suppose that observations are made on a finite $r \times c$ regular array with $r = p_1 \times m$, $c = p_2 \times m$ and lattice spacing $1/m$. We can then approximate the Laplace differenced operator D_m by the discrete (graph) Laplacian matrix of this grid. However, note that the matrix $-\mathbf{W}$ in (2) is the Laplacian matrix on this finite grid. This leads to saying that the spatial lattice process $Z_{u,v}^{(m)}$ on a finite integer lattice are approximately zero

mean Gaussian random variables with precision matrix $\sigma_m^{-2}\mathbf{W}^\nu$. This gives rise to (2) with $\lambda_\psi = \sigma_m^{-2}$.

Furthermore, suppose for $k = r$ or $k = c$ that $\mathbf{M}_k$ is a $k \times k$ matrix corresponding to the discrete cosine transformation with entries

$$m_{1,j} = k^{-\frac{1}{2}},\ m_{i,j} = (2/k)^{-\frac{1}{2}} \cos\{\pi(i-1)(j-\tfrac{1}{2})/k\}, \quad i = 2, \ldots, k,\ j = 1, \ldots, k.$$

Suppose also that $\mathbf{D}_k$ is a diagonal matrix with ith diagonal entry equal to

$$d_{k,i} = 2[1 - \cos\{\pi(i-1)/k\}].$$

It then follows that $\mathbf{M}_k$ is orthogonal and $\mathbf{M} = \mathbf{M}_c \otimes \mathbf{M}_r$ diagonalizes $\mathbf{W}$. Specifically,

$$\mathbf{M}\mathbf{W}\mathbf{M}^T = \tfrac{1}{4}\mathbf{I}_c \otimes \mathbf{D}_r + \tfrac{1}{4}\mathbf{D}_c \otimes \mathbf{I}_r = \mathbf{D}_{01} + \mathbf{D}_{10},$$

where $\mathbf{D}_{01} = \frac{1}{4}\mathbf{I}_c \otimes \mathbf{D}_r$ and $\mathbf{D}_{10} = \frac{1}{4}\mathbf{D}_c \otimes \mathbf{I}_r$. Thus $\mathbf{W}^\nu$ has the spectral decomposition

$$\mathbf{W}^\nu = \mathbf{M}^T(\mathbf{D}_{01} + \mathbf{D}_{10})^\nu \mathbf{M}.$$

It is now obvious that, for any $m = 1, 2, \ldots$, the eigenvalues of $\mathbf{R}^-$, the Moore–Penrose inverse of $\mathbf{R}$, exactly trace out the spectral density (3) at the discrete cosine frequencies, further justifying the validity of (2). It is also obvious that with the increase of m, i.e., as the lattice spacing diminishes, the eigenvalues of $\mathbf{R}^-$ trace out the limiting continuum spectral density (4) better.

The spatial formulation in (2) as representation of the infinite lattice fractional random fields in (3) is enhanced further if we allow a few extra layers of conceptual 'border' grid cells when we embed sampling locations. The idea was put forward by Besag and Higdon (1999), purely as a computational ploy, to alleviate edge effects, and, in practice, its implementation requires just a straightforward adjustment to $\mathbf{F}$.

3. Estimation for the mixed model

The singularity of $\mathbf{R}$ presents some difficulty in interpreting the mixed model in (1). Put another way, (1) is interpreted in terms of contrasts of $\mathbf{y}$, i.e., $\mathbf{C}_0\mathbf{y} = \mathbf{C}_0\mathbf{T}\boldsymbol{\tau} + \mathbf{C}_0\mathbf{F}\boldsymbol{\psi} + \mathbf{C}_0\boldsymbol{\epsilon}$, where rows of $\mathbf{C}_0$ are vector of contrasts and $\text{rank}(\mathbf{C}_0)$ is equal to $\text{rank}(\mathbf{F}\mathbf{R}^-\mathbf{F}^T)$. Thus, contrasts of $\boldsymbol{\tau}$ are estimable and contrasts of $\boldsymbol{\psi}$ are predictable. In such settings, typical REML estimation goes as follows. We estimate $\boldsymbol{\tau}$ from marginal distribution of $\mathbf{C}_0\mathbf{y}$. We estimate BLUPs of the contrasts of $\boldsymbol{\psi}$ using conditional mean formula given data differences $\mathbf{C}_0\mathbf{y}$. Separate from the estimation of $\boldsymbol{\tau}$, we obtain estimates of λ_y, λ_ψ and ν from maximizing the log REML function. However, unless carefully done, such estimation procedures based on data differencing do not lead to any simplification nor do they provide any insight into the REML estimation problem. Thus, in what follows, we draw upon the work of Henderson (1950, 1975), Lee and Nelder (1996, 2001), and Dutta and Mondal (2015a) and derive exact REML estimation using h-likelihood formulation.

3.1. Estimation of fixed effects and prediction of spatial effects

Let $\mathbf{B}$ denote the matrix of orthogonal contrasts formed by the last $rc-1$ eigenvectors of $\mathbf{W}$, and let $\mathbf{G}$ denote the diagonal matrix formed by the corresponding eigenvalues of $\mathbf{W}$, i.e.,

$$\mathbf{G} = \text{diag}\{(d_{01,i} + d_{10,i})^{\nu},\ i = 2, \ldots, rc\}$$

where $d_{01,i}$ and $d_{10,i}$ are the ith diagonal entries of $\mathbf{D}_{01}$ and $\mathbf{D}_{10}$ respectively. It then follows that $\mathbf{B}\boldsymbol{\psi}$ has an $rc-1$ dimensional Gaussian distribution with zero mean and covariance matrix $\lambda_{\psi}^{-1}\mathbf{G}^{-1}$. Extending Henderson (1950, 1975), Lee and Nelder (1996, 2001) and Dutta and Mondal (2015a), we represent the mixed model equations in (1)-(3) by

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{T} & \mathbf{F} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{\psi} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\zeta} \end{pmatrix}, \tag{5}$$

where $\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$ are independent centered Gaussian distributions random vectors with covariance matrices $\lambda_y^{-1}\mathbf{I}$ and $\lambda_{\psi}^{-1}\mathbf{G}^{-1}$ respectively. For notational brevity, we further define

$$\mathbf{z} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{T} & \mathbf{F} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{\psi} \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\zeta} \end{pmatrix} \quad \text{and}$$
$$\mathbf{Q} = \begin{pmatrix} \lambda_y\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \lambda_{\psi}\mathbf{G} \end{pmatrix}.$$

Then we can express the mixed model in (1) as a linear regression model

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}$$

where $\boldsymbol{\eta} \sim \mathcal{N}_{n+rc-1}(\mathbf{0}, \mathbf{Q}^{-1})$. Under the assumption that $\mathbf{T1} = \mathbf{F1} = \mathbf{1}$, the design matrix $\mathbf{X}$ has one rank deficiency; that is there is a non-zero vector $\boldsymbol{\ell}$ such that $\mathbf{X}\boldsymbol{\ell} = \mathbf{0}$. For fixed parameters λ_y, λ_{ψ} and ν, we then estimate $\boldsymbol{\beta}$ by solving the normal equation

$$\begin{pmatrix} \lambda_y\mathbf{T}^T\mathbf{T} & \lambda_y\mathbf{T}^T\mathbf{F} \\ \lambda_y\mathbf{F}^T\mathbf{T} & \lambda_y\mathbf{F}^T\mathbf{F} + \lambda_{\psi}\mathbf{W}^{\nu} \end{pmatrix} \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{\psi} \end{pmatrix} = \begin{pmatrix} \lambda_y\mathbf{T}^T\mathbf{y} \\ \lambda_y\mathbf{F}^T\mathbf{y} \end{pmatrix}$$

or in more succinct notation

$$\mathbf{X}^T\mathbf{Q}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Q}\mathbf{z} \quad \text{or,} \quad \mathbf{A}\boldsymbol{\beta} = \mathbf{b} \tag{6}$$

subject to the constraint $\mathbf{A}\boldsymbol{\ell} = \mathbf{0}$. Let $\widehat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ that we obtain solving the normal equation (6). Let $\widehat{\boldsymbol{\tau}}$ and $\widehat{\boldsymbol{\psi}}$ be the corresponding estimates of $\boldsymbol{\tau}$ and $\boldsymbol{\psi}$, which we obtain from $\widehat{\boldsymbol{\beta}}$.

Extending the argument given in Dutta and Mondal (2015a), one can then see that the contrasts of $\widehat{\boldsymbol{\tau}}$ coincide with the best linear unbiased estimators of the contrasts of $\boldsymbol{\tau}$. Similarly, the contrasts of $\widehat{\boldsymbol{\psi}}$ coincide with the best linear unbiased predictors of the contrasts of $\boldsymbol{\psi}$. Furthermore, even if the symmetric coefficient matrix $\mathbf{A} = \mathbf{X}^T\mathbf{Q}\mathbf{X}$ is not sparse, we shall see that the above normal equation is solved using a novel and efficient matrix-free Lanczos algorithm.

3.2. REML estimation for precision parameters

Extending Lee and Nelder (1996, 2001) and Dutta and Mondal (2015a), the log-likelihood function of the residuals obtained from the regression model (5) takes the form of

$$2\ell_r(\boldsymbol{\lambda}, \nu) = \log\det \mathbf{Q} - \log|\mathbf{X}^T\mathbf{Q}\mathbf{X}|_+ - (\mathbf{z} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T\mathbf{Q}(\mathbf{z} - \mathbf{X}\widehat{\boldsymbol{\beta}}). \tag{7}$$

The derivations provided in Dutta and Mondal (2015a) further show that the traditional REML log-likelihood arising from $\mathbf{C}_0\mathbf{y}$ and (7) is actually same up to an additive constant that depends only on the contrast matrix $\mathbf{C}_0$. Thus, maximizing (7) gives rise to REML precision parameter estimate of $\boldsymbol{\theta} = (\lambda_y, \lambda_\psi, \nu)$.

Traditional maximization of the log REML function uses score equations which is obtained by equating the gradient of ℓ_r to zero. Thus suppose $\mathbf{Q}_1 = \partial\mathbf{Q}/\partial\lambda_y$, $\mathbf{Q}_2 = \partial\mathbf{Q}/\partial\lambda_\psi$, and $\mathbf{Q}_3 = \partial\mathbf{Q}/\partial\nu$. The score equations that maximize the log–REML function in (7) are then given by

$$\tfrac{1}{2}\mathrm{Tr}\left(\mathbf{Q}^{-1}\mathbf{Q}_i\right) - \tfrac{1}{2}\mathrm{Tr}\ \left((\mathbf{X}^T\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}_i\mathbf{X}\right) - \tfrac{1}{2}(\mathbf{z} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T\mathbf{Q}_i(\mathbf{z} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 0$$

for $i = 1, 2$ and 3. It is not difficult to see that the above score equations can also be written as

$$\tfrac{1}{2}\mathrm{Tr}\left(\mathbf{Q}^{-1}\mathbf{Q}_i\right) - \tfrac{1}{2}\mathrm{Tr}\ \mathbf{H}\mathbf{Q}^{-1}\mathbf{Q}_i - \tfrac{1}{2}(\mathbf{z} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T\mathbf{Q}_i(\mathbf{z} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 0, \quad i = 1, 2, 3, \tag{8}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}$ denotes the 'hat'-matrix of the linear regression model (5). Typically, Fisher's scoring method is used to solve the score equations and to obtain REML estimates. However, this also requires computation of the second derivatives of the log REML function or the information matrix $\boldsymbol{\mathfrak{I}}$ whose (i, j)th entry is equal to

$$\mathfrak{I}(i, j) = \tfrac{1}{2}\mathrm{Tr}\left\{(\mathbf{I} - \mathbf{H})\mathbf{Q}^{-1}\mathbf{Q}_i(\mathbf{I} - \mathbf{H})\mathbf{Q}^{-1}\mathbf{Q}_j\right\}, \tag{9}$$

and whose inverse is used to produce estimates for dispersion of the REML estimators of the precision parameter.

