Accepted Manuscript

Flexible and efficient estimating equations for variogram estimation

Ying Sun, Xiaohui Chang, Yongtao Guan

PII: S0167-9473(18)30001-X
DOI: https://doi.org/10.1016/j.csda.2017.12.006
Reference: COMSTA 6541

To appear in: *Computational Statistics and Data Analysis*

Received date : 23 January 2017
Revised date : 20 December 2017
Accepted date : 30 December 2017

Please cite this article as: Sun Y., Chang X., Guan Y., Flexible and efficient estimating equations for variogram estimation. *Computational Statistics and Data Analysis* (2018), https://doi.org/10.1016/j.csda.2017.12.006

Computational Statistics & Data Analysis **0** (2018), 000-000

FLEXIBLE AND EFFICIENT ESTIMATING EQUATIONS FOR VARIOGRAM ESTIMATION

Ying Sun, Xiaohui Chang * and Yongtao Guan

King Abdullah University of Science and Technology,

Oregon State University and University of Miami

Abstract:

Variogram estimation plays a vastly important role in spatial modeling. Different methods for variogram estimation can be largely classified into least squares methods and likelihood based methods. A general framework to estimate the variogram through a set of estimating equations is proposed. This approach serves as an alternative approach to likelihood based methods and includes commonly used least squares approaches as its special cases. The proposed method is highly efficient as a low dimensional representation of the weight matrix is employed. The statistical efficiency of various estimators is explored and the lag effect is examined. An application to a hydrology dataset is also presented.

Key words and phrases: Estimating equations; Lag effect; Low rank approxima-

*Correspondence to: Xiaohui Chang, College of Business, Oregon State University, Corvallis, OR 97331, USA. Tel.: +1 541 737 2551; fax: +1 541 737 4890. E-mail address: xiaohui.chang@oregonstate.edu.

tion; Statistical efficiency.

1. Introduction

The variogram is a fundamental tool in the modeling of spatial processes. For a spatial stochastic process $Z(\mathbf{s})$ defined on $\mathcal{D} \subset \mathbb{R}^2$, if it has a constant mean: $\mathrm{E}(Z(\mathbf{s})) = \mu$ for all $\mathbf{s} \in \mathcal{D}$, and its variogram $2\gamma(\mathbf{s_1} - \mathbf{s}_2) \equiv \mathrm{Var}(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))$ for $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}$ only depends on the displacement between the locations, $\mathbf{s}_1 - \mathbf{s}_2$, then Z is *intrinsically stationary*. A process is *isotropic* if its variogram is dependent on the absolute distance, not the relative direction, between locations, i.e., $2\gamma(\mathbf{s}_1 - \mathbf{s}_2) = 2\gamma(||\mathbf{s}_1 - \mathbf{s}_2||)$ where $||\cdot||$ denotes the Euclidean distance, and *geometric anisotropic* if $Z(\mathbf{Hs})$ is isotropic for some invertible matrix $\mathbf{H} \neq \mathbf{I}$. The variogram is widely used to quantify the spatial variability of many physical processes and is essential for obtaining accurate spatial predictions. Some examples are from mining (Matheron, 1963; Journel and Huijbregts, 1978), forestry (Moeur, 1993), hydrogeology (Kitanidis, 1997; Yu, Harvey, and Harvey, 2003), soil science (Heuvelink and Webster, 2001; McGrath, Zhang, and Carton, 2004), epidemiology (Kleinschmidt et al. (2000); Wong, Yuan, and Perlin (2004)), meteorology (Haylock et al., 2008) and many others (Isaaks and Srivastava, 1989; Cressie, 1993; Webster and Oliver, 2007). In practice, the variogram is usually unknown and needs to be estimated from the data.

To model the spatial relationships, especially for spatial predictions, it is crucial to obtain variogram estimates with high statistical efficiency and also low variance.

A graphical assessment of the variogram estimate can be achieved using a variogram cloud, that is a plot of squared differences of observations versus pair-wise distances. The variogram cloud has the advantage of displaying the individual contributions to the overall variogram from all pair-wise distances (Müller, 1999). Compared to a variogram cloud, the empirical variogram is widely used in variogram estimation because it is more robust to modeling assumptions and provides more information in identifying an appropriate model form for the variogram. The classical empirical variogram estimator is based on the method of moments (Matheron, 1962),

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_1) - Z(\mathbf{s}_2))^2,$$

where $N(\mathbf{h}) = \{(\mathbf{s}_1, \mathbf{s}_2) \in \mathcal{D} \subset \mathbb{R}^2 : \mathbf{s}_1 - \mathbf{s}_2 = \mathbf{h}\}$ with cardinality $|N(\mathbf{h})|$. To eliminate the constraint of considering pairs of sites with exactly $\mathbf{h}$ lag apart, some modifications of the classical empirical variogram are proposed to use a tolerance region or group distances into bins (Cressie, 1993). There are also other variations of the empirical variogram to take care of the susceptibility to outliers through robust variogram estimators based on $|Z(\mathbf{s}_1) - Z(\mathbf{s}_2)|^{1/2}$ (Cressie and Hawkins, 1980) or other highly robust scale estimators Genton

(1998a).

Fitting the empirical variogram directly using nonparametric methods may not necessarily produce a valid variogram. In practice, variogram estimation is usually done by first assuming a parametric form that is known to be conditionally negative definite and then estimating the parameters to ensure that the estimated variogram is close to the empirical variogram of the data. Refer to Cressie (1993) for a list of popular choices of variogram models.

Least squares and likelihood based methods are the two approaches that are commonly adopted for fitting variogram models to spatial data. The least squares method fits a parametric model by minimizing a quadratic distance measure between the empirical variogram $2\hat{\gamma}(\mathbf{h})$ (or some other nonparametric variogram estimators) and the parametric model $2\gamma(\mathbf{h}, \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^p$ denotes p unknown parameters. More specifically, for a variogram estimated at K discrete lags, the parameter vector $\boldsymbol{\theta}$ is estimated by minimizing

$$\{2\hat{\boldsymbol{\gamma}} - 2\boldsymbol{\gamma}(\boldsymbol{\theta})\}' V(\boldsymbol{\theta})^{-1} \{2\hat{\boldsymbol{\gamma}} - 2\boldsymbol{\gamma}(\boldsymbol{\theta})\},$$

where $2\hat{\boldsymbol{\gamma}}$ is a $K \times 1$ vector of the empirical variogram at K lags, $V(\boldsymbol{\theta})$ is a $K \times K$ non-negative definite matrix characterizing the dependence structure and different variabilities of $2\hat{\boldsymbol{\gamma}}$ among the K lags.

The weights assigned to different lags are embedded in the inverse matrix $V(\boldsymbol{\theta})^{-1}$, and how these weights are allocated is in fact directly linked with the type of least squares estimation. For example, when the empirical variogram values are assumed to be uncorrelated with same variance, $V(\boldsymbol{\theta})$ reduces to the identity matrix, corresponding to having equal weight to all lags and leading to the ordinary least squares (OLS) estimate. The weighted least squares (WLS) estimate is when $V(\boldsymbol{\theta})$ is chosen to be some diagonal matrix with specified variances along the diagonal; see Cressie (1985) and Genton (1998c), among others, for different choices of weights. The generalized least squares (GLS) estimate is when $V(\boldsymbol{\theta})$ is set to be the asymptotic covariance matrix of the empirical variogram, i.e., $V(\boldsymbol{\theta}) = \text{Var}(2\hat{\boldsymbol{\gamma}})$, and only assigning equal weight to all pairs with equal lag. Zhu and Stein (2002) proposed to define generalized variograms using linear filters where the horizontal, vertical and diagonal increments can be treated separately under a variogram setting. Under an increasing domain regime, all least squares estimates, including the OLS, WLS and GLS estimates, are consistent and asymptotically normal, with the GLS estimates being asymptotically statistically efficient among all (Lahiri, Lee, and Cressie, 2002). Despite the statistical efficiency of the GLS estimates, the facts that GLS requires the full covariance matrix that is difficult to obtain and consists of a nonlinear

optimization make GLS less appealing to implement in practice. WLS is often employed as an alternative as it is a compromise between OLS and GLS.

Likelihood based methods are convenient tools for variogram estimation when a distributional assumption, usually normality, is valid. Maximum likelihood (ML) methods have been developed for stationary isotropic Gaussian process models for regularly and irregularly spaced locations, and large datasets problem using different approximation methods, many of which have been reviewed by Sun et al. (2012). When the mean parameter of the process is unknown and needs to be estimated, ML usually underestimates the variability of the process as it assumes the mean parameter is known (Stein (1999), Section 6.6). Restricted maximum likelihood (REML) is useful for this problem as it maximizes the likelihood for linear combinations of the observations whose means are independent of the unknown mean parameter (Patterson and Thompson, 1971; Kitanidis, 1983). REML produces estimates with less bias compared to that from the ML estimation, especially when the number of parameters is large relative to the sample size (McGilchrist (1989); Tunnicliffe-Wilson (1989), Kang, Shin, and Lee (2003)), and is also less computationally involved in practice. Compared to likelihood based methods, an alternative is via estimating equations. For ex-

ample, for stationary Gaussian processes, Kaufman, Schervish, and Nychka (2008), Stein (2013) and Sun and Stein (2016) proposed different types of biased or unbiased estimating equations for covariance function estimation. In this paper, we develop new and flexible unbiased estimating equations to fit variogram models. The proposed method provides an alternative approach to the likelihood based methods, and includes the commonly used OLS, WLS and GLS as its special cases. Our method is highly efficient as a low dimensional representation of the weight matrix is adopted. The asymptotic properties of the estimators and the effect of lag set are explored. We illustrate our methodology for both lattice data and irregularly spaced data.

