On the Theoretical Guarantees for Parameter Estimation of Gaussian Random Field Models: A Sparse Precision Matrix Approach

Sam Davanloo Tajbakhsh   DAVANLOO.1@OSU.EDU
Department of Integrated Systems Engineering
The Ohio State University
Columbus, OH 43210-1273

Necdet Serhat Aybat   NSA10@PSU.EDU
Department of Industrial and Manufacturing Engineering
The Pennsylvania State University
University Park, 16802-4400

Enrique Del Castillo   EXD13@PSU.EDU
Department of Industrial and Manufacturing Engineering and Department of Statistics
The Pennsylvania State University
University Park, 16802-4400

Editor: David Wipf

Abstract

Iterative methods for fitting a Gaussian Random Field (GRF) model via maximum likelihood (ML) estimation requires solving a nonconvex optimization problem. The problem is aggravated for anisotropic GRFs where the number of covariance function parameters increases with the dimension. Even evaluation of the likelihood function requires $\mathcal{O}(n^3)$ floating point operations, where n denotes the number of data locations. In this paper[1], we propose a new two-stage procedure to estimate the parameters of second-order stationary GRFs. First, a *convex* likelihood problem regularized with a weighted ℓ_1-norm, utilizing the available distance information between observation locations, is solved to fit a sparse *precision* (inverse covariance) matrix to the observed data. Second, the parameters of the covariance function are estimated by solving a least squares problem. Theoretical error bounds for the solutions of stage I and II problems are provided, and their tightness are investigated.

Keywords: nonconvex optimization, Gaussian Markov random fields, kernel methods, hyperparameter optimization, covariance selection, spatial statistics

1. Introduction

Gaussian Random Field (GRF) models[2] are widely used in several fields, e.g., Machine Learning, Geostatistics, Computer Experiments (metamodeling) and Industrial Metrology. Traditional methods for fitting a GRF model to given sample data rely on computing the maximum likelihood estimate (MLE) of the parameters of an assumed spatial covariance

---

1. This research is based on the dissertation of the first author, supervised by Drs. Aybat and del Castillo.
2. A GRF is also called a Gaussian Process (GP) model or simply a Gaussian Field (GF).

function belonging to a known parametric family. As it is well-known in Spatial Statistics (Warnes and Ripley, 1987), the log-likelihood function for the covariance parameters of a GRF is *non-concave*, which leads to numerical difficulties in solving the optimization problem (computing a global optimal solution) for MLE, yielding suboptimal estimates that do not possess the desirable properties of MLE. Even though parametric GRF covariance matrices for *isotropic* processes require estimation of a small number of parameters, finding the MLEs is challenging due to nonconvexity of the negative loglikelihood function. Furthermore, each evaluation of the negative loglikelihood function requires $O(n^3)$ operations due to covariance matrix inversions (where n is the number of distinct data locations). Since n is typically large in GRF modeling, the computational issues due to large n is called the "big n problem" (Banerjee et al., 2008). The problem is much more significant for *anisotropic* processes where the number of parameters scales with the dimension.

To overcome these difficulties, we propose a new method, Sparse Precision matrix Selection (SPS), for fitting a GRF model, and establish the recovery guarantees and theoretical error bounds for the proposed estimator. In the *first stage* of the SPS method, a sparse precision (inverse covariance) matrix is fitted to the GRF data observations by solving a *convex* likelihood problem regularized with a weighted ℓ_1-norm, utilizing the available distance information among observation locations. This precision matrix is not parameterized in any form and constitutes a Gaussian Markov Random Field (GMRF) approximation to the GRF. The first-stage problem is solved using a variant of the Alternating Direction Method of Multipliers (ADMM) with a *linear* convergence rate. Suppose the covariance function has q parameters ($q \ll n$). In the *second stage*, these parameters are estimated via a least-squares (LS) problem in $\mathbb{R}^q$, resulting in more consistent estimates than the suboptimal solutions of the non-convex MLE problem. Although the second stage LS problem is non-convex in general, it is still numerically much easier to solve when compared to the non-convex MLE problem. In particular, the solution to the second stage LS problem can be computed via a line search in the range parameter for isotropic GRFs. Empirical evidence suggests that the first stage optimization "zooms-in" to a region in the covariance parameter space that is close to the true parameter values, alleviating the non-convexity to a certain degree. We next provide some preliminaries, including a brief review of other state-of-the-art methods.

Preliminaries

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $y(\mathbf{x})$ denote the value of a latent GRF $f : \mathcal{X} \to \mathbb{R}$ observed with additive Gaussian noise at location $\mathbf{x} \in \mathcal{X}$: $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$, where $f(\mathbf{x})$ has a mean function $m_f(\mathbf{x})$ and covariance function $c_f(\mathbf{x}, \mathbf{x}') = \mathrm{cov}\,(f(\mathbf{x}), f(\mathbf{x}'))$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and $\epsilon \sim \mathcal{N}(0, \theta_0^*)$ models the "nugget" error, assumed independent of $f(\mathbf{x})$. We assume the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i^{(r)}) : i = 1, ..., n, \ r = 1, ..., N\}$ contains N realizations of the GRF at each of n distinct locations in $\mathcal{D}^x \triangleq \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$. Let $\mathbf{y}^{(r)} = [y_i^{(r)}]_{i=1}^n \in \mathbb{R}^n$ denote the vector of r-th realization values for locations in $\mathcal{D}^x$. Given a new location $\mathbf{x}_0 \in \mathcal{X}$, the goal in GRF modeling is to predict $f_0 \triangleq f(\mathbf{x}_0)$. We assume that the GRF has a constant mean equal to zero, i.e., $m_f(\mathbf{x}) = 0$. Since any countable collection of observations from a GRF follows a multivariate normal distribution, the joint distribution of $(\mathbf{y}^\top, f_0)^\top$ is given by

$$\begin{pmatrix} \mathbf{y}^{(r)} \\ f_0 \end{pmatrix} \sim \mathcal{N} \left(\mathbf{0}_{n+1}, \begin{bmatrix} C_f + \theta_0^* \mathbf{I}_n & \mathbf{c}_0 \\ \mathbf{c}_0^\top & c_{00} \end{bmatrix} \right), \text{ for all } r = 1, \ldots, N, \text{ where } c_{00} = c_f(\mathbf{x}_0, \mathbf{x}_0),$$

$\mathbf{c}_0^\top = [c_f(\mathbf{x}_1, \mathbf{x}_0), ..., c_f(\mathbf{x}_n, \mathbf{x}_0)]$, and the covariance matrix $C_f \in \mathbb{R}^{n \times n}$ is formed such that its $(i, j)^{th}$ element is equal to $c_f(\mathbf{x}_i, \mathbf{x}_j)$. Therefore, the conditional distribution of f_0 given $\{\mathbf{y}^{(r)}\}_{r=1}^N$, i.e., the predictive distribution of f_0 denoted by $p(\cdot \mid \{\mathbf{y}^{(r)}\}_{r=1}^N)$, is given as

$$p(f_0 \mid \{\mathbf{y}^{(r)}\}_{r=1}^N) = \mathcal{N}\left(\mathbf{c}_0^\top (C_f + \theta_0^* \mathbf{I}_n)^{-1} \sum_{r=1}^N \mathbf{y}^{(r)}/N, \ c_{00} - \mathbf{c}_0^\top (C_f + \theta_0^* \mathbf{I}_n)^{-1} \mathbf{c}_0 \right). \tag{1}$$

The mean of this predictive distribution is a point estimate (known as the *Kriging* estimate in Geostatistics) and its variance measures the uncertainty of this prediction.

It is clear from (1) that the prediction performance can be made significantly robust by correctly estimating the unknown covariance function c_f, which is typically assumed to belong to some parametric family $\{c_f(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}_f) : \ \boldsymbol{\theta}_f = [\boldsymbol{\theta}_\rho^\top, \theta_v]^\top \in \Theta_f\}$, where $\Theta_f = \{(\boldsymbol{\theta}_\rho, \theta_v) \in \mathbb{R}^q \times \mathbb{R} : \ \boldsymbol{\theta}_\rho \in \Theta_\rho, \ \theta_v \geq 0\}$ is a set that contains the true parameters $\boldsymbol{\theta}_f^*$ of the f-process—this practice is common in both the Geostatistics and in the Machine Learning literature, e.g., Cressie (1993); Rasmussen and Williams (2006). In particular, we consider the following parametric form for the covariance function of the f-process:

$$c_f(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}_f) \triangleq \theta_v r(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}_\rho) \tag{2}$$

for $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $\boldsymbol{\theta}_f \in \Theta_f$, where $r(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}_\rho)$ is a parametric *correlation function*, and $\boldsymbol{\theta}_\rho$ and θ_v denote the spatial correlation and variance parameters, respectively. A covariance function is called valid if it leads to a positive definite covariance matrix for any finite set of fixed locations $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$.

For *isotropic* correlation functions we have $q = 1$; for instance, the Squared-Exponential (SE) $\exp\left(\frac{-\|\mathbf{x}-\mathbf{x}'\|^2}{\theta_\rho^2}\right)$ and the Matern-$\frac{3}{2}$ function, $\left(1 + \frac{\sqrt{3}\|\mathbf{x}-\mathbf{x}'\|}{\theta_\rho}\right) \exp\left(\frac{-\sqrt{3}\|\mathbf{x}-\mathbf{x}'\|}{\theta_\rho}\right)$. In the isotropic setting, $\boldsymbol{\theta}_\rho \in \mathbb{R}$ is the range parameter, and $\Theta_\rho = \mathbb{R}_+$. In second-order stationary *anisotropic* random fields, the correlation between two points is a function of the vector connecting the two locations rather than the distance, e.g., the anisotropic squared exponential correlation function has the following form:

$$r(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}_\rho) = \exp\left(-(\mathbf{x} - \mathbf{x}')^\top M(\boldsymbol{\theta}_\rho)(\mathbf{x} - \mathbf{x}') \right), \tag{3}$$

where $M(\boldsymbol{\theta}_\rho) \in \mathbb{S}^d$ is a symmetric matrix, e.g., $q = d$, $M(\boldsymbol{\theta}_\rho) = \mathrm{diag}(\boldsymbol{\theta}_\rho^{-2})$ and $\Theta_\rho = \mathbb{R}_+^d$.

Let $\boldsymbol{\theta}^* = [\boldsymbol{\theta}_f^{*\top}, \theta_0^*]^\top \in \Theta$ denote the unknown true parameters of the y-process, where $\Theta \triangleq \Theta_f \times \mathbb{R}_+$. Thus, the parametric covariance function of the y-process can be written as

$$c(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}) \triangleq c_f(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}_f) + \theta_0 \delta(\mathbf{x}, \mathbf{x}') \tag{4}$$

for $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$, where $\delta(\mathbf{x}, \mathbf{x}') = 1$ if $\mathbf{x} = \mathbf{x}'$ and equals 0 otherwise.

Definition 1 *Given a set of locations $\mathcal{D}^x = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$, and $\boldsymbol{\theta} = [\boldsymbol{\theta}_f^\top, \theta_0]^\top \in \Theta$, let $C_f(\boldsymbol{\theta}_f) \in \mathbb{R}^{n \times n}$ be such that its $(i, j)^{th}$ element is $c_f(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta}_f)$, and define $C(\boldsymbol{\theta}) \triangleq C_f(\boldsymbol{\theta}_f) + \theta_0 \mathbf{I}_n$, i.e., $(i, j)^{th}$ element is equal to $c(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta})$ as defined in (4). Hence, $C_f(\boldsymbol{\theta}_f^*)$ and $C^* \triangleq C(\boldsymbol{\theta}^*)$ denote true covariance matrices of the f-process and y-process, respectively, corresponding to locations $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$.*

The log marginal likelihood $\ell(\boldsymbol{\theta} \mid \mathcal{D}) \triangleq \frac{1}{N} \sum_{r=1}^N \log p\left(\mathbf{y}^{(r)} \mid \boldsymbol{\theta}, \mathcal{D}^x\right)$ can be written as

$$\ell(\boldsymbol{\theta} \mid \mathcal{D}) = -\frac{1}{2} \log \det(C(\boldsymbol{\theta})) - \frac{1}{2N} \sum_{r=1}^N \mathbf{y}^{(r)\top} C(\boldsymbol{\theta})^{-1} \mathbf{y}^{(r)} - \frac{n}{2} \log(2\pi).$$

3

Let $S = \frac{1}{N} \sum_{r=1}^{N} \mathbf{y}^{(r)} \mathbf{y}^{(r)\top}$. Given $X, Y \in \mathbb{R}^{n \times n}$, let $\langle X, Y \rangle = \mathbf{Tr}(X^\top Y)$. Hence, finding the MLE of the y-process parameters requires solving

$$\hat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \left\langle S, C(\boldsymbol{\theta})^{-1} \right\rangle + \log \det(C(\boldsymbol{\theta})) \tag{5}$$

over a set $\Theta = \Theta_f \times \mathbb{R}_+$ containing the true unknown parameters $\boldsymbol{\theta}^* = [\boldsymbol{\theta}_f^{*\top}, \theta_0^*]^\top$.

The log-likelihood function $\ell(\boldsymbol{\theta}|\mathcal{D})$ is *not* concave in $\boldsymbol{\theta}$ for many important parametric families of covariance functions. Therefore, the MLE problem in (5) is *non-convex*, which causes standard optimization routines to be trapped in local minima (Mardia and Watkins, 1989; Rasmussen and Williams, 2006). The main result of this paper given in Theorem 9, and empirical evidence from our numerical experiments indicate the reason why our *two-step* SPS approach works better compared to well-known one-step *non-convex* log-likelihood maximization approaches. SPS defers dealing with the non-convexity issue to a later stage, and first obtains a regularized log-likelihood estimate of the precision matrix solving a convex problem. At the second stage, a non-convex least-squares problem is solved, in which the global minimum is shown to be "close" to the unknown true values; moreover, it is also shown that the objective function is strongly convex in a neighborhood of the global minimum—see Figure 4.

Several other methods have been proposed in the literature to deal with the non-convex "Big n" problem in GRF estimation. These approaches can be broadly classified in six main classes: **1)** *Likelihood approximation methods* approximate the likelihood function in the spectral domain (Fuentes, 2007; Stein, 1999), or approximate it as a product of conditional densities (Vecchia, 1988; Stein et al., 2004); **2)** *Covariance tapering* provides a sparse covariance matrix in which the long range (usually weak) covariance elements are set to zero. Sparse matrix routines are then used to efficiently find the inverse and determinant of the resulting matrix (Furrer et al., 2006); **3)** *Low-rank process approximation methods* are based on a truncated basis expansion of the underlying GRF which results in reducing the computational complexity from $O(n^3)$ to $O(p^3)$, where p is the number of basis functions used to approximate the process (Higdon, 2002; Cressie and Johannesson, 2008; Banerjee et al., 2008; Nychka et al., 2015); **4)** *Sampling-based stochastic approximation* draws m sample data points ($m \ll n$) at each iteration, and the model parameters are updated according to a stochastic approximation technique until the convergence is achieved (Liang et al., 2013); **5)** *Localized GRFs* split the input domain into different segments, and the covariance parameters are estimated via ML locally on each segment (Gramacy and Apley, 2013)—this approach requires further formulation to avoid discontinuities on the predicted surface over the full domain (Park et al., 2011); and finally **6)** *Markov random field approximations*, related to our proposed SPS method, will be discussed in more detail in Section 2. There are also methods in the intersection of two classes: Sang and Huang (2012) combined low-rank process approximation with covariance tapering; Snelson and Ghahramani (2007) proposed a mix of likelihood approximation and localized GRF; Hensman et al. (2013) used stochastic variational inference for GRF models and extended their approach to non-Gaussian and latent variable models; Rodner et al. (2012) considered using GRF models for multi-class classification under the big n scenario and used parametrized histogram intersection kernels.

The rest is organized as follows: in Section 2, we motivate the proposed method. In Sections 3 and 4, we discuss the two-stage SPS method in detail and prove the statistical

properties of the SPS estimator. From a computational perspective, it is shown that the first stage has linear convergence rate, and that the second stage problem is strongly convex around the estimator, which can be solved efficiently via a line search for isotropic GRFs. Next, in Section 5, we assess the prediction performance of the proposed method comparing it to alternative methods on both synthetic and real data. Finally, in Section 6 we conclude by providing some summarizing remarks and directions for further research.

2. Motivation for the Proposed SPS Method

Let $\boldsymbol{\theta}^* \in \Theta$ be the true covariance parameters, and $C^* = C(\boldsymbol{\theta}^*) \in \mathbb{R}^{n \times n}$ be the true covariance matrix of the y-process corresponding to given locations $\mathcal{D}^x = \{\mathbf{x}_i\}_{i=1}^n$. The proposed method can be motivated by providing four interrelated remarks: **a)** the precision matrix $P^* \triangleq (C^*)^{-1}$ of a stationary GRF can be approximated with a sparse matrix; **b)** powerful convex optimization algorithms exist to solve Sparse Covariance Selection (SCS) problems to find a sparse approximation to P^*; **c)** the past and recent work on directly approximating a GRF with a GMRF also involves determining a sparse precision matrix; **d)** the available distance information among given locations can be incorporated into the GRF estimation. In the rest, let $\mathcal{I} \triangleq \{1, \ldots, n\}$ denote the location index set.

a) First motivation: Our method is motivated by the observation that $P^* = (C^*)^{-1}$ of a stationary GRF can be approximated by a sparse matrix. The *off-diagonal* element P_{ij}^*, i.e., $i \neq j$, is determined by the conditional covariance (*partial covariance*) between $y(\mathbf{x}_i)$ and $y(\mathbf{x}_j)$ given the rest of the variables; indeed, for any $i \neq j$, $|P_{ij}^*| \to 0$ as $\mathrm{Cov}\left(y(\mathbf{x}_i),\ y(\mathbf{x}_j) \mid \{y(\mathbf{x}_k)\}_{k \in \mathcal{I}_{ij}^c}\right) \to 0$, where $\mathcal{I}_{ij}^c \triangleq \mathcal{I} \setminus \{i, j\}$, because for $i \neq j$ we have

$$
P_{ij}^* = \frac{- \mathrm{Cov}\left(y(\mathbf{x}_i),\ y(\mathbf{x}_j) \mid \{y(\mathbf{x}_k)\}_{k \in \mathcal{I}_{ij}^c}\right)}{\mathrm{Var}(y(\mathbf{x}_i)|\{y(\mathbf{x}_k)\}_{k \in \mathcal{I}_{ij}^c}) \mathrm{Var}(y(\mathbf{x}_j)|\{y(\mathbf{x}_k)\}_{k \in \mathcal{I}_{ij}^c}) - \mathrm{Cov}(y(\mathbf{x}_i),y(\mathbf{x}_j)|\{y(\mathbf{x}_k)\}_{k \in \mathcal{I}_{ij}^c})^2}.
$$

In particular, *conditionally independent* variables lead to a zero entry in the precision matrix (Whittaker, 2009). This is why sparse precision matrices are common in graphical models and Bayesian networks (Whittaker, 2009) when most of the variable pairs are conditionally independent. The fact that the precision matrix of a GRF is close to sparse is related to the interesting behavior of the so-called *screen effect* in a spatial process (Cressie (1993), p. 133; Journel and Huijbregts (1978), p. 346). The *screen effect* is complete in $\mathbb{R}^1$, i.e., for given three data points on a line, the two outer points are conditionally independent (in time series models, the partial (auto)correlation function "cuts off" after lag k for a Markovian AR(k) process—see Box et al. (2008)). However, for a GRF in $\mathbb{R}^d$ with $d > 1$, the screen effect is *not* complete; hence, the corresponding precision matrix is not sparse for any *finite* set of variables pertaining to the GRF.

In addition, existing results from numerical linear algebra demonstrate that if the elements of a matrix show a decay property, then the elements of its inverse also show a similar behavior—see Benzi (2016); Jaffard (1990). In particular, consider the two decay classes defined in Jaffard (1990):

Definition 2 *Given $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ and a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, a matrix $A \in \mathbb{R}^{n \times n}$ belongs to the class $\mathcal{E}_\gamma$ for some $\gamma > 0$, if for all $\gamma' < \gamma$ there exists a constant $K_{\gamma'}$ such that*

$$
|A_{ij}| \leq K_{\gamma'}\ \exp\left(-\gamma' d(\mathbf{x}_i, \mathbf{x}_j)\right), \quad \forall\ i, j \in \mathcal{I}. \tag{6}
$$

Moreover, A belongs to the class $\mathcal{Q}_\gamma$ for some $\gamma > 1$ if there exists a constant K such that

$$|A_{ij}| \le K \left(1 + d(\mathbf{x}_i, \mathbf{x}_j)\right)^{-\gamma}, \quad \forall \ i, j \in \mathcal{I}. \tag{7}$$

Theorem 3 *Given $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ and a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, let $A \in \mathbb{R}^{n \times n}$ be an invertible matrix. If $A \in \mathcal{E}_\gamma$ for some $\gamma > 0$, then $A^{-1} \in \mathcal{E}_{\gamma'}$ for some $\gamma' > 0$. Moreover, if $A \in \mathcal{Q}_\gamma$ for some $\gamma > 0$, then $A^{-1} \in \mathcal{Q}_\gamma$.*

Proof See Proposition 2 and Proposition 3 in (Jaffard, 1990). ■

This fast decay structure in the precision matrix, as also observed in Figure 1, makes it a *"compressible signal"* (Candès et al., 2006), i.e., although not sparse, due to quick decay in its entries when sorted by their magnitude, it can be well-approximated by a sparse matrix—see Corollary 10. For all stationary GRFs tested, we observed that for a finite set of locations, the magnitudes of the off-diagonal elements of the *precision* matrix decay to 0 much *faster* than the elements of the covariance matrix. To illustrate this behavior, we compared the decay in covariance and precision elements in Figure 1 for data generated from GRFs with Matern-$\frac{3}{2}$, Squared Exponential, and Exponential covariance functions. Clearly, the precision matrix can be *better* approximated than the covariance matrix by using a sparse matrix.
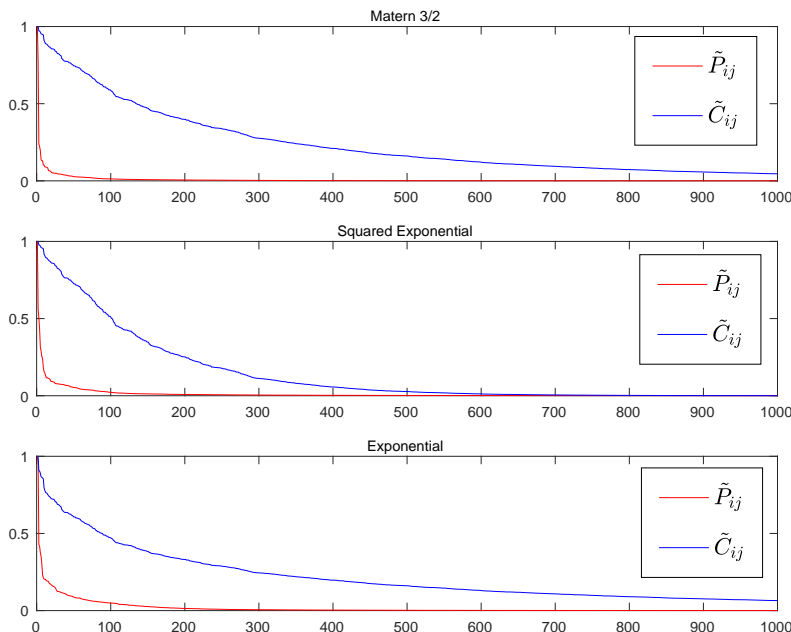

Figure 1: Decaying behavior of elements of the precision and covariance matrices for GRFs. For $(i,j) \in \mathcal{S}$, $\tilde{P}_{ij} \triangleq |P_{ij}^*|/\max\{|P_{ij}^*| : (i,j) \in \mathcal{S}\}$ and $\tilde{C}_{ij} \triangleq C_{ij}^*/\max\{C_{ij}^* : (i,j) \in \mathcal{S}\}$, where $\mathcal{S} = \{(i,j) \in \mathcal{I} \times \mathcal{I} : i \ne j\}$. The largest 1000 elements of $\{\tilde{P}_{ij}\}_{(i,j) \in \mathcal{S}}$ and $\{\tilde{C}_{ij}\}_{(i,j) \in \mathcal{S}}$ are plotted after being sorted in descending order. The underlying GRF was evaluated over a set of 100 points generated uniformly at random from $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^2 : -50 \le \mathbf{x} \le 50\}$ for three different covariance functions with range and variance parameters equal to 10, and 1, resp.

% of elements s.t. $|\tilde{P}_{ij}| > \epsilon$			
$\epsilon \setminus n$	10	100	1000
0.1	10.49	0.29	0.00
0.01	26.94	2.52	0.03
0.001	46.46	9.34	0.28

% of elements s.t. $\tilde{C}_{ij} > \epsilon$			
$\epsilon \setminus n$	10	100	1000
0.1	16.82	13.00	12.94
0.01	36.84	32.31	32.18
0.001	56.86	53.04	52.86

Table 1: Effect of increasing density in a fixed spatial domain on the near-sparsity of precision (**left**) and covariance (**right**) matrices for Matern GRF ($\nu = 3/2$) over 100×100 fixed spatial domain.