3.3. Nonconvexity in REML estimation

Neither the log REML function (7), nor the scoring equations (8) give insight into the exact non-convex nature of the optimization problem. Here, instead of maximizing the REML function directly, we consider an alternative approach that allows us to characterize the REML optimization problem in terms of a non-linear gamma regression. This alternative approach follows the work of Lee and Nelder (1996, 2001) and Dutta and Mondal (2015a). The basic idea is as follows. Rather than using traditional REML computations that separate estimations of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, we can consider an iterative approach. Thus, starting with an initial estimate $\widehat{\boldsymbol{\beta}}$, we compute the residuals $\mathbf{e} = \mathbf{z} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ and use them

to estimate the precision parameters λ_y, λ_ψ and ν. Then, once estimates $\widehat{\lambda}_y, \widehat{\lambda}_\psi$ and $\widehat{\nu}$ are obtained, we update $\boldsymbol{\beta}$ by computing the normal equation from (6). The process continues until the estimates converge numerically. It is just that the final estimates are actually REML estimates. Specifically, suppose h_{jj} is the jth diagonal element of the hat matrix $\mathbf{H}$, and q_{jj} is the jth diagonal entry of $\mathbf{Q}$. Then

$$q_{jj} = \lambda_y v_{y,j} + \lambda_\psi (v_{01,j} + v_{10,j})^\nu$$

where

$$v_{y,j} = \begin{cases} 1 & \text{if } 1 \le j \le n \\ 0 & \text{if } n+1 \le j \le n + rc - 1 \end{cases},$$

$$v_{01,j} = \begin{cases} 0 & \text{if } 1 \le j \le n \\ d_{01,j-n+1} & \text{if } n+1 \le j \le n + rc - 1 \end{cases}$$

and

$$v_{10,j} = \begin{cases} 0 & \text{if } 1 \le j \le n \\ d_{10,j-n+1} & \text{if } n+1 \le j \le n + rc - 1. \end{cases}.$$

Accordingly the score equations in (8) then can be written as

$$\sum_{j=1}^{n+rc-1} (1 - h_{jj}) q_{jj}^{(i)} \left(\frac{e_j^2}{1 - h_{jj}} - \frac{1}{q_{jj}} \right) = 0, \quad i = 1, \ldots, 3, \tag{10}$$

where $q_{jj}^{(i)}$ denote the jth diagonal entry of $\mathbf{Q}_i$. The above score equations coincide with the estimating equations of a nonlinear gamma regression, where we assume that the adjusted residuals $e_j^2/(1 - h_{jj})$'s are the responses variables that are distributed independently as Gamma random variables, and we have an inverse link, prior weights $(1 - h_{jj})$'s and nonlinear predictors $q_{jj} = \lambda_y v_{y,j} + \lambda_\psi (v_{01,j} + v_{10,j})^\nu$.

The above provides a precise characterization of the non-convex nature of REML estimation in that non-convexity arises through the estimation of the dependence parameter ν. In particular, when ν is fixed and known, the optimization becomes a convex optimization in which case any local maximum is also the global maximum.

4. Matrix-free computations

Typical statistical computations such as REML estimation from Matérn mixed models of spatial data with n irregularly distributed sampling locations require at least $O(n^3)$ computations and $O(n^2)$ storage space. The main objective here is to reduce computations for REML estimation of our mixed linear model in (1) to $O(Kn(\log n)^2)$ operations, where $K << n$ (and effectively a constant) and storage to $O(n)$ space. We do this without losing any efficiency properties of the REML estimators. To this end, we divide our task into three parts. First, we

discuss the two dimensional discrete cosine transformation that is essential for various matrix-free matrix-vector statistical computations for our mixed model. Second, we indicate how we can compute BLUEs and BLUPs of contrasts using a novel matrix-free Lanczos algorithm. Here we also develop a novel preconditioning method to reduce the order of computations. Third, we develop fast matrix-free ways to obtain REML estimates using stochastic trace approximations and a trust region method. In this paper we do not pursue equation (8) for estimating the precision parameters, which would require computing the diagonal entries of the hat matrix. In a future paper, we shall take up matrix-free computations of such quantities. Furthermore, in this section, we provide some preliminary matrix-free computations of the log REML function based on Taylor series approximations. In our limited experience, it appears that these computations work better than those by Aune et al. (2014). More accurate scalable matrix-free computations of the log REML function will be a matter of future investigation.

4.1. DCT and matrix–vector multiplications

For any $r \times c$ matrix $\mathbf{E} = (e_{i,j})$, its two dimensional discrete cosine transformation gives rise to a $r \times c$ matrix whose (s,t)th entry is given by

$$c_s c'_t \sum_{i=1}^{r} \sum_{j=1}^{c} e_{i,j} \cos(\pi(i-1/2)(s-1)/r) \cos(\pi(j-1/2)(t-1)/c)$$

where $c_s = \sqrt{1/r}$ if $s = 1$ and $\sqrt{2/r}$ otherwise, and $c'_t = \sqrt{1/c}$ if $t = 1$ and $\sqrt{2/c}$ otherwise. On the contrary, the two dimensional inverse discrete cosine transformation on $\mathbf{E}$ produces another $r \times c$ matrix with (s,t)th entry:

$$\sum_{i=1}^{r} \sum_{j=1}^{c} c_i c'_j e_{i,j} \cos(\pi(s-1/2)(i-1)/r) \cos(\pi(t-1/2)(j-1)/c).$$

The above transformation is very closely related to the two dimensional fast Fourier transformation (FFT). Indeed, following the works of Cooley and Tukey (1965) and Rao and Yip (1990), one can factorize the above computations further, as done in the case of FFT. These factorizations reduce the computational complexities of DCT to $O(rc\log(rc))$. Alternatively, it is also possible to use FFT to compute the DCTs with additional $O(rc)$ pre- and post-processing steps, as shown by Makhoul (1980). Over the decades, highly optimized algorithms for FFT have been developed for various machine architectures. We have found that these algorithms in conjunction with $O(rc)$ pre- and post-processing steps are more time efficient to compute DCTs than directly implementing the Cooley-Tuckey and the Rao-Yip algorithms. In this paper, we follow Frigo and Johnson (2005) that show how to eliminate the redundant operations of the FFT algorithms to compute the DCTs. The codes from Frigo and Johnson (2005) are freely available on the web: `http://www.fftw.org/`.

In what follows, the two dimensional DCT is used in various matrix-vector multiplications. We encounter these matrix-vector multiplications in solving the normal equation (6), and in computing the REML score equations and in computing the entries of the Fisher information matrix. Essentially, these matrix-vector multiplications would require us to compute either $\mathbf{W}^{\nu}\mathbf{x}$ or $\mathbf{W}^{-\nu}\mathbf{x}$ for certain candidate vector $\mathbf{x}$. The matrix $\mathbf{W}^{\nu}$ and its pseudo-inverse $\mathbf{W}^{-\nu}$ are non sparse. However, spectral decomposition allows us to write $\mathbf{W}^{\nu} = \mathbf{M}^{T}(\mathbf{D}_{01} + \mathbf{D}_{10})^{-\nu}\mathbf{M}$ and $\mathbf{W}^{-\nu} = \mathbf{M}^{T}(\mathbf{D}_{01} + \mathbf{D}_{10})^{-\nu}\mathbf{M}$, where $(\mathbf{D}_{01} + \mathbf{D}_{10})^{-\nu}$ is obtained by inverting all the non-zero entries of $(\mathbf{D}_{01} + \mathbf{D}_{10})^{\nu}$. Thus, multiplying a vector with $\mathbf{W}^{\nu}$ or $\mathbf{W}^{-\nu}$ essentially reduces to multiplying a vector with $\mathbf{M}$ or $\mathbf{M}^{T}$. At this point, we note that obtaining $\mathbf{Mv}$ is same as performing a two-dimensional DCT on the $r \times c$ matrix formed using elements of $\mathbf{v}$ filling each column at a time and then converting the resulting matrix back into a vector by stacking the columns. Similarly computing $\mathbf{M}^{T}\mathbf{v}$ is same as obtaining a two-dimensional inverse DCT.

It must be noted that, for very large arrays, one can further improve the speed of the two-dimensional DCTs using distributed computing. The main idea comes from the fact that a two-dimensional DCT consists of many one-dimensional DCTs along the columns and then along the rows. Thus, these one-dimensional DCT computations can be distributed equally or parallely among the cores of the processors. This division of computations is particularly useful on a graphical processing unit (GPU) which typically has thousands of processing cores and this division of computations can provide substantial gain in computational time. For further discussions on parallel implementation of DCT, we refer to the website: http://www.fftw.org/parallel/parallel-fftw.html.

4.2. Solving linear equations with Lanczos algorithm

First, note that the Lanczos algorithm in Dutta and Mondal (2015a) for solving the sparse system of equations $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ with $\nu = 1$ can be extended to solve a dense system structured by general values of $\nu > 0$ in matrix-free way. This matrix-free extension of the algorithm from sparse to the dense case is possible because all that the Lanczos algorithm uses are matrix vectors products of the form $\mathbf{Ax}$ which can be computed via the DCT and the inverse DCT described in the previous section. The Lanczos algorithm proceeds as follows. It computes a set of orthonormal vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots$ called the Lanczos vectors, from the span of $\mathbf{b}, \mathbf{Ab}, \mathbf{A}^2\mathbf{b}, \ldots$. At the kth iteration, it then obtain an approximate tridiagonal factorization of $\mathbf{AV}_k \approx \mathbf{V}_k\boldsymbol{\Delta}_k$ where $\mathbf{V}_k$ has columns $\mathbf{v}_1, \ldots, \mathbf{v}_k$ and $\boldsymbol{\Delta}_k$ is a $k \times k$ positive definite tridiagonal matrix. In the implemented version of the algorithm none of these matrices $\mathbf{V}_k$ and $\boldsymbol{\Delta}_k$ is stored but the solution is iteratively updated on the fly by progressively computing the lower bidiagonal Cholesky factorization of $\boldsymbol{\Delta}_k$. We refer to Dutta and Mondal (2015a) for all technical details on the algorithm including computing the norm of the residuals in almost no additional cost that gives a practical stopping criterion. The algorithm stores only few vectors of length $m + rc$ and a few scalars thus requiring only $O(rc)$ memory. With the use of the discrete cosine transformation,

a matrix-vector multiplication of the form $\mathbf{Av}$ incurs a $O(rc\log(rc))$ computational cost. Thus, if the Lanczos algorithm takes K_0 iterations to converge, then the total computational cost of solving $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ becomes $O(K_0 rc\log(rc))$.

Second, we do not solve $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ directly, but rather we carefully construct a preconditioning matrix $\mathbf{P}$ and solve a well-conditioned system of linear equations $\mathbf{PAP}^T\boldsymbol{\beta}' = \mathbf{Pb}$. We then obtain the final solution as $\boldsymbol{\beta} = \mathbf{P}^T\boldsymbol{\beta}'$. The idea is to reduce the number of iterations K_0 as much as possible. The matrix $\mathbf{P}$ has the following requirements: 1) we want $\mathbf{P}$ such that the matrix-vector products $\mathbf{Px}$ and $\mathbf{P}^T\mathbf{x}$ can be computed in matrix–free way at a computational cost of $O(rc\log(rc))$, and 2) the condition number of $\mathbf{PAP}^T$ will be finite so that Lanczos algorithm for solving $\mathbf{PAP}^T\boldsymbol{\beta}' = \mathbf{Pb}$ converges geometrically fast, i.e., in $O(\log(rc))$ iterations.