The remainder of the paper is organized as follows. In Section 2, we first describe the estimating equations approach, present our low rank approximation procedure, and then demonstrate how this framework can be used to generalize many widely used estimators. We end this section by presenting their theoretical properties. We investigate the statistical efficiency of different estimators constructed using our approach and examine the lag effect using a series of numerical studies and simulations in Section 3. The application of the proposed method is illustrated using a hydrology data set in Section 4. We conclude with a discussion of the proposed method.

2. Methodology

2.1. Estimating Equations

Let $\{Z(\mathbf{s}) : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^{\mathbf{2}}\}$ be an intrinsically stationary and isotropic Gaussian process. Assume the mean of the process is zero for simplicity. The semivariogram (half of the variogram) is

$$\gamma(\|\mathbf{s}_1 - \mathbf{s}_2\|; \boldsymbol{\theta}) = \frac{1}{2} E\{Z(\mathbf{s}_1) - Z(\mathbf{s}_2)\}^2, \quad \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D},$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a $p \times 1$ vector of the unknown parameters that are to be estimated from the data. Suppose data are observed on a regularly-spaced rectangular grid $\{(i, j) : i = 1, \ldots, \ell; j = 1, \ldots, \omega\}$, we define the squared increments process as $Y_{ij,i'j'} = [Z(i, j) - Z(i', j')]^2$ where $(i, j) \neq (i', j')$ and the semivariogram as $\gamma_{ij,i'j'} = \gamma(\|(i, j) - (i', j')\|, \boldsymbol{\theta})$. The corresponding vector representations are $\mathbf{Y} = [Y_{ij,i'j'}]$ and $\boldsymbol{\gamma}(\boldsymbol{\theta}) = [\gamma_{ij,i'j'}]$, where both $\mathbf{Y}$ and $\boldsymbol{\gamma}$ are $n \times 1$ vectors, and n is the total number of increments used to form $\mathbf{Y}$.

Since $E(\mathbf{Y}) = 2\boldsymbol{\gamma}(\boldsymbol{\theta})$, an estimator of $\boldsymbol{\theta}$ can be obtained by solving the unbiased estimating equations

$$\mathbf{W}(\boldsymbol{\theta})^{\mathrm{T}}[\mathbf{Y} - 2\boldsymbol{\gamma}(\boldsymbol{\theta})] = \mathbf{0}_p, \tag{2.1}$$

where $\mathbf{W}(\boldsymbol{\theta})$ is an $n \times p$ weight matrix dependent on $\boldsymbol{\theta}$ and $\mathbf{0}_p$ is a $p \times 1$ vector of zeros. Let $\mathbf{D}(\boldsymbol{\theta})$ be the first order partial derivative of $2\boldsymbol{\gamma}(\boldsymbol{\theta})$

with respect to $\boldsymbol{\theta}$, and $\mathbf{V}(\boldsymbol{\theta})$ be the covariance matrix of $\mathbf{Y}$, then the most efficient form of $\mathbf{W}(\boldsymbol{\theta})$ is given by

$$\mathbf{W}(\boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{D}(\boldsymbol{\theta}). \tag{2.2}$$

Finding $\mathbf{W}(\boldsymbol{\theta})$ can be challenging because it involves computing the inverse of an $n \times n$ matrix $\mathbf{V}(\boldsymbol{\theta})$ where n can be quite large, or solving $\mathbf{V}(\boldsymbol{\theta})\mathbf{W}(\boldsymbol{\theta}) = \mathbf{D}(\boldsymbol{\theta})$. We propose to use $\tilde{\mathbf{W}}(\boldsymbol{\theta})$ to approximate $\mathbf{W}$, where $\tilde{\mathbf{W}}(\boldsymbol{\theta})$ is an $m \times p$ weight matrix, such that

$$\mathbf{W}(\boldsymbol{\theta}) = \mathbf{A}\tilde{\mathbf{W}}(\boldsymbol{\theta}),$$

where $\mathbf{A}$ is an $n \times m$ binary matrix and $m \ll n$. Thus, the estimating equations become

$$\tilde{\mathbf{W}}(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}[\mathbf{Y} - 2\boldsymbol{\gamma}(\boldsymbol{\theta})] = \mathbf{0}_p,$$

and

$$\tilde{\mathbf{W}}(\boldsymbol{\theta}) = [\mathbf{A}^{\mathrm{T}}\mathbf{V}(\boldsymbol{\theta})\mathbf{A}]^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{D}(\boldsymbol{\theta}),$$

where the dimension of the matrix to be inverted, $\mathbf{A}^{\mathrm{T}}\mathbf{V}(\boldsymbol{\theta})\mathbf{A}$, is $m \times m$, much smaller than before. With a predetermined $\mathbf{A}$ that is explained in more details below, we can solve the estimating equations using nonlinear solver. We used `nleqslv` to solve systems of nonlinear equations in R.

The method adopted is a Broyden method that is a quasi-newton method. It begins by computing a Jacobian of the function and then updates the Jacobian after each successful iteration. This method demonstrates linear convergence towards a solution (Hasselman, 2015).

The structure of the low dimensional representation $\mathbf{A}\tilde{\mathbf{W}}(\boldsymbol{\theta})$ implies which increments and how the increments are used to estimate the variogram. Increments corresponding to identical rows in $\mathbf{A}$ have equal weight as these increments will be multiplied by the same weight in $\tilde{\mathbf{W}}$. We call these increments *belong to the same category*. Let Ω be the collection of all categories considered and the total number of categories $|\Omega| = m$. The vector $\mathbf{Y}^{\mathrm{T}} = [\mathbf{Y}_1^{\mathrm{T}} \dots \mathbf{Y}_m^{\mathrm{T}}]$ contains all the squared increments ordered by category. For $i = 1, \dots, n$, and $j = 1, \dots, m$, the $n \times m$ matrix $\mathbf{A}$ is defined as follows:

$$A_{ij} = \begin{cases} 1, & \text{if the } i\text{th increment belongs to the } j\text{th category,} \\ 0, & \text{otherwise.} \end{cases}$$

Different situations can be handled by varying the classification criteria and categories as presented in the examples below. An illustration of some segments on a 7×7 grid is presented in Figure 1 to facilitate the understanding of these examples.

2.2. Examples

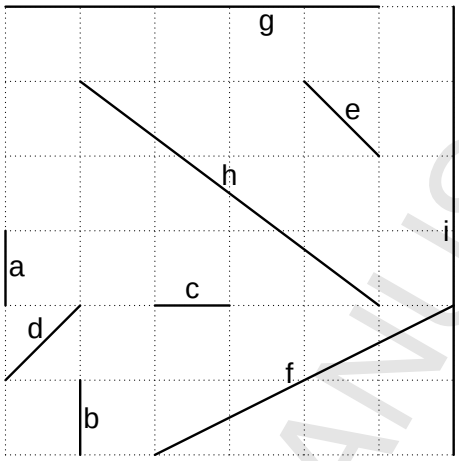

Figure 1: Illustration of some increments on a 7×7 grid.

Example 1 The simplest example to demonstrate this idea is when we classify increments using their lags, so Ω is the collection of all the different lags considered (hereafter referred as the *lag set* following Zhu and Stein (2002)). We consider all M lags, denoted by $k_1, \ldots, k_M$, where M is the size of the lag set. We sort all lags such that $k_1 < \cdots < k_M$, and then the number of categories $m = M$. Note that it is not necessary to include all unique lags into the model as shown in Examples 2 and 3. The definition of $\mathbf{A}$ is as follows.

$$A_{ij} = \begin{cases} 1, & \text{if the } i\text{th increment has lag } k_j, \\ 0, & \text{otherwise.} \end{cases}$$

Using the 7×7 grid given in Figure 1, increments that are labeled as a, b and c all have the shortest possible lag, $k_1 = 1$, and their corresponding rows in $\mathbf{A}$ are all $[1, 0, 0, \ldots, 0]$. Both rows in $\mathbf{A}$ that correspond to increments d and e are $[0, 1, 0, \ldots, 0]$. Similarly, since increments g and h have lag 5, that specific entries in their rows are 1 and the rest are all 0s. This scheme corresponds to the situation where we assign equal weight to increments with equal lag and is in fact the GLS.

Example 2 Pairs of observations that are further apart or with longer increments are less correlated compared to pairs that are closer to each other. It is reasonable to assume that after a certain threshold lag, the increments contribute little to the variogram estimation. That means we may assign 0 weights to increments corresponding to lags longer than the threshold lag, or leave out these increments in our classification procedure. All unique lags are ordered such that $k_1 < \ldots < k_m < \ldots < k_M$, and let the threshold lag be k_m, then $\mathbf{A}$ is defined as follows:

$$A_{ij} = \begin{cases} 1, & \text{if the } i\text{th increment has lag } k_j \text{ and } k_j \leq k_m, \\ 0, & \text{otherwise.} \end{cases}$$

Increments corresponding to lags that are longer than k_m do not have a corresponding row in $\mathbf{A}$. Using the illustration in Figure 1, if we

assume the threshold lag to be 4, then increments f, g, h and i will not have a row to represent them in $\mathbf{A}$ because their lags are longer than the threshold.

Example 3 Similar to Example 2, instead of assigning all 0 weights to increments corresponding to lags that are longer than the threshold lag k_m, we could also put some constant weight to all pairs with increments with lags that are at least as long as the threshold. In other words, we combine all increments corresponding to lags that are longer or equal to the threshold in the same category. The matrix $\mathbf{A}$ is the following:

$$\begin{cases} A_{ij} = 1 & \text{if the } i\text{th increment has lag } k_j \text{ and } k_j < k_m, \\ A_{im} = 1 & \text{if the } i\text{th increment has lag } k_j \text{ and } k_j \geq k_m, \\ A_{..} = 0 & \text{otherwise.} \end{cases}$$

Using the same threshold lag of 4 as in Example 2, the rows for increments f, g, h and i in $\mathbf{A}$ will be $[0, \ldots, 0, 1]$ to indicate that the increments corresponding to lags that are longer than the threshold will have the same weights.