By fixing the domain of the process and increasing n (increasing the density of the data points), the *screen effect* becomes stronger, i.e., off-diagonal entries decay to 0 faster. Hence, the precision matrices can be better approximated with a sparse matrix as n increases in a fixed domain. To illustrate this phenomenon numerically, we calculate the precision matrices corresponding to a Matern GRF ($\nu = 3/2$) with variance and range parameters equal to 1 and 10, resp., for $n \in \{10, 100, 1000\}$ over a *fixed* square domain $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^2 : -50 \le \mathbf{x} \le 50\}$. Then, as a measure of near-sparsity, we computed the percentage of scaled off-diagonal elements in the precision matrix greater in absolute value than a certain threshold $\epsilon \in \{0.1, 0.01, 0.001\}$, i.e., $\text{card}\left(\{(i,j) \in \mathcal{S} : |\tilde{\mathrm{P}}_{ij}| > \epsilon\}\right)/(n^2 - n)$, and for comparison, we report the same quantities for the covariance matrices in Table 1, where $\{\tilde{P}_{ij}\}_{(i,j) \in \mathcal{S}}$ and $\{\tilde{C}_{ij}\}_{(i,j) \in \mathcal{S}}$ are defined in the caption of Figure 1. This shows the effect of *infill asymptotics* (Cressie, 1993) on the *screen effect*: in a fixed domain as n increases, precision matrices get closer to sparse matrices, while the covariance matrices are much less sensitive to increasing density.

b) Second motivation: Our work is also motivated by the recent optimization literature on the Sparse Covariance Selection (SCS) problem (Dempster, 1972) in (8)—compare it with (5). Given a sample covariance matrix $S \in \mathbb{R}^{n \times n}$ of a zero-mean multivariate Gaussian $\mathbf{y} \in \mathbb{R}^n$, d'Aspremont et al. (2008) proposed to estimate the corresponding precision matrix by solving a regularized *maximum likelihood problem*: $\min_{P \succ 0} \langle S, P \rangle - \log \det(P) + \alpha \, \text{card}(P)$, where $\text{card}(P)$ denotes the cardinality of non-zero elements of P, $P \succ 0$ denotes the cone of symmetric, positive definite (PD) matrices. This problem is combinatorially hard due to the cardinality operator in the objective function. Since the ℓ_1-norm, defined as $\|P\|_1 \triangleq \sum_{1 \le i,j \le n} |P_{ij}|$, is the tightest convex envelope of $\text{card}(\cdot)$, a convex approximation problem can be formulated as

$$\min_{P \succ 0} \quad \langle S, P \rangle - \log \det(P) + \alpha \|P\|_1. \tag{8}$$

The growth of interest in SCS in the last decade is mainly due to development of first-order algorithms that can efficiently deal with large-scale ℓ_1-regularized convex problems (Yuan, 2012; Friedman et al., 2008; Mazumder and Hastie, 2012; Honorio and Jaakkola, 2013; Hsieh et al., 2013; Lu, 2010, 2009; Scheinberg et al., 2010).

c) Third motivation: Further motivation comes from prior work on approximating a GRF with a Gaussian Markov Random Field (GMRF) to obtain computational gains using sparsity. A GRF process on a *lattice* is a GMRF under the conditional independence assumption, i.e., a variable is conditionally independent of the other variables on the lattice given its "neighbors" (Rue and Held, 2005). While the index set is countable for the lattice data, the index set $\mathcal{X}$ for a GRF is *uncountable*; hence, in general GMRF models cannot represent GRFs *exactly*. For a very special class, Lindgren et al. (2011) recently established

that the Matern GRFs are Markovian; in particular, when the smoothing parameter ν is such that $\nu - d/2 \in \mathbb{Z}_+$, where d is the dimension of the input space—see Lindgren et al. (2011); Simpson et al. (2012) for using this idea in the approximation of anisotropic and non-stationary GRFs. Rather than using a triangulation of the input space as proposed by Lindgren et al. (2011), or assuming a lattice process, we let the data determine the near-conditional independence pattern between variables through the precision matrix estimated via a weighted ℓ_1-regularization similar to that used in the SCS problem.

d) Fourth motivation: Since the spatial locations of the observations are known, i.e., $\mathcal{D}^x$, these data can be utilized to improve the estimation even when the number of realizations at each location is low. As established in Theorem 3, since $|C_{ij}^*|$ decreases to 0 exponentially as $\|\mathbf{x}_i - \mathbf{x}_j\|_2$ increases, $|P_{ij}^*|$ decays to 0 exponentially fast as well. In fact, for all stationary covariance functions tested, we observed this behavior—see Figure 2. Therefore, this information can be utilized for regularizing the likelihood function (see Section 3.1).
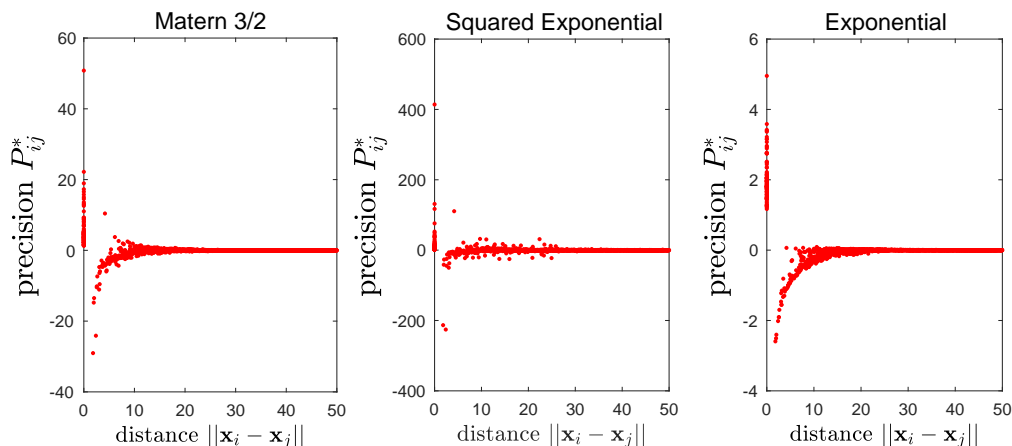

Figure 2: Elements of precision matrices from three covariance functions as a function of the Euclidean distance between the data points. The variance, range, and nugget parameters of the covariance functions are 1, 10, and 0, respectively.

3. The SPS Algorithm for Fitting a GRF Model

The proposed method for fitting a GRF is composed of two stages: 1) the true precision matrix corresponding to the training data set is approximated with a sparse matrix by solving a convex maximum likelihood problem regularized with a weighted ℓ_1-norm; 2) after inverting the fitted precision matrix from the first stage, a least-squares problem is solved to estimate the unknown covariance function parameters.

3.1 STAGE-I: Estimation of Precision Matrices

Consider the *unknown* parameter vector $\boldsymbol{\theta}^* = [\boldsymbol{\theta}_f^{*\top}, \theta_0^*]^\top \in \Theta \triangleq \Theta_f \times \mathbb{R}_+$ and suppose $\theta_0 \geq 0$ and $\boldsymbol{\theta}_f^* \in \text{int}(\Theta_f)$, i.e., $\boldsymbol{\theta}_\rho^* \in \text{int}(\Theta_\rho)$, $\boldsymbol{\theta}_v^* > 0$. Let $C^* = C(\boldsymbol{\theta}^*)$ be the covariance matrix of a *zero-mean* GRF corresponding to $\mathcal{D}^x = \{\mathbf{x}_i\}_{i \in \mathcal{I}} \subset \mathcal{X}$, and $P^* = (C^*)^{-1}$, where $\mathcal{I} \triangleq \{1, \ldots, n\}$. We define $1/0 = \infty$ and $1/\infty = 0$. Fix $0 \leq a^* \leq b^* \leq \infty$ satisfying $0 \leq \frac{1}{b^*} \leq \sigma_{\min}(C^*) \leq \sigma_{\max}(C^*) \leq \frac{1}{a^*}$ where $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denote minimum and maximum

singular values, respectively; hence, $a^* \leq \sigma_{\min}(P^*) \leq \sigma_{\max}(P^*) \leq b^*$. Given $\mathcal{D} = \{(\mathbf{x}_i, y_i^{(r)}) : i \in \mathcal{I}, \ r = 1, ..., N\}$, compute the unbiased estimator of C^*, $S = \frac{1}{N} \sum_{r=1}^N \mathbf{y}^{(r)} \mathbf{y}^{(r)\top} \in \mathbb{S}^n$, where $\mathbb{S}^n$ denotes the set of $n \times n$ symmetric matrices, and $\mathbf{y}^{(r)} = [y_i^{(r)}]_{i=1}^n \in \mathbb{R}^n$. Furthermore, we form the distance matrix $\tilde{G} \in \mathbb{S}^n$ as follows: for all $i \in \mathcal{I}$, let

$$\tilde{G}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|, \quad \forall j \in \mathcal{I} \setminus \{i\}, \quad \tilde{G}_{ii} = \min\{\|\mathbf{x}_i - \mathbf{x}_j\| : \ j \in \mathcal{I} \setminus \{i\}\}. \tag{9}$$

Let $\tilde{G}_{\min} \triangleq \min_{i \in \mathcal{I}} \tilde{G}_{ii}$. Next, we define the weight matrix as follows:

$$G_{ij} = \tilde{G}_{ij}/\tilde{G}_{\min} \quad \forall \, (i,j) \in \mathcal{I} \times \mathcal{I}. \tag{10}$$

To approximate the true precision matrix with a sparse matrix, we propose to solve the following convex problem:

$$\hat{P} \triangleq \operatorname{argmin}\{\langle S, P \rangle - \log \det(P) + \alpha \langle G, |P| \rangle : \ a^*\mathbf{I} \preceq P \preceq b^*\mathbf{I}\}, \tag{11}$$

where $|\cdot|$ is the element-wise absolute value operator; hence, the last term is a weighted ℓ_1-norm with weights equal to the normalized distances G_{ij}—compare it with (5) and (8). The choice of weighted ℓ_1-norm with weights equal to the pairwise distances between the points is supported by the trends observed in Figure 2. Choosing $G_{ii} > 0$ as in (10) controls the magnitude of diagonal elements of $\hat{P}$ relative to off-diagonals – if we did not penalize the diagonal elements, i.e., G_{ii} was set to 0, then some diagonal entries might get unreasonably large. Note that $\hat{P}$ is always a full rank matrix due to the $\log \det(\cdot)$ term in the objective function. Furthermore, having non-trivial bounds $0 < a^* \leq b^* < \infty$ is useful in practice to control the condition number of the estimator, which is also argued for in (d'Aspremont et al., 2008; Rothman et al., 2008).

If there is no prior information on the process to obtain non-trivial $0 < a^* \leq b^* < \infty$, then setting $a^* = 0$, and $b^* = \infty$ trivially satisfies the condition on a^* and b^*. For this case, (11) reduces to $\min\{\langle S, P \rangle - \log \det(P) + \alpha \langle G, |P| \rangle : \ P \succ \mathbf{0}\}$. On the other hand, when there is prior information on the process, one can also exploit it to obtain non-trivial bounds $a^* > 0$ and $b^* < \infty$. For instance, let $C^* = C(\boldsymbol{\theta}^*) = C_f(\boldsymbol{\theta}_f^*) + \theta_0^*\mathbf{I}$ be the true covariance matrix corresponding to locations in $\mathcal{D}^x$, where $\boldsymbol{\theta}_f^* = [\boldsymbol{\theta}_\rho^{*\top}, \theta_v^*]^\top$, $\boldsymbol{\theta}_\rho^* \in \operatorname{int}(\Theta_\rho)$ and $\theta_v^* > 0$ denote the true spatial correlation and variance parameters of the f-process. The common structure of the covariance functions implies that $\operatorname{diag}(C_f(\boldsymbol{\theta}_f^*)) = \theta_v^*\mathbf{1}$, where $\mathbf{1}$ denotes the vector of ones. Therefore, $\sigma_{\min}(P^*) \geq 1/\operatorname{\mathbf{Tr}}(C(\boldsymbol{\theta}^*)) = \frac{1}{n(\theta_0^* + \theta_v^*)}$. Hence, if upper bounds on θ_0^* and θ_v^* are known a priori, then one can obtain non-trivial lower bounds.

In comparison to the SCS problem in (8), the proposed formulation (11) penalizes each element of the estimator $\hat{P}_{ij}$ for $i \neq j$ with a different weight proportional to $\|\mathbf{x}_i - \mathbf{x}_j\|_2$, i.e., the distance between the corresponding locations. This model assumes that the off-diagonal precision magnitudes decrease with distance, for which there is an empirical evidence as shown in Figure 2. More importantly, Theorem 3 shows that this assumption indeed always holds when the covariance elements decay with increasing $\|\mathbf{x}_i - \mathbf{x}_j\|_2$.

The form of STAGE-I problem in (11) is well-studied in the optimization literature. Indeed, for any $0 \leq a^* \leq b^* \leq \infty$, one can use the Alternating Direction Method of Multipliers (ADMM) to generate a sequence of iterates that Q-linearly[3] converges to $\hat{P}$, where

---

3. Let $\{X_\ell\}$ converge to X^* for a given norm $\|.\|$. The convergence is called Q-linear if $\frac{\|X_{\ell+1} - X^*\|}{\|X_\ell - X^*\|} \leq c$, for some $c \in (0,1)$

$\hat{P}$ is the unique optimal solution to STAGE-I problem given in (11). For the sake of completeness, in the online-only supplementary material, we provide an ADMM algorithm and state its convergence properties—see Figure 11 and Theorem 12 in the online supplement.

3.2 STAGE-II: Estimation of Covariance Function Parameters

After estimating the precision matrix in the first stage according to (11), a least-squares problem is solved in the second stage to fit a parametric covariance function to $\hat{P}$. Although this is a non-convex problem for parametric covariance functions in general, our main result, Theorem 9, and empirical evidence from our numerical experiments suggest that non-convexity of this problem is much less serious than that of the likelihood function in (5). In STAGE-II, we propose to estimate the covariance parameters by solving

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \|C(\boldsymbol{\theta}) - \hat{P}^{-1}\|_F^2, \tag{12}$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_\rho^\top, \theta_v, \theta_0]^\top$, $\Theta \triangleq \{\boldsymbol{\theta} : \boldsymbol{\theta}_\rho \in \Theta_\rho, \ \theta_v \geq 0, \ \theta_0 \geq 0\}$, and $C(\boldsymbol{\theta})$ is the parametric covariance matrix, as given in Definition 1, corresponding to the locations of the training data $\mathcal{D}^x$. Here $\boldsymbol{\theta}_\rho \in \mathbb{R}^q$ denotes the spatial parameters of the *correlation* function, θ_v is the variance parameter, and θ_0 is the nugget, which in some applications is set equal to zero. Indeed, $C_{ij}(\boldsymbol{\theta}) = c(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta})$ for $(i,j) \in \mathcal{I} \times \mathcal{I}$. The SPS method is summarized in Figure 3.

Algorithm SPS $(\mathcal{D})$
1: **input** : $\mathcal{D} = \{(\mathbf{x}_i, \ y_i^{(r)}) : \ r = 1, \ldots, N, \ i \in \mathcal{I}\} \subset \mathcal{X} \times \mathbb{R}$
2: /* Compute sample covariance and the distance penalty matrices */
3: $\mathbf{y}^{(r)} \leftarrow [y_i^{(r)}]_{i=1}^n, \ S \leftarrow \frac{1}{N} \sum_{r=1}^N \mathbf{y}^{(r)} \mathbf{y}^{(r)\top}$
4: $\tilde{G}_{ij} \leftarrow \|\mathbf{x}_i - \mathbf{x}_j\|_2, \ \forall (i,j) \in \mathcal{I} \times \mathcal{I}, \ \tilde{G}_{ii} \leftarrow \min\{\|\mathbf{x}_i - \mathbf{x}_j\|_2 : \ j \in \mathcal{I} \setminus \{i\}\}, \ \forall i \in \mathcal{I}$
5: $\tilde{G}_{\min} \leftarrow \min_{i \in \mathcal{I}} \tilde{G}_{ii}, \ G_{ij} \leftarrow \tilde{G}_{ij}/\tilde{G}_{\min}, \ \forall (i,j) \in \mathcal{I} \times \mathcal{I}$
6: /* Compute the fitted precision matrix—See Section 3.1 */
7: $\hat{P} \leftarrow \operatorname{argmin}\{\langle S, P\rangle - \log\det(P) + \alpha\langle G, |P|\rangle : \ a^*\mathbf{I} \preceq P \preceq b^*\mathbf{I}\}$
8: /* Estimate covariance function parameters—See Section 3.2 */
9: $\hat{\boldsymbol{\theta}} \leftarrow \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \|C(\boldsymbol{\theta}) - \hat{P}^{-1}\|_F^2$
10: **return** $\hat{\boldsymbol{\theta}}$

Figure 3: Sparse Precision matrix Selection (SPS) method

Solution to the STAGE-II problem. Let $\hat{C} \triangleq \hat{P}^{-1}$, where $\hat{P}$ is defined in (11). Consider sequentially solving (12): Fixing $\boldsymbol{\theta}_\rho$, the objective in (12) is first minimized over θ_v and θ_0 in closed form (inner optimization); hence, it can be written as a function of $\boldsymbol{\theta}_\rho$ *only*. Next, the resulting function is minimized over $\boldsymbol{\theta}_\rho$ (outer optimization), i.e.,

$$\min_{\boldsymbol{\theta}_\rho \in \Theta_\rho} \left\{ \min_{\theta_v \geq 0, \ \theta_0 \geq 0} \ \frac{1}{2} \sum_{i,j \in \mathcal{I}} \left(\theta_v \ r(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta}_\rho) + \theta_0 \ \delta(\mathbf{x}_i, \mathbf{x}_j) - \hat{C}_{ij} \right)^2 \right\}, \tag{13}$$

where $\delta(\mathbf{x}_i, \mathbf{x}_j) = 1$ if $\mathbf{x}_i = \mathbf{x}_j$, and equals 0 otherwise. Consider the inner optimization problem written as follows:

$$\psi(\boldsymbol{\theta}_\rho; \hat{\mathbf{c}}) \triangleq \min_{\theta_0 \geq 0, \ \theta_v \geq 0} \ \frac{1}{2} \|\theta_v \mathbf{r}(\boldsymbol{\theta}_\rho) + \theta_0 \mathbf{d} - \hat{\mathbf{c}}\|^2, \tag{14}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\hat{\mathbf{c}}$, $\mathbf{r}$, and $\mathbf{d}$ are *long vectors* in $\mathbb{R}^{n^2}$ such that $\hat{\mathbf{c}}_{ij} = \hat{C}_{ij}$, $\mathbf{r}_{ij}(\boldsymbol{\theta}_\rho) = r(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta}_\rho)$, and $\mathbf{d}_{ij} = \delta(\mathbf{x}_i, \mathbf{x}_j)$ for $(i,j) \in \mathcal{I} \times \mathcal{I}$. We write $\mathbf{r}(\boldsymbol{\theta}_\rho)$ as $\mathbf{r}$ for short when we do not need to emphasize the dependence of $\mathbf{r}$ on $\boldsymbol{\theta}_\rho$.

Theorem 4 *For any given $\boldsymbol{\theta}_\rho \in \Theta_\rho$, the minimization problem in (14) has a unique global optimal solution $(\hat{\theta}_v, \hat{\theta}_0)$ that can be computed as*

$$(\hat{\theta}_v, \hat{\theta}_0) = \begin{cases} (0, \ \mathbf{d}^\top \hat{\mathbf{c}}/n) & \text{if } \ \mathbf{r}^\top \hat{\mathbf{c}} \leq \mathbf{d}^\top \hat{\mathbf{c}}, \\ \left(\frac{\mathbf{r}^\top \hat{\mathbf{c}} - \mathbf{d}^\top \hat{\mathbf{c}}}{\|\mathbf{r}\|^2 - n}, \frac{\mathbf{d}^\top \hat{\mathbf{c}} \ \|\mathbf{r}\|^2/n - \mathbf{r}^\top \hat{\mathbf{c}}}{\|\mathbf{r}\|^2 - n} \right) & \text{if } \ \mathbf{d}^\top \hat{\mathbf{c}} < \mathbf{r}^\top \hat{\mathbf{c}} < \mathbf{d}^\top \hat{\mathbf{c}} \ \|\mathbf{r}\|^2/n, \\ (\mathbf{r}^\top \hat{\mathbf{c}}/\|\mathbf{r}\|^2, 0) & \text{if } \ \mathbf{d}^\top \hat{\mathbf{c}} \ \|\mathbf{r}\|^2/n \leq \mathbf{r}^\top \hat{\mathbf{c}}. \end{cases} \qquad (15)$$

Corollary 5 *In the absence of the nugget parameter, i.e., $\theta_0^* = 0$, the unique global minimizer to the problem $\min_{\theta_v \geq 0} \ \frac{1}{2}\|\theta_v \mathbf{r} - \hat{\mathbf{c}}\|^2$ is $\hat{\theta}_v \triangleq \max\{0, \mathbf{r}^\top \hat{\mathbf{c}}/\|\mathbf{r}\|^2\}$.*

Using Theorem 4 or Corollary 5, the solution to the inner problem can be computed as a function of the outer optimization variable, $\boldsymbol{\theta}_\rho$, in (13). In Lemma 7, we show that under certain conditions, the outer problem objective, $\psi(\boldsymbol{\theta}_\rho; \hat{\mathbf{c}})$, is strongly convex in $\boldsymbol{\theta}_\rho$ around the global minimum. Moreover, for isotropic covariance functions, $\hat{\boldsymbol{\theta}}_\rho = \operatorname{argmin}\{\psi(\boldsymbol{\theta}_\rho; \hat{\mathbf{c}}) : \boldsymbol{\theta}_\rho \in \mathbb{R}_+\}$ can be simplified to a one-dimensional line search over $[0, D_{\max}]$, where $D_{\max}$ is an upper bound on $\hat{\boldsymbol{\theta}}_\rho$. This is illustrated in Figure 4 which displays the STAGE-II objective as a function of $\boldsymbol{\theta}_\rho$. One can compute $\psi(\boldsymbol{\theta}_\rho; \hat{\mathbf{c}})$ as described in Section 3.2 using the sample data $\mathcal{D}$ coming from an *isotropic* GRF with true parameters $(\theta_\rho^* = 4, \theta_\nu^* = 8, \theta_0^* = 4)$ for a SE covariance function. Note $\psi(\cdot; \hat{\mathbf{c}})$ is *unimodal* with global minimum close the true $\boldsymbol{\theta}_\rho^*$ value. The univariate minimization is performed via bisection; hence, after $\log_2(D_{\max}/\epsilon)$ iterations, the search reaches a point within ϵ-ball of $\hat{\boldsymbol{\theta}}_\rho$.
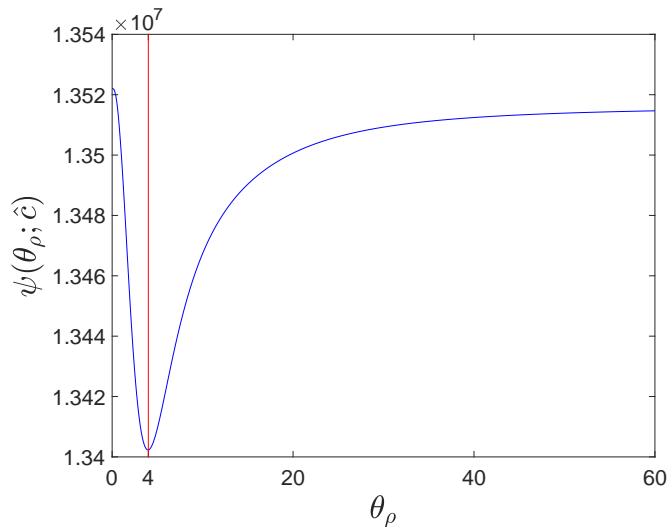

Figure 4: STAGE-II outer optimization objective $\psi(\boldsymbol{\theta}_\rho; \hat{\mathbf{c}})$. The red line shows the true parameter $\boldsymbol{\theta}_\rho^* = 4$.

4. Statistical Analysis of the SPS Estimator

We now focus on the statistical convergence properties of the parameter estimates obtained by the SPS algorithm displayed in Figure 3. Given $S, G \in \mathbb{S}^n$, the SPS estimator of the precision matrix, defined in (11), can be computed efficiently using an ADMM algorithm—see Theorem 12 in the online supplement. Throughout this section, we assume that non-trivial bounds $0 < a^* \leq b^* < \infty$ are given. This same assumption is also made by Rothman et al. (2008) to analyze the statistical properties of an estimator related to $\hat{P}$ in (11)—see

Assumptions A2 and A3 in (Rothman et al., 2008). Moreover, these bounds are useful in practice to control the condition number of the estimator (d'Aspremont et al., 2008).

4.1 Learning the Precision Matrix

Theorem 6 *Let $\{\mathbf{y}^{(r)}\}_{r=1}^N \subset \mathbb{R}^n$ be independent realizations of a GRF with zero-mean and stationary covariance function $c(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}^*)$ observed over n distinct locations $\{\mathbf{x}_i\}_{i \in \mathcal{I}}$ with $\mathcal{I} \triangleq \{1, ..., n\}$; furthermore, let $P^* \triangleq (C^*)^{-1}$ be the corresponding true precision matrix for these observations. Finally, let $\hat{P}$ be the SPS estimator computed as in (11) for some $G \in \mathbb{S}^n$ such that $G_{ij} \geq 0$ for all $(i, j) \in \mathcal{I} \times \mathcal{I}$. Then for any given $M > 0$ and $N \geq N_0 \triangleq \lceil 2(M+2)\log n + \log 16 \rceil$, we have*

$$Pr\Big(\|\hat{P} - P^*\|_F \leq 2b^{*2}(n + \|G\|_F)\alpha\Big) \geq 1 - n^{-M},$$

for all b^* and α such that $\sigma_{\max}(P^*) \leq b^*$ and $40(\theta_v^* + \theta_0^*)\sqrt{N_0/N} \leq \alpha \leq 40(\theta_v^* + \theta_0^*)$.