To this end, we apply a block preconditioner for $\mathbf{A}$ where the first block is a Cholesky factor of the small dimensional matrix $\lambda_y\mathbf{T}^T\mathbf{T}$ and the second block is a preconditioner for the matrix

$$\mathbf{C} = \lambda_y\mathbf{F}^T\mathbf{F} + \lambda_\psi\mathbf{W}^\nu.$$

When $\nu = 1$ (or some other positive integer), the above matrix is sparse, in which case Dutta and Mondal (2015a) proposed using its incomplete Cholesky factor as preconditioner. For a general non-integer value of ν, we construct here a practical sparse approximation $\widetilde{\mathbf{C}}$ of $\mathbf{C}$ by thresholding or tapering in such a way that the small eigenvalues of $\widetilde{\mathbf{C}}$ stay at the same order of the small eigenvalues of $\mathbf{C}$. We then apply an incomplete Cholesky factor of $\widetilde{\mathbf{C}}$ to obtain an effective matrix-free preconditioner of $\mathbf{C}$.

In order to construct the sparse approximation $\widetilde{\mathbf{C}}$ of $\mathbf{C}$, we use the following decomposition of the matrix $\mathbf{W}$:

$$\mathbf{W} = \mathbf{I}_c \otimes \mathbf{W}_r + \mathbf{W}_c \otimes \mathbf{I}_r,$$

where $4\mathbf{W}_k = \mathbf{M}_k^T\mathbf{D}_k\mathbf{M}_k$ and k is either r or c. In the above the matrix $\mathbf{W}_k$ is a tridiagonal matrix and $\mathbf{W}_k^\nu$ is a non-sparse matrix, but its entries far away from the diagonal are very small in magnitude compared to the ones on or near the diagonal. Let $\widetilde{\mathbf{W}}_k^\nu$ be a sparse matrix approximation of $\mathbf{W}_k^\nu$ obtained by suitable thresholding or tapering. Then a candidate for the matrix $\widetilde{\mathbf{C}}$ is given by

$$\widetilde{\mathbf{C}} = \lambda_y\mathbf{F}^T\mathbf{F} + \lambda_\psi\left[\mathbf{I}_c \otimes \widetilde{\mathbf{W}}_r^\nu \;+\; \widetilde{\mathbf{W}}_c^\nu \otimes \mathbf{I}_r\right].$$

Thus the block diagonal preconditioner for $\mathbf{A}$ becomes $\mathbf{P} = \mathrm{diag}\{\mathbf{L}_1^{-1}, \mathbf{L}_2^{-1}\}$ where $\mathbf{L}_1$ is the lower triangular Cholesky factor of $\lambda_y\mathbf{T}^T\mathbf{T}$ and $\mathbf{L}_2$ is the lower triangular incomplete Cholesky factor of $\widetilde{\mathbf{C}}$. This effectively makes the condition number of the matrix $\mathbf{PAP}^T$ bounded. It can seen that, if r and c are of the same order, the overall cost of constructing this matrix $\mathbf{P}$ is $O(rc\log(rc))$. Furthermore, since the inverse of $\mathbf{P}$ sparse, the matrix vector multiplications $\mathbf{Pv}$ and $\mathbf{P}^T\mathbf{v}$ incurs a computational cost of at most $O(rc\log(rc))$.

The above preconditioner effectively brings down the computation cost of calculating the BLUEs of the contrasts of $\boldsymbol{\tau}$ and the BLUPs of the contrasts of $\boldsymbol{\psi}$ to $O(rc(\log(rc))^2)$ operations.

Alternatively, one can obtain a preconditioner of $\mathbf{A}$ using Chebyshev polynomial approximations, see e.g., (Saad, 1985), but we did not pursue its use in this work.

4.3. Stochastic trace approximation

Equations (8) and (9) require computations of the trace terms $\text{Tr } \mathbf{H}\mathbf{Q}^{-1}\mathbf{Q}_i$ and $\text{Tr}(\mathbf{I}-\mathbf{H})\mathbf{Q}^{-1}\,\mathbf{Q}_i(\mathbf{I}-\mathbf{H})\mathbf{Q}^{-1}\mathbf{Q}_j$, which are expensive as computing individual diagonal elements requires at least $O(rc(\log(rc))^2)$ calculations even with the use of discrete cosine transformation and the Lanczos algorithm. In other words, overall trace computations using this method require at least $O((rc)^2(\log(rc))^2)$ operations. Here, instead we apply Hutchinson's method that stochastically approximates the trace of a symmetric matrix using a Monte Carlo average of its quadratic forms in random vectors with zero mean and identity covariance matrix. Thus, for a symmetric matrix $\mathbf{E}$, the Hutchinson's method approximates the trace of $\mathbf{E}$ by

$$\text{Tr } \mathbf{E} \quad \widehat{=} \frac{1}{K}\sum_{t=1}^{K}\mathbf{u}_t^T\mathbf{E}\mathbf{u}_t$$

where $\mathbf{u}_t$'s are either i.i.d Rademacher or Gaussian random variables and K is an integer that is much smaller than the dimension of $\mathbf{E}$. Throughout, we only consider i.i.d. Rademacher variables because they minimize the variance of the above estimate among all i.i.d random variables of size K from a distribution with zero mean vector and identity covariance matrix. Although further reduction in variances of the trace approximations are possible if we allow dependent random variables in trace estimation, we do not pursue such stochastic approximations in this paper. It then follows that the REML score equations reduce to the following unbiased estimation equations

$$g_i(\boldsymbol{\theta}) := \frac{1}{2K}\sum_{t=1}^{K}\mathbf{u}_t^T\mathbf{Q}^{-1}\mathbf{Q}_i(\mathbf{I}-\mathbf{H})\mathbf{u}_t - \tfrac{1}{2}(\mathbf{z}-\mathbf{X}\widehat{\boldsymbol{\beta}})^T\mathbf{Q}_i(\mathbf{z}-\mathbf{X}\widehat{\boldsymbol{\beta}}) = 0. \qquad (11)$$

And the (i,j)th entry of the Hessian matrix becomes

$$\frac{1}{2K}\sum_{t=1}^{K}\mathbf{u}_t^T\Big((\mathbf{I}-\mathbf{H})\mathbf{Q}^{-1}\mathbf{Q}_i(\mathbf{I}-\mathbf{H})\mathbf{Q}^{-1}\mathbf{Q}_j\Big)\mathbf{u}_t.$$

Overall the computations of the trace terms, REML score equations and the information matrix require $O(Krc(\log(rc))^2)$ flops and $O(rc)$ storage and allow us to pursue numeric optimization of the log REML function via the Newton-Raphson line search method or the trust region methods.

4.4. Trust region method

Finding a solution to the REML score equation that minimizes the negative log REML function is a non-convex optimization problem, and in such settings, the traditional Netwon-Rhapson algorithm with line search methods are often perceived to be susceptible in practice. Thus, following Powell (1970, 1984) and Nocedal and Wright (1999), we adopt here a trust-region method that can be considered as a global line search method and that allows us to obtain solutions to the REML score equations in a numerically stable way. At any iteration, the trust region method approximates the objective function by a suitable quadratic function, identifies a region within which it perceives the quadratic approximation to be good and then minimizes the quadratic approximation over the identified region around the current value of the variable to obtain the next iterate. Understandably, this method avoids directly computing the negative log REML function $\ell_R(\theta)$. Instead, it sets the objective function to be $(1/2)\|\nabla \mathbf{g}(\boldsymbol{\theta})\|^2$, where $\nabla \mathbf{g}(\boldsymbol{\theta})$ denotes gradient of the function $\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), g_3(\boldsymbol{\theta}))$, namely, the approximate REML score equations in (11) . Thus, the objective function is minimized at $\mathbf{x} = \mathbf{c}$ if and only if $\nabla \mathbf{g}(\mathbf{c}) = \mathbf{0}$ provided that the Hessian $\nabla^2 \mathbf{g}(\boldsymbol{\theta})$ is positive definite. Furthermore, this method uses the following the quadratic function to approximate $(1/2)\|\nabla \mathbf{g}(\boldsymbol{\theta})\|^2$ around a point $\boldsymbol{\theta}_k$.

$$\mathbf{p}_k(\boldsymbol{\delta}) = \tfrac{1}{2}\|\nabla \mathbf{g}(\boldsymbol{\theta}_k)\|^2 + \boldsymbol{\delta}^T[\nabla^2 \mathbf{g}(\boldsymbol{\theta}_k)]^T \nabla \mathbf{g}(\boldsymbol{\theta}_k) + \tfrac{1}{2}\boldsymbol{\delta}^T[\nabla^2 \nabla \mathbf{g}(\boldsymbol{\theta}_k)]^2 \boldsymbol{\delta}.$$

It must be emphasized that the above quadratic function can computed in a matrix free way, as we can do so for computing the gradient and the Hessian using the stochastic trace approximations. Next, the trust region method picks the step size $\boldsymbol{\delta}_k$ by minimizing $\mathbf{p}_k(\boldsymbol{\delta})$ in a suitable region. The choice of the suitable region is critical here. If the region is too large then $\mathbf{p}_k(\boldsymbol{\delta})$ could be a poor approximation to $(1/2)\|\nabla \mathbf{g}(\boldsymbol{\theta}_k)\|^2$ and minimizing $\mathbf{p}_k(\boldsymbol{\delta})$ could become unsuitable. On the other hand, if the region is too small then the algorithm will pick a tiny step size resulting in little effective gain. Thus, the trust region method adaptively chooses the radius of the region based on the performance of the previous iteration. If there was a significant gain in the previous iteration then the algorithm becomes optimistic and expands the radius of the trust region; a loss on the other hand results in shrinking the radius. Overall, the key steps of the algorithm are then given by:

Trust-Region algorithm:

- Set a maximum possible trust radius $\rho > 0$, an initial radius $0 < \rho_0 < \rho$ and $0 < \phi_0 < 1/4$.
- Iterate: For $k \geq 0$
    1. Compute steepest descend direction:

$$\boldsymbol{\delta}_{1k} = \frac{\nabla \mathbf{g}(\boldsymbol{\theta}_k)^T[\nabla^2 \mathbf{g}(\boldsymbol{\theta}_k)]^2 \nabla \mathbf{g}(\boldsymbol{\theta}_k)}{\nabla \mathbf{g}(\boldsymbol{\theta}_k)^T[\nabla^2 \mathbf{g}(\boldsymbol{\theta}_k)]^4 \nabla \mathbf{g}(\boldsymbol{\theta}_k)} \nabla^2 \mathbf{g}(\boldsymbol{\theta}_k) \nabla \mathbf{g}(\boldsymbol{\theta}_k)$$

2. Compute the Newton-Raphson direction:

$$\boldsymbol{\delta}_{2k} = -[\nabla^2 \mathbf{g}(\boldsymbol{\theta}_k)]^{-1} \nabla \mathbf{g}(\boldsymbol{\theta}_k)$$

3. Compute the trust region step size by:

$$\boldsymbol{\delta}_k = \boldsymbol{\delta}_{1k} + \lambda(\boldsymbol{\delta}_{2k} - \boldsymbol{\delta}_{1k})$$

where λ is the largest number in $[0, 1]$ such that $\|\boldsymbol{\delta}_k\| \leq \rho_k$ and this can be computed analytically.

4. Compute the effectiveness of the step:

$$\phi_k = \frac{\|\nabla \mathbf{g}(\boldsymbol{\theta}_k)\|^2 - \|\nabla \mathbf{g}(\boldsymbol{\theta}_k + \boldsymbol{\delta}_k)\|^2}{2(\mathbf{p}_k(\mathbf{0}) - \mathbf{p}_k(\boldsymbol{\delta}_k))}$$

5. Decide whether the trust region radius shrinks, expands or stays the same:

   If $\phi_k < 1/4$, set $\rho_{k+1} = \|\boldsymbol{\delta}_k\|/4$
   Else if $\phi_k > 3/4$ and $\|\boldsymbol{\delta}_k\| = \rho_k$, set $\rho_{k+1} = \min(2\rho_k, \rho_0)$.
   Else set $\rho_{k+1} = \rho_k$.