Example 4 For a process defined on a grid, the characteristics of increments on the edges of the grid can be quite different from that of increments not on the edges. This edge effect poses another estimation problem

for spatial processes as the number of boundary points increases with the size of the grid (Guyon (1982); Dahlhaus and Künsch (1987); Genton (1998b)). The edge effect can be accounted for in our model by defining the categories according to both the lag set and whether the increment is on one of the edges or not. Then $m = m_1 + m_2$ where m_1 is the size of the lag set on the edges and m_2 is the size of the lag set not on the edges, and the rest of the setup is similar to that in Examples 1, 2 and 3 depending on what lag set is used for the analysis. In Figure 1, the row for increment c in $\mathbf{A}$ will be different from that for a and b as the latter two are on the edges while c is not. Increments g and h will also be treated differently under this setup.

Example 5 For anisotropic processes, the increments can be grouped according to both the lags and their directions. Then m becomes the number of these directed lags, and

$$A_{ij} = \begin{cases} 1, & \text{if the } i\text{th increment has the } j\text{th directed lag,} \\ 0, & \text{otherwise.} \end{cases}$$

For this example, increments a and b in Figure 1 will have identical rows, which are different from that of increment c. The rows for g and h will also be different.

The efficiency of estimating equations can be evaluated using the Godambe information matrix (Stefanski and Boos, 2002; Varin, Reid, and Firth, 2011). For a set of unbiased estimated equations $\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathbf{0}$, the Godambe information matrix is $\mathbf{G}(\boldsymbol{\theta}) = \big(\mathrm{E}\dot{\boldsymbol{\psi}}(\boldsymbol{\theta})\big)^{\mathrm{T}}\big(\mathrm{E}\boldsymbol{\psi}(\boldsymbol{\theta})\boldsymbol{\psi}(\boldsymbol{\theta})^{\mathrm{T}}\big)^{-1}\big(\mathrm{E}\dot{\boldsymbol{\psi}}(\boldsymbol{\theta})\big)$ where $\dot{\boldsymbol{\psi}}(\boldsymbol{\theta})$ is the $p \times p$ matrix whose jth column is $\partial\boldsymbol{\psi}(\boldsymbol{\theta})/\partial\theta_j$ (Varin, Reid, and Firth, 2011). For our estimating equations given in (2.2), the Godambe information matrix is

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbf{D}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{W}(\boldsymbol{\theta})\big(\mathbf{W}(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{W}(\boldsymbol{\theta})\big)^{-1}\mathbf{W}(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{D}(\boldsymbol{\theta}).$$

The inverse matrix $\mathbf{G}(\boldsymbol{\theta})^{-1}$ represents the approximate covariance matrix of the estimates obtained from these estimating equations with its diagonal elements as the approximate variances. The resulting unbiased estimator is asymptotically normally distributed with approximate covariance matrix $\mathbf{G}(\boldsymbol{\theta})^{-1}$.

For our proposed approximation framework, the Godambe information matrix is dependent on the matrix $\mathbf{A}$ as

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbf{D}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{A}\big(\mathbf{A}^{\mathrm{T}}\mathbf{V}(\boldsymbol{\theta})\mathbf{A}\big)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{D}(\boldsymbol{\theta}).$$

Since different methods (WLS, GLS, GLS with edge effect and others) and different lag sets are associated with different $\mathbf{A}$'s, statistical efficiencies of these methods and the effect of lag set can be examined and compared

through the matrix $\mathbf{G}(\boldsymbol{\theta})^{-1}$, the approximate asymptotic covariance matrix. Since the estimators have relatively simple forms, it is feasible to derive the general forms of their corresponding $\mathbf{G}(\boldsymbol{\theta})^{-1}$'s. In the numerical studies carried out in Section 3.1, we are able to numerically compute $\mathbf{G}(\boldsymbol{\theta})^{-1}$'s to compare the statistical efficiencies of these different methods. Thus, the proposed mechanism provides us with a new framework to conveniently study the edge effect and lag set effect.

2.3. Theoretical Properties

Let $\hat{\boldsymbol{\theta}}$ be the exact estimator obtained from the estimation equations (2.1). In this section, we study the asymptotic efficiency of the estimators with different choices of $\mathbf{A}$ under an increasing-domain setting. Specifically, let $\hat{\boldsymbol{\theta}}_t$ be the estimator obtained by the GLS, and $\hat{\boldsymbol{\theta}}_e$ be the one considering the edge effect. We show that under some mild conditions, the squared increments process is weakly dependent. In particular, $\hat{\boldsymbol{\theta}}_t$ and $\hat{\boldsymbol{\theta}}_e$ are asymptotically as efficient as $\hat{\boldsymbol{\theta}}$. We present the results for an $m-$dependent process. In general, a sequence of random variables $\{X_t\}$ is said to be $m-$dependent if $|t-s| > m$ implies X_t is independent of X_s (Shumway and Stoffer, 2011). In our case, this is related to if min $(|d_{ij,kl}|, |d_{i'j',k'l'}|, |d_{ij,k'l'}|, |d_{i'j',kl}|) > m$, where $d_{ij,kl}$ represents the distance between $Z(i,j)$ and $Z(k,l)$, then $\mathrm{Cov}(Y_{ij,i'j'}, Y_{kl,k'l'}) = 0$. Although pro-

cesses in general are not necessarily $m-$dependent, this theorem provides some useful insights on the statistical properties of the estimators.

Theorem 1. *Let $Z(\boldsymbol{s})$ be an intrinsically stationary process on a regularly-spaced grid and $\mathbf{Y}$ be the corresponding squared increments process, as defined in Section 2.1. If $\mathbf{Y}$ is an $m-$dependent process, then $\hat{\boldsymbol{\theta}}_t$ and $\hat{\boldsymbol{\theta}}_e$ are asymptotically efficient.*

The intuition behind the theorem is that a weakly dependent process defined on $\mathbb{R}^2$ can be well approximated by its counterpart on a torus. It is straightforward to show that on a torus, $\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}$ without approximation, and consequently $\hat{\boldsymbol{\theta}}_t$ has the same statistical efficiency as $\hat{\boldsymbol{\theta}}$. Theorem 1 also suggests that the optimal weights defined by $\mathbf{W}$ can be well approximated by their lower dimensional counterparts for weakly dependent processes. In practice, this can greatly reduce the computational burden while retaining the statistical efficiency.

3. Numerical and Simulation Studies

3.1. Numerical Study

We perform a set of numerical studies to explore the statistical efficiency of our proposed estimating equations approach. Many physical processes observed in real life can be modeled using fractional Brownian fields, which are a class of nonstationary Gaussian random fields with stationary incre-

ments. Fractional Brownian fields for spatial processes defined on $\mathbb{R}^2$ are just fractional Brownian surfaces. It is worth noting that fractional Brownian fields are self-similar processes, meaning their statistical properties are the same from both the increasing-domain and fixed-domain perspectives.

Let $Z(\mathbf{s})$ be a fractional Brownian surface with constant mean and semivariogram $\gamma(\|\mathbf{s}_1 - \mathbf{s}_2\|; \boldsymbol{\theta}) = \beta\|\mathbf{s}_1 - \mathbf{s}_2\|^\alpha$, $\mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^2$, where $\boldsymbol{\theta} = (\alpha, \beta)$, $\alpha \in (0, 2)$ and $\beta > 0$. The parameter α determines the smoothness of the surface with larger values corresponding to smoother processes, and β is a scale parameter. The covariance $\mathbf{V}(\boldsymbol{\theta})$ of the squared increments process $\mathbf{Y}$ is determined by

$$\begin{aligned}\mathrm{Cov}\Big(Y_{ij,i'j'}, Y_{kl,k'l'}\Big) &= 2\Big\{\mathrm{Cov}\Big(Z(i,j) - Z(i',j'), Z(k,l) - Z(k',l')\Big)\Big\}^2 \\ &= 2\beta^2\Big\{\Big|d_{ij,kl}\Big|^\alpha + \Big|d_{i'j',k'l'}\Big|^\alpha - \Big|d_{ij,k'l'}\Big|^\alpha - \Big|d_{i'j',kl}\Big|^\alpha\Big\}^2,\end{aligned}$$

where $d_{ij,kl}$ denotes the distance between $Z(i,j)$ and $Z(k,l)$.

We consider a fractional Brownian surface observed on an $l \times l$ regular lattice laid over a square window $[0,1]\times[0,1]$, with neighboring observations l^{-1} apart and sample size $N = l^2$ for $N \in \{100, 400, 1600\}$. The scale parameter β is set to 1, and smoothness parameter $\alpha \in \{0.1, 0.5, 1.5\}$. We considered two cases, $\mathcal{K}_2$ (up to 2 lags) and $\mathcal{K}_4$ (up to 4 lags), to explore the effect of the lag set. For each case, we examine the effect of the three

different weight matrices $\mathbf{W}(\boldsymbol{\theta})$: (i) $\mathbf{W}_d(\boldsymbol{\theta}) = \mathbf{V}_d(\boldsymbol{\theta})^{-1}\mathbf{D}(\boldsymbol{\theta})$ where $\mathbf{V}_d(\boldsymbol{\theta})$ is a diagonal matrix, (ii) $\mathbf{W}_t(\boldsymbol{\theta}) = \mathbf{A}_t[\mathbf{A}_t^{\mathrm{T}}\mathbf{V}(\boldsymbol{\theta})\mathbf{A}_t]^{-1}\mathbf{A}_t^{\mathrm{T}}\mathbf{D}(\boldsymbol{\theta})$ where $\mathbf{A}_t$ is as described in Example 1 of Section 2.2, and (iii) $\mathbf{W}_e(\boldsymbol{\theta})$ which considers the edge effect. The first method corresponds to the WLS procedure, the second is the GLS, and the last one is the GLS with edge effect. We summarize the asymptotic standard deviations of the estimators at the true values using the diagonal elements of $\mathbf{G}(\boldsymbol{\theta})^{-1}$ in Table 1.