Proof Through the change of variables $\Delta \triangleq P - P^*$, we can write (11) in terms of Δ:

$$\hat{\Delta} = \operatorname{argmin}\{F(\Delta) \triangleq \langle S, \Delta + P^* \rangle - \log\det(\Delta + P^*) + \alpha\langle G, |\Delta + P^*| \rangle : \Delta \in \mathcal{F}\},$$

where $\mathcal{F} \triangleq \{\Delta \in \mathbb{R}^{n \times n} : \Delta = \Delta^\top, a^*\mathbf{I} \preceq \Delta + P^* \preceq b^*\mathbf{I}\}$. Note that $\hat{\Delta} = \hat{P} - P^*$. Define $g(\Delta) \triangleq -\log\det(\Delta + P^*)$ on $\mathcal{F}$. $g(\cdot)$ is strongly convex over $\mathcal{F}$ with modulus $1/b^{*2}$; hence, for any $\Delta \in \mathcal{F}$, it follows that $g(\Delta) - g(\mathbf{0}) \geq -\langle (P^*)^{-1}, \Delta \rangle + \frac{1}{2b^{*2}}\|\Delta\|_F^2$. Let $H(\Delta) \triangleq F(\Delta) - F(\mathbf{0})$ and $S_\Delta \triangleq \{\Delta \in \mathcal{F} : \|\Delta\|_F > 2b^{*2}(n + \|G\|_F)\alpha\}$.

Under probability event $\Omega = \{|S_{ij} - C_{ij}^*| \leq \alpha, \ \forall(i, j) \in \mathcal{I} \times \mathcal{I}\}$, for any $\Delta \in S_\Delta \subset \mathcal{F}$,

$$H(\Delta) \geq \langle S, \Delta \rangle - \langle (P^*)^{-1}, \Delta \rangle + \frac{1}{2b^{*2}}\|\Delta\|_F^2 + \alpha\langle G, |\Delta + P^*| \rangle - \alpha\langle G, |P^*| \rangle \tag{16}$$

$$\geq \frac{1}{2b^{*2}}\|\Delta\|_F^2 + \langle \Delta, S - C^* \rangle - \alpha\langle G, |\Delta| \rangle \tag{17}$$

$$\geq \frac{1}{2b^{*2}}\|\Delta\|_F^2 - \alpha(n + \|G\|_F)\|\Delta\|_F > 0, \tag{18}$$

where the second inequality follows from the triangle inequality, the third one holds under the probability event Ω and follows from the Cauchy-Schwarz inequality, and the final strict one follows from the definition of S_Δ. Since $F(\mathbf{0})$ is a constant, $\hat{\Delta} = \operatorname{argmin}\{H(\Delta) : \Delta \in \mathcal{F}\}$. Hence, $H(\hat{\Delta}) \leq H(\mathbf{0}) = 0$. Therefore, $\hat{\Delta} \notin S_\Delta$ under the probability event Ω. It is important to note that $\hat{\Delta} \in \mathcal{F}$; hence, $\hat{\Delta} \notin S_\Delta$ implies $\|\hat{\Delta}\|_F \leq 2b^{*2}(n + \|G\|_F)\alpha$ whenever the probability event Ω is true. Thus,

$$\begin{aligned}
\Pr\Big(\|\hat{P} - &P^*\|_F \leq 2b^{*2}(n + \|G\|_F)\alpha\Big) \\
&\geq \Pr\big(|S_{ij} - C_{ij}^*| \leq \alpha, \ \forall(i, j) \in \mathcal{I} \times \mathcal{I}\big) \\
&= 1 - \Pr\Big(\max_{i,j \in \mathcal{I}} |S_{ij} - C_{ij}^*| > \alpha\Big) \geq 1 - \sum_{i,j \in \mathcal{I}} \Pr\big(|S_{ij} - C_{ij}^*| > \alpha\big).
\end{aligned}$$

Recall that $S = \frac{1}{N}\sum_{r=1}^N \mathbf{y}^{(r)}\mathbf{y}^{(r)\top}$ and $\mathbf{y}^{(r)} = [y_i^{(r)}]_{i \in \mathcal{I}}$ for $r = 1, \ldots, N$. Since $y_i^{(r)}/\sqrt{C_{ii}^*} \sim \mathcal{N}(\mu = 0, \sigma = 1)$, i.e., standard normal, for all i and r, Lemma 1 in (Ravikumar et al., 2011)

implies $\Pr\left(|S_{ij} - C^*_{ij}| > \alpha\right) \leq B_\alpha$ for any $(i,j) \in \mathcal{I} \times \mathcal{I}$ and $\alpha \in (0, 40 \max_{i \in \mathcal{I}} C^*_{ii})$, where

$$B_\alpha \triangleq 4 \exp\left(\frac{-N}{2}\left(\frac{\alpha}{40 \max_{i \in \mathcal{I}} C^*_{ii}}\right)^2\right).$$

Hence, given any $M > 0$, by requiring $N \geq \left(\frac{40 \max_{i \in \mathcal{I}} C^*_{ii}}{\alpha}\right)^2 N_0$, we get $B_\alpha \leq \frac{1}{n^2} n^{-M}$. Thus, for any $N \geq N_0$, we have $\sum_{i,j \in \mathcal{I}} \Pr\left(|S_{ij} - C^*_{ij}| > \alpha\right) \leq n^{-M}$ for all $40(\theta^*_v + \theta^*_0)\sqrt{\frac{N_0}{N}} \leq \alpha \leq 40(\theta^*_v + \theta^*_0)$ since $C^*_{ii} = \theta^*_v + \theta^*_0$ for all $i \in \mathcal{I}$; and this completes the proof. ∎

After setting α to its lower bound in Theorem 6, our bound conforms with the one provided by Rothman et al. (2008) (Theorem 1) which states $\|\hat{P}_\alpha - P^*\|_F = \mathcal{O}\left(\sqrt{\frac{\operatorname{card}(P^*)\log n}{N}}\right)$, where $\hat{P}_\alpha$ is the solution to (8). Furthermore, our proof uses the strong convexity property of the objective function, and does not require second-order differentiability of the objective as in Rothman et al. (2008); hence, ours is a simpler proof leading to a similar conclusion.

4.2 Learning the Hyperparameters

To ease the notational burden in the proofs, we assume the nugget parameter $\theta^*_0 = 0$ in the rest of the discussion. Hence, $\boldsymbol{\theta}^* = [\boldsymbol{\theta}^{*\top}_\rho, \theta^*_v]^\top \in \mathbb{R}^{q+1}$.

Recall that $\mathbf{r}$ in Theorem 4 and Corollary 5 depends on $\boldsymbol{\theta}_\rho$, i.e., $\mathbf{r}(\boldsymbol{\theta}_\rho)$, where $\boldsymbol{\theta}_\rho$ is the decision variable of the outer optimization in (13). After substituting the optimal solution from Corollary 5, the objective of the outer problem in (13) is obtained by evaluating

$$\psi(\boldsymbol{\theta}_\rho; \mathbf{c}) = \frac{1}{2}\|\max\left\{0, \frac{\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \mathbf{c}}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2}\right\}\mathbf{r}(\boldsymbol{\theta}_\rho) - \mathbf{c}\|^2, \tag{19}$$

at $\mathbf{c} = \hat{\mathbf{c}}$, where $\mathbf{r}(\boldsymbol{\theta}_\rho)$ and $\hat{\mathbf{c}}$ are defined as in (14). In the rest, $\mathbb{R}_{++} \triangleq \{t \in \mathbb{R} : t > 0\}$.

Lemma 7 *Let* $\boldsymbol{\theta}^* = [\boldsymbol{\theta}^{*\top}_\rho, \theta^*_v]^\top \in \operatorname{int}(\Theta_\rho) \times \mathbb{R}_{++}$ *be the true parameters, and* $\mathbf{c}^* = \theta^*_v \mathbf{r}(\boldsymbol{\theta}^*_\rho) \in \mathbb{R}^{n^2}$ *be the true covariance matrix* C^* *in vector form, i.e.,* $\mathbf{c}^* \in \mathbb{R}^{n^2}$ *such that* $\mathbf{c}^*_{ij} = C^*_{ij}$ *for* $(i,j) \in \mathcal{I} \times \mathcal{I}$. *Suppose the correlation function* $r(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}_\rho)$ *is twice continuously differentiable in* $\boldsymbol{\theta}_\rho$ *over* Θ_ρ *for all* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, *then there exists* $\gamma^* > 0$ *such that* $\nabla^2 \psi(\boldsymbol{\theta}^*_\rho; \mathbf{c}^*) \succeq \gamma^* \mathbf{I}$ *if and only if* $\{\mathbf{r}(\boldsymbol{\theta}^*_\rho), \mathbf{r}'_1(\boldsymbol{\theta}^*_\rho), \ldots, \mathbf{r}'_q(\boldsymbol{\theta}^*_\rho)\} \subset \mathbb{R}^{n^2}$ *are linearly independent, where* $\mathbf{r}'_j(\boldsymbol{\theta}^*_\rho)$ *is the j-th column of* $\mathbf{Jr}(\boldsymbol{\theta}^*_\rho)$, *i.e., the Jacobian of* $\mathbf{r} : \mathbb{R}^q \to \mathbb{R}^{n^2}$ *at* $\boldsymbol{\theta}^*_\rho$.

Proof Recall $\mathbf{r} : \mathbb{R}^q \to \mathbb{R}^{n^2}$ such that $\mathbf{r}_{ij}(\boldsymbol{\theta}_\rho) = r(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta}_\rho)$ for $(i,j) \in \mathcal{I} \times \mathcal{I}$. Let $\boldsymbol{\theta}_\rho = [\xi_1, \ldots, \xi_q]^\top$. Define $g : \mathbb{R}^q \times \mathbb{R}^{n^2} \to \mathbb{R}$ such that

$$g(\boldsymbol{\theta}_\rho; \mathbf{c}) \triangleq \frac{1}{2}\|\left(\frac{\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \mathbf{c}}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2}\right)\mathbf{r}(\boldsymbol{\theta}_\rho) - \mathbf{c}\|^2. \tag{20}$$

Note the objective of the outer problem in (13), i.e., $\psi(\boldsymbol{\theta}_\rho; \hat{\mathbf{c}})$ defined in (19), is equal to $g(\boldsymbol{\theta}_\rho; \hat{\mathbf{c}})$ whenever $\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \hat{\mathbf{c}} \geq 0$. Let $\mathbf{z} : \mathbb{R}^q \times \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ such that $\mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c}) \triangleq \left(\frac{\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \mathbf{c}}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2}\right)\mathbf{r}(\boldsymbol{\theta}_\rho) -$

$\mathbf{c}$ and define $p(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{x}\|^2$, where $\boldsymbol{\theta}_\rho$ and $\mathbf{c}$ are the variable and parameter vectors of function $\mathbf{z}$, respectively. Hence, $g(\boldsymbol{\theta}_\rho; \mathbf{c}) = p(\mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c}))$. In the rest, all the derivatives for $\mathbf{z}$ and g are written with respect to $\boldsymbol{\theta}_\rho$ only, not $\mathbf{c}$. Applying the chain rule we obtain:

$$\nabla g(\boldsymbol{\theta}_\rho; \mathbf{c}) = \mathbf{Jz}(\boldsymbol{\theta}_\rho; \mathbf{c})^\top \nabla p\left(\mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c})\right) = \mathbf{Jz}(\boldsymbol{\theta}_\rho; \mathbf{c})^\top \mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c}), \tag{21}$$

where $\mathbf{Jz}(\boldsymbol{\theta}_\rho; \mathbf{c}) \in \mathbb{R}^{n^2 \times q}$ denotes the Jacobian matrix, i.e., for $(i,j) \in \mathcal{I} \times \mathcal{I}$, and $k \in \{1, \ldots, q\}$, $(\mathbf{Jz}(\boldsymbol{\theta}_\rho; \mathbf{c}))_{(i,j),k} = \frac{\partial}{\partial \xi_k} \mathbf{z}_{ij}(\boldsymbol{\theta}_\rho; \mathbf{c})$. Let $\mathbf{Hz}(\boldsymbol{\theta}_\rho; \mathbf{c}) \in \mathbb{R}^{n^2 q \times q}$ be the matrix of second-order derivatives of $\mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c})$, i.e., $\mathbf{Hz}(\boldsymbol{\theta}_\rho; \mathbf{c}) = [w_{k_1, k_2}]_{k_1, k_2 \in \{1, \ldots, q\}}$ and $w_{k_1, k_2} = \frac{\partial^2}{\partial \xi_{k_1} \partial \xi_{k_2}} \mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c}) \in \mathbb{R}^{n^2}$. Let $\mathbf{I}_q$ denote $q \times q$ identity matrix. Then the Hessian of g can be written as follows:

$$\nabla^2 g(\boldsymbol{\theta}_\rho; \mathbf{c}) = \mathbf{Hz}(\boldsymbol{\theta}_\rho; \mathbf{c})^\top \left(\mathbf{I}_q \otimes \nabla p\left(\mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c})\right)\right) + \mathbf{Jz}(\boldsymbol{\theta}_\rho; \mathbf{c})^\top \nabla^2 p\left(\mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c})\right) \mathbf{Jz}(\boldsymbol{\theta}_\rho; \mathbf{c})$$

$$= \mathbf{Hz}(\boldsymbol{\theta}_\rho; \mathbf{c})^\top \left(\mathbf{I}_q \otimes \mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c})\right) + \mathbf{Jz}(\boldsymbol{\theta}_\rho; \mathbf{c})^\top \mathbf{Jz}(\boldsymbol{\theta}_\rho; \mathbf{c}). \tag{22}$$

Note $\mathbf{c}^* = \theta_v^* \mathbf{r}(\boldsymbol{\theta}_\rho^*)$; hence, $\theta_v^* = \frac{\mathbf{r}(\boldsymbol{\theta}_\rho^*)^\top \mathbf{c}^*}{\|\mathbf{r}(\boldsymbol{\theta}_\rho^*)\|^2}$. Therefore, $\mathbf{z}(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*) = \mathbf{0}$, and the definition of g in (20) implies that $g(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*) = 0$, and $\nabla g(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*) = \mathbf{0}$. Thus,

$$\nabla^2 g(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*) = \mathbf{Jz}(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*)^\top \mathbf{Jz}(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*), \tag{23}$$

which is clearly a positive semidefinite matrix. Next, we investigate the condition under which $\nabla^2 g(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*)$ is *positive definite*. $\mathbf{Jz}(\boldsymbol{\theta}_\rho; \mathbf{c})$ can be explicitly written as

$$\mathbf{Jz}(\boldsymbol{\theta}_\rho; \mathbf{c}) = \frac{\mathbf{r}(\boldsymbol{\theta}_\rho)}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2} \left(\mathbf{c} - 2\frac{\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \mathbf{c}}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2} \mathbf{r}(\boldsymbol{\theta}_\rho)\right)^\top \mathbf{Jr}(\boldsymbol{\theta}_\rho) + \frac{\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \mathbf{c}}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2} \mathbf{Jr}(\boldsymbol{\theta}_\rho). \tag{24}$$

Plugging in $\boldsymbol{\theta}_\rho^*$ and $\mathbf{c}^*$, and using $\mathbf{c}^* = \theta_v^* \mathbf{r}(\boldsymbol{\theta}_\rho^*)$ and $\theta_v^* = \frac{\mathbf{r}(\boldsymbol{\theta}_\rho^*)^\top \mathbf{c}^*}{\|\mathbf{r}(\boldsymbol{\theta}_\rho^*)\|^2}$, we get

$$\mathbf{Jz}(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*) = \frac{\mathbf{r}(\boldsymbol{\theta}_\rho^*)}{\|\mathbf{r}(\boldsymbol{\theta}_\rho^*)\|^2} \left(\mathbf{c}^* - 2\theta_v^* \mathbf{r}(\boldsymbol{\theta}_\rho^*)\right)^\top \mathbf{Jr}(\boldsymbol{\theta}_\rho^*) + \theta_v^* \mathbf{Jr}(\boldsymbol{\theta}_\rho^*)$$

$$= -\theta_v^* \frac{\mathbf{r}(\boldsymbol{\theta}_\rho^*)}{\|\mathbf{r}(\boldsymbol{\theta}_\rho^*)\|^2} \mathbf{r}(\boldsymbol{\theta}_\rho^*)^\top \mathbf{Jr}(\boldsymbol{\theta}_\rho^*) + \theta_v^* \mathbf{Jr}(\boldsymbol{\theta}_\rho^*)$$

$$= \theta_v^* \left(\mathbf{I} - \left(\frac{\mathbf{r}(\boldsymbol{\theta}_\rho^*)}{\|\mathbf{r}(\boldsymbol{\theta}_\rho^*)\|}\right) \left(\frac{\mathbf{r}(\boldsymbol{\theta}_\rho^*)}{\|\mathbf{r}(\boldsymbol{\theta}_\rho^*)\|}\right)^\top\right) \mathbf{Jr}(\boldsymbol{\theta}_\rho^*).$$

Let the q columns of the Jacobian matrix $\mathbf{Jz}(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*)$ be denoted by $[\mathbf{z}_1', ..., \mathbf{z}_q']$, and the q columns of $\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)$ be denoted by $[\mathbf{r}_1', ..., \mathbf{r}_q']$. Define $\mathbf{r}^* \triangleq \mathbf{r}(\boldsymbol{\theta}_\rho^*)$ and $\tilde{\mathbf{r}} \triangleq \mathbf{r}^*/\|\mathbf{r}^*\|$, then

$$\mathbf{z}_j' = \theta_v^* (\mathbf{I} - \tilde{\mathbf{r}}\tilde{\mathbf{r}}^\top)\mathbf{r}_j', \quad \forall\, j = 1, ..., q. \tag{25}$$

For $\nabla^2 g(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*)$ to be positive definite, the matrix $\mathbf{Jz}(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*)$ should be full rank, i.e., $\{\mathbf{z}_1', ..., \mathbf{z}_q'\}$ should be linearly independent. Note $\{\mathbf{z}_1', ..., \mathbf{z}_q'\}$ are linearly dependent if and only if there exists $\boldsymbol{\beta} \neq \mathbf{0}$ such that $\sum_{j=1}^q \beta_j \mathbf{z}_j' = \mathbf{0}$, which is equivalent to the condition $\sum_{j=1}^q \beta_j \mathbf{r}_j' = \bar{\beta}\, \tilde{\mathbf{r}}$ due to (25), where $\bar{\beta} \triangleq \tilde{\mathbf{r}}^\top \left(\sum_{j=1}^q \beta_j \mathbf{r}_j'\right)$. If $\bar{\beta} = 0$, then the set of vectors

14

$\{\mathbf{r}'_1, \ldots, \mathbf{r}'_q\}$ are linearly dependent; otherwise, $\bar{\boldsymbol{\beta}} \neq 0$ implies that $\tilde{\mathbf{r}} = \sum_{j=1}^{q}(\beta_j/\bar{\boldsymbol{\beta}})\mathbf{r}'_j$; thus, $\{\mathbf{r}, \mathbf{r}'_1, \ldots, \mathbf{r}'_q\}$ are linearly dependent. Therefore, $\{\mathbf{z}'_1, \ldots, \mathbf{z}'_q\}$ are linearly independent if and only if $\{\mathbf{r}, \mathbf{r}'_1, \ldots, \mathbf{r}'_q\}$ are linearly independent. Finally, the function

$$s(\boldsymbol{\theta}_\rho, \mathbf{c}) \triangleq \frac{\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \mathbf{c}}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2}$$

is continuous in $(\boldsymbol{\theta}_\rho, \mathbf{c})$; hence, the preimage $s^{-1}(\mathbb{R}_{++})$ is an open set. Moreover, $s(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*) = \theta_v^* > 0$; hence, $(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*) \in s^{-1}(\mathbb{R}_{++})$. Therefore, there exists $\delta_1 > 0$ such that $B_{\delta_1}(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*)$, the open ball around $(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*)$ with radius δ_1, satisfies $B_{\delta_1}(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*) \subseteq s^{-1}(\mathbb{R}_{++})$, and the objective of the outer problem in (13), i.e., $\psi(\boldsymbol{\theta}_\rho; \mathbf{c})$ defined in (19), is equal to $g(\boldsymbol{\theta}_\rho; \mathbf{c})$ on $B_{\delta_1}(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*)$. Thus, $\nabla^2 \psi(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*)$ exists, and it satisfies $\nabla^2 \psi(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*) = \nabla^2 g(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*) \succ \gamma^* \mathbf{I}$. ∎

Remark 8 *We here comment on the linear independence condition stated in Lemma 7. For instance, consider the* anisotropic exponential correlation *function* $r(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}_\rho) = \exp\big(-(\mathbf{x} - \mathbf{x}')^\top \mathrm{diag}(\boldsymbol{\theta}_\rho)(\mathbf{x} - \mathbf{x}')\big)$, *where* $q = d$, *and* $\Theta_\rho = \mathbb{R}_+^d$. *Let* $\mathcal{X} = [-\beta, \beta]^d$ *for some* $\beta > 0$, *and suppose* $\{\mathbf{x}_i\}_{i \in \mathcal{I}}$ *is a set of independent identically distributed* uniform *random samples inside* $\mathcal{X}$. *Then it can be shown that for this correlation function, the condition in Lemma 7 holds with probability 1, i.e.,* $\{\mathbf{r}(\boldsymbol{\theta}_\rho^*), \mathbf{r}'_1(\boldsymbol{\theta}_\rho^*), \ldots, \mathbf{r}'_q(\boldsymbol{\theta}_\rho^*)\}$ *are linearly independent w.p. 1.*

The next result shows the convergence of the SPS estimator as the number of samples per location, $N \in \mathbb{Z}_+$, increases.

Theorem 9 *Let* $\boldsymbol{\theta}^* = [\boldsymbol{\theta}_\rho^{*\top}, \theta_v^*]^\top \in \mathrm{int}(\Theta_\rho) \times \mathbb{R}_{++}$ *be the true parameters. Suppose* $\mathbf{r} : \mathbb{R}^q \to \mathbb{R}^{n^2}$ *is twice continuously differentiable, and vectors in* $\{\mathbf{r}(\boldsymbol{\theta}_\rho^*), \mathbf{r}'_1(\boldsymbol{\theta}_\rho^*), \ldots, \mathbf{r}'_q(\boldsymbol{\theta}_\rho^*)\}$ *are linearly independent. For any given* $M > 0$ *and* $N \geq N_0 \triangleq \lceil 2(M+2)\log n + \log 16 \rceil$, *let* $\hat{\boldsymbol{\theta}}^{(N)} = [\hat{\boldsymbol{\theta}}_\rho^\top, \hat{\theta}_v]^\top$ *be the SPS estimator of* $\boldsymbol{\theta}^*$, *i.e.,* $\hat{\boldsymbol{\theta}}_\rho \in \mathrm{argmin}_{\boldsymbol{\theta}_\rho \in \Theta_\rho} \psi(\boldsymbol{\theta}_\rho; \hat{\mathbf{c}})$, *and* $\hat{\theta}_v$ *be computed as in Corollary 5. Then for any* $\epsilon > 0$, *there exists* $N \geq N_0$ *satisfying* $N = \mathcal{O}(N_0/\epsilon^2)$ *such that setting* $\alpha = 40\theta_v^* \sqrt{\frac{N_0}{N}}$ *in (11) implies* $\|\hat{\boldsymbol{\theta}}^{(N)} - \boldsymbol{\theta}^*\| \leq \epsilon$ *with probability at least* $1 - n^{-M}$; *moreover, the STAGE-II function* $\psi(\cdot; \hat{\mathbf{c}})$ *is strongly convex around* $\hat{\boldsymbol{\theta}}_\rho$.

Proof From the hypothesis, $\{\mathbf{r}(\boldsymbol{\theta}_\rho^*), \mathbf{r}'_1(\boldsymbol{\theta}_\rho^*), \ldots, \mathbf{r}'_q(\boldsymbol{\theta}_\rho^*)\}$ are linearly independent; hence, Lemma 7 implies there exists $\gamma^* > 0$ such that $\nabla^2 g(\boldsymbol{\theta}_\rho^*; \mathbf{c}^*) \succeq \gamma^* \mathbf{I}$ for g defined in (20)—throughout the proof, all the derivatives of g are written with respect to $\boldsymbol{\theta}_\rho$ only, not $\mathbf{c}$. Recall from the proof of Lemma 7 that the function $s(\boldsymbol{\theta}_\rho, \mathbf{c}) \triangleq \frac{\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \mathbf{c}}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2}$ is continuous in $(\boldsymbol{\theta}_\rho, \mathbf{c})$; therefore, there exists $\delta_1 > 0$ such that $B_{\delta_1}(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*) \subseteq s^{-1}(\mathbb{R}_{++})$. Hence, the objective of the outer problem in (13), i.e., $\psi(\boldsymbol{\theta}_\rho; \mathbf{c})$ defined in (19), is equal to $g(\boldsymbol{\theta}_\rho; \mathbf{c})$ on $B_{\delta_1}(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*)$. Moreover, since $\boldsymbol{\theta}_\rho^* \in \mathrm{int}(\Theta_\rho)$, δ_1 can be chosen to satisfy $B_{\delta_1}(\boldsymbol{\theta}_\rho^*) \subset \mathrm{int}(\Theta_\rho)$.