6. Decide whether to update the solution:

   If $\phi_k > \phi_0$, set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \boldsymbol{\delta}_k$.
   Else keep $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k$.

- End for.

The algorithm stops when either there is no significant change in the $\boldsymbol{\theta}_k$'s and the value of the objective function is sufficiently close to zero. Sometimes the algorithm does not converge in the sense that the radius of the trust region becomes very small and yet the value of the objective function remains far from zero. In such a case the algorithm must be restarted with a different initial value.

Furthermore, this algorithm computes the Jacobian matrix of the score equation, which is nothing but the negative of the observed information matrix. Thus, as a by-product of the trust region algorithm, we can obtain standard errors of the $\widehat{\boldsymbol{\theta}}$ as the square roots of the diagonal entries of the inverse of the observed information matrix (of the stochastic score equations).

Powell (1984) proves convergence to stationary points for $\phi_0 = 0$, i.e. when a step is taken whenever the value of the objective function goes down under some weak assumptions. Moré and Sorensen (1983) and Nocedal and Wright (1999, ch. 4) provide convergence results for $\phi_0 \in (0, 1/4)$ under stronger assumptions that the objective function is Lipschitz, continuously differentiable and the corresponding Hessian matrix is bounded.

4.5. Approximate computations of log REML likelihood function

Some preliminary computations of REML log-likelihood function can be obtained in a matrix-free way using Taylor series expansions. We indicate these

computations in the no covariate case (i.e., with $\mathbf{T} = \mathbf{0}$). First, in order to compute the log REML function, we need to find the log determinant of $\mathbf{C} = \lambda_y \mathbf{F}^T\mathbf{F} + \mathbf{R}$, where $\mathbf{R} = \lambda_\psi \mathbf{W}^\nu$ as given in (2). Using $\mathbf{C} = (\lambda_y \mathbf{I} + \mathbf{R}) - \lambda_y(\mathbf{I} - \mathbf{F}^T\mathbf{F})$ the log determinant of $\mathbf{C}$ can then be expressed as:

$$\log\det \mathbf{C} = \sum_{i=1}^{rc} \log\{\lambda_y + \lambda_\psi (d_{01,i} + d_{10,i})^\nu\} + \operatorname{Tr}\log\{\mathbf{I} - (\mathbf{I} + \lambda_y^{-1}\mathbf{R})^{-1}(\mathbf{I} - \mathbf{F}^T\mathbf{F})\}.$$

Next, we use the Taylor series expansion and write the second term as

$$\operatorname{Tr}\log\{\mathbf{I} - (\mathbf{I} + \lambda_y^{-1}\mathbf{R})^{-1}(\mathbf{I} - \mathbf{F}^T\mathbf{F})\} = \sum_{j=1}^{J} \frac{(-1)^j}{j} \operatorname{Tr}[\{(\mathbf{I} + \lambda_y^{-1}\mathbf{R})^{-1}(\mathbf{I} - \mathbf{F}^T\mathbf{F})\}^j] + o_J,$$

where o_J is a negligible term. The above is possible because the matrix $(\mathbf{I} + \lambda_y^{-1}\mathbf{R})^{-1}(\mathbf{I} - \mathbf{F}^T\mathbf{F})$ has spectral radius less than 1. In particular, note that the spectral radii of each of the matrices $\mathbf{I} - \mathbf{F}^T\mathbf{F}$ and $(\mathbf{I} + \lambda_y^{-1}\mathbf{R})^{-1}$ are exactly equal to 1. However, 1 is not an eigenvalue of their product because the matrix $\lambda_y \mathbf{F}^T\mathbf{F} + \mathbf{R}$ is non-singular. This also implies that the above Taylor series converges geometrically, a value of J can be set carefully to make the approximation sufficiently accurate.

Next we approximate the trace terms using Huchinson's estimators. Let $\tilde{\mathbf{u}}_k$'s be i.i.d Rademacher vectors. We then approximate the trace as

$$\operatorname{Tr}[\{(\mathbf{I} + \lambda_y^{-1}\mathbf{R})^{-1}(\mathbf{I} - \mathbf{F}^T\mathbf{F})\}^j] = \frac{1}{K'} \sum_{k=1}^{K'} \tilde{\mathbf{u}}_k^T \{(\mathbf{I} + \lambda_y^{-1}\mathbf{R})^{-1}(\mathbf{I} - \mathbf{F}^T\mathbf{F})\}^j \tilde{\mathbf{u}}_k$$

Furthermore, using the two-dimensional DCT, we can compute all of the above quadratic forms in a matrix-free way. This allows us to approximately evaluate the REML log-likelihood function in a matrix-free way.

In practice, negative log REML likelihood values can be substantially large, and can cause some numerical instability. However, in practice, we are mostly required to evaluate the difference of the log REML functions at two different parameter values, in which case we can write the difference of the log REML function in terms of term by term differences of quadratic forms. Computing these series of term by term differences is numerically more stable and they converge faster than the individual series of quadratic forms.

5. Simulation studies

5.1. Local maxima in REML estimations

The non-linear gamma score equations derived in Section 3.3 characterize the non-convex nature of the REML computations. The trust region method considered in Section 4.4 guarantees convergence to a local maximum, but its convergence to the global maximum requires a more careful investigation, as such convergences are often proved under strong regularity assumptions. Here we present

through simulations the performance of the REML estimators, computed using matrix-free trust region method and summarized for two different values of the dependence parameter. Results for further values of the dependence parameter and for varying proportions of missing observations are reported in the supplement to the paper. However, the overall conclusions of these more extensive simulation studies do not change from what we cover here. Thus the simulations summarized below are largely representative of what we expect to see for the entire range of parameter space and provide some critical understanding of the strength and weakness of the REML computations.

The specific details of the simulation runs are as follows. We generate realizations from the spatial mixed linear model (1) on 512×512 arrays by setting $\mathbf{T} = \mathbf{0}$, $\boldsymbol{\tau} = \mathbf{0}$, $\lambda_\psi = 8$, $\lambda_y = 4$ and using $\nu = 1.25$ and 1.50. This is done in several steps. First, for a fixed value of the dependence parameter ν, a random effect $\boldsymbol{\psi}$ is generated on a 512×512 array from a Gaussian distribution with mean $\mathbf{0}$ and a precision matrix $8\mathbf{W}^\nu$ with the sum constraint $\boldsymbol{\psi}^T\mathbf{1} = 0$. Then, each $\psi_{u,v}$ value, independent of others, is removed randomly with probability $p = 0.8$, and a Gaussian white noise with precision $\lambda_y = 4$ is added to the remaining ones. This produces a random incidence matrix $\mathbf{F}$ and a vector of realization $\mathbf{y}$ with an expected sample size of $n = 52428$. The procedure is then repeated with different values of ν.

Tables 1 and 2 display results of REML computations done on each of the simulated dataset using our matrix-free trust region method with 9 different starting values of the dependence parameter, namely, $\nu^{(0)} = 0.8, 0.9, 1.0, \ldots, 1.6$. Throughout we consider $K = 50$ Rademacher vectors to stochastically approximate the traces in the score equation (11). We then pick the values of $\lambda_y^{(0)}$ and $\lambda_\psi^{(0)}$ by finding the solutions to the REML score equation with fixed $\nu = \nu^{(0)}$. Using this triplet $(\nu^{(0)}, \lambda_y^{(0)}, \lambda_\psi^{(0)})$ as the starting point we next run the trust region algorithm and obtain final solutions $\widehat{\nu}, \widehat{\lambda}_y$ and $\widehat{\lambda}_\psi$ to the REML score equation. In some instances, the estimate of λ_y grew increasingly large with successive iterations and we mark it by '∞' in the tables. Thus, in a typical trust-region run '∞' signifies a value that is larger than $10,000$. While the first three columns of each table record the initial values of $\nu^{(0)}, \lambda_y^{(0)}$ and $\lambda_\psi^{(0)}$, the last three columns of each table comprise of the final estimated values $\widehat{\nu}, \widehat{\lambda}_y$ and $\widehat{\lambda}_\psi$.

As we glean through the numbers in these Tables, several conclusions can be drawn: (a) When the initial value $\nu^{(0)}$ is close to the true ν value, we find that the estimates of $\lambda_y^{(0)}$ and $\lambda_\psi^{(0)}$ are very reasonable, and the final estimates $\widehat{\nu}, \widehat{\lambda}_y$ and $\widehat{\lambda}_\psi$ are indeed very close to the true parameter values. (b) As we locally move the value of $\nu^{(0)}$ away from the true ν value, the estimates of $\lambda_y^{(0)}$ and $\lambda_\psi^{(0)}$ get further away from the true parameter values, but the procedure still converges to the same final estimates $\widehat{\nu}, \widehat{\lambda}_y$ and $\widehat{\lambda}_\psi$. There is an interesting pattern here, namely, if we increase the value of $\nu^{(0)}$ locally from the true ν value, we force a slightly smoother spatial process, and as a compensation we end up with a larger value of $\lambda_\psi^{(0)}$ but a smaller value of $\lambda_y^{(0)}$. However, as we

TABLE 1

Starting values and final REML estimates from trust region methods. True values of ν are as follows. Top: $\nu = 1.25$; bottom: $\nu = 1.5$. The true values of other parameters are kept at $\lambda_\psi = 8$ and $\lambda_y = 4$.

$\nu^{(0)}$	$\lambda_y^{(0)}$	$\lambda_\psi^{(0)}$	$\widehat{\nu}$	$\widehat{\lambda}_y$	$\widehat{\lambda}_\psi$
0.8	∞	2.419	0.988	∞	2.908
0.9	∞	2.688	0.988	∞	2.908
1.0	12.038	3.549	1.253	4.032	7.941
1.1	6.181	4.876	1.253	4.032	7.941
1.2	4.511	6.696	1.253	4.032	7.941
1.3	3.729	9.209	1.253	4.032	7.941
1.4	3.281	12.708	1.253	4.032	7.941
1.5	2.993	17.611	1.253	4.032	7.941
1.6	2.794	24.528	1.253	4.032	7.941

$\nu^{(0)}$	$\lambda_y^{(0)}$	$\lambda_\psi^{(0)}$	$\widehat{\nu}$	$\widehat{\lambda}_y$	$\widehat{\lambda}_\psi$
0.8	∞	1.542	1.229	∞	2.988
0.9	∞	1.932	1.229	∞	2.988
1.0	∞	2.287	1.229	∞	2.988
1.1	∞	2.611	1.505	4.004	8.030
1.2	17.534	3.259	1.505	4.004	8.030
1.3	7.325	4.385	1.505	4.004	8.030
1.4	5.043	5.887	1.505	4.004	8.030
1.5	4.043	7.902	1.505	4.004	8.030
1.6	3.485	10.620	1.505	4.004	8.030

run the trust region iterations, the algorithm makes subsequence adjustments to get close to the true values. (c) When the initial value of $\nu^{(0)}$ is really far from the true ν value, the non-convex nature of the REML likelihood function takes over, and in some instances, we see that trust region algorithm converges to final values that are far from the true parameter values. Furthermore, it is just that only the global maximum is in the interior of the parameter space and is very close to the true parameter values. However, the second local maximum, found from poor starting points, falls on the boundary of the parameter space with $\widehat{\lambda}_y = \infty$.