For all cases, a large sample size and a large lag set reduce the asymptotic variability of the estimates. As the surface becomes smoother, *i.e.* α becomes bigger, the asymptotic standard deviation of the estimate increases. It typically becomes more difficult to estimate the parameters for a smooth process as noted in Sun and Stein (2016) where the Whittle covariance model is compared with the exponential model. Among the three methods using different weight matrices, the approach using a weight matrix incorporating the edge effect performs the best, and the one analogous to the WLS performs the worst with the highest asymptotic standard deviations of the estimates as seen in Table 1. For smoother processes, the reduction in the asymptotic standard deviations of the estimates when we account for the edge effect using the GLS is also generally higher compared to coarser processes. A set of simulation study is also carried out to com-

Table 1: Properties of estimates of $\boldsymbol{\theta} = (\alpha, \beta)$ in the variogram model $\beta\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|^\alpha$, $\boldsymbol{s}_1, \boldsymbol{s}_2 \in [0,1]^2$ when $\alpha = \{0.1, 0.5, 1.5\}$ and $\beta = 1$. Entries are asymptotic standard deviations ($\times 1000$) of $(\hat{\alpha}, \hat{\beta})$ using three different weight matrices: $\mathbf{W}_d$ (WLS), $\mathbf{W}_t$ (GLS) and $\mathbf{W}_e$ (GLS with edge effect). Lag sets $\mathcal{K}_2$ and $\mathcal{K}_4$ contain up to 2 and 4 lags respectively.

N	Weight Matrix	Lag set	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.5$
100	$\mathbf{W}_d$	$\mathcal{K}_2$	(72.423, 70.209)	(85.548, 72.768)	(106.874, 164.151)
		$\mathcal{K}_4$	(41.813, 69.551)	(85.303, 73.278)	(126.144, 161.442)
	$\mathbf{W}_t$	$\mathcal{K}_2$	(72.316, 70.059)	(85.459, 72.571)	(102.621, 163.766)
		$\mathcal{K}_4$	(41.336, 68.948)	(79.196, 72.108)	(89.578, 146.890)
	$\mathbf{W}_e$	$\mathcal{K}_2$	(70.627, 67.993)	(83.680, 70.411)	(98.666, 150.967)
		$\mathcal{K}_4$	(40.737, 65.905)	(76.767, 69.574)	(69.468, 111.178)
400	$\mathbf{W}_d$	$\mathcal{K}_2$	(34.349, 34.570)	(41.824, 35.952)	(71.168, 114.805)
		$\mathcal{K}_4$	(19.637, 34.114)	(42.712, 35.981)	(84.272, 112.318)
	$\mathbf{W}_t$	$\mathcal{K}_2$	(34.339, 34.545)	(41.820, 35.933)	(67.822, 112.786)
		$\mathcal{K}_4$	(19.503, 33.982)	(39.453, 35.712)	(51.662, 88.728)
	$\mathbf{W}_e$	$\mathcal{K}_2$	(29.384, 29.497)	(35.842, 30.678)	(58.937, 96.208)
		$\mathcal{K}_4$	(16.817, 28.970)	(33.892, 30.452)	(41.972, 71.180)
1600	$\mathbf{W}_d$	$\mathcal{K}_2$	(16.753, 17.155)	(20.814, 17.892)	(48.959, 80.807)
		$\mathcal{K}_4$	(9.469, 16.906)	(21.549, 17.847)	(57.782, 78.902)
	$\mathbf{W}_t$	$\mathcal{K}_2$	(16.752, 17.150)	(20.814, 17.891)	(44.814, 75.902)
		$\mathcal{K}_4$	(9.413, 16.867)	(19.782, 17.778)	(28.485, 49.855)
	$\mathbf{W}_e$	$\mathcal{K}_2$	(13.074, 13.377)	(16.251, 13.955)	(35.255, 59.410)
		$\mathcal{K}_4$	(7.362, 13.153)	(15.457, 13.863)	(21.887, 38.169)

pute the actual standard deviations of the estimates. A similar trend is observed as that in the asymptotic standard deviations of the estimates.

Table 1 demonstrates that the larger lag set $\mathcal{K}_4$ produces smaller asymptotic standard deviations of the estimates compared to the lag set $\mathcal{K}_2$ for all three estimating equations methods. To understand whether increasing the lag set is always able to improve the performance of the estimators, we compute the asymptotic standard deviations for all possible lag sets when $N = 100$. More specifically, we consider the respective standard deviations of estimates for $\mathcal{K} = \{k_1, \ldots, k_M\}, k_1 < \ldots < k_M$, and we select $M = 50$. The asymptotic standard deviations of the estimates of the GLS and WLS (given in Figure 2) show drastically different patterns. For the GLS method, it is best to select the largest possible lag set as the asymptotic standard deviations of the estimates decrease with the lag size. On the other hand, the WLS performs best when the lag size is somewhere between 2 and 4 lags, and works poorly when the lag size becomes very large, especially for smooth processes with high α. Our observation is consistent with other works in the literature. Using simulation studies, Zhu and Stein (2002) demonstrated that the optimal lag size for the GLS estimator is the largest possible lag set. Both Zhu and Stein (2002) and Davies and Hall (1999) noted that the lags beyond 2 yield poor estimates for the OLS estimators,

which is a simplified version of the WLS estimators considered in this paper.

The performance of the GLS estimators improves when the lag size increases, but it is not realistic to consider the largest possible lag set for computational reasons. Therefore, we would like to study how using a partial set of all possible lags affects the weight associated with each lag. We denote by $w_1, \ldots, w_M$ the corresponding weights for lags $k_1, \ldots, k_M$ when all possible lags are considered. Let $\mathring{w}_1, \ldots, \mathring{w}_m$ be the weights for lags $k_1, \ldots, k_m$ when using lags up to k_m. We use $\tilde{w}_1, \ldots, \tilde{w}_{m-1}, \tilde{w}_{\geq m}$ to represent the weights when using all lags, where $\tilde{w}_i$ is the weight for lag k_i, $i = 1, \ldots, m-1$, and $\tilde{w}_{\geq m}$ is the total weight for all pairs with lags $\geq k_m$. Figure 3 compares these three types of weights when we only consider the first 16 lags ($m = 16$, $k_{16} = 4\sqrt{2}$), and the first 11 lags ($m = 11$, $k_{11} = 3\sqrt{2}$).

Using a long partial set ($m = 16$ versus $m = 11$) improves the efficiency of the GLS estimators regardless of the type of weighting schemes we use. When $m = 16$, compared to the weights $\{\mathring{w}_1, \ldots, \mathring{w}_{16}\}$ that use only the first 16 lags, the weights $\{\tilde{w}_1, \ldots, \tilde{w}_{15}, \tilde{w}_{\geq 16}\}$ are closer to that obtained using all possible lags as seen in the left panel of Figure 3, and their corresponding asymptotic standard deviations are also closer to that found using all lags as seen in Table 2. Therefore, in this case, combining all lags at least as long as the threshold in one category improves the quality of approximation,