Since $\mathbf{r}(\boldsymbol{\theta}_\rho)$ is assumed to be twice continuously differentiable in $\boldsymbol{\theta}_\rho$, it follows from (22) that $\nabla^2 g(\boldsymbol{\theta}_\rho; \mathbf{c})$ is continuous in $(\boldsymbol{\theta}_\rho, \mathbf{c})$ on $B_{\delta_1}(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*)$. Moreover, eigenvalues of a matrix are continuous functions of matrix entries; hence, $\lambda_{\min}\big(\nabla^2 g(\boldsymbol{\theta}_\rho; \mathbf{c})\big)$ is continuous in $(\boldsymbol{\theta}_\rho, \mathbf{c})$ on $B_{\delta_1}(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*)$ as well. Thus, ψ is strongly convex around $(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*)$. Indeed, there exists $\delta_2 > 0$ such that $\delta_2 \leq \delta_1$ and $\nabla^2 g(\boldsymbol{\theta}_\rho; \mathbf{c}) \succ \frac{\gamma^*}{2}\mathbf{I}$ for all $(\boldsymbol{\theta}_\rho, \mathbf{c}) \in B_{\delta_2}(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*)$. Define

$$\mathcal{C} \triangleq \{\mathbf{c} : \|\mathbf{c} - \mathbf{c}^*\| \leq \frac{\delta_2}{\sqrt{2}}\}, \qquad \Theta'_\rho \triangleq \{\boldsymbol{\theta}_\rho \in \Theta_\rho : \|\boldsymbol{\theta}_\rho - \boldsymbol{\theta}_\rho^*\| \leq \frac{\delta_2}{\sqrt{2}}\};$$

and for all $\mathbf{c} \in \mathcal{C}$, let

$$\boldsymbol{\theta}_\rho(\mathbf{c}) \triangleq \operatorname{argmin}\{g(\boldsymbol{\theta}_\rho; \mathbf{c}) : \boldsymbol{\theta}_\rho \in \Theta'_\rho\}$$

be the unique minimizer as $g(\cdot; \mathbf{c})$ is strongly convex in $\boldsymbol{\theta}_\rho$ over Θ'_ρ for $\mathbf{c} \in \mathcal{C}$. Furthermore, since Θ'_ρ is a convex compact set and $g(\boldsymbol{\theta}_\rho; \mathbf{c})$ is jointly continuous in $(\boldsymbol{\theta}_\rho, \mathbf{c})$ on $\Theta'_\rho \times \mathcal{C}$, by Berge's Maximum Theorem (Ok, 2007), $\boldsymbol{\theta}_\rho(\mathbf{c})$ is continuous at $\mathbf{c}^*$ and $\boldsymbol{\theta}_\rho(\mathbf{c}^*) = \boldsymbol{\theta}_\rho^*$. Hence, given any $0 < \eta < \frac{\delta_2}{\sqrt{2}}$, there exists $\delta(\eta) > 0$ such that $\delta(\eta) \leq \frac{\delta_2}{\sqrt{2}}$, and $\|\boldsymbol{\theta}_\rho(\mathbf{c}) - \boldsymbol{\theta}_\rho^*\| \leq \eta$ for all $\|\mathbf{c} - \mathbf{c}^*\| \leq \delta(\eta)$. It follows from Theorem 6 that by setting an appropriate $\alpha(\eta)$ in STAGE-I problem (11), it is guaranteed with high probability that $\|\hat{\mathbf{c}} - \mathbf{c}^*\| \leq \delta(\eta)$—we will revisit this claim at the end. Thus, $\|\boldsymbol{\theta}_\rho(\hat{\mathbf{c}}) - \boldsymbol{\theta}_\rho^*\| \leq \eta < \frac{\delta_2}{\sqrt{2}}$, which implies that $\boldsymbol{\theta}_\rho(\hat{\mathbf{c}}) = \operatorname{argmin}\{g(\boldsymbol{\theta}_\rho; \hat{\mathbf{c}}) : \boldsymbol{\theta}_\rho \in \Theta_\rho\}$, i.e., it is equal to the solution to the outer problem in (13) of STAGE-II: $\hat{\boldsymbol{\theta}}_\rho = \boldsymbol{\theta}_\rho(\hat{\mathbf{c}})$. Hence, it follows that $\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}_\rho^*\|^2 + \|\hat{\mathbf{c}} - \mathbf{c}^*\|^2 < \delta_2^2 \leq \delta_1^2$, which implies that $(\hat{\boldsymbol{\theta}}_\rho, \hat{\mathbf{c}}) \in B_{\delta_1}(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*) \subseteq s^{-1}(\mathbb{R}_{++})$; and since $\hat{\theta}_v = \max\{0, s(\hat{\boldsymbol{\theta}}_\rho, \hat{\mathbf{c}})\}$, we also have $\hat{\theta}_v = s(\hat{\boldsymbol{\theta}}_\rho, \hat{\mathbf{c}}) > 0$. Moreover, $\hat{\boldsymbol{\theta}}_\rho \in B_{\delta_1}(\boldsymbol{\theta}_\rho^*)$ implies $\hat{\boldsymbol{\theta}}_\rho \in \operatorname{int}(\Theta_\rho)$. Note that $\hat{\boldsymbol{\theta}}_\rho \in \Theta'_\rho$ and $\hat{\mathbf{c}} \in \mathcal{C}$; hence, $g(\cdot; \hat{\mathbf{c}})$ is strongly convex at $\hat{\boldsymbol{\theta}}_\rho$ with modulus $\frac{\gamma^*}{2}$.

Next we establish a relation between $\delta(\eta)$ and η by showing $\hat{\boldsymbol{\theta}}_\rho(\mathbf{c})$ is Lipschitz around $\mathbf{c}^*$. Since, for $\mathbf{c} \in \mathcal{C}$, $g(\cdot; \mathbf{c})$ is strongly convex in $\boldsymbol{\theta}_\rho$ over Θ'_ρ with convexity modulus $\gamma^*/2$, for any $\mathbf{c} \in \mathcal{C}$, and $\boldsymbol{\theta}_\rho^i \in \Theta'_\rho$ for $i = 1, 2$, we have

$$\langle \boldsymbol{\theta}_\rho^2 - \boldsymbol{\theta}_\rho^1, \ \nabla g(\boldsymbol{\theta}_\rho^2; \mathbf{c}) - \nabla g(\boldsymbol{\theta}_\rho^1; \mathbf{c}) \rangle \geq \frac{\gamma^*}{2} \|\boldsymbol{\theta}_\rho^2 - \boldsymbol{\theta}_\rho^1\|^2.$$

Suppose $\mathbf{c}^1, \mathbf{c}^2 \in \mathcal{C}$. Since $\boldsymbol{\theta}_\rho(\mathbf{c}^i) = \operatorname{argmin}\{g(\boldsymbol{\theta}_\rho; \mathbf{c}^i) : \boldsymbol{\theta}_\rho \in \Theta'_\rho\}$ for $i = 1, 2$, it follows from the first-order optimality conditions that

$$\langle \boldsymbol{\theta}_\rho - \boldsymbol{\theta}_\rho(\mathbf{c}^i), \ \nabla g(\boldsymbol{\theta}_\rho(\mathbf{c}^i); \mathbf{c}^i) \rangle \geq 0, \quad \forall \, \boldsymbol{\theta}_\rho \in \Theta'_\rho, \quad i \in \{1, 2\}. \tag{26}$$

Strong convexity and (26) imply $\langle \boldsymbol{\theta}_\rho(\mathbf{c}^2) - \boldsymbol{\theta}_\rho(\mathbf{c}^1), \ \nabla g(\boldsymbol{\theta}_\rho(\mathbf{c}^2); \mathbf{c}^1) \rangle \geq \frac{\gamma^*}{2} \|\boldsymbol{\theta}_\rho(\mathbf{c}^2) - \boldsymbol{\theta}_\rho(\mathbf{c}^1)\|^2$. Adding and subtracting $\nabla g(\boldsymbol{\theta}_\rho(\mathbf{c}^2); \mathbf{c}^2)$, and using (26) again, we get $\frac{\gamma^*}{2} \|\boldsymbol{\theta}_\rho(\mathbf{c}^2) - \boldsymbol{\theta}_\rho(\mathbf{c}^1)\|^2 \leq \langle \boldsymbol{\theta}_\rho(\mathbf{c}^2) - \boldsymbol{\theta}_\rho(\mathbf{c}^1), \ \nabla g(\boldsymbol{\theta}_\rho(\mathbf{c}^2); \mathbf{c}^1) - \nabla g(\boldsymbol{\theta}_\rho(\mathbf{c}^2); \mathbf{c}^2) \rangle$. Thus, from Cauchy-Schwarz,

$$\|\nabla g(\boldsymbol{\theta}_\rho(\mathbf{c}^2); \mathbf{c}^2) - \nabla g(\boldsymbol{\theta}_\rho(\mathbf{c}^2); \mathbf{c}^1)\| \geq \frac{\gamma^*}{2} \|\boldsymbol{\theta}_\rho(\mathbf{c}^2) - \boldsymbol{\theta}_\rho(\mathbf{c}^1)\|. \tag{27}$$

Moreover, (21) and (24) imply that

$$\nabla g(\boldsymbol{\theta}_\rho; \mathbf{c}) = \frac{1}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2} \mathbf{J}\mathbf{r}(\boldsymbol{\theta}_\rho)^\top v(\boldsymbol{\theta}_\rho; \mathbf{c}), \tag{28}$$

$$v(\boldsymbol{\theta}_\rho; \mathbf{c}) \triangleq \mathbf{r}(\boldsymbol{\theta}_\rho)^\top (\mathbf{c} - \mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c})) \mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c}) - \frac{(\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \mathbf{c})(\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c}))}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2} \mathbf{r}(\boldsymbol{\theta}_\rho),$$

$$= \left(\frac{1}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2} \mathbf{r}(\boldsymbol{\theta}_\rho) \mathbf{r}(\boldsymbol{\theta}_\rho)^\top - \mathbf{I} \right) \mathbf{c}\mathbf{c}^\top \mathbf{r}(\boldsymbol{\theta}_\rho),$$

where $\mathbf{z} : \mathbb{R}^q \times \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ is defined as in the proof of Lemma 7, i.e., $\mathbf{z}(\boldsymbol{\theta}_\rho; \mathbf{c}) \triangleq \left(\frac{\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \mathbf{c}}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2} \right) \mathbf{r}(\boldsymbol{\theta}_\rho) - \mathbf{c}$. Observe $\|\mathbf{I} - \frac{1}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2} \mathbf{r}(\boldsymbol{\theta}_\rho) \mathbf{r}(\boldsymbol{\theta}_\rho)^\top\|_2 = 1$ and $\|\mathbf{r}(\boldsymbol{\theta}_\rho)\| \geq \sqrt{n}$; therefore, given $\mathbf{c}^1, \mathbf{c}^2 \in C$, we have

$$\|\nabla g(\boldsymbol{\theta}_\rho; \mathbf{c}^2) - \nabla g(\boldsymbol{\theta}_\rho; \mathbf{c}^1)\| = \frac{1}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2} \|\mathbf{J}\mathbf{r}(\boldsymbol{\theta}_\rho)^\top \left(\mathbf{I} - \frac{1}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2} \mathbf{r}(\boldsymbol{\theta}_\rho) \mathbf{r}(\boldsymbol{\theta}_\rho)^\top \right) (\mathbf{c}^1 \mathbf{c}^{1\top} - \mathbf{c}^2 \mathbf{c}^{2\top}) \mathbf{r}(\boldsymbol{\theta}_\rho)\|,$$

$$\leq \frac{1}{\sqrt{n}} \|\mathbf{J}\mathbf{r}(\boldsymbol{\theta}_\rho)\|_2 (\|\mathbf{c}^1\| + \|\mathbf{c}^2\|) \|\mathbf{c}^2 - \mathbf{c}^1\|, \tag{29}$$

16

where (29) follows from $\|\mathbf{c}^1 \mathbf{c}^{1\top} - \mathbf{c}^2 \mathbf{c}^{2\top}\|_2 \leq (\|\mathbf{c}^1\| + \|\mathbf{c}^2\|)\|\mathbf{c}^2 - \mathbf{c}^1\|$. Since $\boldsymbol{\theta}_\rho^* = \boldsymbol{\theta}_\rho(\mathbf{c}^*)$ and $\hat{\boldsymbol{\theta}}_\rho = \boldsymbol{\theta}_\rho(\hat{\mathbf{c}})$, by setting $\mathbf{c}^1 = \hat{\mathbf{c}}$ and $\mathbf{c}^2 = \mathbf{c}^*$ within (27) and (29), we get

$$\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}_\rho^*\| \leq \frac{2}{\gamma^* \sqrt{n}} \|\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)\|_2 \left(2\|\mathbf{c}^*\| + \frac{\delta_2}{\sqrt{2}}\right) \|\hat{\mathbf{c}} - \mathbf{c}^*\|, \tag{30}$$

where we have used the fact that $\|\hat{\mathbf{c}} - \mathbf{c}^*\| \leq \frac{\delta_2}{\sqrt{2}}$. Thus, given any $0 < \eta < \frac{\delta_2}{\sqrt{2}}$, it follows that $\|\hat{\mathbf{c}} - \mathbf{c}^*\| \leq \delta(\eta)$ implies $\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}_\rho^*\| \leq \eta$, for $\delta(\eta)$ chosen as

$$\delta(\eta) \triangleq \min\left\{\frac{\gamma^* \sqrt{n}}{2\|\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)\|_2}(2\|\mathbf{c}^*\| + \frac{\delta_2}{\sqrt{2}})^{-1}\eta, \ \frac{\delta_2}{\sqrt{2}}\right\}. \tag{31}$$

Now, we show that $|\hat{\theta}_v - \theta_v^*|$ can be made arbitrarily small. Let $\mathbf{t} : \mathbb{R}^q \to \mathbb{R}^{n^2}$ such that $\mathbf{t}(\boldsymbol{\theta}_\rho) = \mathbf{r}(\boldsymbol{\theta}_\rho)/\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2$; hence, $\mathbf{Jt}(\boldsymbol{\theta}_\rho) = \left(\mathbf{I} - \frac{2}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2}\mathbf{r}(\boldsymbol{\theta}_\rho)\mathbf{r}(\boldsymbol{\theta}_\rho)^\top\right) \frac{\mathbf{Jr}(\boldsymbol{\theta}_\rho)}{\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2}$. Since $\mathbf{r}$ is twice continuously differentiable, there exists $0 < U \in \mathbb{R}$ such that

$$U \triangleq \max\{\|\mathbf{Jr}(\boldsymbol{\theta}_\rho)\|_2 : \ \|\boldsymbol{\theta}_\rho - \boldsymbol{\theta}_\rho^*\| \leq \delta_2/\sqrt{2}\}.$$

Therefore, $\|\mathbf{Jt}(\boldsymbol{\theta}_\rho)\|_2 \leq U/\|\mathbf{r}(\boldsymbol{\theta}_\rho)\|^2 \leq U/n$ for any $\boldsymbol{\theta}_\rho$ belonging to the line segment connecting $\boldsymbol{\theta}_\rho^*$ and $\hat{\boldsymbol{\theta}}_\rho$. Furthermore, $\mathbf{t}(\hat{\boldsymbol{\theta}}_\rho) = \mathbf{t}(\boldsymbol{\theta}_\rho^*) + \left(\int_0^1 \mathbf{Jt}(\boldsymbol{\theta}_\rho^* + \ell \ (\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}_\rho^*)) \ d\ell\right)(\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}_\rho^*)$; hence, $\|\mathbf{t}(\hat{\boldsymbol{\theta}}_\rho) - \mathbf{t}(\boldsymbol{\theta}_\rho^*)\| \leq \frac{U}{n}\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}_\rho^*\| \leq \frac{U}{n}\eta$ as $\eta > 0$ is chosen as $\eta < \delta_2/\sqrt{2}$. Since $\hat{\theta}_v = s(\hat{\boldsymbol{\theta}}_\rho, \hat{\mathbf{c}}) > 0$ and $\theta_v^* = s(\boldsymbol{\theta}_\rho^*, \mathbf{c}^*) > 0$, it follows

$$
\begin{aligned}
|\hat{\theta}_v - \theta_v^*| &= \left| \frac{\mathbf{r}(\hat{\boldsymbol{\theta}}_\rho)^\top \hat{\mathbf{c}}}{\|\mathbf{r}(\hat{\boldsymbol{\theta}}_\rho)\|^2} - \frac{\mathbf{r}(\boldsymbol{\theta}_\rho^*)^\top \mathbf{c}^*}{\|\mathbf{r}(\boldsymbol{\theta}_\rho^*)\|^2} \right| \\
&= \left| \left\langle \frac{\mathbf{r}(\hat{\boldsymbol{\theta}}_\rho)}{\|\mathbf{r}(\hat{\boldsymbol{\theta}}_\rho)\|^2} - \frac{\mathbf{r}(\boldsymbol{\theta}_\rho^*)}{\|\mathbf{r}(\boldsymbol{\theta}_\rho^*)\|^2}, \mathbf{c}^* \right\rangle + \left\langle \frac{\mathbf{r}(\hat{\boldsymbol{\theta}}_\rho)}{\|\mathbf{r}(\hat{\boldsymbol{\theta}}_\rho)\|^2}, \hat{\mathbf{c}} - \mathbf{c}^* \right\rangle \right| \\
&\leq \|\mathbf{t}(\hat{\boldsymbol{\theta}}_\rho) - \mathbf{t}(\boldsymbol{\theta}_\rho^*)\| \|\mathbf{c}^*\| + \frac{1}{\|\mathbf{r}(\hat{\boldsymbol{\theta}}_\rho)\|}\|\mathbf{c}^* - \hat{\mathbf{c}}\| \\
&\leq \frac{U}{n}\|\mathbf{c}^*\|\eta + \frac{1}{\sqrt{n}}\delta(\eta).
\end{aligned}
$$

Therefore, using the identity $(a + b)^2 \leq 2(a^2 + b^2)$ for any $a, b \in \mathbb{R}$, we get

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|^2 = \|\boldsymbol{\theta}_\rho^* - \hat{\boldsymbol{\theta}}_\rho\|^2 + |\theta_v^* - \hat{\theta}_v|^2 \leq \left(\frac{2U^2}{n^2}\|\mathbf{c}^*\|^2 + 1\right)\eta^2 + \frac{2}{n}(\delta(\eta))^2.$$

From (31) and $\|\mathbf{c}^*\| = \|C^*\|_F \leq \mathbf{Tr}(C^*) \leq n\theta_v^*$, let $\kappa \triangleq (\sqrt{2}U\theta_v^* + 1)(2n\theta_v^* + \frac{\delta_2}{\sqrt{2}}) + \frac{\gamma^*}{\sqrt{2}\|\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)\|_2}$; hence, choosing $\eta_\epsilon \triangleq \min\left\{\frac{1}{\kappa}(2\|\mathbf{c}^*\| + \frac{\delta_2}{\sqrt{2}}) \ \epsilon, \ \frac{\delta_2}{\sqrt{2}}\right\}$ implies that $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\| \leq \epsilon$. Thus, for all sufficiently small $\epsilon > 0$, having $\|\hat{\mathbf{c}} - \mathbf{c}^*\| \leq \delta_\epsilon$ implies that $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\| \leq \epsilon$, where $\delta_\epsilon \triangleq \delta(\eta_\epsilon) = \frac{\gamma^* \sqrt{n}}{2\|\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)\|_2} \frac{\epsilon}{\kappa}$. Next, according to Theorem 6, given $\epsilon > 0$ and $M > 0$, for all N such that $\sqrt{N/N_0} \geq 160\frac{\kappa}{\gamma^*}\|\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)\|_2 \left(\frac{b^*}{a^*}\right)^2 (n + \|G\|_F)\theta_v^*\frac{1}{\epsilon}$, choosing $\alpha = 40\theta_v^* \sqrt{\frac{N_0}{N}}$ in (11) guarantees

$$\|\hat{\mathbf{c}} - \mathbf{c}^*\| = \|\hat{C} - C^*\|_F \leq \sqrt{n}\|\hat{C} - C^*\|_2 \leq \frac{\sqrt{n}}{a^{*2}}\|\hat{P} - P^*\|_2 \leq \frac{\sqrt{n}}{a^{*2}}\|\hat{P} - P^*\|_F \leq \delta_\epsilon, \tag{32}$$

with probability at least $1 - n^{-M}$, where the equality follows from the definitions of $\mathbf{c}^*$ and $\hat{\mathbf{c}}$, the second inequality follows from $\hat{C} = (\hat{P})^{-1}$, $C^* = (P^*)^{-1}$, and the fact that $P \mapsto P^{-1}$ is Lipschitz continuous on $P \succeq a^*\mathbf{I}$. This completes the proof. ∎

4.2.1 TIGHTER BOUNDS THROUGH EXPLOITING THE DECAY PROPERTY

In the GRF setting considered in this paper $\text{card}(P^*) = n^2$; hence, the bound obtained in Theorem 6 may seem loose. On the other hand, Theorem 3 shows that the elements of P^* exhibit a fast decay similar to the elements of the covariance matrix. Therefore, although $\text{card}(P^*) = n^2$, significant portion of these elements are close to 0. In this section, we argue that by exploiting this property, the $\mathcal{O}(1)$ constant in Theorem 6 can be significantly improved. More precisely, the main reason we get $\sqrt{n^2 \log n / N}$ bound on $\|\hat{P} - P^*\|_F$ is that the analysis provided for Theorem 6 is similar to Rothman et al. (2008) (Theorem 1) and it does not exploit the fast decay in the elements of P^*, leading us to $\sqrt{\frac{\text{card}(P^*) \log n}{N}}$ and $\text{card}(P^*) = n^2$. Instead in Corollary 10, we were able to replace $\text{card}(P^*)$ with the cardinality of some set $S_{\bar{\epsilon}} \subset \{(i,j) \in \mathcal{I} \times \mathcal{I} : |P^*_{ij}| \geq \bar{\epsilon}\}$, defined in (34) for some $\bar{\epsilon} > 0$. Clearly, $|S_{\bar{\epsilon}}| \leq \text{card}(P^*)$ for all $\bar{\epsilon} \geq 0$.

The fast decay of the elements in inverse of the covariance matrix (Theorem 3) is the motivation behind sparse approximation of the inverse covariance matrix in the STAGE-I of SPS as Jaffard's decay algebra clearly applies to the inverse covariance matrix. The proposed sparse approximation of P^* leads to a $|S_{\bar{\epsilon}}|$ term in the bound on $\|\hat{P} - P^*\|_F$. More precisely, intuitively, given a reasonable threshold value $\bar{\epsilon} > 0$, due to fast decay seen in the elements of P^*, one expects $|S_{\bar{\epsilon}}| \ll n^2$. Unfortunately, even for very simple deterministic designs such as a d-dimensional lattice, computing $|S_{\bar{\epsilon}}|$ in closed-form is a hard combinatorial problem—indeed, we spend quite some time computing it in the closed form for lattice designs, only to get some partial results. Therefore, we numerically investigated how $|S_{\bar{\epsilon}}|$ compares to n for different values of $\bar{\epsilon} > 0$ and it turns out that $|S_{\bar{\epsilon}}|$ behaves similar to $\mathcal{O}(n)$, rather than $\mathcal{O}(n^2)$—In Remark 11, we plot $|S_{\bar{\epsilon}}|/n$ and compare it to n^2/n for (squared) exponential and Matern covariance functions for different values of $\bar{\epsilon}$ ranging between 10^{-1} and 10^{-15}.

In the rest of this section, we assume that the true covariance matrix C^* belongs to the class $\mathcal{E}_\gamma$ for some $\gamma > 0$; hence, it follows from Theorem 3 that P^* belongs to the class $\mathcal{E}_{\gamma'}$ for some $\gamma' > 0$. Thus, for all $(i,j) \in \mathcal{I} \times \mathcal{I}$, $|P^*_{ij}| \leq K_{\gamma'} \exp(-\gamma' d(\mathbf{x}_i, \mathbf{x}_j))$, where $d(\mathbf{x}_i, \mathbf{x}_j)$ is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Given $\bar{\epsilon} > 0$, we define

$$q(\bar{\epsilon}) \triangleq \tfrac{1}{\gamma'} \log(K_{\gamma'}/\bar{\epsilon}). \tag{33}$$

If $d(\mathbf{x}_i, \mathbf{x}_j) \geq q(\bar{\epsilon})$, then we have $|P^*_{ij}| \leq K_{\gamma'} \exp(-\gamma' d(\mathbf{x}_i, \mathbf{x}_j)) \leq \bar{\epsilon}$. Next, we define

$$S_{\bar{\epsilon}} \triangleq \{(i,j) \in \mathcal{I} \times \mathcal{I} : d(\mathbf{x}_i, \mathbf{x}_j) \leq q(\bar{\epsilon})\} \subseteq \{(i,j) \in \mathcal{I} \times \mathcal{I} : |P^*_{ij}| \geq \bar{\epsilon}\} \tag{34}$$

and let $S_{\bar{\epsilon}}^c$ be its complement, i.e., $S_{\bar{\epsilon}}^c \triangleq \mathcal{I} \times \mathcal{I} \backslash S_{\bar{\epsilon}}$.