5.2. Numeric consistency towards geostatistical models

The purpose of our next computer experiment is to demonstrate that using lattice-based approximations we can obtain inference for the continuum Matérn dependence plus the white noise structures, even when observations are made at irregular sampling locations. The basic idea is to embed the study region and irregular sampling locations into finer and finer lattice arrays by diminishing the lattice spacing and then check for numeric consistency (in terms of convergence in distribution of the estimators of dependence parameters) as we fit the approximations of the continuum Matérn dependence with nugget effect on finer and finer lattices. To this end, we consider the unit square $(0, 1) \times (0, 1)$ as

TABLE 2
Estimates of the precision parameters for different array sizes. The standard errors are shown in parenthesis.

r	$\widehat{\nu}$	$\widehat{\lambda}_y$	$\widehat{\lambda}_\psi$
100	1.301 (0.044)	1.391 (0.068)	3.395 (0.374)
200	1.224 (0.030)	1.040 (0.039)	3.540 (0.355)
300	1.233 (0.028)	0.987 (0.035)	4.322 (0.464)
400	1.226 (0.026)	0.973 (0.034)	4.946 (0.552)
500	1.242 (0.025)	0.956 (0.033)	5.848 (0.691)

the study region and pick $20,000$ points uniformly within this study region as sampling locations. At these randomly generated irregular sampling locations, we then draw a realization from the intrinsic Matérn process with $\nu = 1.25$ and $\sigma^{-2} = 2/(4\pi^2)$. We also add a random noise to each observations by generating i.i.d. standard normal random variables. Next, we embed the unit square and the irregular sampling points in an $r \times r$ regular square array and estimate parameters using methods detailed in Sections 2, 3 and 4. Table 2 provides these estimates along with their standard errors for various values of r.

The results vindicate the theoretical findings in Section 2. First, the estimates of ν hover around the true value 1.25 as the lattice spacing decreases. Second, the sample size is fixed, but with diminishing lattice spacing, standard errors for $\widehat{\nu}$ are essentially constant, which endorse that the procedure is converging in distribution and approximating geostatistical inference. Third, as lattice spacing decreases, $\widehat{\lambda}_y$ gets closer to the true value 1. The slight difference between the lattice-based nugget effect $\widehat{\lambda}_y$ and the geostatistical nugget 1 occurs because of (essentially) constant difference between the variograms of the intrinsic Matérn process and the fractionally differenced process. Fourth, the estimates $\widehat{\lambda}_\psi$ increase with the diminishing lattice spacing. But this increase and its exact nature can be explained by the approximation theory detailed in Section 2. Specifically, it follows from the scaling limit equations (3) that at lattice spacings $1/m$ and $1/m'$ the ratio $(\sigma_m^2/\sigma_{m'}^2)(m/m')^{2\nu-2}$ should be close to 1. Thus, for example, between arrays of size 400×400 and 500×500, this ratio is seen to be $(4.946/5.848) \times (500/400)^{0.5} = 0.946$. Overall, the estimates and the standard errors are indicative of numeric convergences of estimators in distribution, and, it can be seen that after appropriate rescaling of parameters, we can obtain geostatistical inference from the lattice-based approximations.

5.3. Efficiency of matrix-free REML estimators

In classical statistics, efficiency of an estimator is often judged based on the variance of the estimator and both the Fisher information matrix and the Cramėr–Rao lower bound play a central role in assessing the efficiency of an estimator. In particular, for linear mixed models, it is typical that REML estimators are asymptotically efficient and achieve the Cramėr–Rao lower bound. The question that we ask here is how efficient are our matrix-free estimators. We address

TABLE 3
Monte Carlo standard deviation and average of the standard errors of matrix-free REML estimates when $\nu = 1.25$.

Estimate	Mean	Monte Carlo SD	Average of SE
$\widehat{\nu}$	1.251	0.022	0.021
$\widehat{\lambda}_y$	3.997	0.157	0.151
$\widehat{\lambda}_\psi$	8.055	0.539	0.517

TABLE 4
Monte Carlo standard deviation and average of the standard errors of matrix-free REML estimates when $\nu = 1.75$.

Estimate	Mean	Monte Carlo SD	Average of SE
$\widehat{\nu}$	1.751	0.018	0.018
$\widehat{\lambda}_y$	3.998	0.119	0.119
$\widehat{\lambda}_\psi$	8.032	0.380	0.372

this question using another set of computer experiments where we generate 1000 Monte Carlo samples from the spatial mixed linear model (1) on 256×256 arrays by setting $\mathbf{T} = \mathbf{0}$, $\boldsymbol{\tau} = \mathbf{0}$, $\lambda_\psi = 8$, $\lambda_y = 4$ and using $\nu = 1.25$. As in Section 5.1, this is done in several steps. First, for each Monte Carlo sample, a random effect $\boldsymbol{\psi}$ is generated on a 256×256 array from a Gaussian distribution with mean $\mathbf{0}$ and a precision matrix $8\mathbf{W}^\nu$ with the sum constraint $\boldsymbol{\psi}^T\mathbf{1} = 0$. Then, each $\psi_{u,v}$ value, independent of others, is removed randomly with probability $p = 0.6$, and a Gaussian white noise with precision $\lambda_y = 4$ is added to the remaining ones. For each Monte Carlo sample, this produces a random incidence matrix $\mathbf{F}$ and a vector of realization $\mathbf{y}$ with an expected sample size of $n = 39322$. For each Monte Carlo sample, we then apply our matrix free computations and obtain the corresponding REML estimates $\widehat{\nu}$, $\widehat{\lambda}_y$ and $\widehat{\lambda}_\psi$ and their standard errors by calculating the inverse of observed Fisher information matrix in a matrix-free way. Table 3 provides the Monte Carlo summaries of these estimates and their standard errors. It can be seen that the estimates are very accurate, and the Monte Carlo standard deviations of the estimators match very well with the corresponding Monte Carlo averages of the standard errors of estimates. (There is a slight discrepancy which occurs because only 50 Rademacher vectors are used in approximating the trace terms and this introduces a very slight additional variation in the estimates).

We next repeat the above simulation experiment with $\nu = 1.75$ and report the summaries in Table 4. Again we see that Monte Carlo standard deviations of the estimators match very well with the corresponding Monte Carlo averages of the standard errors of estimates. These confirm that there is little loss of statistical efficiency when computations are implemented in a matrix-free way.

The above summaries of computer simulations along with those from Section 5.2 now suggest that we can use lattice-based fractional random fields as proxies for continuum Matérn random fields and when we do so we can still achieve geostatistical inference with little loss of statistical efficiency.

TABLE 5

Run times in seconds of our REML method (with 30 Rademacher vectors) and INLA. The results of REML and INLA are obtained by running INLA (version 0.0-1420281647) on a laptop with Intel® Core i7-4700MQ processor and 8GB RAM. The results for INLA** are obtained by running the same version of INLA on a workstation with two Intel® Xeon® E5430 cpus and 32GB RAM. The results are rounded to the next second.*

Array size ($r \times r$)	REML	INLA*	INLA**
200×200	26	23	48
300×300	54	75	148
400×400	92	261	471
500×500	143	Out-of-memory	858
600×600	206	Out-of-memory	1962
700×700	277	Out-of-memory	3735

5.4. Computational times and practical gains

We now demonstrate a practical gain of our matrix-free method by comparing our actual run times with those of Lindgren et al. (2011) based on iterated nested Laplacian approximations (INLA). It must be noted that the current version of INLA (numbered 0.0-1420281647) can only fit a Matérn dependence model for the first few integer values of the dependence parameter ν. Furthermore, INLA is a Bayesian method where the estimates of parameters can be very sensitive to the choice of prior. Thus, in order to keep the comparison fair, we only consider the run times for the case $\nu = 1$. In a fashion identical to Section 5.1, we simulate observations on $r \times r$ arrays with $\nu = 1$, $\lambda_y = 4$ and $\lambda_\psi = 8$, and, discard 40% as missing. We then estimate these parameters using our matrix-free method and INLA.

Table 5 provides the average run times in seconds of these methods. Specifically, the second column indicates run times for our matrix-free method, and the third and the fourth columns show run times for INLA. These are obtained respectively by running INLA on (1) a laptop with a Intel®i7-4700MQ processor and 8GB of RAM on Linux operating system and (2) a workstation with two Intel®Xeon®E5430 processors and 32GB RAM on a Linux operating system. Furthermore, our matrix-free REML computations are done with some preliminary codes running on MATLAB 8.4 using only a single core of the laptop, while INLA is granted to use four cores on either machine.

We see that the computation time of our matrix-free REML algorithm scales very well with increasing dimension and does so at an approximate rate $O(n \log n)$ as suggested by the theory in Section 4. Furthermore because RAM requirement of our matrix-free algorithms scales only linearly with the array size, we find that these computations are possible on an ordinary laptop for moderately large array dimension. In contrast, we see that the computation times of INLA algorithm scales poorly with the array size. And to make things worse, INLA runs out of memory on the laptop even for a moderate array size. Only when it is run on the workstation, it can handle moderately large arrays. Thus, we conclude that we save both time and storage space.

Although we used only a single core to carry out the matrix-free REML optimization, our method can be made much faster with parallel implementation. Furthermore, we can deal with non-integer values of ν, which the sparse computations of Lindgren et al. (2011) can not.

6. Arsenic mapping in Bangladesh

Arsenic contamination of the groundwater in Bangladesh (and in West Bengal, India) is a serious problem with approximately one in five of the wells used for drinking water currently contaminated with arsenic above the government's drinking water standard (50 mg/L). Following an agreement with the Government of Bangladesh, the British Geological Survey (BGS) and the Department of Public Health Engineering (DPHE) conducted an extensive investigation of the arsenic problem during the period 1998 to 2001 and their website http://www.bgs.ac.uk/research/groundwater/health/arsenic/Bangladesh/data.html provides a rich treasury of groundwater arsenic concentration data. See also BGS and DPHE (2001) for a full report of the survey. The primary focus of the survey was to assess the scale of the groundwater arsenic contamination so that appropriate arsenic mitigation program could be developed. Consequently, a key component of the study was to construct an extensive geographic map of arsenic contamination. Over the years, numerous important works on the cause and damage of arsenic contamination and on the mechanism of arsenic mobilization came out of this study. These include Nickson et al. (2000), Chowdhury et al. (2000), McArthur et al. (2001), Smith et al. (2000), Harvey et al. (2002) and Neumann et al. (2009). Although till date the exact cause of arsenic contamination remained sketchy, scientists speculate that it is the supply of excess oxygen during pumping the tube wells that accelerates hydrolysis of precipitated arsenate and that releases soluble arsenous acid into groundwater.