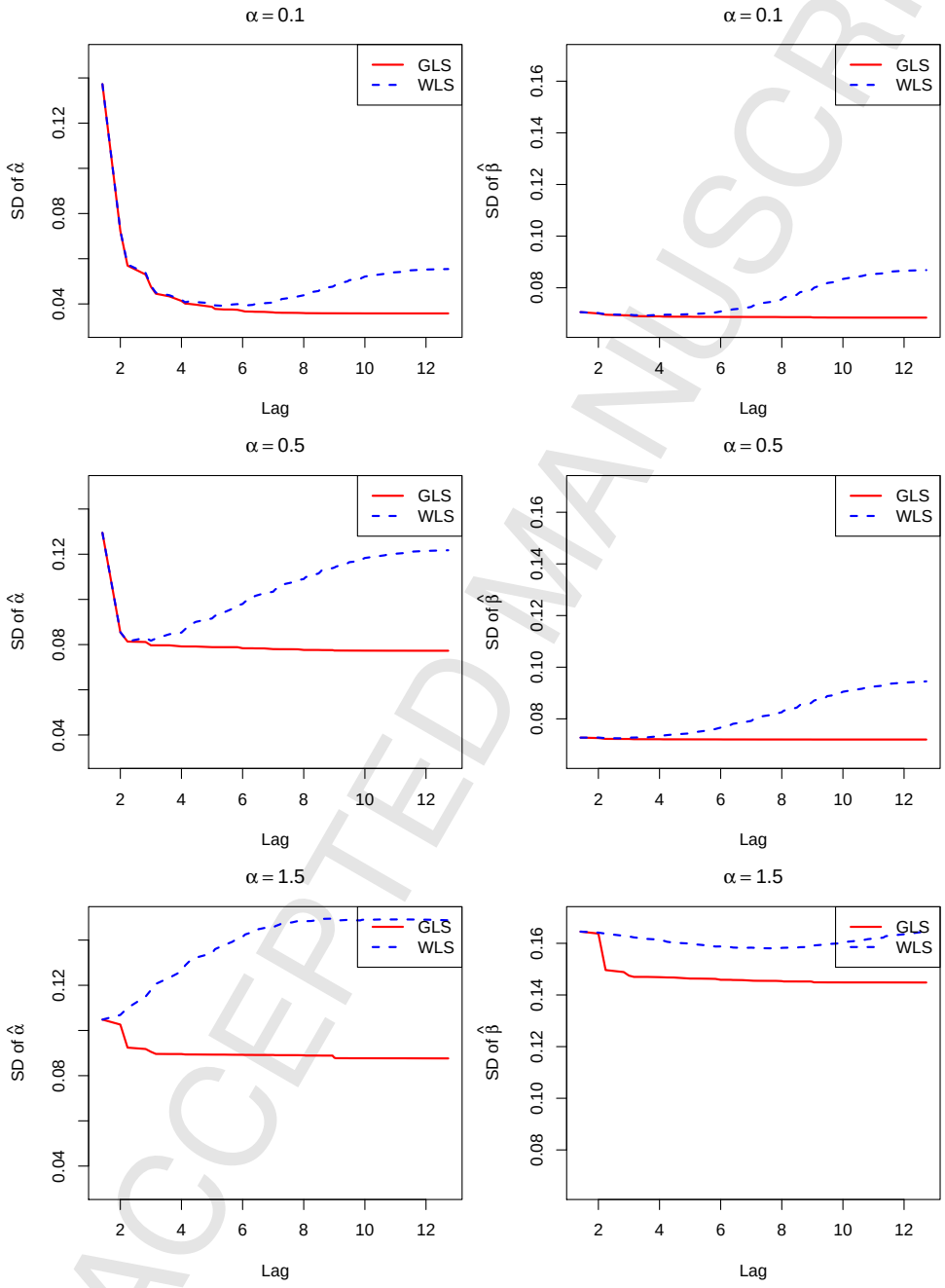

Figure 2: Asymptotic standard deviations of $\hat{\alpha}$ (left panel) and that of $\hat{\beta}$ (right panel) when we compare the statistical efficiency of GLS and WLS for different lag sets.

Table 2: Properties of estimates of $\boldsymbol{\theta} = (\alpha, \beta)$ in the semivariogram model $\beta\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|^\alpha$ for $N = 100$ observations in $[0,1]^2$ when $\alpha = \{0.1, 0.5, 1.5\}$ and $\beta = 1$. Entries are asymptotic standard deviations ($\times 1000$) of $(\hat{\alpha}, \hat{\beta})$ calculated from $\mathbf{G}(\boldsymbol{\theta})^{-1}$ for three types of weights: w, $\tilde{w}$, $\mathring{w}$ with $m = 16, 11$.

m	Weight Type	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.5$
50	w	$(35.803, 68.444)$	$(77.280, 71.991)$	$(87.638, 144.823)$
16	$\tilde{w}$	$(35.924, 68.600)$	$(77.572, 72.006)$	$(88.283, 145.287)$
	$\mathring{w}$	$(37.526, 68.730)$	$(78.864, 72.049)$	$(89.256, 146.273)$
11	$\tilde{w}$	$(36.246, 68.695)$	$(79.048, 72.273)$	$(91.032, 149.644)$
	$\mathring{w}$	$(40.026, 68.827)$	$(79.165, 72.063)$	$(89.373, 146.782)$

while the computational costs of these two methods are about the same. However, it is also worth pointing out that including more lags in this way is not always beneficial. For example, when $m = 11$, both the right panel of Figure 3 and Table 2 show that $\{\mathring{w}_1, \ldots, \mathring{w}_{11}\}$ provides better approximations and leads to smaller standard deviations of $\hat{\boldsymbol{\theta}}$ for $\alpha = 0.5$ and 1.5, compared to $\{\tilde{w}_1, \ldots, \tilde{w}_{10}, \tilde{w}_{\geq 11}\}$. This is due to the reason that $m = 11$ is not sufficient for such strongly dependent processes, so that k_{11} is still too significant to be combined with other lags in one category.

3.2. Simulation Study

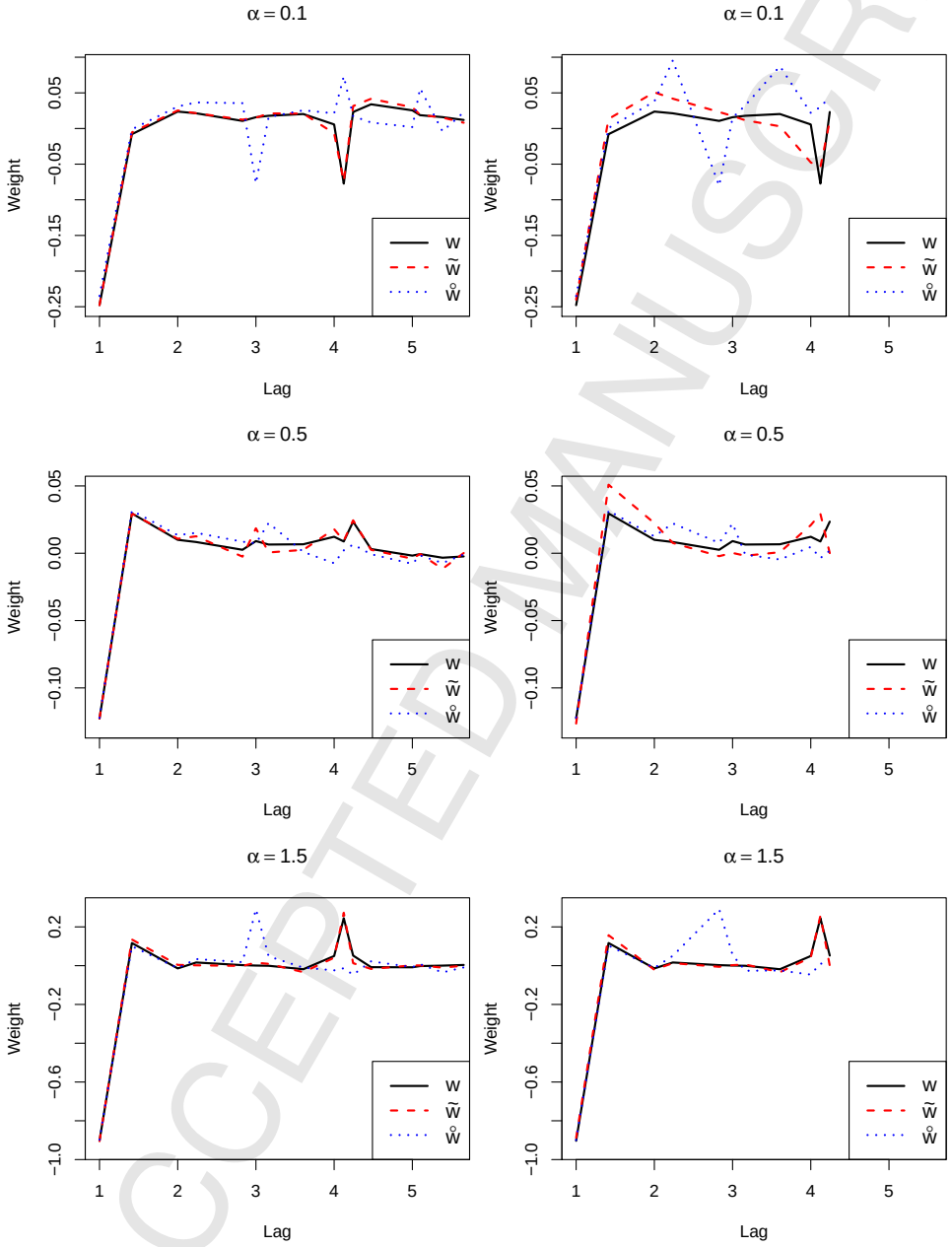

Figure 3: Weights from three different methods when $m = 16$ (left panel) and $m = 11$ (right panel).

A simulation study is also carried out to compare the efficiency of the three different methods. Using the same setup as in the previous subsection, we simulate fractional Brownian surfaces with constant mean, $\alpha = \{0.1, 0.5, 1.5\}$ and $\beta = 1$ on a $l \times l$ regular grid with the sample size $N = l^2 = \{100, 400, 1600\}$ for 500 replications for each combination. The boxplots and the mean squared errors of parameter estimates are given in Figure 4 and Table 3 respectively. Among the three methods, the WLS performs the worst with both the largest bias and the largest variance, and fails to converge including the cases when $N = 400$ and 600 for $\alpha = 0.5$ due to its incorrect weighting scheme. Taking the edge effect into consideration improves the efficiency of the GLS, especially when the sample size N is small and the dependence of the process is strong, *i.e.*, when α is large.

For irregularly spaced observations, the results are very similar to that for observations on a grid: the GLS with edge effect works much better than the other two. Certainly how we choose the border to define pairs on the edge is very important. For example, consider 100 irregularly spaced points in a unit square. When the lag set $\mathcal{K}_2$ (up to 2 lags) is selected, we classify all pairs to three categories with lags corresponding to $1, \sqrt{2}$ and 2, respectively. We then take 0.1 as the width of the border and treat pairs with at least one point in the border area as pairs on the edge. The

Table 3: MSEs ($\times 1000$) of $(\hat{\alpha}, \hat{\beta})$ using three different weight matrices: $\mathbf{W}_d$ (WLS), $\mathbf{W}_t$ (GLS) and $\mathbf{W}_e$ (GLS with edge effect) for lag set $\mathcal{K}_2$ with 500 replications.

N	Weight Matrix	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.5$
100	$\mathbf{W}_d$	$(538.26, 58.37)$	$(182.11, 35.68)$	$(108.51, 129.77)$
	$\mathbf{W}_t$	$(11.31, 21.38)$	$(26.95, 20.82)$	$(43.64, 117.50)$
	$\mathbf{W}_e$	$(11.10, 20.67)$	$(26.79, 20.58)$	$(36.60, 115.90)$
400	$\mathbf{W}_d$	$(284.15, 27.87)$	–	$(69.19, 61.41)$
	$\mathbf{W}_t$	$(3.47, 5.42)$	$(7.19, 4.82)$	$(15.66, 48.71)$
	$\mathbf{W}_e$	$(3.42, 5.34)$	$(7.08, 4.79)$	$(13.87, 47.23)$
1600	$\mathbf{W}_d$	$(51.82, 3.97)$	–	$(51.69, 30.80)$
	$\mathbf{W}_t$	$(1.31, 1.67)$	$(1.60, 1.25)$	$(6.64, 22.23)$
	$\mathbf{W}_e$	$(1.29, 1.64)$	$(1.55, 1.25)$	$(5.98, 21.15)$

resulting $\mathbf{A}$ matrix has the same size as the regularly spaced case, and the GLS method with the edge effect corrected also outperforms the GLS without the edge effect and the WLS with similar numerical results seen in Table 1 and Figure 4.