Next, in Corollary 10, we establish $\mathcal{O}(\log(1/\bar{\epsilon})\sqrt{|S_{\bar{\epsilon}}| \log n / N})$ bound on $\|\hat{P} - P^*\|_F$ and show that $N = \mathcal{O}(1/\epsilon^2)$ rate result of Theorem 9 still holds for all $\epsilon > 0$.

Corollary 10 *Let $\{\mathbf{y}^{(r)}\}_{r=1}^N \subset \mathbb{R}^n$ be independent realizations of a GRF with zero-mean and stationary covariance function $c(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}^*)$ observed over n distinct locations $\{\mathbf{x}_i\}_{i\in\mathcal{I}}$ with*

$\mathcal{I} \triangleq \{1, ..., n\}$. Suppose there exists some $K_\gamma, \gamma > 0$ such that $c(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}^*) \leq K_\gamma \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|)$ for $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Then, under the premise of Theorem 9, for any given $\epsilon > 0$ and $M > 0$, there exists $\bar{\epsilon} > 0$ and $N_\epsilon = \mathcal{O}(N_0/\epsilon^2)$ such that with probability $1 - n^{-M}$, we have

$$\|\hat{P} - P^*\|_F \leq \mathcal{O}\left(\log(1/\bar{\epsilon}) \sqrt{|S_{\bar{\epsilon}}| \log n / N_\epsilon} \right) = \mathcal{O}(\epsilon), \tag{35}$$

where $\hat{P}$ is the SPS estimator of P^* computed as in (11) with $\alpha = 40\theta_v^* \sqrt{N_0/N_\epsilon}$. Moreover, (35) implies that the STAGE-II solution satisfies $\|\hat{\boldsymbol{\theta}}^{(N_\epsilon)} - \boldsymbol{\theta}^*\| \leq \epsilon$.

Proof Given any $\Delta \in \mathbb{R}^{n \times n}$, we define $\Delta_{S_{\bar{\epsilon}}}$ and $\Delta_{S_{\bar{\epsilon}}^c}$ as follows: $(\Delta_{S_{\bar{\epsilon}}})_{ij}$ is equal to Δ_{ij} if $(i, j) \in S_{\bar{\epsilon}}$ and zero otherwise ($\Delta_{S_{\bar{\epsilon}}^c}$ is defined *similarly*). Furthermore, $\|\Delta\|_1 \triangleq \sum_{i,j \in \mathcal{I}} |\Delta_{ij}|$ and $\|\Delta\|_\infty \triangleq \max_{i,j \in \mathcal{I}} |\Delta_{ij}|$.

For any given $\bar{\epsilon} > 0$, define $T \triangleq \{\Delta \in \mathcal{F} : \|\Delta\|_F > \alpha b^{*2}(1 + q(\bar{\epsilon})/\tilde{G}_{\min})|S_{\bar{\epsilon}}|^{1/2} + \left(\left(\alpha b^{*2}(1 + q(\bar{\epsilon})/\tilde{G}_{\min})|S_{\bar{\epsilon}}|^{1/2} \right)^2 + 4\alpha b^{*2}\|G\|_\infty \sum_{(i,j) \in S_{\bar{\epsilon}}^c} |P_{ij}^*| \right)^{1/2} \}$ where $\mathcal{F}$ is defined as in the proof of Theorem 6. From (16), for any $\bar{\epsilon} > 0$ and $\Delta \in T$, we have

$$\begin{aligned} H(\Delta) &\geq \frac{1}{2b^{*2}} \|\Delta\|_F^2 + \langle \Delta, S - C^* \rangle + \alpha \langle G, |\Delta + P^*| - |P^*| \rangle, \\ &\geq \frac{1}{2b^{*2}} \|\Delta\|_F^2 + \langle \Delta, S - C^* \rangle - \alpha \sum_{(i,j) \in S_{\bar{\epsilon}}} G_{ij}|\Delta_{ij}| + \alpha \sum_{(i,j) \in S_{\bar{\epsilon}}^c} G_{ij}(|\Delta_{ij}| - 2|P_{ij}^*|), \end{aligned} \tag{36}$$

$$\geq \frac{1}{2b^{*2}} \|\Delta\|_F^2 - \alpha\|\Delta_{S_{\bar{\epsilon}}}\|_1 - \alpha\|\Delta_{S_{\bar{\epsilon}}^c}\|_1 - \alpha \frac{q(\bar{\epsilon})}{\tilde{G}_{\min}}\|\Delta_{S_{\bar{\epsilon}}}\|_1 + \alpha \frac{\tilde{G}_{\min}}{\tilde{G}_{\min}}\|\Delta_{S_{\bar{\epsilon}}^c}\|_1 - 2\alpha \sum_{(i,j) \in S_{\bar{\epsilon}}^c} G_{ij}|P_{ij}^*|, \tag{37}$$

$$= \frac{1}{2b^{*2}} \|\Delta\|_F^2 - \alpha(1 + \frac{q(\bar{\epsilon})}{\tilde{G}_{\min}})\|\Delta_{S_{\bar{\epsilon}}}\|_1 - 2\alpha \sum_{(i,j) \in S_{\bar{\epsilon}}^c} G_{ij}|P_{ij}^*|, \tag{38}$$

$$\geq \frac{1}{2b^{*2}} \|\Delta\|_F^2 - \alpha(1 + \frac{q(\bar{\epsilon})}{\tilde{G}_{\min}})|S_{\bar{\epsilon}}|^{1/2}\|\Delta\|_F - 2\alpha\|G\|_\infty \sum_{(i,j) \in S_{\bar{\epsilon}}^c} |P_{ij}^*| > 0, \tag{39}$$

where (36) follows from the triangle inequality, (37) follows under the probability event Ω as defined in the proof of Theorem 6 and the definition of $q(\bar{\epsilon})$ in (34) and $G_{ij} = \tilde{G}_{ij}/\tilde{G}_{\min} \geq 1$, as defined in (10), and finally (39) holds for any $\Delta \in S_\Delta$. Hence, following a similar argument within the proof of Theorem 6, under the probability event Ω, we have

$$\|\hat{\Delta}\|_F \leq \alpha b^{*2}(1 + q(\bar{\epsilon})/\tilde{G}_{\min})|S_{\bar{\epsilon}}|^{1/2} + \left(\left(\alpha b^{*2}(1 + q(\bar{\epsilon})/\tilde{G}_{\min})|S_{\bar{\epsilon}}|^{1/2} \right)^2 + 4\alpha b^{*2}\|G\|_\infty \sum_{(i,j) \in S_{\bar{\epsilon}}^c} |P_{ij}^*| \right)^{1/2}. \tag{40}$$

Now, suppose the inequality

$$4b^{*-2}\|G\|_\infty (1 + q(\bar{\epsilon})/\tilde{G}_{\min})^{-2} \frac{1}{|S_{\bar{\epsilon}}|} \sum_{(i,j) \in S_{\bar{\epsilon}}^c} |P_{ij}^*| \leq \alpha \tag{41}$$

holds, which is equivalent to

$$\left(4\alpha b^{*2}\|G\|_\infty \sum_{(i,j) \in S_{\bar{\epsilon}}^c} |P_{ij}^*| \right)^{1/2} \leq \alpha b^{*2}(1 + q(\bar{\epsilon})/\tilde{G}_{\min})|S_{\bar{\epsilon}}|^{1/2}.$$

Therefore, provided that (41) holds, we have

$$\|\hat{P} - P^*\|_F \leq (1 + \sqrt{2})b^{*2}(1 + q(\bar{\epsilon})/\tilde{G}_{\min})|S_{\bar{\epsilon}}|^{1/2}\alpha. \tag{42}$$

19

It follows from the proof of Theorem 6 that setting $\alpha = 40\theta_v^*\sqrt{N_0/N}$ implies

$$\|\hat{P} - P^*\|_F \leq \mathcal{O}\Big(\log(1/\bar{\epsilon})|S_{\bar{\epsilon}}|^{1/2}\sqrt{N_0/N} \Big) \tag{43}$$

with high probability, i.e., w.p. $1 - n^{-M}$. Hence, using the definition of N_0 and by taking $N \geq N_0$, then with probability $1 - n^{-M}$, we get the bound in (35). Provided the sufficient condition (41), from (32) in the proof of Theorem 9, we get $\|\hat{\boldsymbol{\theta}}^{(N)} - \boldsymbol{\theta}^*\| \leq \epsilon$ whenever

$$\frac{\sqrt{n}}{a^{*2}}\|\hat{P} - P^*\|_F \leq \delta_\epsilon \triangleq \frac{\gamma^*\sqrt{n}\epsilon}{2\kappa\|\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)\|_2},$$

which is equivalent to

$$\|\hat{P} - P^*\|_F \leq \frac{\gamma^* a^{*2}\epsilon}{2\kappa\|\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)\|_2}. \tag{44}$$

Therefore, using the bound for $\|\hat{P} - P^*\|_F$ provided in (42) for $\alpha = 40\theta_v^*\sqrt{N_0/N}$, the condition in (44) holds for all large N such that

$$(1 + \sqrt{2})b^{*2}(1 + q(\bar{\epsilon})/\tilde{G}_{\min})|S_{\bar{\epsilon}}|^{1/2}40\theta_\nu^*\sqrt{\frac{N_0}{N}} \leq \frac{\gamma^* a^{*2}}{2\kappa\|\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)\|_2}\epsilon. \tag{45}$$

Define N_ϵ such that (45) holds with equality. Next, we discuss how $\bar{\epsilon} > 0$ should be set so that the sufficient condition in (41) holds for $\alpha_\epsilon \triangleq 40\theta_v^*\sqrt{N_0/N_\epsilon}$, i.e.,

$$4b^{*-2}\|G\|_\infty(1 + q(\bar{\epsilon})/\tilde{G}_{\min})^{-2}\frac{1}{|S_{\bar{\epsilon}}|}\sum_{(i,j)\in S_{\bar{\epsilon}}^c}|P_{ij}^*| \leq \alpha_\epsilon = \frac{(a^*/b^*)^2}{2(1+\sqrt{2})}\frac{\gamma^*}{\kappa\|\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)\|_2}\frac{(1 + q(\bar{\epsilon})/\tilde{G}_{\min})^{-1}}{\sqrt{|S_{\bar{\epsilon}}|}}\epsilon, \tag{46}$$

where the equality follows from the definitions of N_ϵ and α_ϵ. Since $\kappa = \mathcal{O}(n)$ and $\|\mathbf{Jr}(\boldsymbol{\theta}_\rho^*)\|_2 = \mathcal{O}(n)$, we define

$$g_n(\bar{\epsilon}) \triangleq \frac{n^2\sum_{(i,j)\in S_{\bar{\epsilon}}^c}|P_{ij}^*|}{|S_{\bar{\epsilon}}|^{1/2}(1 + q(\bar{\epsilon})/\tilde{G}_{\min})}, \tag{47}$$

which is clearly an increasing function of $\bar{\epsilon}$. Hence, for any $\epsilon > 0$, there exists an $\bar{\epsilon} > 0$ small enough such that

$$\frac{g_n(\bar{\epsilon})}{\epsilon} \leq \frac{a^{*2}\gamma^*}{8(1 + \sqrt{2})\|G\|_\infty}. \tag{48}$$

Therefore, it follows from (46) that for $\bar{\epsilon}$ satisfying (48), the sufficient condition in (41) is true for $\alpha_\epsilon = 40\theta_\nu^*\sqrt{\frac{N_0}{N_\epsilon}}$, where $N_\epsilon = \mathcal{O}(1/\epsilon^2)$. This completes the proof. ∎

Remark 11 *Note that $g_n(\bar{\epsilon})$ in (48) is an increasing function of $\bar{\epsilon}$ since $|S_{\bar{\epsilon}}|$ is nonincreasing (hence $|S_{\bar{\epsilon}}^c|$ is nondecreasing), and $q(\bar{\epsilon})$ is a decreasing function of $\bar{\epsilon}$—see also the left plots in Figures 5, 6, and 7. Hence, for any $\epsilon > 0$ (the STAGE-II bound), there exists $\bar{\bar{\epsilon}} > 0$ such that the condition (48) holds for all $\bar{\epsilon} \in (0, \bar{\bar{\epsilon}}]$. The plots on the left in Figures 5, 6, and 7 show decay of $g_n(\bar{\epsilon})$ with decreasing $\bar{\epsilon}$ for exponential, Matern ($\nu = 3/2$), and squared exponential covariance functions over a two-dimensional uniform design, respectively. The*

corresponding plots on the right illustrate the growth of $|S_{\bar{\epsilon}}|$ in n for different $\bar{\epsilon}$ values. More precisely, to visually compare the growth of $|S_{\bar{\epsilon}}|$ and n^2, we plotted $|S_{\bar{\epsilon}}|/n$ and n in log-scale against n in the x-axis. Note $|S_{\bar{\epsilon}}|/n$ behaves as a constant that depends on $\bar{\epsilon}$; hence, $|S_{\bar{\epsilon}}|$ scales as $\mathcal{O}(n)$ compared to $\mathrm{card}(P^*) = \mathcal{O}(n^2)$ appearing in the bound provided in Theorem 6—see also Section 5.1.
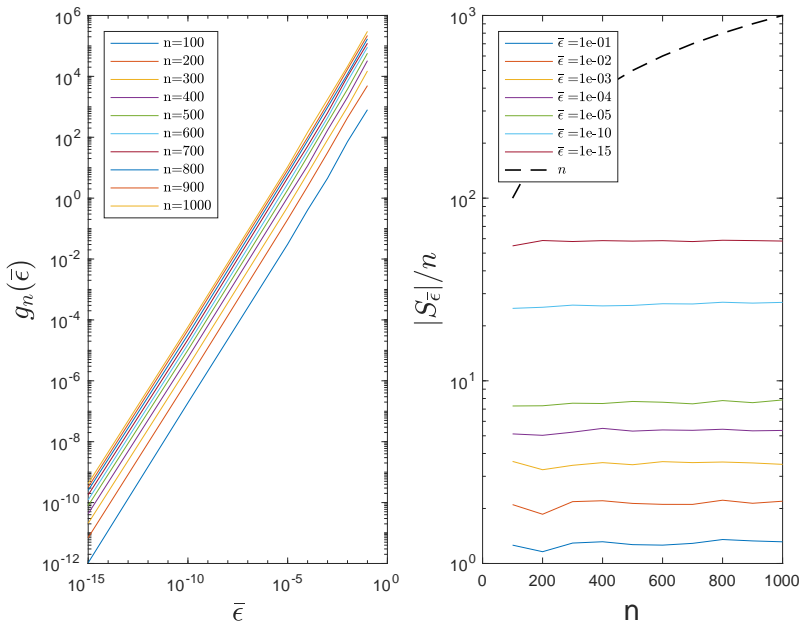
Figure 5: Exponential covariance function

5. Numerical Performance of the Proposed SPS Method

In this section, the performance of the proposed algorithms is reported for both simulated and real data sets.

To solve the STAGE-I problem, the sparsity parameter α in (11) was set to $c\sqrt{\log(n)/N}$ (except for Section 5.1) with $c = 10^{-3}$. Furthermore, to solve STAGE-I problem, we used a particular ADMM implementation displayed in Figure 11 of the online supplement, for which the penalty sequence $\{\rho_\ell\}$ is set to a geometrically increasing sequence $\rho_{\ell+1} = 1.05\rho_\ell$ with $\rho_0 = n$. In the simulation studies R denotes the number of simulation replications. The numerical tests were carried on computer with an Intel Xeon x5650 CPU and 12.0 GB memory. Implementation of the proposed SPS algorithm in MATLAB is available at `https://github.com/samdavanloo/SPS`. The package also includes the extension of the SPS algorithm to fit *multivariate* GRF models – see Tajbakhsh et al. (2018).

5.1 Actual Error vs. the Theoretical Bound Based on Corollary 10

We simulate an isotropic GRF with exponential covariance function with $\boldsymbol{\theta}_\rho^* = 10$, $\theta_\nu^* = 1$, and $\theta_0^* = 0$ in a 2-dimensional square $\mathcal{X} = [0, 100] \times [0, 100]$ over a uniform design with $n =$
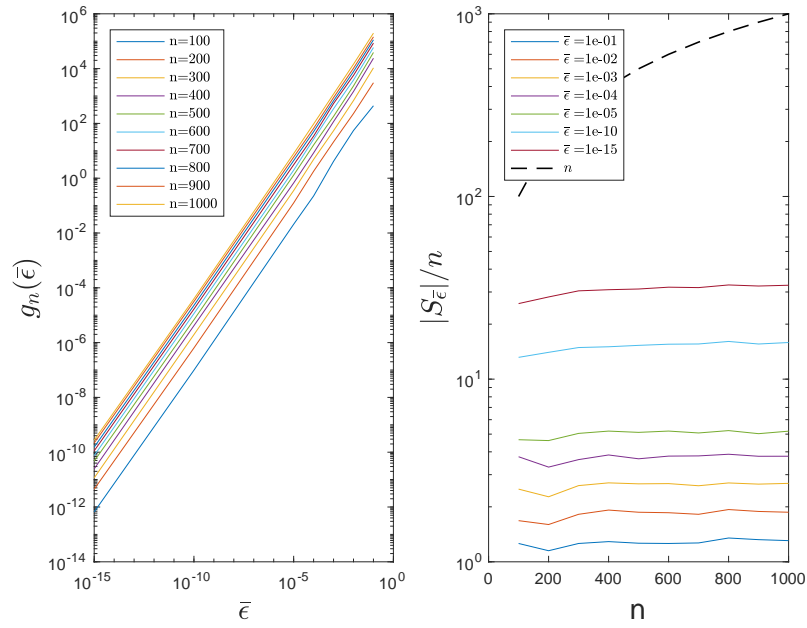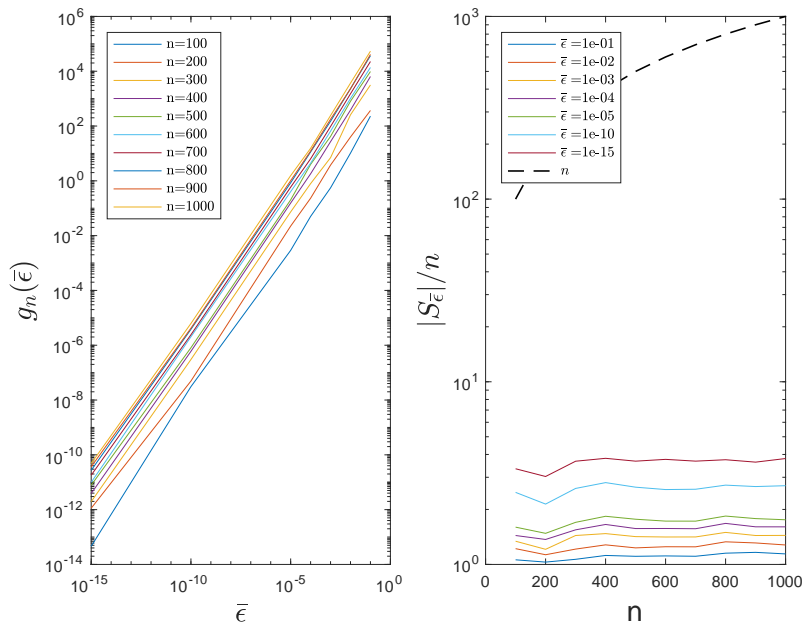
Figure 6: Matern covariance function

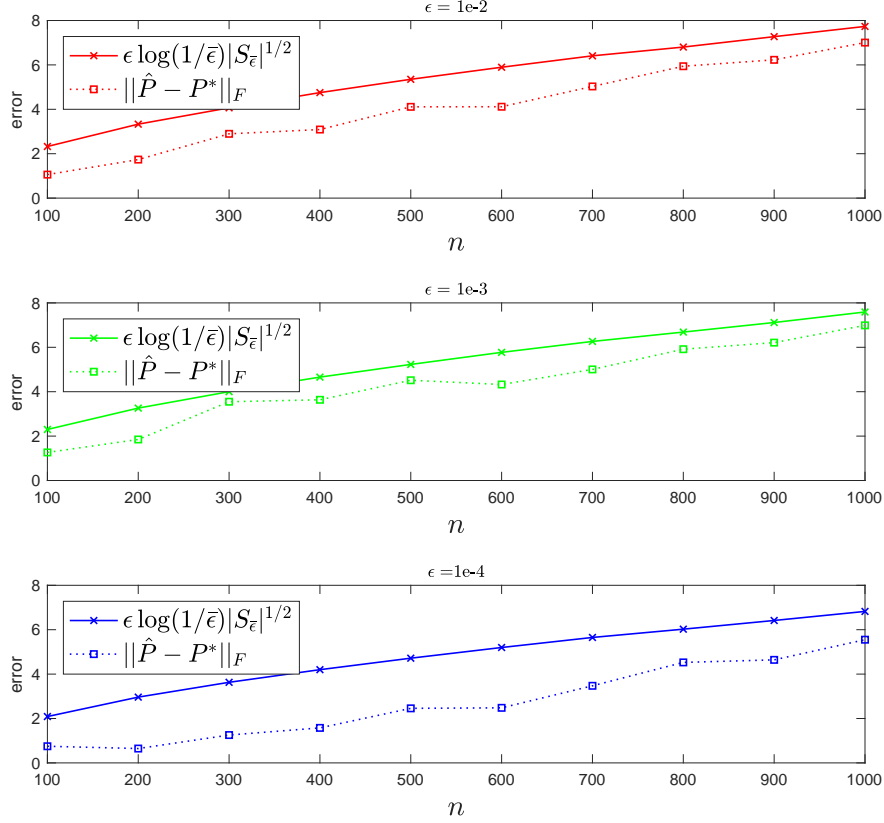Figure 7: Squared exponential covariance function

Figure 8: Actual error $\|\hat{P} - P^*\|_F$ versus the theoretical bound $\epsilon \log(1/\bar{\epsilon})|S_{\bar{\epsilon}}|^{1/2}$

$100, 200, ..., 1000$ points, and calculate C^* and P^* accordingly. Next, we sample $N = N_0/\epsilon^2$ realizations from the resulting GRFs for $\epsilon \in \{1e-2, 1e-3, 1e-4\}$, where we set $M = 1$ in the definition of N_0.

To obtain the *numerical error*, the STAGE-I problems are solved using $\alpha = 40\theta^*_\nu\sqrt{N_0/N}$ with $N = N_0/\epsilon^2$, i.e., $\alpha = 40\theta^*_\nu\epsilon$ to obtain $\hat{P}$. The numerical error is then calculated as $\|\hat{P} - P^*\|_F$. The *theoretical error bound* is computed based on (35) of Corollary 10; in particular, using (43) in the proof Corollary 10, i.e., for any fixed $\epsilon > 0$ we set the bound to $\log(1/\bar{\epsilon})|S_{\bar{\epsilon}}|^{1/2}\sqrt{N_0/N}$ for $N = N_0/\epsilon^2$ and $\bar{\epsilon} > 0$ satisfying (48); hence, it is given as $\epsilon \log(1/\bar{\epsilon})|S_{\bar{\epsilon}}|^{1/2}$. Given ϵ, to compute $\bar{\epsilon}$ satisfying (48), one needs $a^* = \lambda_{\min}(P^*)$ and γ^*, which is the strong convexity modulus of the $g(\cdot)$ function defined in (20) and is calculated using (23). For any $\epsilon \in \{1e-2, 1e-3, 1e-4\}$, to calculate $g_n(\bar{\epsilon})$ small enough satisfying (48), we start from $\bar{\epsilon} = 1$ and keep dividing by 10 until the condition is satisfied. The first $\bar{\epsilon}$ that satisfies (48) is then used to calculate the theoretical error bound. Results are displayed in Figure 8, which shows that the numerical error nicely matches with the theoretical bound for all values of $\epsilon \in \{1e-2, 1e-3, 1e-4\}$.

23

5.2 SPS vs ML Parameter Estimates

An *anisotropic* zero-mean GRF with a squared exponential correlation function, i.e., (3) with $M(\boldsymbol{\theta}_\rho) = \mathrm{diag}(\boldsymbol{\theta}_\rho^{-2})$ and $\boldsymbol{\theta}_\rho \in \Theta_\rho = \mathbb{R}_+^d$, was simulated R times in a hypercube domain $\mathcal{X} = [0, 10]^d$ where the variance and nugget parameters are fixed at $\theta_\nu^* = 1$, $\theta_0^* = 0.1$, and $\boldsymbol{\theta}_\rho^* \in \mathbb{R}^d$ is sampled randomly from the intersection of a hypersphere having radius 10 within the positive quadrant in each replication—$\boldsymbol{\theta}_{\rho,l}^*$ denotes the true correlation parameter vector for the l-th replication. Table 2 compares the quality of the SPS and ML parameter estimates in terms of the mean and standard deviation of $\{\|\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l^*\|\}_{l=1}^R$ for $R = 5$ repeated model fits as the dimension d, the numbers of locations n, and process realizations N change, where $\boldsymbol{\theta}_l^* = [\boldsymbol{\theta}_{\rho,l}^{*\top}, \theta_v^*, \theta_0^*]^\top$ and $\hat{\boldsymbol{\theta}}_l$ is the SPS estimate for the l-th replication. To deal with the nonconcavity of the likelihood function, the ML method is initialized from 1, 10, and 100 random starting points, and the *stationary* solutions with the best objective values among those corresponding to 1, 10, and 100 initializations are denoted as MLE-1, MLE-10, and MLE-100, respectively. To avoid numerical inconsistencies when solving the likelihood problem, the numerically stable approach mentioned in Lophaven et al. (2002) is used. Table 2 also includes the timing comparison between the two methods in seconds. Since the run times do not change much as N changes, we reported the run times for each (d, n) setting averaged over changing values of N. We highlighted in the blue color the regimes where SPS performs better than *all* MLE methods in the mean. According to these empirical findings, *generally* in higher dimensions, e.g., $d \in \{5, 10\}$ and specifically for $N \geq 5$, SPS works as well as (or even better than) all MLE methods. Comparing the estimation times, SPS almost always beats MLE-10 and finishes an order of magnitude sooner than MLE-100. In the lower dimensional case $(d = 2)$, SPS has a reasonable estimation performance, better than MLE-1; but is dominated by MLE-10 and MLE-100. On the other hand, it is worth emphasizing the fact that the better performance of MLE-10 and MLE-100 comes at the cost of considerably longer computation time.