Here, using methods developed in Sections 3, 4 and 5, we explore mapping groundwater arsenic concentration data collected by BGS and DPHE during 1998 and 1999. These data are from 3534 tube wells located at irregularly distributed sites at 61 out of 64 districts in Bangladesh and are measured in parts per billion (ppb which is same as micrograms per liter). The three districts in the south-eastern Bangladesh that are left out of the study are known to be arsenic safe. Overall the arsenic concentration measurements in the study region have a large range with values varying from less than 0.5 ppb to 1660 ppb, and a fraction of measurements are truncated on the right at the instrumental detection limit. The left panel of Figure 2 displays these data after grouping into different categories. Notice that, 0.5ppb is the machine detection limit, 10ppb is the WHO permissible limit, 50ppb is the Bangladesh Governtment permissible limit and 150ppb is believed to be the threshhold above which cancer mortality appears (Lamm et al., 2006). We see arsenic contamination is endemic in the middle southern part of Bangladesh. The south-west and the parts of northern Bangladesh region also have a patchy high arsenic concentration values. Overall, about 42% wells sampled had arsenic concentration more than the World

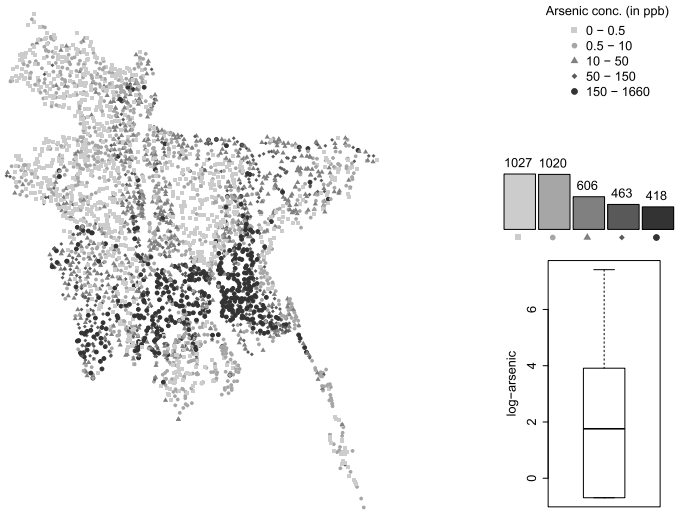

FIG 2. *Plot of the raw data. Group legends are described at the top right corner. The barplot gives the number of observations in each group and the box-plot of the log-arsenic concentration is shown at the bottom right panel.*

Health Organization's permissible limit of 10 ppb and about a quarter of the wells sampled had arsenic more than the Bangladesh Government permissible limit of 50 ppb.

In order to map groundwater log arsenic concentration, we then embed the study region with latitudes between 20 and 27 degrees north (approximately 778 km in length) and longitudes between 88 to 93 degrees east (approximately between 495 and 522 km in width) into a 500×300 array. Thus each array cell is approximately 2.64 square kilometers[1] in area and we tally arsenic concentrations in $n = 3224$ of these cells, (which is only about 2.15% of total array cells). Moreover, among these $n = 3224$ array cells, 2941 contain only one sampling location each, 258 contain two sampling locations each, 23 contain three sampling locations each and only 2 contain four sampling locations. Thus, due to clustering nature of the sampling locations, some array cells contained two or more sampling locations even after placing a such large array. To rectify this problem partially, we then take the average of logarithm of the arsenic concentrations over each cell and obtain our data vector $\mathbf{y}$. However, since a small fraction of cells contained multiple sampling locations, in our preliminary analysis we decided against adjusting the residual or nugget vector $\boldsymbol{\epsilon}$ using the number of sampling locations over which these averages are calculated. We then fit the mixed model (1) with $\mathbf{T} = \mathbf{0}$ to obtain BLUP of $\boldsymbol{\psi} - \bar{\psi}\mathbf{1}$ where $\mathbf{1}$ is the $rc-$vector of all ones, and to obtain REML estimates of λ_y, λ_ψ and ν.

First, we fit the data with $\nu = 1$. This is mainly because $\nu = 1$ corresponds to fitting a Gaussian autoregression, which is often the standard practice in spatial

[1]Computed from http://www.nhc.noaa.gov/gccalc.shtml.

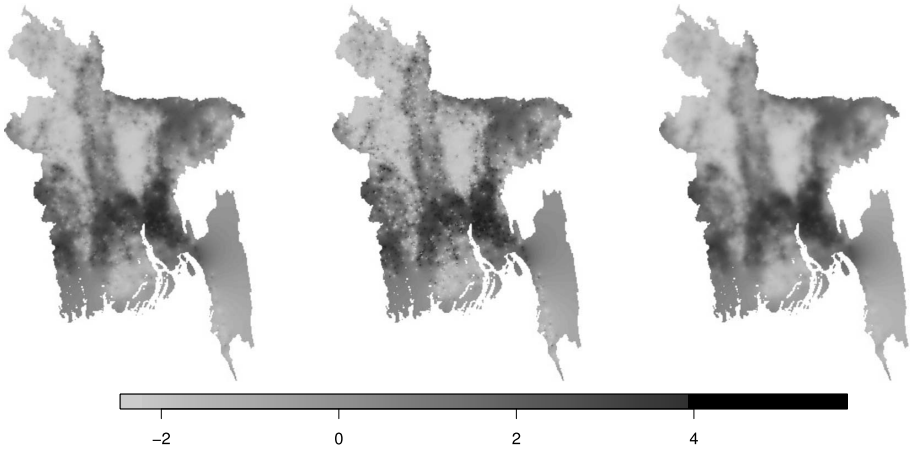

FIG 3. *The BLUPs of* $\boldsymbol{\psi}$. *Left: for* $\nu = 1$, *Center: for* $\widehat{\nu} = 0.858$ *(no nugget), Right: for* $\widehat{\nu} = 1.240$.

TABLE 6
Estimates of the precision parameters for the ground water log arsenic concentration data.

Initial value			Final Solution		
$\nu^{(0)}$	$\lambda_y^{(0)}$	$\lambda_\psi^{(0)}$	$\widehat{\nu}$	$\widehat{\lambda}_y$	$\widehat{\lambda}_\psi$
0.7	892.470	0.533	0.858	∞	0.660
0.9	1.449	0.983	1.240	0.596	4.607
1.1	0.715	2.399	1.240	0.596	4.607
1.3	0.567	6.143	1.240	0.596	4.607
1.5	0.508	16.688	1.240	0.596	4.607
1.7	0.477	48.634	1.240	0.596	4.607

statistics. To this end we fix a sample of 20 Rademacher variables and use the Lanczos algorithm and trust-region algorithm described in Section 3 and obtain global REML estimates $\widehat{\lambda}_y = 0.906$ (with a s.e. 0.074) and $\widehat{\lambda}_\psi = 1.527$ (with a s.e. 0.106). The left panel in Figure 3 provide the image plots of the BLUP of $\boldsymbol{\psi} - \bar{\psi}\mathbf{1}$ and the estimates of the residuals. These plots suggests that an intrinsic autoregression model captures the local variation of arsenic log-concentration fairly well. Overall the BLUP of $\boldsymbol{\psi} - \bar{\psi}\mathbf{1}$ highlights the high arsenic concentration areas in northern Bangladesh, central and southwest Bangladesh.

Next we fit the data with an arbitrary dependence parameter $\nu > 0$. Here, we use the same 20 Rademacher variables, but pursue REML computations using 11 different starting values of ν as in the simulation study. We summarize both the initial set of estimates and the final REML estimates in Table 6. As was the case with our simulation study, we found two local maxima; one corresponds to no nugget model with $\widehat{\lambda}_y = \infty$, $\widehat{\nu} = 0.858$ (with a s.e. 0.016) and $\widehat{\lambda}_\psi = 0.660$ (with a s.e. 0.020); and other with nugget effect where $\widehat{\lambda}_y = 0.596$ (with a s.e. 0.040), $\widehat{\nu} = 1.24$ (with a s.e. 0.054) and $\widehat{\lambda}_\psi = 4.607$ (with a s.e. 1.260). The center

TABLE 7

Final estimates (with nugget) of the precision parameters for the arsenic data on different resolutions. Here r denotes the number of grid cells along the latitude and c denotes the number of grid cells along the longitude.

r	c	$\widehat{\nu}$	$\widehat{\lambda}_y$	$\widehat{\lambda}_\psi$
350	210	1.315 (0.065)	0.595 (0.039)	4.477 (1.266)
400	240	1.143 (0.047)	0.737 (0.073)	2.392 (0.515)
450	270	1.379 (0.065)	0.547 (0.028)	7.688 (2.411)
500	300	1.240 (0.054)	0.596 (0.040)	4.607 (1.260)
550	330	1.322 (0.064)	0.606 (0.043)	5.310 (1.647)
600	360	1.329 (0.071)	0.588 (0.039)	6.661 (2.371)

and right panels in Figure 3 provide corresponding image plots of the BLUP of $\boldsymbol{\psi} - \bar{\psi}\mathbf{1}$ at these local maxima. Next we compute the difference of log REML function at these two local maxima and found this difference to be -49.82. This confirms that second set of the estimates are the global REML estimates and they corresponds to a model that includes nugget effects or residual values to capture the small scale variations.

Next we reanalyze the data at different lattice resolutions and check for numeric consistency of the results. In particular, we embed the study region and irregular sampling locations into finer and finer lattice arrays by diminishing the lattice spacing but by keeping the aspect ratio (r/c) of the arrays fixed. We then fit the lattice approximations of continuum Matérn dependence with nugget effect to the observed arsenic contamination data. Table 7 provides the estimates of the precision parameters, and their standard errors for various lattice sizes. We find that, at different lattice resolutions, there are little changes in the estimates of ν and λ_y after accounting for their uncertainties. The estimates of λ_ψ, however, increases with diminishing lattice spacings, but this increase can further be explained by the approximation theory laid out in Section 2. Specifically, the scaling limit equation (3) suggests that at lattice spacings $1/m$ and $1/m'$ the ratio $(\sigma_m^2/\sigma_{m'}^2) \times (m/m')^{2\nu-2}$ should be close to 1. Thus, for example, by inserting $\widehat{\nu} = 1.322$, in arrays of size 500×300 and 550×330, we find that this ratio is equal to 0.923. Overall, these results are largely consistent with what saw in computer generated experiments in Section 5. and we can conclude that the inference drawn from lattice-based approximations actually mimic the inference from the usual continuum intrinsic Matérn dependence structures with nugget effects.

We now extend the arsenic mapping problem further to study various covariate effects. It is believed that water from old and/or shallow tube wells tend to have higher arsenic concentration than water from new and deeper tube-wells. The plots of arsenic concentration against depth (in meters) and age (in years) shown in Figure 4 graphically supports this view. In other words, there is an apparent negative correlation between the arsenic concentration and the depth of the tube-wells and an apparent positive correlation between the arsenic concentration and the age of these tube-wells. Thus, we include the depth and the age of the tube-wells as covariate information into our spatial mixed linear model and make some preliminary investigation on the significance of these covariate

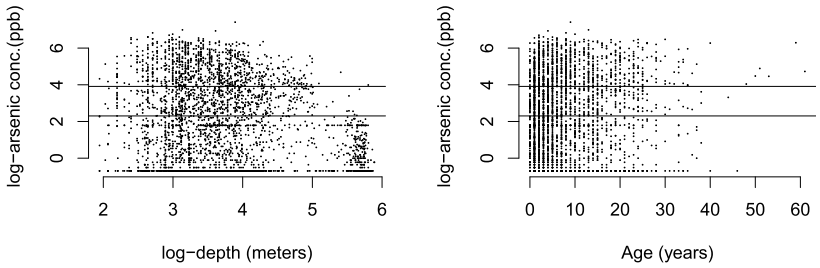

FIG 4. *Plot of arsenic log-concentration versus log-depth (left) and age (right) of tubewells in Bangladesh. The horizontal lines denote the permissible limits by WHO and Bangladesh Government (*$\log 10$ *and* $\log 50$*).*

TABLE 8
BLUEs of the covariate effects and REML estimates of the precision parameters for the ground water log arsenic concentration data.

r	c		log-depth	Age	ν	λ_y	λ_ψ
300	180	Estimate	-0.620	0.019	1.177	0.719	2.675
		Standard error	0.048	0.003	0.051	0.072	0.051
500	300	Estimate	-0.659	0.019	1.151	0.663	3.612
		Standard error	0.046	0.003	0.047	0.066	0.047
600	360	Estimate	-0.646	0.019	1.134	0.787	2.683
		Standard error	0.044	0.003	0.044	0.102	0.397

effects and robustness of such analyses to the changes of scales. To this end, Table 8 provides the BLUEs of the age and depth effects, REML estimates of the precision parameters, and the corresponding standard errors for two array sizes. These estimates are obtained by using the same sample of 20 Rademacher variables that we used before for the same array sizes to obtain our REML estimates for the kriging problem. As expected, we find that these covariate effects are significant. But the most striking observation here is that the BLUEs of the covariate effects appear to be robust to any changes of scales. This exact same phenomenon was observed also in Dutta and Mondal (2015a) in the context of agricultural variety trial, and, in the current context this observation reaffirms that even for a moderately small lattice spacings, a fractionally differenced random field plus a white noise on a lattice grid is a very good approximation to the intrinsic Matérn process plus a little white noise. This offers further justifications for the use of lattice-based approximations.