4. Application

4.1. Data Description

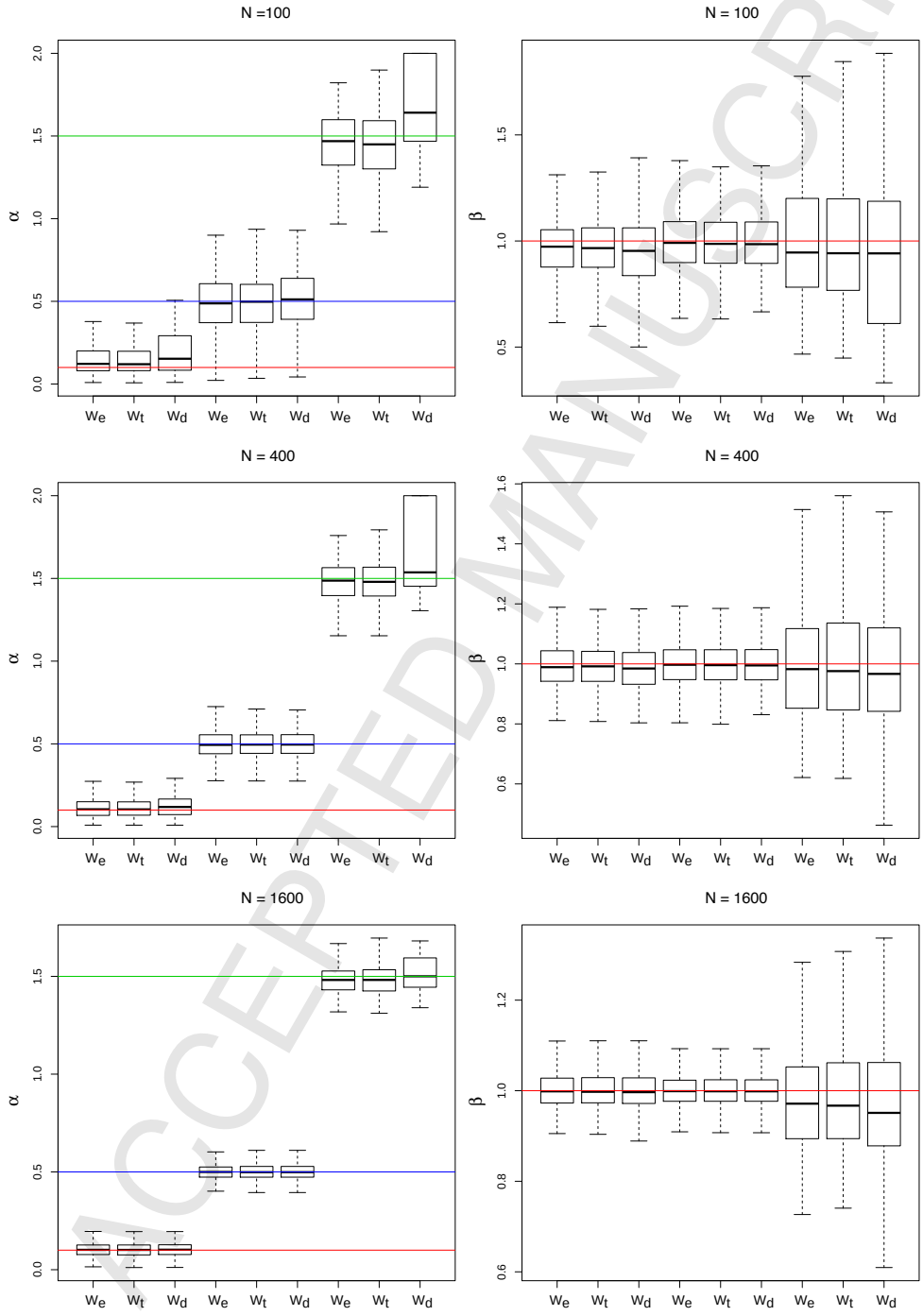

Figure 4: Boxplots of $\hat{\alpha}$ and $\hat{\beta}$ for $\mathcal{K}_2$.

As a major component of soil hydrology, soil moisture is critical in shaping the ecosystem response to the physical environment. Soil moisture helps to understand and predict variations of surface temperature, drought, flood, weather forecasting and climate change (Robock et al., 2000). Thus, it is the interest of many to characterize and predict any spatial features of the soil moisture, especially near the surface soil moisture, that is to be able to estimate the variogram accurately.

The Mississippi River has the fourth largest drainage basin in the world and also the largest in the North America. The Mississippi River basin covers more than 1,245,000 miles^2 (3,220,000 km^2), including all or parts of 33 U.S. states and two Canadian provinces. We obtained daily soil moisture data generated using numerical models at the top layer of the Mississippi River basin on January 1, 2014, from a spatial region from 92.475° to 107.646° W longitude and 32.446° to 43.363° N latitude. About 11% of the observations are missing, mostly scattered on the edges of this region (see Figure 5). This dataset has been also studied by Huang and Sun (2017) for fitting a Gaussian process model. We take the largest region with the least amount of missing observations, *i.e.*, longitudes between 93.142° and 107.058° W and latitudes between 33.029° and 42.946° N, and use linear interpolation to infill any remaining missing values. We are able to collect

two sets of data at two different resolutions: (1) the averaged soil moisture data on a grid of size 21×15 with spatial resolution of 0.6667° in both the longitude and latitude, and (2) the averaged data on a 42×30 grid with spatial resolution of 0.3333° in both directions. Since the data are on the globe, we would have to map the region, which is actually a trapezoid, onto a regular rectangle using the one-to-one mapping procedure explained in Chang and Stein (2014) to ensure the data is in a grid format. Since 1° change in the longitude is not the same as 1° change in the latitude, we used the actual distances between two grid points instead of the coordinate representation to account for the difference in length for the rest of the analysis

4.2. Variogram Estimation

We assume the averaged daily soil moisture data to be a fractional Brownian surface with a constant mean and a semivariogram $\gamma(\|\mathbf{h}\|; \alpha, \beta) = \beta\|\mathbf{h}\|^\alpha$, where $\|\mathbf{h}\|$ is the distance between two grid points of the Mississippi River basin. An exploratory analysis demonstrates that the averaged daily soil moisture data is isotropic. Three different methods are applied to both datasets to estimate the parameters of the variogram, and the lag set $\mathcal{K}$ consists of all lags that are shorter than 7.35 km.

Among the three methods, the WLS fails to achieve a reasonable esti-

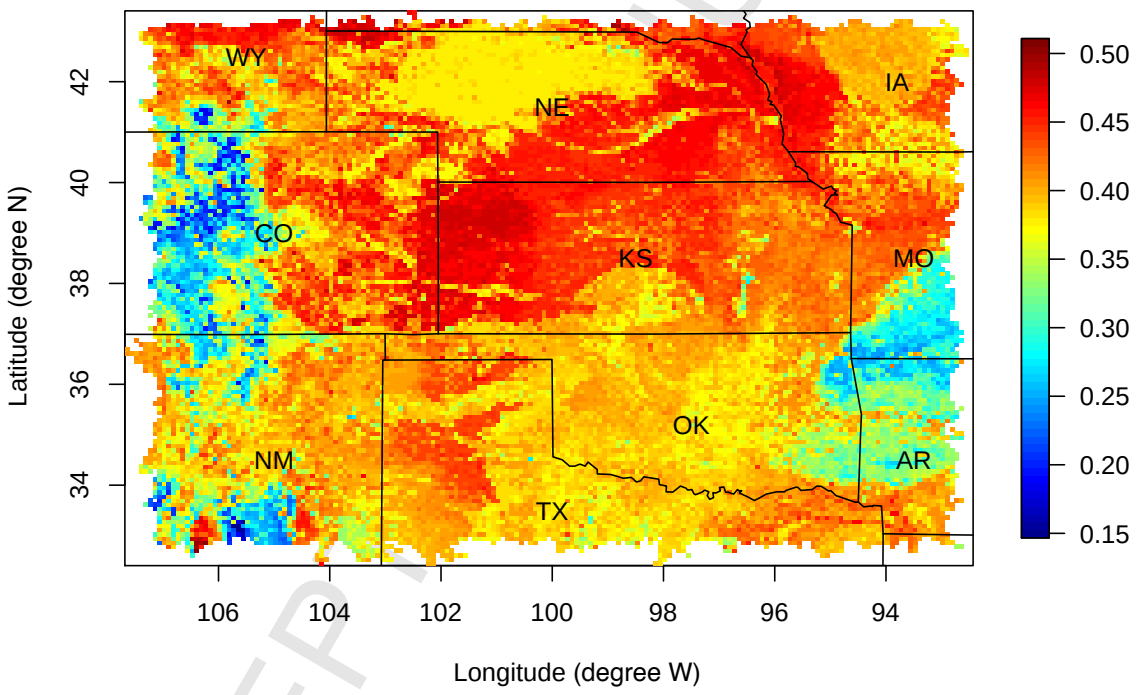

Figure 5: Map of the daily soil moisture data at the top layer of the Mississippi River Basin on January 1, 2014.

mate for the smoothness parameter α due to the incorrect weighting scheme associated with the WLS. The estimates for (α, β) for the GLS and GLS with edge effect are given in Table 4. The estimated smoothness parameter, $\hat{\alpha}$ for the low resolution data set is greater than the corresponding estimate from the high resolution data set due to the high degree of smoothness associated with the former. Based on these parameter estimates, we plot the estimated variogram using the GLS, GLS with edge effect, and the smoothed empirical variogram on Figure 6. The smoothed empirical variogram is obtained using the kernel smoothing in *geoR* package. It is not surprising to observe that for both data sets at different resolutions, accounting for the edge effect increases the smoothness parameter estimates. Although the estimated variograms of the GLS with and without the edge effect do not differ much due to the moderate sizes of the grid we studied, a method that takes the edge effect into consideration is probably the more appropriate choice among the two.