Next, we discuss the effect of parameter estimation quality on the process predictions. When n locations are dense in the domain, interpolating predictions may still be adequate even if the parameter estimates are biased; however, when the location density is low, poor estimates will result in weak prediction performance. This issue is further aggravated in *extrapolation* scenarios. To show the extrapolation behavior, we sample $n = 1000$ training data within a 10-dimensional hypersephere with radius 10 from a zero-mean isotropic GRF with variance, nugget, and range parameters equal to 1, 0.1, and $\sqrt{2}$, respectively. Next, we sample 10,000 test data of which distance to the center is between 10 and 15, i.e., from a hyper-ring. The left graph on Figure 9 shows the design locations projected on the $x_1 - x_2$ plane. The training data is used to fit GRF models using MLE-1, MLE-10, MLE-100 (see the paragraph above for their definitions), and SPS. The graph on the right of Figure 9 shows the prediction errors as a function of the distance between the test point and the convex hull of the training data set. Prediction performance of SPS is better than MLE-1 and MLE-10; but slightly worse than MLE-100, while estimation time is greatly in favor of SPS compared to the MLE-100 method.

---

4. For $d = 10$ and $n = 1000$, one replicate of MLE-100 was not finished in 24hrs.

d	n	Method	N				Time in seconds
			1	5	10	40	
2	100	SPS	2.9 (2.1)	2.5 (1.7)	1.6 (0.9)	1.1 (0.6)	6.3 (0.9)
		MLE-1	2.6 (1.7)	2.1 (1.5)	1.4 (0.9)	1.0 (0.7)	1.2 (0.2)
		MLE-10	2.6 (1.7)	2.1 (1.5)	1.4 (0.9)	1.0 (0.7)	11.1 (1.1)
		MLE-100	2.6 (1.7)	2.1 (1.5)	1.4 (0.9)	1.0 (0.7)	119.2 (7.1)
	500	SPS	2.7 (1.8)	1.8 (1.1)	1.5 (0.9)	1.0 (0.6)	274.3 (8.1)
		MLE-1	2.3 (1.5)	1.4 (1.0)	1.0 (0.6)	1.0 (0.4)	49.7 (25.1)
		MLE-10	2.3 (1.5)	1.4 (1.0)	1.0 (0.6)	1.0 (0.4)	407.6 (72.2)
		MLE-100	2.3 (1.5)	1.4 (1.0)	1.0 (0.6)	1.0 (0.4)	3931.5 (419.4)
	1000	SPS	2.1 (1.4)	1.6 (0.9)	1.3 (0.6)	0.9 (0.3)	1793.9 (41.6)
		MLE-1	1.9 (1.3)	1.2 (0.8)	0.9 (0.5)	0.8 (0.2)	405.3 (342.4)
		MLE-10	1.9 (1.3)	1.2 (0.8)	0.9 (0.5)	0.8 (0.2)	2858.6 (434.9)
		MLE-100	1.9 (1.3)	1.2 (0.8)	0.9 (0.5)	0.8 (0.2)	24771.1 (2413.1)
5	100	SPS	3.2 (1.6)	2.3 (1.4)	1.8 (1.2)	1.3 (0.8)	7.6 (2.4)
		MLE-1	3.1 (1.9)	2.5 (1.7)	2.1 (1.6)	1.8 (1.4)	1.9 (0.4)
		MLE-10	3.1 (1.9)	2.5 (1.7)	1.9 (1.6)	1.7 (1.4)	20.7 (0.8)
		MLE-100	2.9 (1.5)	2.3 (1.6)	1.9 (1.5)	1.6 (1.2)	233.7 (68.6)
	500	SPS	2.9 (1.7)	1.9 (1.4)	1.8 (1.1)	1.3 (0.8)	359.3 (45.3)
		MLE-1	2.8 (1.7)	1.9 (1.6)	1.8 (1.5)	1.6 (1.4)	72.3 (12.2)
		MLE-10	2.8 (1.7)	1.8 (1.6)	1.8 (1.5)	1.3 (1.1)	782.0 (116.1)
		MLE-100	2.5 (1.7)	1.7 (1.4)	1.6 (1.4)	1.2 (1.0)	7924.2 (1720.7)
	1000	SPS	2.8 (1.6)	1.6 (1.1)	1.3 (0.7)	0.9 (0.4)	2050.8 (137.7)
		MLE-1	2.7 (1.4)	1.8 (1.3)	1.5 (1.0)	1.1 (0.9)	520.9 (133.4)
		MLE-10	2.5 (1.4)	1.7 (1.3)	1.3 (1.0)	1.0 (0.7)	4485.7 (686.0)
		MLE-100	2.3 (1.3)	1.5 (1.3)	1.2 (0.9)	1.0 (0.6)	49587.4 (1099.7)
10	100	SPS	5.8 (2.3)	4.1 (1.6)	3.2 (1.4)	1.9 (1.0)	9.8 (5.3)
		MLE-1	6.6 (2.6)	5.1 (2.4)	5.0 (2.4)	4.3 (2.1)	3.9 (2.2)
		MLE-10	6.2 (2.3)	5.1 (2.1)	4.9 (1.9)	4.8 (1.9)	48.9 (10.7)
		MLE-100	6.2 (2.1)	5.0 (2.0)	4.4 (1.9)	3.9 (1.8)	532.1 (121.8)
	500	SPS	4.9 (2.0)	3.9 (1.5)	2.9 (1.3)	1.6 (0.9)	284.6 (8.7)
		MLE-1	5.8 (2.4)	4.9 (2.2)	4.3 (2.0)	3.7 (1.7)	230.4 (154.7)
		MLE-10	5.0 (2.3)	4.8 (2.0)	3.9 (1.9)	3.6 (1.7)	2873.7 (1042.7)
		MLE-100	4.8 (2.0)	4.3 (1.9)	3.4 (1.8)	2.8 (1.5)	42739.0 (21786.0)
	1000	SPS	4.8 (1.9)	3.6 (1.3)	2.6 (1.3)	1.4 (0.8)	3544.1 (171.2)
		MLE-1	5.8 (2.3)	4.8 (2.0)	4.1 (1.7)	3.6 (1.6)	644.8 (180.3)
		MLE-10	4.8 (2.1)	4.5 (1.8)	3.7 (1.4)	3.4 (1.3)	5535.6 (926.5)
		MLE-100[4]	– (–)	– (–)	– (–)	– (–)	– (–)

Table 2: SPS vs MLE methods. Numbers are the mean (standard deviation) of $\{\|\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l^*\|\}_{l=1}^5$. Blue color is used for the SPS method and red color highlights the regimes where MLE performs poorer than SPS *in the mean*.

5.3 Dealing with the "Big n" Problem

To solve the STAGE-I problem efficiently for large n, we propose to segment the set of training locations $\mathcal{D}^x \triangleq \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ into K segments (or blocks) $\{\mathcal{D}_k^x\}_{k=1}^K$ of size $n_k \triangleq \mathrm{card}(\mathcal{D}_k^x)$ such that the number of data points in each segment, n_k, is less than n_B (in our experiments we set n_B to 1000). If $n \leq n_B$, then segmentation is not needed; otherwise, we used the following two segmentation schemes in our numerical experiments.

Spatial Segmentation (SS): This scheme is based on partitioning the spatial domain $\mathcal{X}$ into K non-overlapping hypercubes, and considering the training data points in each hypercube $\mathcal{X}_k$ as one block. Let $\mathcal{X}_k \triangleq \{\mathbf{x} \in \mathbb{R}^d : \bar{\ell}_k \leq \mathbf{x} \leq \bar{u}_k\}$ for some $\{\bar{\ell}_k, \bar{u}_k\}_{k=1}^K$ such that $\cup_{k=1}^K \mathcal{X}_k = \mathcal{X}$ and $\mathcal{X}_{k_1} \cap \mathcal{X}_{k_2} = \emptyset$ for all $k_1 \neq k_2$. Then we define $\mathcal{D}_k^x \triangleq \mathcal{D}^x \cap \mathcal{X}_k$ for all k. Assuming that data point locations, $\mathcal{D}^x$, are uniformly distributed within $\mathcal{X}$, each block $\mathcal{D}_k^x$ will contain $\frac{n}{K}$ observations in expectation.
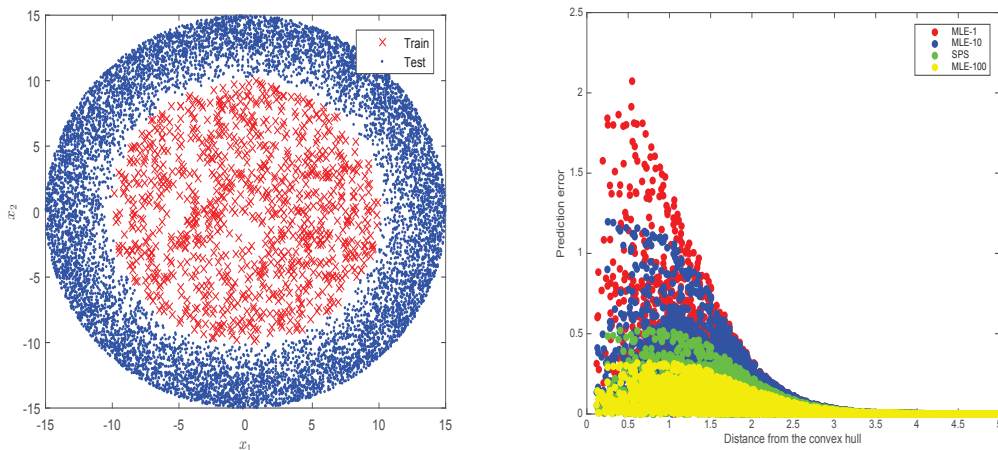
Figure 9: **Left:** Hyper-sphere design locations projected on the $x_1 - x_2$ plane. **Right:** Prediction performance

Random Segmentation **(RS):** The set of training data locations $\mathcal{D}^x$ is partitioned *uniformly at random* into K blocks $\{\mathcal{D}^x_k\}^K_{k=1}$ such that the first $K - 1$ blocks contain $\lfloor \frac{n}{K} \rfloor$ data locations and the last block contains $n - (K-1)\lfloor \frac{n}{K} \rfloor$ many. Let $\mathcal{D}_k = \{(\mathbf{x}_i, y_i^{(r)}) : r = 1, \ldots, N, \ i \in \mathcal{I}_k\}$ denote the subset of training data corresponding to the k-th block, where the index set $\mathcal{I}_k$ is defined as $\mathcal{I}_k \triangleq \{1 \leq i \leq n : \ \mathbf{x}_i \in \mathcal{D}^x_k\}$. Hence, $n_k = |\mathcal{I}_k|$. Note that the SS segmentation scheme, but not the RS scheme, can handle non-stationary GRFs.

For both segmentation schemes, we solve STAGE-I problem *for each segment k* separately using the sample covariance S_k and the matrix of pairwise distances G_k corresponding to observations in segment k, i.e., $\hat{P}_k = \mathrm{argmin}_{P \succ 0} \langle S_k, P \rangle - \log \det(P) + \alpha_k \langle G_k, |P| \rangle$, for $k = 1, \ldots, K$, for some $\alpha_k > 0$. Under the assumption that the underlying stochastic process is second-order stationary, i.e., the covariance function parameters are fixed across the domain $\mathcal{X}$ (Stein, 1999), one can fit a single covariance function for the whole domain. In this case, for both schemes we propose to estimate the covariance function parameters by solving the following least squares problem: $\hat{\boldsymbol{\theta}} \in \mathrm{argmin}_{\boldsymbol{\theta} \in \Theta} \sum^K_{k=1} \|\hat{P}_k^{-1} - C_k(\boldsymbol{\theta})\|^2_F$. Note that this method generates a predicted surface with *no discontinuities* along the boundary between segments. This is in contrast to other methods that partition large data sets for fitting a GRF but require further formulation to achieve continuity, see Park et al. (2011). Finally, in case the process cannot be assumed to be stationary, the second stage optimization is solved separately for each segment k, which has its own covariance parameter estimates. These estimates are computed by solving $\hat{\boldsymbol{\theta}}_k \in \mathrm{argmin}_{\boldsymbol{\theta}_k \in \Theta} \|\hat{P}_k^{-1} - C_k(\boldsymbol{\theta}_k)\|^2_F$, for each segment $k = 1, \ldots, K$.

SPS-Fitted GRF for Small and Large Data Sets. We simulated two data sets of sizes $n = 1,000$ and $n = 64,000$ points from a GRF with zero mean and isotropic SE covariance function with parameters: range $\boldsymbol{\theta}^*_\rho = 4$, variance $\theta^*_v = 8$, and nugget $\theta^*_0 = 4$ over a square domain $\mathcal{X} = [0, \ 100] \times [0, \ 100]$. In the simulation with $n = 1000$, the results are based on R (number of simulation replications) equal to 100, while for the simulation with $n = 64,000$, given that each run of the simulation takes around 3-4 hours, results are given for $R = 5$ replicates. The number of realizations N is set to 1 in these simulations. Let $\hat{\boldsymbol{\theta}}_l$ denote the covariance parameter estimates obtained in the l-th replication. In all the

tables, $\bar{\boldsymbol{\theta}} \triangleq \sum_{l=1}^{R} \hat{\boldsymbol{\theta}}_l / R$ and $\mathbf{stdev}_{\boldsymbol{\theta}} \triangleq \sqrt{\frac{1}{R} \sum_{l=1}^{R} (\hat{\boldsymbol{\theta}}_l - \bar{\boldsymbol{\theta}})^2}$ denote the sample mean and the standard deviation of the parameter estimates, respectively. For benchmarking, we compare our estimates with those obtained by *Domain Decomposition* method (DDM) of Park et al. (2011) using a rectangular mesh. The number of control points on the boundaries and the number of constraining degrees of freedom equal to 3. DDM is selected since it shows the best performance among the other big-n methods considered in Section 5.4.

Method	n=1000 (R=100 replicates)		n=64000 (R=5 replicates)	
	$\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{\theta}}_\rho, \bar{\theta}_v, \bar{\theta}_0)$	$\mathbf{stdev}_{\boldsymbol{\theta}}$	$\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{\theta}}_\rho, \bar{\theta}_v, \bar{\theta}_0)$	$\mathbf{stdev}_{\boldsymbol{\theta}}$
SPS-SS	(3.98, 7.77, 4.87)	(0.41, 1.01, 0.76)	(4.01, 8.16, 4.75)	(0.43, 0.78, 0.29)
SPS-RS	(4.01, 8.11, 4.22)	(0.90, 1.16, 0.85)	(3.98, 7.97, 4.83)	(0.06, 0.11, 0.11)
DDM	(1.73, 7.34, 6.11)	(1.23, 1.98, 2.67)	(1.43, 8.67, 3.65)	(0.87, 0.95, 0.53)

Table 3: Parameter estimate of SPS under the two segmentation schemes vs. DDM (Domain Decomposition method of Park et al. (2011)) for simulated data with N=1 realization. The covariance function is squared-exponential and the true parameter values are $\boldsymbol{\theta}_\rho^* = 4$, $\theta_v^* = 8$, and $\theta_0^* = 4$.

Both segmentation schemes were used for comparison. When $n = 1000$, for the SS segmentation scheme, the domain was split into $3 \times 3 = 9$ equal size square segments; for the RS segmentation scheme $\mathcal{D}^x$ was randomly partitioned into 9 equal cardinality sets. Similarly, when $n = 64,000$, for the SS segmentation scheme, the domain was split into $8 \times 8 = 64$ equal size square segments; for the RS scheme, $\mathcal{D}^x$ was randomly partitioned into 64 equal cardinality sets. Table 3 shows the model fitting results.

When $n = 1000$, the parameter estimates using either segmentation scheme appear unbiased. Computing $\hat{\boldsymbol{\theta}} \in \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{k=1}^{K} \|\hat{P}_k^{-1} - C_k(\boldsymbol{\theta})\|_F^2$, we explicitly ignore the correlation of process values for any two points in different blocks. Over the fixed domain when n is large, i.e., $n = 64,000$, the larger data location density results in more observations close to boundaries; hence, correlations between blocks for the SS scheme may not be ignored anymore. Empirical results show lower $\mathbf{stdev}_{\boldsymbol{\theta}}$ under the RS scheme when n is large. DDM underestimates the range parameter in all scenarios.

Effect of range and nugget parameters. To analyze the effect of the range, $\boldsymbol{\theta}_\rho^*$, and nugget, θ_0^*, parameters on the performance of the proposed method, we setup another simulation with n=64,000 points with results shown in Table 4. When the range parameter increases, the standard deviations of the parameter estimates increase under both segmentation schemes. Furthermore, the RS scheme appears to be less sensitive to changes in the nugget parameter. In general, for big-n scenarios and given the high point density, the RS scheme results in more robust parameter estimates. DDM highly underestimates the range parameter, especially when $\boldsymbol{\theta}_\rho^* = 30$.

5.4 SPS vs. State-of-the-Art Methods for Fitting GRFs to Big Data Sets

We compare the SPS method against the *Partial Independent Conditional* (PIC) method by Snelson and Ghahramani (2007), the *Domain Decomposition* (DDM) method of Park et al. (2011), and the *Full Scale covariance Approximation* (FSA) of Sang and Huang (2012). Park et al. (2011) provided computer codes for PIC and DDM, and we coded the FSA method. We simulated a data set of size $n = 64,000$ generated from a zero mean GRF with

Nugget	Method	$\bar{\boldsymbol{\theta}} = (\bar{\theta}_\rho, \bar{\theta}_v, \bar{\theta}_0)$ $\boldsymbol{\theta}_\rho^* = 4$	$\mathbf{stdev}_{\boldsymbol{\theta}}$	$\bar{\boldsymbol{\theta}} = (\bar{\theta}_\rho, \bar{\theta}_v, \bar{\theta}_0)$ $\boldsymbol{\theta}_\rho^* = 30$	$\mathbf{stdev}_{\boldsymbol{\theta}}$
$\theta_0^* = 4$	SPS-SS	(4.01, 8.16, 4.75)	(0.43, 0.78, 0.29)	(29.03, 7.94, 4.80)	(1.15, 1.76, 0.11)
	SPS-RS	(3.98, 7.97, 4.83)	(0.06, 0.11, 0.11)	(29.24, 7.95, 4.79)	(1.93, 0.26, 0.19)
	DDM	(1.43, 8.67, 3.65)	(0.87, 0.95, 0.53)	(19.32, 8.65, 3.69)	(0.97, 0.64, 0.73)
$\theta_0^* = 8$	SPS-SS	(4.03, 8.08, 8.77)	(0.47, 0.86, 0.36)	(28.39, 7.97, 8.77)	(1.35, 1.89, 0.13)
	SPS-RS	(3.98, 7.98, 8.83)	(0.07, 0.12, 0.15)	(28.93, 7.87, 8.78)	(1.65, 0.31, 0.18)
	DDM	(1.27, 7.35, 9.10)	(0.08, 0.08, 0.25)	(11.45, 7.35, 10.10)	(1.89, 0.70, 0.49)

Table 4: SPS vs. DDM (Domain Decomposition method of Park et al. (2011)) estimates for simulated data sets with $n = 64,000$, $N = 1$ and $R = 5$. The covariance function is squared-exponential with variance parameter $\theta_v^* = 8$.

isotropic Squared-Exponential (SE) covariance function with $N = 1$ realization using the following parameter values: $\boldsymbol{\theta}_\rho^* = 4$, $\theta_v^* = 8$, and $\theta_0^* = 4$.

For each replication, 90% of the simulated data was allocated for training (i.e., for estimating the parameters), and 10% for testing (prediction). The Mean Square Prediction Error (MSPE) is computed on the *test data*. The MSPE corresponding to the l-th replication is computed as follows: $\text{MSPE}_l \triangleq \frac{1}{n_l^t} \|\mathbf{y}_l^t - \hat{\mathbf{y}}_l^t\|^2$, where n_l^t denotes the number of *test data* points in the l-th replication, $\mathbf{y}_l^t \in \mathbb{R}^{n_l^t}$ and $\hat{\mathbf{y}}_l^t \in \mathbb{R}^{n_l^t}$ are vectors of true and predicted function values, respectively. Since the true parameter values are known for the simulated data set, the true function values $\mathbf{y}^t$ are taken to be the predictions obtained via (1) using the *true* parameter values. Calculating the MSPE this way shows the specific error due to the discrepancies between the estimated and true parameter values. Finally, $\overline{\text{MSPE}}$ and $\mathbf{stdev}_{\text{MSPE}}$ are defined similar to $\bar{\boldsymbol{\theta}}$ and $\mathbf{stdev}_{\boldsymbol{\theta}}$.

In the SPS method, we used SS segmentation scheme with 64 equal-size blocks as described in Section 5.3. In the PIC method, the number of local regions was set to 64, and the number of pseudo inputs was set to 100. In the DDM method, a rectangular mesh was selected with both the number of control points on the boundaries and the number of constraining degrees of freedom equal to 3. In the FSA method, the number of knots was set to 50 on a regular grid, the tapering function used was spherical with taper range set to 10. These settings are based on the guidance provided in the corresponding papers. In the training phases for PIC, DDM, and FSA methods, the initial values for each covariance function parameters were randomly selected from the uniform distribution over $(0, 10]$ in each replicate. The reason is that these methods attempt to solve non-convex problem in (5); hence, the local minima generated by the optimization solvers highly depend on the initial point. Therefore, to be fair to these methods, we run them starting from many randomly generated initial solutions for each replicate. The mean and standard deviation of MSPE and parameter estimates for $R = 5$ replications are reported in Table 5.

The SPS method provides the least biased estimates for all three covariance parameters, with the degree of bias provided by the other methods being much more substantial. Furthermore, the mean MSPE for the SPS method is considerably lower than that of the other

Method	$\overline{\boldsymbol{\theta}} = (\overline{\boldsymbol{\theta}}_\rho, \overline{\theta}_v, \overline{\theta}_0)$	$\mathbf{stdev}_{\boldsymbol{\theta}}$	$\overline{\mathrm{MSPE}}$	$\mathbf{stdev}_{\mathrm{MSPE}}$
PIC	(5.11, 6.22, 5.01)	(2.03, 2.82, 2.532)	2.87	1.03
DDM	(0.83, 8.23, 4.03)	(0.09, 0.96, 0.73)	2.23	0.44
FSA	(3.31, 2.58, 0.65)	(2.97, 0.76, 0.14)	4.47	1.35
SPS	(4.14, 7.82, 4.64)	(0.65, 1.06, 0.57)	0.42	0.35

Table 5: Comparison of the SPS method against PIC, DDM, and FSA for $n = 64,000$, $N = 1$ and $R = 5$ on data sets generated from a GRF with zero mean and SE covariance function with true parameters $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_\rho^*, \theta_v^*, \theta_0^*) = (4, 8, 4)$.

alternatives (one order of magnitude less), and has the least variability. The CPU times required by each method in the learning and prediction stages are displayed in Figure 10.
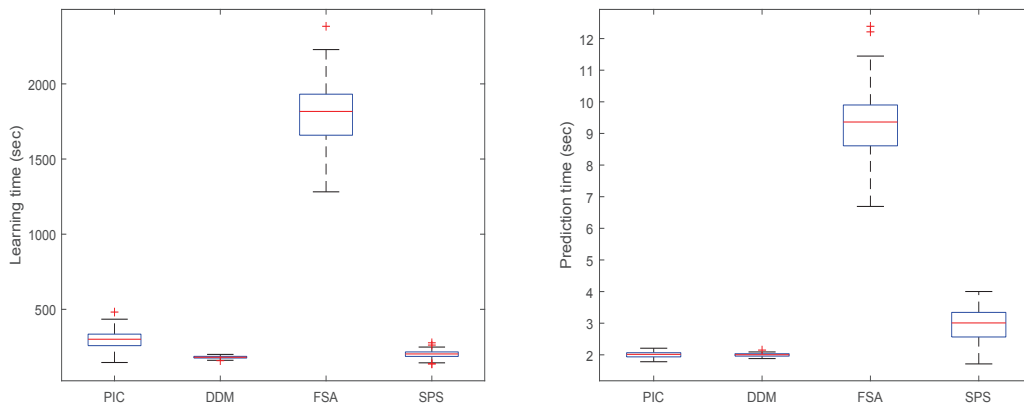
Figure 10: Learning times **(Left)** and prediction times **(Right)** of PIC, DDM, FSA, and SPS methods.