Finally, some words should be added on why we did not fit a stationary Matérn covariance to the arsenic concentration data and instead fit a limiting intrinsic version. This question of stationary vs. intrinsic has dominated many discussions of spatial statistics in the past; see e.g., Beran (1992), Besag and Kooperberg (1995), Besag (2002), McCullagh and Clifford (2006) and many others that address the inadequacy of stationary models in various spatial applications. In the arsenic example, we first need to acknowledge that these observations are not point-wise measurements. Rather they corresponds to average arsenic concentrations over water intake areas of the wells. This adds

difficulties to fitting a stationary Matérn covariance model. Second, even if we assume that data are point-referenced, we could not estimate MLEs of the covariance parameters as fitting a stationary Matérn model to these data runs into boundary and other numerical problems. Similar to the computer experiments in Section 5.1, we tried to explore basins of attraction to local maxima by doing a grid search on few parameters. But these efforts proved futile and the log-likelihood surface revealed a long flat ridge. Presence of such flat ridges not only complicate the numerical optimization problem but is also indicative of parameter non-interpretability and redundancy.

7. Discussion

We open the discussion by noting the practical benefits of implementing our lattice-based approximations of continuum spatial models. As we have seen through simulations and REML analysis of arsenic contamination, there is little loss in turning a geo-statistical dataset into an areal one by embedding the study region on to a fine regular grid. In this context, we must recognize that we never measure a spatial random variable at a point in space, but as an average value (or an integral) over a small non-null, perhaps infinitesimal, region. Thus, conceptually, there is no problem in discretizing the space, and treating the observed values as averages over discretized regions. Furthermore, in most spatial applications, the scale of sampling is neither infinitesimally small nor infinitely great. Thus, in a range of applications, it should not be difficult to implement a lattice-based model at a reasonably fine grid and still obtain the same inference that we would have obtained had we fit the corresponding continuum geostatistical model.

As statisticians our rule of computations is simple. We should pursue exact computations when possible. When we can not pursue exact computations directly (for example if the sample size is large), we can look for alternatives that will give us answers that are as good as exact answers. In this context, our matrix-free scalable REML computations are relevant, as it is difficult to pursue exact REML computations in large datasets. Numerical experiments in Section 5.3 show that there is little loss of statistical efficiency due to our matrix-free scalable computations and we can obtain answers that are as good as exact answers.

We believe that the fractional Laplacian differenced random fields on regular arrays (that are presented in this paper) are of interest on their own. Lattice based models, particularly Markov random fields, have played a fundamental role in the development of spatial statistics and the use of fractional differencing will widen the overall scope of lattice models.

In the current paper, we did not consider any stationary models, but they deserve some attention here. Typically, for stationary models, one can pursue REML computations first by embedding the sampling locations in a finer rectangular grid and then embedding the rectangular grid into a much larger torus lattice using block circulant embedding, as proposed in Dietrich and Newsam (1993, 1997) and Wood and Chan (1994). This allows for the use of station-

ary models on a torus lattice, where one can take computational advantages of fast Fourier transform (Besag and Moran, 1975) and derive matrix-free scalable REML computations within our h-likelihood framework. From a computational point of view, this h-likelihood framework will be easier to implement than directly maximizing the REML function by adapting the works of Anitescu et al. (2012). Furthermore, the h-likelihood framework will allow us to characterize non-convexity in the optimization in terms of gamma non-linear models and one will be able to obtain useful circulant or other preconditioners. However, the main advantage here is that these computations will be not just matrix-free and scalable but also statistically efficient (in terms of achieving Cramer–Rao lower bound) and thus will yield qualitatively better estimates than those by Fuentes (2007).

In the above context, we must note that, even for stationary models, to date there is no known Whiitle likelihood method that can deal with irregularly sampled observations and produce asymptotically efficient estimators. Thus, the work of Anitescu et al. (2012), Dutta and Mondal (2015a), this paper and the computations mentioned above for the stationary case are a significant step forward.

Furthermore, if, in some applications, stationary Matérn model is of interest, one can also pursue lattice-based approximations. Specifically, as in Mondal (2011), we can consider a sequence of Gaussian random fields on $\mathcal{Z}_m^2$ with spectral densities

$$\tilde{f}_m(\omega,\eta) = \frac{\sigma_m^2}{m^2\left[1-4\beta_m+4\beta_m\{\sin^2(\frac{1}{2m}\omega)+\sin^2(\frac{1}{2m}\eta)\}\right]^\nu},$$

with $\omega,\eta \in (-m\pi, m\pi]$. In this case, we need to take, as $m\to\infty$, $\beta_m \uparrow 1/4$, $4m^2(1-4\beta_m)\to\kappa^2>0$ and $m^{\nu-1}\sigma_m \to \sigma/2^\nu$. It then follows that $\tilde{f}_m(\omega,\eta)$ converges to

$$\tilde{f}(\omega,\eta) = \frac{\sigma^2}{(\kappa^2+\omega^2+\eta^2)^\nu}, \quad \sigma>0, \quad \kappa>0,$$

which is the spectral density formula of the continuum Gaussian stationary Matérn random field. On finite rectangular lattice, we can then approximate the precision matrix by $\lambda_\psi\{(1-4\beta)I+4\beta W\}^\nu$ and pursue our matrix-free REML computations. Notice that this precision matrix has a finite condition number and hence the computational complexity becomes only $O(n(\log n)^2)$, even without a preconditioner. The strength and weakness of such approximations will be a matter of future research.

In arsenic data example, Gaussian distribution is used as an approximation, as only a small fraction of the data is truncated to the right. If there is a concern that this right-censoring can affect the Gaussianity assumption, one can further consider a non-linear model using truncated Gaussian distribution and obtain appropriate statistical estimation and inference. In this regard, it must be noted that REML estimation applies only for Gaussian models and not for truncated Gaussian models. Thus, although REML estimation can not be extended to a

non-linear truncated Gaussian model, one can still derive meaningful statistical estimation and inference extending the computations we have developed in this paper.

It is worthwhile to mention that there are many interesting directions in which we can take our research forward in the future. One direction is to develop matrix-free methods for conditional simulations of the spatial effect $\boldsymbol{\psi}$. The works of Borici (2000), Schneider and Willsky (2003), Parker and Fox (2012) and Aune et al. (2013) have paved ways for matrix-free Lanczos methods for solving equations of the form $\mathbf{A}^{1/2}\mathbf{x} = \mathbf{b}$, where $\mathbf{A}$ is a positive definite matrix that can be multiplied with a vector in matrix-free way in less than $O(n^2)$ computations. By applying these works, we can thus advance conditional simulations of $\boldsymbol{\psi}$. Specifically note that the conditional precision matrix of $\boldsymbol{\psi}$ is $\lambda_y \mathbf{F}^T\mathbf{F} + \lambda_\psi \mathbf{W}^\nu$, which can be multiplied with a vector in matrix-free way in $O(n \log n)$ computations. Thus such conditional simulations will allow us to make various statistical inference about the underlying latent spatial random field. At the same time, they would help us develop novel Bayesian computations, as such simulations can be implemented for block Gibbs updates or within Matropolis-Hasting steps or other Markov chain Monte Carlo computations.

Another direction will be to combine the current work with that of Mondal (2013) to develop fractionally differenced conditional autoregressive spatial models. This will allow for advancement of Box–Jenkins type methodology in spatial statistics.

A third direction is to develop more complex spatial models that can accommodate anisotropy and heterogeneity. In reality, there are several reasons why substantial anisotropy and heterogeneity may be present in arsenic concentration. For example, aquifers may not be homogeneous, the ages and the depths of the wells can vary from place to place. Similarly, demographic variables such as population density, agricultural practices and industrial variables may also affect local arsenic contamination. Furthermore, it is typical that older or deeper wells supply more oxygen to the aquifers and accelerate hydrolysis arsenates to a greater extent. It is also typical that the underground water usage is high where population density is high or where the land is used for agriculture and these factors have an adverse effect on contamination. Thus it would be of interest to think how we can collect more covariate information and how we can develop a more elaborate spatial model that can accommodate different sources of anisotropy and heterogeneity.

Appendix: Variogram calculations when $\nu \geq 2$

Consider a zero mean stationary Gaussian Matérn process $Z(x)$ on $\mathcal{R}^2$ with the covariance function $\{2\pi\sigma^2/\Gamma(\nu)\}\{\|\mathbf{x}\|/2\kappa)\}^{\nu-1}K_{\nu-1}(\kappa\|\mathbf{x}-\mathbf{y}\|)$, for some $\nu \geq 2$, $\kappa > 0$ and $\sigma > 0$. Take μ_1 and μ_2 to be finitely supported signed measures such that

$$\int x_1^{i_1} x_2^{i_2} \mu_1(dx) = \int x_1^{i_1} x_2^{i_2} \mu_2(dx) = 0, \quad i_1, i_2 \geq 0,\ i_1 + i_2 \leq \lfloor \nu - 1 \rfloor.$$

Next, define higher-order contrasts or differences by $Z(\mu_j) = \int Z(\mathbf{x})\mu_j(\mathbf{dx})$. Consequently, we obtain

$$\operatorname{cov}(Z(\mu_1), Z(\mu_2)) = \int\int \frac{2\pi\sigma^2}{\Gamma(\nu)}\Big(\frac{\|\mathbf{x}\|}{2\kappa}\Big)^{\nu-1} K_{\nu-1}(\kappa\|\mathbf{x}-\mathbf{y}\|)\mu_1(\mathbf{dx})\mu_2(\mathbf{dy}). \tag{12}$$

As $x \to 0+$, equations (9.6.2) and (9.6.10) of Abramowitz and Stegun (1972) give

$$x^{\nu-1}K_{\nu-1}(x) = \frac{2^{\nu-1}\pi}{-\sin(\nu\pi)}\sum_{j=0}^{k}\frac{(x/2)^{2j}}{j!\Gamma(j+2-\nu)} + \frac{2^{-\nu}\pi}{\sin(\nu\pi)\Gamma(\nu)}x^{2(\nu-1)} + \mathcal{O}(x^{2k+2}),$$

where $k = \lfloor \nu - 1 \rfloor$, when ν is not an integer. Similarly, when ν is an integer, equation (9.6.11) of Abramowitz and Stegun (1972) give

$$\begin{aligned} x^{\nu-1}K_{\nu-1}(x) = {} & 2^{\nu-2}\sum_{j=0}^{\nu-2}\frac{(\nu-j-2)!}{j!}(-\tfrac{1}{2}x)^{2j} \\ & + (-1)^{\nu-1}\frac{\Psi(1)+\Psi(\nu)+\log 4}{2^{\nu}(\nu-1)!}x^{2(\nu-1)} \\ & + \frac{(-1)^{\nu}}{2^{\nu-1}(\nu-1)!}x^{2(\nu-1)}\log x + o(x^{2\nu}), \quad \text{as } x \to 0, \end{aligned}$$

where Ψ is the digamma function. It then follows that, as $\kappa \to 0$, the covariance in (12) converges to

$$\frac{4^{1-\nu}\pi^2\sigma^2}{\Gamma(\nu)^2\sin(\nu\pi)}\int\int \|\mathbf{x}-\mathbf{y}\|^{2\nu-2}\mu_1(\mathbf{dx})\mu_2(\mathbf{dy})$$

when $\nu > 2$ is not an integer, and to

$$(-1)^{\nu}\frac{4^{1-\nu}\pi\sigma^2}{\Gamma(\nu)^2}\int\int \|\mathbf{x}-\mathbf{y}\|^{2\nu-2}\log\|\mathbf{x}-\mathbf{y}\|^2\mu_1(\mathbf{dx})\mu_2(\mathbf{dy})$$

when $\nu \geq 2$ is a positive integer. We refer to Matheron (1973) further discussions on intrinsic spatial models. Since the limit $\kappa \to 0+$ corresponds to an intrinsic Matérn random field with spectral density (4), we can apply the above results to compute variogram of higher-order contrasts of $Z(u, v)$. Specifically, if $2 < \nu < 4$, we can exactly compute

$$\gamma_D(s,t) = \operatorname{var}\{DZ(s,t) - DZ(0,0)\}$$

by appropriately defining two finitely supported signed measures μ_1 and μ_2. Furthermore, we can compute

$$\gamma_{D,m}(s,t) = \operatorname{var}\{DZ^{(m)}(s,t) - DZ^{(m)}(0,0)\}$$

by numerically computing

$$\sigma_m^2 \int_{-m\pi}^{m\pi} \int_{-m\pi}^{m\pi} \left(\sin^2(\tfrac{1}{2}\omega) + \sin^2(\tfrac{1}{2}\eta)\right)^2 \frac{1-\cos(s\omega)\cos(t\eta)}{4\pi^2 m^2 \left(\sin^2 \frac{\omega}{2m} + \sin^2 \frac{\eta}{2m}\right)^\nu} d\omega d\eta.$$

We refer to the supplement of the paper for numeric comparisons of $\gamma_D(s,t)$ and $\gamma_{D,m}(s,t)$.