4.3. Prediction

The estimated variograms using various methods can be used for predictions. Out of the total of 315 grid points from the low resolution data set, we randomly leave out 30 points, fit both the GLS and the GLS with edge effect, and then predict soil moisture values at these 30 points. For the high

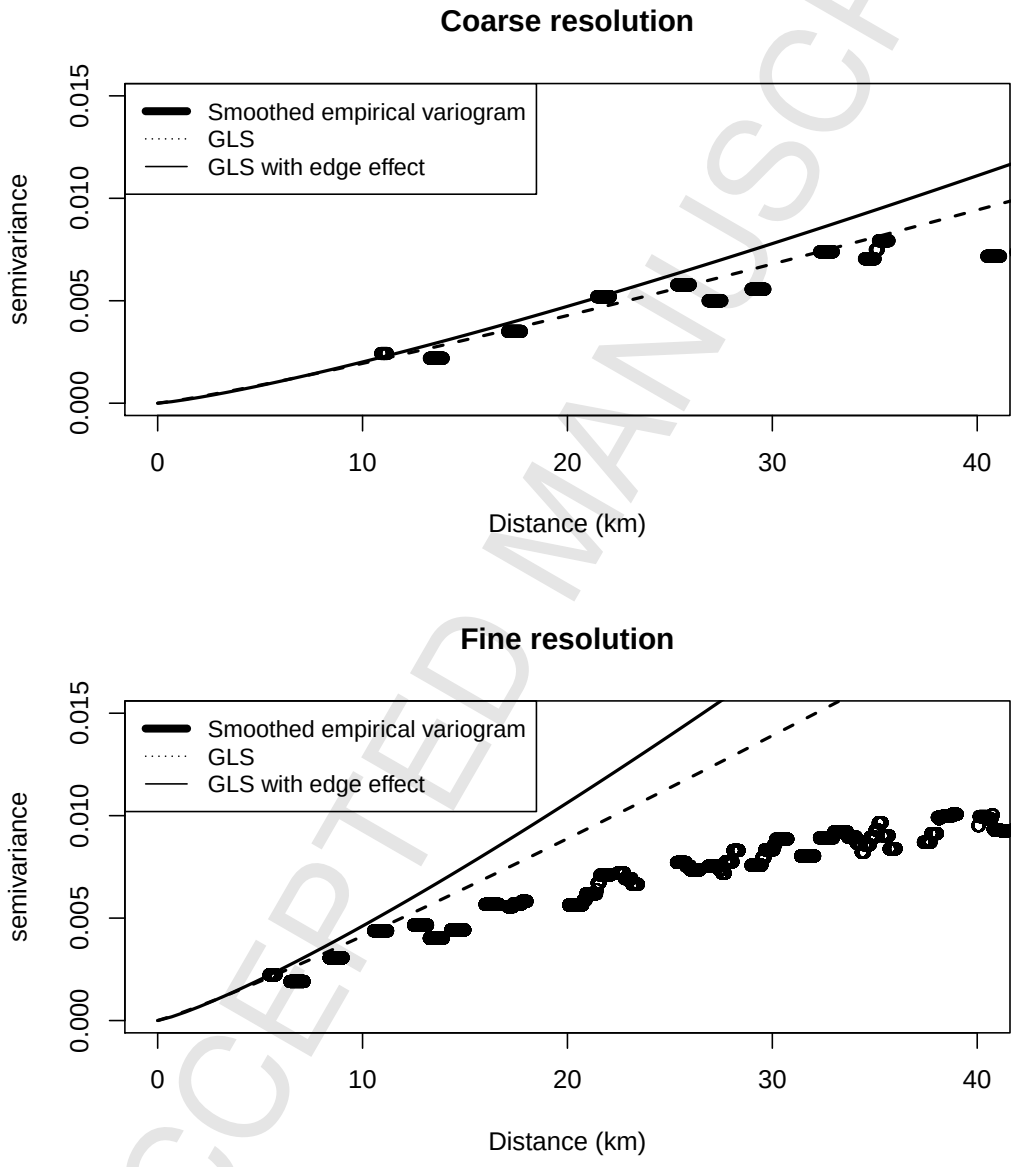

Figure 6: The empirical variogram and the estimated variogram using the GLS and the GLS with edge effect when we consider the lag set $\mathcal{K} = \{1, \sqrt{2}, 2\}$ where 1 unit is approximately 7.35 km.

Table 4: Parameter estimates $(\hat{\alpha}, \hat{\beta})$ obtained for the soil moisture data using two different methods.

Data	Method	$\hat{\alpha}$	$\hat{\beta}$ ($\times$1000)
Low resolution	GLS	1.140	0.200
	GLS with edge effect	1.232	0.172
High resolution	GLS	1.107	0.453
	GLS with edge effect	1.205	0.417

resolution data set, a similar set up is adopted except that we randomly leave out 120 points from the total of 1260 grid points. We repeat the same procedure for 1000 times, and compute the means, standard deviations of the prediction errors and the continuous ranked probability score (CRPS, Gneiting and Raftery, 2007) of different models. For both the low and high resolution data, the means and standard deviations of the prediction errors are fairly small. The CRPS is around 0.23 for the low resolution data set and 0.26 for the high resolution data, suggesting that the prediction is more accurate and precise for the low resolution data. When the edge effect is accounted for, the CRPS improves by about 0.01% for both cases. The improvement is expected to be more significant for processes with much stronger dependence.

5. Discussion

We proposed a new framework for the variogram estimation using estimating equations. Many frequently adopted models in the literature are special cases of the proposed method, implying a wide application of the results presented here. Using a low rank approximation for the weight matrix, our method achieves high computational efficiency. This approach allows users to easily change lag sets, conveniently adjust how the weights are assigned to long lags and readily incorporate the edge effect, offering flexibility to satisfy a wide range of needs.

For spatial modeling, when the lag set effect for spatial modeling is concerned, we recommend to assign a fixed weight to all the lags that are at least as long as the threshold rather than a complete omission. The gain in statistical efficiency outweighs the marginal increase in the computational cost, as seen in our numerical studies. It is intuitive to learn that the GLS outperforms the WLS, and accounting for the edge effect further improves the efficiency of the GLS, especially for smooth processes, where the dependence is strong, on a small grid. Several authors (Zhu and Stein, 2002; Sun and Stein, 2016) also noted that parameter estimation becomes challenging for smooth processes. Our approach perhaps offers a solution to this problem. Moreover, in this paper, we have not considered variogram estimation

in the presence of trend (Bliznyuk et al., 2012) using our framework. This is worthy of further study.

Appendix: Proof of Theorem 1

We use the same notations as in Sections 2.1 and 3.1. The original $n \times p$ weight matrix is $\mathbf{W}$ while the weight matrix of the same dimension corresponding to the GLS procedure is $\mathbf{W}_t$.

Suppose we have a regular square grid on $[-\frac{1}{2}, \frac{1}{2}]^2$ (see Figure 7), the resulting covariance matrix for the $m-$dependent squared increments process $\mathbf{Y}$ is denoted by $\mathbf{V}$, which is the covariance matrix we are interested in. The matrix $\mathbf{V}$ is a $n \times n$ banded matrix from the $m-$dependence. Here, m denotes the spatial dependence range, or the support of the covariance. The correlation is exactly zero beyond the m distance, leading to all the zero entries in the $\mathbf{V}$ matrix as shown in Equation (5.1).

Now we consider the process on a torus, meaning the covariance function is periodic. The resulting covariance matrix is (block) circulant, denoted by $\mathbf{V}_\tau$, with the dimension of $2n \times 2n$:

$$\mathbf{V}_\tau = \begin{bmatrix} \mathbf{V} & \mathbf{V}_* \\ \mathbf{V}_*^T & \mathbf{V}^T \end{bmatrix} = \begin{bmatrix} & & \\ & 0 & \\ 0 & & \end{bmatrix}, \qquad (5.1)$$

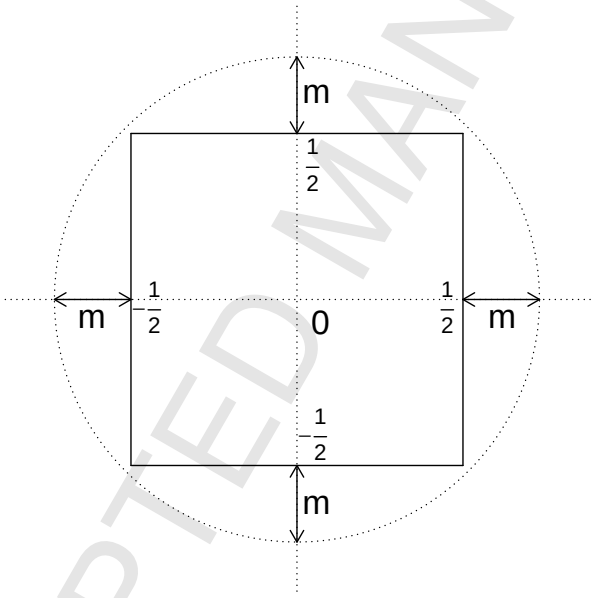

Figure 7: A regular square grid on $[-\frac{1}{2}, \frac{1}{2}]^2$.

where $\mathbf{V}^T = \mathbf{V}$ and ${\mathbf{V}_*}^T = \mathbf{V}_*$. For each row and each column of $\mathbf{V}_\tau$, there are $\mathcal{O}(m)$ non-zero elements that are bounded.