For both learning and prediction the DDM method is the fastest, and the FSA method the slowest. However, note that while PIC, DDM and SPS use segmentation (64 blocks in this example), there is no segmentation in FSA but instead one should select the number of knots. Hence computing times of FSA against the other three methods are not completely comparable. Of the remaining two methods contrasted, PIC is faster than SPS in the prediction phase but SPS is faster than PIC in learning. In view of the prediction performance of all the methods compared, the slight speed advantage of DDM over SPS is not a demerit of our method: DDM is unable to provide an unbiased estimate of the range parameter $\boldsymbol{\theta}_\rho^*$, crucial in spatial modeling, and this naturally results in considerably worse predictions.

5.5 Implementation of the SPS Method for Real Data Sets

Finally, we tested the SPS method on two real data sets. The first data set contains ozone and air quality data as recorded by the Total Ozone Mapping Spectrometer (TOMS) instrument on board the Nimbus-7 satellite. The data set contains 48,331 Total Column Ozone (TCO) measurements over the globe on October 1, 1988 and is available at NASA's

website[5]. The second data set is the Day/Night Cloud Fraction (CF) from January to September 2009 (size $n = 64,800$ points) collected by the Moderate Resolution Imaging Spectrometer (MODIS) instrument on board the Terra satellite, a data set also available at NASA's website[6]. A wrapper Matlab function which can read the TCO data with its specific format and produce the input/output matrices is available with our software package at `https://github.com/samdavanloo/SPS` and also at `http://sites.psu.edu/engineeringstatistics/computer-codes/`.

Data	Covariance Function	Segmentation	$\overline{\boldsymbol{\theta}} = (\overline{\boldsymbol{\theta}}_\rho, \overline{\theta}_v, \overline{\theta}_0)$	$\mathbf{stdev}_{\boldsymbol{\theta}}$	$\overline{\mathrm{MSPE}}$	$\mathbf{stdev}_{\mathrm{MSPE}}$
TCO	Matern	RS	(12.20, 1098, 0.00)	(0.01, 0.54, 0.00)	4.5361	0.0623
CF	Exponential	RS	(10.07, 0.05, 0.82)	(0.04, 0.00, 0.00)	0.0044	0.0000

Table 6: Implementation of the proposed SPS method for two real data sets, TCO and CF, with performance measures computed over R=5 replicated cross-validations (10% of data randomly sampled in each replication for testing).

The mean and standard deviations of the MSPEs and the parameter estimates for $R = 5$ replicates are reported in Table 6. In each replication, 10% of the data is randomly selected for testing, and the remaining 90% is used for learning the covariance parameters. Both segmentation schemes were implemented and RS segmentation resulted in better prediction results for both data sets; therefore, we only report RS results. The RS segmentation scheme is adopted using a random partition of $\mathcal{D}^x$ in each replication. The type of covariance function was selected based on the best MSPE values obtained. Since this is real data, we cannot make any judgment about the quality of parameter estimates; however, the standard deviations are quite small relative to the parameter estimate magnitudes for both data sets.

6. Conclusions and Further Research

A new two-stage method to estimate the parameters of Gaussian Random Field (GRF) models is presented and its theoretical error bound is established. This method, which we named Sparse Precision matrix Selection (SPS), first finds a sparse estimate of the precision (inverse covariance) matrix of the underlying GRF by solving a nonsmooth convex optimization problem, and then estimates the parameters of the GRF model by solving a least-square problem. Numerical studies confirming our theoretical findings are presented; these include numerical comparisons with MLE which requires solving a nonconvex problem. In higher dimensional regimes, especially with higher number of realizations, the SPS method performs better than its MLE counterparts. Moreover, the computational time of SPS scales much better with the number of locations, replicated observations per location, and especially, with the dimension.

The sparse estimation of the precision matrix of the GRF model is well-motivated and theoretically supported by Jaffard's decay algebra (Jaffard, 1990). Indeed, we were able to bound the error of the sparse estimate of the true precision matrix, which is dense but its elements decay rapidly to zero in magnitude. This bound then allowed us to establish the error bound of the parameter estimates in Corollary 10.

---

5. http://ozoneaq.gsfc.nasa.gov/nimbus7Ozone.md

6. http://gdata1.sci.gsfc.nasa.gov/daac-bin/G3/gui.cgi?instance_id=MODIS_MONTHLY_L3

The following are some possible future research directions: **a)** The established error bound for the stage-I problem that involves sparse estimation of the true dense precision matrix includes the cardinality of some set $|S_{\bar\epsilon}|$ given in (34) for some $\bar\epsilon > 0$. Clearly, $|S_{\bar\epsilon}| \leq \mathrm{card}(P^*)$ for all $\bar\epsilon \geq 0$; however, theoretically quantifying the cardinality of this set for even simple designs, e.g. d-dimensional lattice, is a combinatorial challenge and is left for future works. **b)** Our numerical motivations at the beginning of the paper suggest that the decay rate of the precision matrix increases with the density of the points (infill asymptotics). We did not investigate this phenomenon *theoretically* nor did we utilize it in the proposed parameter estimation algorithm, and it is left as a matter for future research.

Appendix A. Proof of Theorem 4

Proof For any given $\boldsymbol{\theta}_\rho \in \Theta_\rho$, note that $\mathbf{d}$ is not parallel to $\mathbf{r} = \mathbf{r}(\boldsymbol{\theta}_\rho)$, i.e., $\mathbf{d} \nparallel \mathbf{r}$. Let h denote the objective function in (14), i.e., $h(\theta_v, \theta_0) \triangleq \frac{1}{2}\|\theta_v \mathbf{r} + \theta_0 \mathbf{d} - \hat{\mathbf{c}}\|^2$, it satisfies

$$\frac{\partial h(\theta_v, \theta_0)}{\partial \theta_v} = \mathbf{r}^\top(\theta_v \mathbf{r} + \theta_0 \mathbf{d} - \hat{\mathbf{c}}), \tag{49a}$$

$$\frac{\partial h(\theta_v, \theta_0)}{\partial \theta_0} = \mathbf{d}^\top(\theta_v \mathbf{r} + \theta_0 \mathbf{d} - \hat{\mathbf{c}}). \tag{49b}$$

The Hessian of h in (14) is $\nabla^2 h = \begin{bmatrix} \mathbf{r}^\top\mathbf{r} & \mathbf{r}^\top\mathbf{d} \\ \mathbf{r}^\top\mathbf{d} & \mathbf{d}^\top\mathbf{d} \end{bmatrix}$. Note $\mathbf{r}^\top\mathbf{r} > n > 0$, and $\det(\nabla^2 h) = \|\mathbf{r}\|^2\|\mathbf{d}\|^2 - (\mathbf{r}^\top\mathbf{d})^2 > 0$ by Cauchy-Schwartz and the fact that $\mathbf{r} \nparallel \mathbf{d}$; thus, $\nabla^2 h$ is positive definite. Therefore, for any given $\boldsymbol{\theta}_\rho \in \Theta_\rho$, h is strongly convex jointly in θ_v and θ_0.

From the definitions of $\mathbf{d}$ and $\mathbf{r}$, we have $\|\mathbf{d}\|^2 = n$, $\mathbf{d}^\top\mathbf{r} = n$, and $\|\mathbf{r}\|^2 > n$ (because $\mathbf{r}(\mathbf{x}, \mathbf{x}, \boldsymbol{\theta}_\rho) = 1$ for any $\mathbf{x}$ and $\boldsymbol{\theta}_\rho \in \Theta_\rho$). Necessary and sufficient KKT conditions imply

$$\nabla h(\theta_v, \theta_0) \geq \mathbf{0}, \tag{50a}$$

$$\theta_v \geq 0, \quad \theta_0 \geq 0, \tag{50b}$$

$$\frac{\partial h(\theta_v, \theta_0)}{\partial \theta_v}\,\theta_v = 0, \quad \frac{\partial h(\theta_v, \theta_0)}{\partial \theta_0}\,\theta_0 = 0. \tag{50c}$$

Below, we consider four possible scenarios for problem (14):

1. $(\theta_v = 0, \theta_0 = 0)$—This solution is optimal if and only if $\mathbf{r}^\top\hat{\mathbf{c}} \leq 0$ and $\mathbf{d}^\top\hat{\mathbf{c}} \leq 0$ (from (49a) (49b), and (50a)). However, since $\hat{C} = \hat{P}^{-1}$ is positive definite, its diagonal elements are strictly positive, $\mathbf{d}^\top\hat{\mathbf{c}} > 0$. Hence, this scenario is not possible.

2. $(\theta_v = 0, \theta_0 > 0)$—From (50a), (50c) and (49b) follows that $(\theta_v = 0, \theta_0 = \mathbf{d}^\top\hat{\mathbf{c}}/n)$ is the optimal solution if and only if $\mathbf{r}^\top\hat{\mathbf{c}} \leq \mathbf{d}^\top\hat{\mathbf{c}}$.

3. $(\theta_v > 0, \theta_0 > 0)$—From (49a), (49b), (50c), and (50a)

$$(\theta_v, \theta_0) = \left(\frac{\mathbf{r}^\top\hat{\mathbf{c}} - \mathbf{d}^\top\hat{\mathbf{c}}}{\|\mathbf{r}\|^2 - n}, \frac{(\mathbf{d}^\top\hat{\mathbf{c}})\|\mathbf{r}\|^2/n - \mathbf{r}^\top\hat{\mathbf{c}}}{\|\mathbf{r}\|^2 - n} \right)$$

is the optimal solution if and only if $(\mathbf{d}^\top\hat{\mathbf{c}})\|\mathbf{r}\|^2/n > \mathbf{r}^\top\hat{\mathbf{c}} > \mathbf{d}^\top\hat{\mathbf{c}}$.

4. $(\theta_v > 0, \theta_0 = 0)$—From (50a), (50c) and (49a) follows that $(\theta_v = \mathbf{r}^\top\hat{\mathbf{c}}/\|\mathbf{r}\|^2, \theta_0 = 0)$ is the optimal solution if and only if $\mathbf{r}^\top\hat{\mathbf{c}} \geq (\mathbf{d}^\top\hat{\mathbf{c}})\|\mathbf{r}\|^2/n$.

■

References

N. S. Aybat and G. Iyengar. An alternating direction method with increasing penalty for stable principal component pursuit. *Computational Optimization and Applications*, 61: 635–668, 2015.

Sudipto Banerjee, Alan E. Gelfand, Andrew O. Finely, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of Royal Statistical Society*, 70: 825–848, 2008.

M. Benzi. Localization in matrix computations: Theory and applications. Technical Report Math/CS Technical Report TR-2016-003, Emory University, May 2016. To appear in M. Benzi and V. Simoncini (Eds.), "Exploiting Hidden Structure in Matrix Computations: Algorithms and Applications (Cetraro, Italy 2015)", Lecture Notes in Mathematics, Springer and Fondazione CIME (95 pages).

G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2008. ISBN 9780470272848.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Emmanuel J Candès et al. Compressive sampling. In *Proceedings of the international congress of mathematicians*, volume 3, pages 1433–1452. Madrid, Spain, 2006.

Noel Cressie and Gardar Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of Royal Statistical Society*, 70:209–226, 2008.

Noel A. C. Cressie. *Statistics for spatial data*. John Wiley and Sons, 1993. ISBN 0471002550.

A. d'Aspremont, O. Banerjee, and Ghaoui L. El. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008. doi: 10.1137/060670985.

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):pp. 157–175, 1972. ISSN 0006341X.

Wei Deng and Wotao Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Montserrat Fuentes. Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, 102(477):321–331, 2007.

Reinhard Furrer, Marc G. Genton, and Douglas W. Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15 (3):502–523, 2006.

Robert B Gramacy and Daniel W Apley. Local Gaussian process approximation for large computer experiments. *arXiv preprint arXiv:1303.0383*, 2013.

Bingsheng He, Li-Zhi Liao, Deren Han, and Hai Yang. A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming*, 92(1):103–118, 2002.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

Dave Higdon. Space and space-time modeling using process convolutions. In CliveW. Anderson, Vic Barnett, PhilipC. Chatwin, and AbdelH. El-Shaarawi, editors, *Quantitative Methods for Current Environmental Issues*, pages 37–56. Springer London, 2002. ISBN 978-1-4471-1171-9.

Jean Honorio and Tommi S. Jaakkola. Inverse covariance estimation for high-dimensional data in linear time and space: Spectral methods for riccati and sparse models. Technical Report arXiv:1309.6838, Massachusetts Institute of Technology, 2013.

Cho-Jui Hsieh, Matyas A Sustik, Inderjit S Dhillon, Pradeep K Ravikumar, and Russell Poldrack. Big & quic: Sparse inverse covariance estimation for a million variables. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3165–3173. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/4923-big-quic-sparse-inverse-covariance-estimation-for-a-million-variables.pdf`.

Stephane Jaffard. Propriétés des matrices ≪bien localisées≫ près de leur diagonale et quelques applications. In *Annales de l'IHP Analyse non linéaire*, volume 7, pages 461–476, 1990.

A. G. Journel and Ch. J. Huijbregts. *Mining Geostatistics*. Academic Press, 1978. ISBN 0-12-391050-1.

S. Kontogiorgis and R. R. Meyer. A variable-penalty alternating direction method for convex optimization. *Mathematical Programming*, 83:29–53, 1998.

A.S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1/2):173–183, 1995.

Faming Liang, Yichen Cheng, Qifan Song, Jincheol Park, and Ping Yang. A resampling-based stochastic approximation method for analysis of large geostatistical data. *Journal of the American Statistical Association*, 108(501):325–339, 2013.

Finn Lindgren, Havard Rue, and Johan Lindstrom. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011. ISSN 1467-9868.

P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16:964–979, 1979.

S. N. Lophaven, H. B. Nielsen, and J. Søndergaard. Aspects of the matlab toolbox DACE. Technical report, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, 2002. URL `http://www2.imm.dtu.dk/pubdb/p.php?1050`.

Zhaosong Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19(4):1807–1827, 2009.

Zhaosong Lu. Adaptive first-order methods for general sparse inverse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2000–2016, 2010.

K. V. Mardia and A. J. Watkins. On multimodality of the likelihood in the spatial linear model. *Biometrika*, 76(2):289–295, 1989.

Rahul Mazumder and Trevor Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research*, 13(Mar): 781–794, 2012.

Douglas Nychka, Soutir Bandyopadhyay, Dorit Hammerling, Finn Lindgren, and Stephan Sain. A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015. doi: 10.1080/10618600.2014.914946.

E.A. Ok. *Real Analysis with Economic Applications.* Number v. 10 in Real analysis with economic applications. Princeton University Press, 2007. ISBN 9780691117683.

Chiwoo Park, Jianhua Z. Huang, and Yu Ding. Domain decomposition approach for fast Gaussian process regression of large spatial data sets. *J. Mach. Learn. Res.*, 12:1697–1728, July 2011. ISSN 1532-4435.

Carl Edward Rasmussen and Christopher K. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006. ISBN 0-262-18253-X.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Erik Rodner, Alexander Freytag, Paul Bodesheim, and Joachim Denzler. Large-scale gaussian process classification with flexible adaptive histogram kernels. In *European Conference on Computer Vision*, pages 85–98. Springer, 2012.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, Ji Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Havard Rue and Leonhard Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, 2005. ISBN 1-58488-432-0.

Huiyan Sang and Jianhua Z. Huang. A full scale approximation of covariance functions for large spatial data sets. *Journal of Royal Statistical Society*, 71:111–132, 2012.

Katya Scheinberg, Shiqian Ma, and Donald Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2101–2109. Curran Associates, Inc., 2010. URL `http://papers.nips.cc/paper/4099-sparse-inverse-covariance-selection-via-alternating-linearization-methods.pdf`.

Daniel Simpson, Finn Lindgren, and Havard Rue. Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics*, 1(0):16 – 29, 2012.

Edward Snelson and Zoubin Ghahramani. Local and global sparse Gaussian process approximations. In *International Conference on Artificial Intelligence and Statistics*, pages 524–531, 2007.

Michael L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series is Statistics. Springer, 1999. ISBN 0-387-98629-4.

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

Sam Davanloo Tajbakhsh, Necdet Serhat Aybat, and Enrique Del Castillo. Generalized sparse precision matrix selection for fitting multivariate gaussian random fields to large data sets. *Statistica Sinica*, pages 941–962, 2018.

A. V. Vecchia. Estimation and model identification for continuous spatial processes. *journal of Royal Statistical Society*, 50(2):297–312, 1988.

J. J. Warnes and B. D. Ripley. Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, 74(3):640–642, 1987.

Joe Whittaker. *Graphical Models in Applied Multivariate Statistics*. John Wiley, 2009. ISBN 978-0470743669.

Xiaoming Yuan. Alternating direction method for covariance selection models. *Journal of Scientific Computing*, 51(2):261–273, 2012. ISSN 0885-7474. doi: 10.1007/s10915-011-9507-1.

Supplementary Material for the paper titled "On the Theoretical Guarantees for Parameter Estimation of Gaussian Random Field Models: A Sparse Precision Matrix Approach"

An ADMM Algorithm for Solving Stage-I Problem (11)

Theorem 12 *Let $0 \leq a^* \leq b^* \leq \infty$. Given arbitrary $Z_0, W_0 \in \mathbb{S}^n$ and $\rho > 0$, let $\rho_\ell = \rho$ for $\ell \geq 0$, and $\{P_\ell, Z_\ell\}_{\ell \geq 1}$ denote the iterate sequence generated by $\boldsymbol{ADMM}(S, G, \alpha, a^*, b^*)$ as shown in Figure 11. Then $\{P_\ell\}$ converges Q-linearly[7] to $\hat{P}$, and $\{Z_\ell\}$ converges R-linearly to $\hat{P}$, where $\hat{P}$ is the unique optimal solution to STAGE-I problem given in (11).*

Algorithm ADMM (S, G, α, a^*, b^*)

1: **input** : $Z_0, W_0 \in \mathbb{S}^n$, $\{\rho_\ell\}_{\ell \geq 0} \subset \mathbb{R}_{++}$, $0 \leq a^* \leq b^* \leq \infty$
2: **if** $a^* > 0$ **and** $b^* < \infty$ **then** $a \leftarrow a^*$, $b \leftarrow b^*$
3: **if** $a^* = 0$ **and** $b^* < \infty$ **then** $a \leftarrow \min\{b^*, \frac{1}{\|S\|_2 + \alpha\|G\|_F}\}$, $b \leftarrow b^*$
4: **if** $a^* > 0$ **and** $b^* = \infty$ **then** $a \leftarrow a^*$, $b \leftarrow \frac{na^*}{\alpha G_{\min}} \max\{\|S\|_2 + \alpha\|G\|_F, \ 1/a^*\}$
5: **if** $a^* = 0$ **and** $b^* = \infty$ **then** $a \leftarrow (\|S\|_2 + \alpha\|G\|_F)^{-1}$, $b \leftarrow n/(\alpha G_{\min})$
6: **while** $\ell \geq 0$ **do**
7:     $P_{\ell+1} \leftarrow \operatorname{argmin}_{P \in \mathbb{S}^n}\{\langle S, P \rangle - \log \det(P) + \frac{\rho_\ell}{2}\|P - Z_\ell + \frac{1}{\rho_\ell}W_\ell\|_F^2 : \ a\mathbf{I} \preceq P \preceq b\mathbf{I}\}$
8:     $Z_{\ell+1} \leftarrow \operatorname{argmin}_{Z \in \mathbb{S}^n}\{\alpha \langle G, |Z| \rangle + \frac{\rho_\ell}{2}\|Z - P_{\ell+1} - \frac{1}{\rho_\ell}W_\ell\|_F^2 : \ \operatorname{diag}(Z) \geq \mathbf{0}\}$
9:     $W_{\ell+1} \leftarrow W_\ell + \rho_\ell(P_{\ell+1} - Z_{\ell+1})$
10: **end while**

Figure 11: ADMM algorithm for STAGE-I

Remark. As shown in the proof of Theorem 12, when $a^* = 0$ and/or $b^* = \infty$, the choice of a, b in Figure 11 satisfies $a \leq \sigma_{\min}(\hat{P}) \leq \sigma_{\max}(\hat{P}) \leq b$ for $\hat{P}$, defined in (11). This technical condition makes sure that the ADMM iterate sequence converges linearly.

The algorithm is terminated at the end of iteration ℓ when both primal and dual residuals (r_ℓ, s_ℓ) are below a given tolerance value, where $r_\ell \triangleq P_{\ell+1} - Z_{\ell+1}$ and $s_\ell \triangleq \rho_\ell(Z_{\ell+1} - Z_\ell)$. From the necessary and sufficient optimality conditions for Step 7 and Step 8 in Figure 11, $r_\ell = s_\ell = 0$ implies $P_{\ell+1} = Z_{\ell+1} = \hat{P}$, i.e., the unique optimal solution to (11). In practice, ADMM converges to an acceptable accuracy within a few tens of iterations, which was also the case in our numerical experiments.

Typically, in ADMM algorithms (Boyd et al., 2011), the penalty parameter is held constant, i.e., $\rho_\ell = \rho > 0$ for all $\ell \geq 1$, for some $\rho > 0$. Although the convergence is guaranteed for all $\rho > 0$, the empirical performance critically depends on the choice of ρ—it deteriorates rapidly if the penalty is set too large or too small (Kontogiorgis and Meyer, 1998). Moreover, Lions and Mercier (1979) discuss that there exists a $\rho^* > 0$ which optimizes the convergence rate bounds for the constant penalty ADMM scheme; however, estimating ρ^* is difficult in practice. In our experiments, we used an increasing penalty sequence $\{\rho_\ell\}_{\ell \geq 1}$. For details on the convergence of variable penalty ADMM, see (He et al., 2002; Aybat and Iyengar, 2015) in addition to the references above.

---

7. Let $\{X_\ell\}$ converge to X^* for a given norm $\|.\|$. The convergence is called Q-linear if $\frac{\|X_{\ell+1} - X^*\|}{\|X_\ell - X^*\|} \leq c$, for some $c \in (0, 1)$; and R-linear if $\|X_\ell - X^*\| \leq c_\ell$, for some $\{c_\ell\}$ converging to 0 Q-linearly.

Next, we show that Steps 7 and 8 of ADMM, displayed in Figure 11, can be computed efficiently. Given a convex function $f : \mathbb{S}^n \to \mathbb{R} \cup \{+\infty\}$ and $\lambda > 0$, the proximal mapping $\mathbf{prox}_{\lambda f} : \mathbb{S}^n \to \mathbb{S}^n$ is defined as $\mathbf{prox}_{\lambda f}(\bar{P}) \triangleq \operatorname{argmin}_{P \in \mathbb{S}^n} \lambda f(P) + \frac{1}{2}\|P - \bar{P}\|_F^2$; and given a set $\mathcal{Q} \subset \mathbb{S}^n$, let $\mathbf{1}_{\mathcal{Q}}(\cdot)$ denote the indicator function of $\mathcal{Q}$, i.e., $\mathbf{1}_{\mathcal{Q}}(P) = 0$ for $P \in \mathcal{Q}$; otherwise equal to $+\infty$. For the proof of Lemma 13, see (Yuan, 2012). The result of Lemma 14 for off-diagonal indices follows from the typical soft-thresholding from the lasso solution, see Friedman et al. (2007).

Lemma 13 *Let $\Psi(P) \triangleq \langle S, P \rangle - \log \det(P) + \mathbf{1}_{\mathcal{Q}}(P)$, and $\mathcal{Q} \triangleq \{P \in \mathbb{S}^n : a\mathbf{I} \preceq P \preceq b\mathbf{I}\}$. In generic form, Step 7 of ADMM can be written as $\mathbf{prox}_{\Psi/\rho}(\bar{P})$ for some $\bar{P} \in \mathbb{S}^n$ and $\rho > 0$. Suppose $\bar{P} - \frac{1}{\rho}S$ has eigen-decomposition $U \operatorname{diag}(\bar{\lambda})U^\top$. Then $\mathbf{prox}_{\Psi/\rho}(\bar{P}) = U \operatorname{diag}(\lambda^*)U^\top$, where*

$$\lambda_i^* = \max\left\{ \min\left\{ \frac{\bar{\lambda}_i + \sqrt{\bar{\lambda}_i^2 + 4/\rho}}{2},\ b \right\},\ a \right\}, \quad i = 1, \ldots, n. \tag{51}$$

Lemma 14 *Let $\Phi(P) \triangleq \alpha \langle G, |P| \rangle + \mathbf{1}_{\mathcal{Q}'}(P)$, and $\mathcal{Q}' \triangleq \{P \in \mathbb{S}^n : \operatorname{diag}(P) \geq \mathbf{0}\}$. In generic form, Step 8 of ADMM can be written as $\mathbf{prox}_{\Phi/\rho}(\bar{P})$ for some $\bar{P} \in \mathbb{S}^n$ and $\rho > 0$, which can be computed as follows:*

$$(\mathbf{prox}_{\Phi/\rho}(\bar{P}))_{ij} = \operatorname{sgn}\left(\bar{P}_{ij}\right)\max\left\{|\bar{P}_{ij}| - \tfrac{\alpha}{\rho}G_{ij}, 0\right\}, \quad \forall(ij) \in \mathcal{I} \times \mathcal{I} \text{ s.t. } i \neq j, \tag{52a}$$

$$(\mathbf{prox}_{\Phi/\rho}(\bar{P}))_{ii} = \max\left\{\bar{P}_{ii} - \tfrac{\alpha}{\rho}G_{ii}, 0\right\}, \quad \forall i \in \mathcal{I}. \tag{52b}$$

The proofs of Theorem 12, Lemma 13 and Lemma 14 follow from the existing results in the literature. For the sake of completeness, these proofs are provided in the supplementary material.