Acknowledgement

The work is supported by NSF Career award DMS-1519890.

Supplementary Material

Supplementary to "REML estimation with intrinsic Matérn dependence in the spatial linear mixed model"
(doi: 10.1214/16-EJS1125SUPP; .pdf).

References

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of mathematical functions*, **1**, New York: Dover.

Anitescu, M., Chen, J. and Wang, L. (2012). A matrix-free approach for solving the Gaussian process maximum likelihood problem, *SIAM Journal on Scientific Computing,* 34, A240–A262. MR2890265

Aune, E., Eidsvik, J. and Pokern, Y. (2013). Iterative numerical methods for sampling from high dimensional Gaussian distributions. *Statistics and Computing*, **23**(4), 501–521. MR3070407

Aune, E., Simpson, D. P. and Eidsvik, J. (2014). Parameter estimation in high dimensional Gaussian distributions. *Statistics and Computing*, **24**(2), 247–263. MR3165552

Beran, J. (1992). Statistical methods for data with long-range dependence. *Statistical science*, 404–416.

Besag, J. E. (2002). Discussion on the paper "What is a Statistical Model?" *The Annals of Statistics,* **30**(5), 1267–1277.

Besag, J. E. and Mondal, D. (2005). First–order intrinsic autoregressions and the de Wijs process. *Biometrika*, **92**, 909–920. MR2234194

Besag, J. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B*, **61**, 691–746. MR1722238

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions, *Biometrika,* 82, 733–746. MR1380811

Besag, J. E. and Moran, P. A. (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, **62**, 555–562. MR0391451

BGS and DPHE. (2001). Arsenic contamination of groundwater in Bangladesh, Kinniburgh, D. G. and Smedley, P. L. (Editors). *British Geological Survey Technical Report WC/00/19.* British Geological Survey: Keyworth.

Borici, A. (2000). A Lanczos approach to the inverse square root of a large and sparse matrix. *Journal of Computational Physics*, **162**, 123–131. MR1772448

Chowdhury, U. K., Biswas, B. K., Chowdhury, T. R., Samanta, G., Mandal, B. K., Basu, G. C., ... and Chakraborti, D. (2000). Groundwater arsenic contamination in Bangladesh and West Bengal, *India. Environmental health perspectives*, 108, 393.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, **19**(90), 297–301. MR0178586

Dahlhaus, R. and Künsch, H. (1987). Edge effects and efficient parameter estimation for stationary random fields. *Biometrika*, **74**, 877–882. MR0919857

Dietrich, C. R. and Newsam, G. N. (1997). Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, **18**, 1088–1107. MR1453559

Dietrich, C. R. and Newsam, G. N. (1993). A fast and exact method for multidimensional Gaussian stochastic simulations. *Water Resources Research*, **29**, 2861–2869.

Diggle, P. J., Ribeiro Jr., P. J. and Christensen, O. F. (2003). An introduction to model-based geostatistics. In *Spatial statistics and computational methods*. Springer New York, 43–86. MR2001385

Diggle, P. J., Menezes, R. and Su, T. L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C*, **59**, 191–232. MR2744471

Duffin, R. J. (1953). Discrete potential theory. *Duke Mathematical Journal*, **20**, 233–251. MR0070031

Duffin, R. J. and Shaffer, D. H. (1960). Asymptotic expansion of double Fourier transforms. *Duke Mathematical Journal*, **27**, 581–596. MR0117501

Dutta, S. and Mondal, D. (2015b). Variogram calculations for random fields on regular lattices using quadrature methods. *To appear in Environmetrics.*

Dutta, S. and Mondal, D. (2015a). An h-likelihood method for spatial mixed linear model based on intrinsic autoregressions, *Journal of the Royal Statistical Society: Series B*, **77**(3), 699–726. MR3351451

Frigo, M. and Johnson, S. G. (2005). The design and implementation of FFTW3. *Proceedings of the IEEE*, **93**(2), 216–231.

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, **102**, 321–331. MR2345545

Gay, R. and Heyde, C. C. (1990). On a class of random field models which allows long range dependence. *Biometrika*, **70**(2), 401–403. MR1064814

Guttorp, P. and Gneiting, T. (2006). Studies in the history of probability and statistics XLIX On the Matérn correlation family. *Biometrika*, **93**(4), 989–995. MR2285084

Guyon, X. (1982). Parameter estimation for a stationary process on a d-

dimensional lattice. *Biometrika*, **69**, 95–105. MR0655674

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Harvey, C. F., Swartz, C. H., Badruzzaman, A. B. M., Keon-Blute, N., Yu, W., Ali, M. A., ... and Ahmed, M. F. (2002). Arsenic mobility and groundwater extraction in Bangladesh. *Science*, **298**(5598), 1602–1606.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**(2), 423–447.

Henderson, C. R. (1950). Estimation of genetic parameters. *Annals of Mathematical Statistics.* **21**, 309–310.

Hosking, J. R. (1981). Fractional differencing. *Biometrika*, **68**(1), 165–176. MR0614953

Kelbert, M. Y., Leonenko, N. N. and Ruiz-Medina, M. D. (2005). Fractional random fields associated with stochastic fractional heat equations. *Advances in Applied Probability*, **37**(1), 108–133. MR2135156

Kent, J. T. and Mardia, K. V. (1996). Spectral and circulant approximations to the likelihood for stationary Gaussian random fields. *Journal of statistical planning and inference*, **50**, 379–394. MR1394139

Lamm, S. H., Engel, A., Penn, C. A., Chen, R. and Feinleib, M. (2006). Arsenic cancer risk confounder in southwest Taiwan data set. *Environ. Health Perspect.* **114**(7): 1077–1082.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion), *Journal of Royal Statistical Society: Sereis B*, **58**, 619–678. MR1410182

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006. MR1872215

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

Makhoul, J. (1980). A fast cosine transform in one and two dimensions. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **28**(1), 27–34.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**, 135–146. MR0738334

Matérn, B. (1960). Spatial variation – stochastic models and their application to some problems in forest surveys and other sampling investigations,*Stockholm: Medd. Statens Skogsforskningsinstitut,* **49**(5), 144. MR0169346

Matheron, G. (1973). The intrinsic random functions and their applications, *Advances in applied probability*, **5**, 439–468. MR0356209

Matheron, G. (1971). *The theory of regionalized variables and its applications.* Vol. 5. Ecole nationale supérieure des mines de Paris.

McCullagh, P. (2002). What is a statistical model? *Annals of statistics,* **30**(5), 1225–1267. MR1936320

McCullagh, P. and Clifford, D. (2006). Evidence for conformal invariance of crop yields. *Proceedings in Royal Society London. Ser. A, mathematical, physical, and engineering sciences*, **462**, 2119–2143.

McArthur, J. M., Ravenscroft, P., Safiulla, S. and Thirlwall, M. F. (2001). Arsenic in groundwater: testing pollution mechanisms for sedimentary aquifers in Bangladesh. *Water Resources Research*, **37**(1), 109–117.

Minasny, B. and McBratney, A. B. (2007). Spatial prediction of soil properties using EBLUP with the Matérn covariance function. *Geoderma,* **140**(4), 324–336.

Mondal, D. (2013). On edge corrections and the local quadratic property of intrinsic autoregressions. Under revision at *Biometrika*.

Mondal, D. (2011). Discussion on the paper An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *Journal of Royal Statistical Society: Series B*, **73**, 483

Mondal, D. (2005). Variogram calculations for first-order intrinsic autoregressions and the de Wijs process. Technical Report 479, Department of Statistics, University of Washington.

Moré, J. J. and Sorensen, D. C. (1983). Computing a trust region step, *SIAM Journal on Scientific and Statistical Computing*, **4**(3), 553–572. MR0723110

Neumann, R. B., Ashfaque, K. N., Badruzzaman, A. B. M., Ali, M. A., Shoemaker, J. K. and Harvey, C. F. (2009). Anthropogenic influences on groundwater arsenic concentrations in Bangladesh. *Nature Geoscience*, **3**(1), 46-52.

Nickson, R. T., McArthur, J. M., Ravenscroft, P., Burgess, W. G. and Ahmed, K. M. (2000). Mechanism of arsenic release to groundwater, Bangladesh and West Bengal. Applied Geochemistry, 15(4), 403-413.

Nocedal, J. and Wright, S. J. (1999). *Numerical optimization*, Springer-Verlag, USA. MR1713114

Pardo-Igúzquiza, E., Mardia, K. V. and Chica-Olmo, M. (2009). ML-MATERN: A computer program for maximum likelihood inference with the spatial Matérn covariance model. *Computers and Geosciences*, **35**, 1139–1150.

Parker, A. and Fox, C. (2012). Sampling Gaussian distributions in Krylov spaces with conjugate gradients. *SIAM Journal of Scientific Computations*, **34**, B312–B334. MR2970281

Powell, M. J. D. (1984). On the global convergence of trust region algorithms for unconstrained minimization. *Mathematical Programming*, **29**(3), 297–303. MR0753758

Powell, M. J. D. (1970). A Fortran subroutine for solving systems of nonlinear algebraic equations, In *Numerical Methods for Nonlinear Algebraic Equations*, (P. Rabinowitz, ed.), Ch.7. MR0343590

Rao, K. R. and Yip, P. (1990). *Discrete cosine transform, algorithms, advantages, applications*, Academic Press. MR1080969

Saad, Y. (1985). Practical use of polynomial preconditionings for the conjugate gradient method. *SIAM Journal on Scientific and Statistical Computing*, **6**(4), 865–881. MR0801178

Schneider, M. K. and Willsky, A. S. (2003). Krylov subspace method for covariance approximation and random processes and fields. *Multidimens. Syst. Signal Process.*, **14**, pp. 295–318. MR1992451

Smith, A. H., Lingas, E. O. and Rahman, M. (2000). Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. *Bulletin of the World Health Organization*, **78**(9), 1093–1103.

Stein, M. L. (1999). *Interpolation of spatial data. Some theory for Kriging.* Springer-Verlag. MR1697409

Wood, A. T. and Chan, G. (1994). Simulation of stationary Gaussian processes in $[0,1]^d$. *Journal of computational and graphical statistics*, **3**, 409–432. MR1323050