We first show the number of the non-zero elements in $\mathbf{W}_t - \mathbf{W}$ is bounded, and these elements go to 0 as $n \to \infty$. We know that

$$\mathbf{VW} = \mathbf{D}, \tag{5.2}$$

$$\mathbf{V}_\tau \mathbf{W}_\tau = \mathbf{D}_\tau = \begin{bmatrix} \mathbf{D} \\ \mathbf{D} \end{bmatrix}_{2n \times p}, \tag{5.3}$$

where

$$\mathbf{W}_\tau = \begin{bmatrix} \mathbf{W}_t \\ \mathbf{W}_t \end{bmatrix}_{2n \times p} \tag{5.4}$$

assigns equal weight to increments with equal lag. Equation (5.3) holds because of a similar idea as circular embedding, where an $n \times n$ matrix can be embedded in a larger matrix of size $2n \times 2n$. For a process defined on a torus, pairs of observations with the same lag are equivalent geometrically. These pairs of observations would be assigned an equal weight in our setup. Suppose we consider all paired increments with lags $\leq k_M$, and $n_{k_1} + \ldots + n_{k_M} = n$ where n_i is the number of increments with lag i for $i \in \mathcal{R}$, then elements of each column of $\mathbf{W}_t$ are of $\mathcal{O}(1/n)$ as each column of $\mathbf{W_t}$ is a set of normalized weights for the n elements in $\mathbf{Y}$. From (5.1), (5.3) and

(5.4), we know $\mathbf{V}\mathbf{W}_t + \mathbf{V}_*\mathbf{W}_t = \mathbf{D}$. Using (5.2), we have

$$\mathbf{V}_*\mathbf{W}_t = \mathbf{V}(\mathbf{W} - \mathbf{W}_t), \tag{5.5}$$

and the elements of each column of $\mathbf{W} - \mathbf{W}_t$ are of $\mathcal{O}(1/n)$.

Next, to prove that $\hat{\boldsymbol{\theta}}_t$ is asymptotically as efficient as $\hat{\boldsymbol{\theta}}$, we show that

$$\mathrm{Var}(\mathbf{W}_t^T\mathbf{Y}) - \mathrm{Var}(\mathbf{W}^T\mathbf{Y}) \to 0, \quad \text{as } n \to \infty.$$

We know that $\mathrm{Var}(\mathbf{W}^T\mathbf{Y}) = \mathbf{W}^T\mathbf{V}\mathbf{W}$, then

$$\begin{aligned}
\mathrm{Var}(\mathbf{W}_t^T\mathbf{Y}) - \mathrm{Var}(\mathbf{W}^T\mathbf{Y}) &= \mathbf{W}_t^T\mathbf{V}\mathbf{W}_t - \mathbf{W}^T\mathbf{V}\mathbf{W} \\
&= \mathbf{W}_t^T\mathbf{V}\mathbf{W}_t - \mathbf{W}_t^T\mathbf{V}\mathbf{W} + \mathbf{W}_t^T\mathbf{V}\mathbf{W} - \mathbf{W}^T\mathbf{V}\mathbf{W} \\
&= \mathbf{W}_t^T\mathbf{V}(\mathbf{W}_t - \mathbf{W}) + (\mathbf{W}_t^T - \mathbf{W}^T)\mathbf{V}\mathbf{W} \\
&= -\mathbf{W}_t^T\mathbf{V}_*\mathbf{W}_t - \mathbf{W}_t^T\mathbf{V}_*\mathbf{W} \qquad \text{(using (5.5))} \\
&= -\mathbf{W}_t^T\mathbf{V}_*(\mathbf{W}_t + \mathbf{W}) \\
&= \mathcal{O}(\frac{m^2 v_{max}}{n}) \qquad \text{(using } m\text{−dependence)} \\
&\to 0 \quad \text{as } n \to \infty,
\end{aligned}$$

where v_{max} is the maximum element of $\mathbf{V}_*$.

Following a similar argument, the proof for the asymptotic efficiency of the estimator obtained from the GLS with edge effect is straightforward.

References

Bliznyuk, N., Carroll, R. J., Genton, M. G., and Wang, Y. (2012). Variogram estimation in the presence of trend. *Statistics and Its Interface* **5,** 159–168.

Chang, X. and Stein, M. L. (2014). Wavelet methods in interpolation of high-frequency spatial–temporal pressure. *Spatial Statistics* **8,** 52–68.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology* **17,** 563–586.

Cressie, N. (1993). *Statistics for Spatial Data.* Wiley, New York.

Cressie, N. and Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology* **12,** 115–125.

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

Davies, S. and Hall, P. (1999). Fractal analysis of surface roughness by using spatial data. *Journal of the Royal Statistical Society, Series B* **61,** 3–37.

Genton, M. G. (1998a). Highly robust variogram estimation. *Mathematical Geology* **30,** 213–221.

Genton, M. G. (1998b). Spatial breakdown point of variogram estimators. *Mathematical Geology* **30,** 853–871.

Genton, M. G. (1998c). Variogram fitting by generalized least squares using an explicit formula for the covariance structure. *Mathematical Geology* **30,** 323–345.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102,** 359–378.

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

Hasselman, B. (2015). *nleqslv: Solve Systems of Nonlinear Equations*. R package version 2.9.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., and New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research: Atmospheres (1984–2012)* **113,** D20119.

Heuvelink, G. and Webster, R. (2001). Modelling soil variation: past, present, and future. *Geoderma* **100,** 269–301.

Huang, H. and Sun, Y. (2017). Hierarchical low rank approximation of likelihoods for large spatial datasets. Forthcoming.

Isaaks, E. H. and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, New York.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic Press, London.

Kang, W., Shin, D. W., and Lee, Y. (2003). Biases of the restricted maximum likelihood estimators for arma processes with polynomial time trend. *Journal of Statistical Planning and Inference* **116,** 163–176.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-

based estimation in large spatial data sets. *Journal of the American Statistical Association* **103,** 1545–1555.

Kitanidis, P. K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research* **19,** 909–921.

Kitanidis, P. K. (1997). *Introduction to Geostatistics: Applications in Hydrogeology.* Cambridge University Press, New York.

Kleinschmidt, I., Bagayoko, M., Clarke, G., Craig, M., and Le Sueur, D. (2000). A spatial statistical approach to malaria mapping. *International Journal of Epidemiology* **29,** 355–361.

Lahiri, S. N., Lee, Y., and Cressie, N. (2002). On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *Journal of Statistical Planning and Inference* **103,** 65–85.

Matheron, G. (1962). *Traité de géostatistique appliquée, Tome I: Mémoires du Bureau de Recherches Géologiques et Minières.* Editions Technip, Paris.

Matheron, G. (1963). Principles of geostatistics. *Economic geology* **58,** 1246–1266.

McGilchrist, C. (1989). Bias of ML and REML estimators in regression models with ARMA errors. *Journal of Statistical Computation and Simulation* **32,** 127–136.

McGrath, D., Zhang, C., and Carton, O. T. (2004). Geostatistical analyses and hazard assessment on soil lead in Silvermines area, Ireland. *Environmental Pollution* **127,** 239–248.

Moeur, M. (1993). Characterizing spatial patterns of trees using stem-mapped data. *Forest science* **39,** 756–775.

Müller, W. G. (1999). Least-squares fitting from the variogram cloud. *Statistics & Probability Letters* **43,** 93–98.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58,** 545–554.

Robock, A., Vinnikov, K. Y., Srinivasan, G., Entin, J. K., Hollinger, S. E., Speranskaya, N. A., Liu, S., and Namkhai, A. (2000). The global soil moisture data bank. *Bulletin of the American Meteorological Society* **81,** 1281–1299.

Shumway, R. H. and Stoffer, D. S. (2011). *Time Series Analysis and Its Applications*. Springer: New York.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician* **56,** 29–38.

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

Stein, M. L. (2013). Statistical properties of covariance tapers. *Journal of Computational and Graphical Statistics* **22,** 866–885.

Sun, Y., Li, B., and Genton, M. G. (2012). Geostatistics for large datasets. In *Advances and challenges in space-time modelling of natural events*, volume 207, chapter 3, pages 55–77.

Springer.

Sun, Y. and Stein, M. L. (2016). Statistically and computationally efficient estimating equations for large spatial datasets. *Journal of Computational and Graphical Statistics* **25,** 187–208.

Tunnicliffe-Wilson, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society, Series B* **15,** 15–27.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21,** 5–42.

Webster, R. and Oliver, M. A. (2007). *Geostatistics for Environmental Scientists*. John Wiley & Sons, Chichester.

Wong, D. W., Yuan, L., and Perlin, S. A. (2004). Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Science and Environmental Epidemiology* **14,** 404–415.

Yu, W. H., Harvey, C. M., and Harvey, C. F. (2003). Arsenic in groundwater in Bangladesh: a geostatistical and epidemiological framework for evaluating health effects and potential remedies. *Water Resources Research* **39,** 1146–1163.

Zhu, Z. and Stein, M. L. (2002). Parameter estimation for fractional brownian surfaces. *Statistica Sinica* **12,** 863–884.

CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia

E-mail: (`ying.sun@kaust.edu.sa`)

College of Business, Oregon State University, U.S.A.

E-mail: (xiaohui.chang@oregonstate.edu)

Department of Management Science, University of Miami, U.S.A.

E-mail: (yguan@bus.miami.edu)