PROOF OF THEOREM 12

Consider a more generic problem of the following form:

$$\min_{P \in \mathbb{S}^n} \Psi(P) + \Phi(P), \tag{53}$$

where $\Psi : \mathbb{S}^n \to \mathbb{R} \cup \{+\infty\}$ and $\Phi : \mathbb{S}^n \to \mathbb{R} \cup \{+\infty\}$ are proper closed convex functions, and $\mathbb{S}^n$ denotes the vector space of n-by-n symmetric matrices. By introducing an auxiliary variable $Z \in \mathbb{S}^n$, (53) can be equivalently written as $\min\{\Psi(P) + \Phi(Z) : P = Z, P, Z \in \mathbb{S}^n\}$. For a given penalty parameter $\rho > 0$, the augmented Lagrangian function is defined as

$$\mathcal{L}_\rho(P, Z, W) \triangleq \Psi(P) + \Phi(Z) + \langle W, P - Z \rangle + \tfrac{\rho}{2}\|P - Z\|_F^2, \tag{54}$$

where $W \in \mathbb{S}^n$ is the dual multiplier for the linear constraint $P - Z = \mathbf{0}$. Given an initial primal-dual point $Z^1, W^1 \in \mathbb{S}^n$, when the ADMM algorithm (Boyd et al., 2011) is implemented on (53), it generates a sequence of iterates $\{P_\ell, Z_\ell\}_{\ell \geq 1}$ according to:

$$P_{\ell+1} = \operatorname*{argmin}_{P \in \mathbb{S}^n} \mathcal{L}_\rho(P, Z_\ell, W_\ell) \quad = \mathbf{prox}_{\Psi/\rho}\left(Z_\ell - \tfrac{1}{\rho}W_\ell\right), \tag{55a}$$

$$Z_{\ell+1} = \operatorname*{argmin}_{Z \in \mathbb{S}^n} \mathcal{L}_\rho(P_{\ell+1}, Z, W_\ell) = \mathbf{prox}_{\Phi/\rho}\left(P_{\ell+1} + \tfrac{1}{\rho}W_\ell\right), \tag{55b}$$

$$W_{\ell+1} = W_\ell + \rho\left(P_{\ell+1} - Z_{\ell+1}\right). \tag{55c}$$

For all $\rho > 0$, convergence of the ADMM iterate sequence $\{P_\ell, Z_\ell\}_{\ell \geq 1}$ is guaranteed. In particular, $\lim_{\ell \geq 1} Z_\ell = \lim_{\ell \geq 1} P_\ell$; moreover, any limit point of $\{P_\ell\}$ is a minimizer of (53).

Recently, Deng and Yin (2016) showed that the ADMM iterate sequence converges linearly if Ψ is strongly convex and has a Lipschitz continuous gradient. In particular, $\{P_\ell, W_\ell\}$ converges[8] Q-linearly to a primal-dual optimal pair $(P^{\mathrm{opt}}, W^{\mathrm{opt}})$, where P^{opt} is the unique primal optimal solution, and $\{Z_\ell\}$ converges R-linearly to P^{opt}.

Returning to the SPS method, note that the precision matrix estimation problem in (11) immediately fits into the ADMM framework by setting $\Psi(P) = \langle S, P \rangle - \log \det(P) + \mathbf{1}_{\mathcal{Q}}(P)$ and $\Phi(P) = \alpha \langle G, |P| \rangle$, where $\mathcal{Q} \triangleq \{P \in \mathbb{S}^n : a^* \mathbf{I} \preceq P \preceq b^* \mathbf{I}\}$ and $\mathbf{1}_{\mathcal{Q}}(\cdot)$ is the indicator function of $\mathcal{Q}$, i.e., $\mathbf{1}_{\mathcal{Q}}(P) = 0$ if $P \in \mathcal{Q}$; and it is equal to $+\infty$, otherwise. Therefore, both Ψ and Φ are closed convex functions. When $0 < a^* \leq b^* < +\infty$, Theorem 12 immediately follows from the convergence properties of ADMM discussed above.

Now consider the case $a^* = 0$ and $b^* = +\infty$. For this scenario, Ψ is strictly convex and differentiable on $\mathcal{Q}$ with $\nabla \Psi(P) = S - P^{-1}$; however, note that $\nabla \Psi(P)$ is not Lipschitz continuous on $\mathcal{Q}$. Therefore, this choice of Ψ and Φ do not satisfy the assumptions in (Deng and Yin, 2016). On the other hand, following the discussion in (d'Aspremont et al., 2008), we will show that by selecting a slightly different $\mathcal{Q}$, one can obtain an equivalent problem to (11) which does satisfy the ADMM convergence assumptions in (Deng and Yin, 2016); hence, linear convergence rate for stage I of the SPS method can be obtained. Noting that $|t| = \max\{ut : |u| \leq 1\}$, one can write (11) equivalently as follows:

$$\hat{F} \triangleq \min_{P \succ 0} \quad \max_{\{U \in \mathbb{S}^n : |U_{ij}| \leq \alpha G_{ij}\}} \mathcal{L}(P, U) \triangleq \langle S + U, P \rangle - \log \det(P), \tag{56}$$

where $\hat{F} \triangleq \left\langle S, \hat{P} \right\rangle - \log \det(\hat{P}) + \alpha \left\langle G, |\hat{P}| \right\rangle$, and $\hat{P}$ is the solution to (11), i.e., $\hat{P} = \mathrm{argmin}\{\langle S, P \rangle - \log \det(P) + \alpha \langle G, |P| \rangle : P \succ \mathbf{0}\}$. Since $\mathcal{L}$ is convex in P, linear in U, and $\{U \in \mathbb{S}^n : |U_{ij}| \leq \alpha G_{ij}\}$ is compact, the strong min-max property holds:

$$\hat{F} = \max_{\{U \in \mathbb{S}^n : |U_{ij}| \leq \alpha G_{ij}\}} \quad \min_{P \succ 0} \mathcal{L}(P, U) = \max_{\{U \in \mathbb{S}^n : |U_{ij}| \leq \alpha G_{ij}\}} n - \log \det \left((S + U)^{-1}\right), \tag{57}$$

where (57) follows from the fact that for a given $U \in \mathbb{S}^n$, $\hat{P}(U) = (S + U)^{-1}$ minimizes the inner problem if $S + U \succ \mathbf{0}$; otherwise, the inner minimization problem is unbounded from below. Therefore, we conclude that $\hat{P}$ is the optimal solution to (11) if and only if there exists $\hat{U} \in \mathbb{S}^n$ such that $\hat{P} = (S + \hat{U})^{-1} \succ \mathbf{0}$, $|\hat{U}_{ij}| \leq \alpha G_{ij}$ for all $(i, j) \in \mathcal{I}$, and $\left\langle S, \hat{P} \right\rangle + \alpha \left\langle G, |\hat{P}| \right\rangle = n$. Since $S, \hat{P} \succeq \mathbf{0}$, we have $\left\langle S, \hat{P} \right\rangle \geq 0$; hence, $\left\langle G, |\hat{P}| \right\rangle \leq n/\alpha$. Hence, we can derive the desired bounds, similar to those derived in (d'Aspremont et al., 2008):

$$a \triangleq \frac{1}{\|S\|_2 + \alpha \|G\|_F} \leq \frac{1}{\|S\|_2 + \|\hat{U}\|_F} \leq \frac{1}{\|S + \hat{U}\|_2} = \sigma_{\min}(\hat{P}), \tag{58}$$

$$b \triangleq \frac{n}{\alpha\, G_{\min}} \geq \frac{\left\langle G, |\hat{P}| \right\rangle}{G_{\min}} \geq \sum_{i,j} |\hat{P}_{ij}| \geq \|\hat{P}\|_F \geq \|\hat{P}\|_2 = \sigma_{\max}(\hat{P}), \tag{59}$$

where $G_{\min} \triangleq \min\{G_{ij} : (i, j) \in \mathcal{I} \times \mathcal{I}, i \neq j\} > 0$. Therefore, (11) is equivalent to

$$\hat{P} = \mathrm{argmin}\{\langle S, P \rangle - \log \det(P) + \alpha \langle G, |P| \rangle : a\mathbf{I} \preceq P \preceq b\mathbf{I}\}, \tag{60}$$

---

8. Q-linear and R-linear convergence were defined in Section A.

for a and b defined in (58) and (59), respectively. Going back to the convergence rate discussion, when ADMM is applied to (60) we can guarantee that the primal-dual iterate sequence converges linearly. In particular, we apply ADMM on (53) with

$$\Psi(P) = \langle S, P \rangle - \log \det(P) + \mathbf{1}_{\tilde{\mathcal{Q}}}(P), \quad \tilde{\mathcal{Q}} \triangleq \{P \in \mathbb{S}^n : a\mathbf{I} \preceq P \preceq b\mathbf{I}\}, \tag{61}$$

$$\Phi(P) = \alpha \langle G, |P| \rangle + \mathbf{1}_{\mathcal{Q}'}(P), \quad \mathcal{Q}' \triangleq \{P \in \mathbb{S}^n : \mathrm{diag}(P) \geq \mathbf{0}\}. \tag{62}$$

Since $\tilde{\mathcal{Q}} \subset \mathbb{S}_+^n \subset \mathcal{Q}'$, the term $\mathbf{1}_{\mathcal{Q}'}(.)$ in the definition of Φ appears redundant. However, defining Φ this way will restrict the sequence $\{Z_\ell\}$ to lie in $\mathcal{Q}'$ rather than in $\mathbb{S}^n$, which leads to faster convergence to feasibility in practice. By resetting $\mathcal{Q}$ to $\tilde{\mathcal{Q}}$ as in (61), we ensure that Ψ is strongly convex with constant $1/b^2$ and $\nabla \Psi$ is Lipschitz continuous with constant $1/a^2$. Indeed, the Hessian of Ψ is a quadratic form on $\mathbb{S}^n$ such that $\nabla^2 \Psi(P)[H, H] = \mathbf{Tr}(P^{-1}HP^{-1}H)$, which implies $\frac{1}{b^2}\|H\|_F^2 \leq \nabla^2 \Psi(P)[H, H] \leq \frac{1}{a^2}\|H\|_F^2$.

The values of $a > 0$ and $b < +\infty$ in the definition of $\tilde{\mathcal{Q}} \triangleq \{P \in \mathbb{S}^n : a\mathbf{I} \preceq P \preceq b\mathbf{I}\}$ for the other cases, i.e., $(a^* = 0, b^* < +\infty)$ and $(a^* > 0, b^* = +\infty)$ are given in Figure 11; these bounds can also be proven very similarly; thus, their proofs are omitted.

PROOF OF LEMMA 13

The $\mathbf{prox}_{\Psi/\rho}$ map can be equivalently written as

$$\mathbf{prox}_{\Psi/\rho}(\bar{P}) = \underset{P \in \mathbb{S}^n}{\mathrm{argmin}}\{-\log \det(P) + \tfrac{\rho}{2}\|P - (\bar{P} - \tfrac{1}{\rho}S)\|_F^2 : a\mathbf{I} \preceq P \preceq b\mathbf{I}\}. \tag{63}$$

Let $U \mathrm{diag}(\bar{\lambda})U^\top$ be the eigen-decomposition of $\bar{P} - \frac{1}{\rho}S$. Fixing $U \in \mathbb{S}^n$, and by restricting the variable $P \in \mathbb{S}^n$ in (63) to have the form $U \mathrm{diag}(\lambda)U^\top$ for some $\lambda \in \mathbb{R}^n$, we obtain the optimization problem (64) over the variable $\lambda \in \mathbb{R}^n$:

$$\min_{\lambda \in \mathbb{R}^n} \left\{ -\sum_{i=1}^n \log(\lambda_i) + \tfrac{\rho}{2}(\lambda_i - \bar{\lambda}_i)^2 : a \leq \lambda_i \leq b, \ i = 1, \ldots, n \right\}. \tag{64}$$

For a given $\bar{t} \in \mathbb{R}$, and $a, b, \gamma > 0$, the unique minimizer of $\min_{t \in \mathbb{R}}\{-\log(t) + \frac{\rho}{2}|t - \bar{t}|^2 : a \leq t \leq b\}$ can be written as $\max\left\{\min\left\{\frac{\bar{t} + \sqrt{\bar{t}^2 + 4/\rho}}{2}, b\right\}, a\right\}$. Hence, $\lambda^* \in \mathbb{R}^n$ given in (51) is the unique minimizer of (64). Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be defined as $h(\lambda) \triangleq -\sum_{i=1}^n \log(\lambda_i) + \mathbf{1}_{\mathcal{H}}(\lambda)$, where $\mathcal{H} \triangleq \{\lambda \in \mathbb{R}^n : a\mathbf{1} \leq \lambda \leq b\mathbf{1}\}$ and $\lambda^* = \mathrm{argmin}_{\lambda \in \mathbb{R}^n}\{h(\lambda) + \frac{\rho}{2}\|\lambda - \bar{\lambda}\|_2^2\}$. From the first-order optimality conditions, it follows that $\bar{\lambda} - \lambda^* \in \frac{1}{\rho} \partial h(\lambda)|_{\lambda = \lambda^*}$.

Let $H : \mathbb{S}^n \to \mathbb{R} \cup \{+\infty\}$ be such that $H(P) = -\log \det(P) + \mathbf{1}_{\mathcal{Q}}(P)$. Definition of $\mathbf{prox}_{\Psi/\rho}(\bar{P})$ implies that $\left(\bar{P} - \frac{1}{\rho}S - \mathbf{prox}_{\Psi/\rho}(\bar{P})\right) \in \frac{1}{\rho}\partial H(P)|_{P = \mathbf{prox}_{\Psi/\rho}(\bar{P})}$. In the rest of the proof, $\sigma : \mathbb{S}^n \to \mathbb{R}^n$ denotes the function that returns the singular values of its argument. Note that $H(P) = h(\sigma(P))$ for all $P \in \mathbb{S}^n$. Since h is *absolutely symmetric*, Corollary 2.5 in Lewis (1995) implies that $P^{\mathrm{prox}} = \mathbf{prox}_{\Psi/\rho}(\bar{P})$ if and only if $\sigma(\bar{P} - \frac{1}{\rho}S - P^{\mathrm{prox}}) \in \frac{1}{\rho}\partial h(\lambda)|_{\lambda = \sigma(P^{\mathrm{prox}})}$ and there exists a simultaneous singular value decomposition of the form $P^{\mathrm{prox}} = U \mathrm{diag}(\sigma(P^{\mathrm{prox}}))U^\top$ and $\bar{P} - \frac{1}{\rho}S - P^{\mathrm{prox}} = U \mathrm{diag}\left(\sigma\left(\bar{P} - \frac{1}{\rho}S_k - P^{\mathrm{prox}}\right)\right)U^\top$. Hence, $\mathbf{prox}_{\Psi/\rho}(\bar{P}) = U \mathrm{diag}(\lambda^*)U^\top$ follows from $\bar{\lambda} - \lambda^* \in \frac{1}{\rho}\partial h(\lambda)|_{\lambda = \lambda^*}$.

Proof of Lemma 14

From the definition of $\mathbf{prox}_{\Phi/\rho}$, we have
$$\mathbf{prox}_{\Phi/\rho} = \operatorname*{argmin}_{P \in \mathbb{S}^n} \Big\{ \sum_{(i,j) \in \mathcal{I} \times \mathcal{I}} \tfrac{\alpha}{\rho} G_{ij} |P_{ij}| + \tfrac{1}{2} |P_{ij} - \bar{P}_{ij}|^2 \; : \; \operatorname{diag}(P) \geq \mathbf{0} \Big\}. \tag{65}$$

For a given $\bar{t} \in \mathbb{R}$, and $\gamma > 0$, the unique minimizer of $\min_{t \in \mathbb{R}} \gamma |t| + \tfrac{1}{2} |t - \bar{t}|^2$ can be written as $\operatorname{sgn}(\bar{t}) \max\{|\bar{t}| - \gamma, \; 0\}$; and the minimizer of $\min_{t \in \mathbb{R}} \{ \gamma t + \tfrac{1}{2} |t - \bar{t}|^2 \; : \; t \geq 0 \}$ can be written as $\max\{\bar{t} - \gamma, \; 0\}$. Hence, (52) follows from the separability of the objective in (65).

Additional Numerical Results

In this section, the importance of STAGE-I in the SPS algorithm is shown numerically. Notice that the STAGE-II can be directly implemented for the sample covariance matrix S. This could be interpreted as direct estimation of the covariance function parameters by fitting the covariogram (Cressie, 1993). For this purpose, the covariance function parameters are estimated from N realizations of a zero-mean GRF simulated over $n = 100$ randomly selected locations over a square domain $\mathcal{X} = [0, \beta] \times [0, \beta]$ with a Matern covariance function with smoothness parameter $3/2$ and the parameter vector $\boldsymbol{\theta}^* = [\boldsymbol{\theta}_\rho^{*\top}, \theta_v^*, \theta_0^*]^\top = [15, 8, 1]^\top$. The SPS and covariogram methods are then compared based on $R = 100$ simulation replications (every time $n = 100$ locations are randomly resampled). Table 7 shows the mean and standard error of the parameter estimates, respectively. With increasing N, we see faster convergence of the SPS parameter estimates to their true values. Compared to the covariogram method, the SPS mean parameter estimates are almost always closer to their true parameter values and their standard errors are lower. The importance of STAGE-I in the SPS algorithm is more evident when $N \ll N_0$. STAGE-I zooms into the region in the parameter space of the nonconvex objective function where the global minimum lies, and this results in better covariance parameter estimates. As expected, increasing the domain size (β) results in a lower point density in the domain, and this deteriorates the performance of both methods.

Table 7: Mean (standard deviation) of parameter estimates from $R = 100$ replications when the process is sampled N times at each of $n = 100$ randomly chosen locations in the domain $\mathcal{X} = [0, \beta] \times [0, \beta]$

$\beta \backslash N$	SPS 1	SPS 5	SPS 20	SPS 40	Covariogram 1	Covariogram 5	Covariogram 20	Covariogram 40
25	10.22 (0.08)	12.81 (0.71)	13.95 (0.34)	14.81 (0.25)	4.14 (1.15)	14.03 (0.73)	14.15 (0.33)	14.63 (0.25)
	8.03 (0.27)	8.10 (0.31)	7.87 (0.16)	8.13 (0.13)	4.01 (0.74)	7.22 (0.34)	7.56 (0.16)	7.99 (0.13)
	0.83 (0.07)	0.92 (0.08)	0.91 (0.05)	0.92 (0.03)	1.05 (0.09)	1.62 (0.07)	1.38 (0.04)	1.09 (0.03)
50	11.84 (0.14)	14.30 (0.71)	14.64 (0.37)	15.19 (0.25)	7.01 (2.04)	14.39 (0.83)	15.63 (0.39)	15.17 (0.26)
	7.93 (0.31)	7.93 (0.21)	8.05 (0.13)	8.02 (0.09)	6.07 (0.45)	7.57 (0.23)	7.92 (0.13)	7.94 (0.09)
	1.13 (0.08)	1.14 (0.11)	1.04 (0.07)	1.05 (0.05)	1.02 (0.13)	1.67 (0.10)	1.54 (0.07)	1.22 (0.05)
75	13.73 (2.14)	15.28 (1.06)	15.37 (0.34)	15.11 (0.21)	21.76 (2.50)	17.65 (1.19)	15.63 (0.35)	15.22 (0.21)
	7.39 (0.41)	7.72 (0.20)	7.88 (0.10)	7.96 (0.07)	7.77 (0.42)	7.69 (0.20)	7.97 (0.10)	8.04 (0.07)
	1.27 (0.14)	1.64 (0.12)	1.64 (0.08)	1.20 (0.05)	1.37 (0.16)	1.29 (0.13)	1.07 (0.08)	0.98 (0.05)
100	19.82 (3.36)	14.87 (0.45)	15.62 (0.27)	15.12 (0.14)	31.56 (3.70)	16.57 (0.58)	15.91 (0.29)	15.11 (0.15)
	7.18 (0.32)	8.21 (0.17)	7.77 (0.09)	7.90 (0.06)	6.88 (0.34)	8.09 (0.17)	7.87 (0.09)	8.01 (0.06)
	1.38 (0.18)	1.57 (0.11)	1.76 (0.07)	1.30 (0.04)	1.91 (0.21)	1.22 (0.11)	1.14 (0.07)	1.01 (0.04)
125	25.28 (4.23)	15.27 (0.76)	15.74 (0.27)	15.03 (0.15)	38.68 (4.50)	16.63 (0.67)	15.88 (0.26)	15.01 (0.15)
	7.70 (0.38)	7.97 (0.18)	7.67 (0.10)	7.95 (0.07)	7.77 (0.38)	7.88 (0.18)	7.79 (0.10)	8.09 (0.07)
	1.90 (0.24)	1.73 (0.13)	1.90 (0.08)	1.28 (0.05)	1.77 (0.24)	1.25 (0.12)	1.25 (0.08)	0.95 (0.05)
150	41.45 (5.74)	17.18 (0.73)	15.26 (0.17)	15.49 (0.14)	46.96 (7.52)	15.82 (0.54)	15.15 (0.16)	15.57 (0.14)
	6.78 (0.35)	7.75 (0.19)	7.85 (0.08)	7.87 (0.06)	6.77 (0.39)	7.83 (0.19)	7.71 (0.08)	7.71 (0.06)
	2.42 (0.27)	1.39 (0.13)	1.09 (0.06)	1.13 (0.06)	2.57 (0.30)	1.91 (0.13)	1.77 (0.06)	1.50 (0.05)
175	47.10 (7.40)	16.32 (0.70)	15.22 (0.16)	14.85 (0.10)	79.66 (10.52)	15.52 (0.67)	15.16 (0.16)	14.92 (0.10)
	6.86 (0.38)	7.76 (0.17)	8.07 (0.08)	8.06 (0.06)	5.87 (0.46)	7.80 (0.17)	7.93 (0.8)	7.88 (0.07)
	2.36 (0.27)	1.28 (0.13)	1.01 (0.06)	0.95 (0.05)	3.67 (0.35)	1.86 (0.13)	1.69 (0.07)	1.35 (0.05)
200	64.18 (9.77)	17.69 (2.01)	15.10 (0.15)	15.02 (0.10)	86.13 (11.73)	17.68 (2.70)	15.04 (0.14)	15.11 (0.10)
	6.12 (0.38)	7.76 (0.17)	7.90 (0.08)	7.99 (0.06)	5.50 (0.42)	7.82 (0.18)	7.75 (0.08)	7.80 (0.06)
	3.01 (0.29)	1.28 (0.13)	1.05 (0.06)	1.00 (0.05)	4.07 (0.35)	1.88 (0.14)	1.75 (0.06)	1.42 (0.05)
250	101.44 (13.90)	22.33 (4.78)	14.97 (0.14)	14.92 (0.09)	144.40 (16.46)	18.42 (3.42)	14.88 (0.14)	14.99 (0.09)
	5.46 (0.40)	7.99 (0.21)	8.10 (0.09)	8.17 (0.06)	4.73 (0.43)	8.17 (0.20)	7.97 (0.09)	7.97 (0.06)
	3.47 (0.33)	1.09 (0.15)	0.95 (0.07)	0.89 (0.05)	4.71 (0.37)	1.56 (0.16)	1.63 (0.07)	1.33 (0.05)
300	172.48 (9.07)	35.91 (8.96)	15.14 (0.14)	15.15 (0.12)	266.24 (19.87)	39.57 (9.77)	15.11 (0.15)	15.22 (0.12)
	4.50 (0.44)	7.51 (0.22)	8.01 (0.08)	7.97 (0.07)	2.92 (0.41)	7.52 (0.25)	7.86 (0.09)	7.77 (0.07)
	4.78 (0.39)	1.47 (0.21)	1.00 (0.07)	1.04 (0.06)	6.89 (0.38)	2.13 (0.24)	1.69 (0.08)	1.49 (0.06)
400	319.69 (22.19)	109.62 (20.91)	26.19 (5.51)	15.03 (0.12)	451.89 (26.09)	119.02 (20.11)	20.55 (7.75)	15.06 (0.11)
	2.62 (0.30)	6.29 (0.36)	7.87 (0.14)	8.01 (0.08)	1.36 (0.36)	6.23 (0.34)	7.86 (0.15)	7.83 (0.07)
	6.18 (0.28)	2.76 (0.35)	1.18 (0.12)	1.03 (0.07)	8.09 (0.32)	3.51 (0.32)	1.74 (0.14)	1.47 (0.06)
500	483.79 (19.57)	227.05 (33.13)	36.48 (13.62)	15.05 (0.12)	645.88 (30.76)	286.89 (31.08)	43.31 (11.85)	15.06 (0.12)
	1.81 (0.24)	5.15 (0.42)	7.48 (0.20)	8.02 (0.08)	0.66 (0.32)	4.71 (0.39)	7.33 (0.18)	7.85 (0.08)
	7.19 (0.25)	4.03 (0.41)	1.49 (0.19)	0.97 (0.08)	9.00 (0.32)	5.15 (0.39)	2.20 (0.17)	1.41 (0.07)