Generalized Sparse Precision Matrix Selection for Fitting Multivariate Gaussian Random Fields to Large Data Sets

S. Davanloo Tajbakhsh¹, N.S. Aybat², and E. del Castillo²

¹Virginia Tech University and ²Pennsylvania State University

Abstract: This paper generalizes the Sparse Precision matrix Selection (SPS) algorithm, proposed by Davanloo et al. (2015) for estimating scalar Gaussian Random Field (GRF) models, to the multivariate, second-order stationary case under a separable covariance function. Theoretical convergence rates for the estimated covariance matrix and for the estimated parameters of the correlation function are established. Numerical simulation results validate our theoretical findings. Data segmentation is used to handle large data sets.

Key words and phrases: Multivariate Gaussian Processes, Gaussian Markov Random Fields, Spatial Statistics, Covariance Selection, Convex Optimization, ADMM.

1. Introduction

Gaussian Random Field (GRF) models have been used extensively for modeling the output of computer experiments, either to approximate complex computer codes (for an early review in this field, see Santner et al. (2003)) or to build a “metamodel” of a stochastic simulation (see, e.g., Ankenman

This research is based on the dissertation of the first author, conducted under the guidance of the second and the third authors, Drs. Aybat and del Castillo.
et al. (2010) and for a review, Kleijnen (2010)). In a GRF model, a key role is played by the covariance function which determines how the covariance between the process values at two locations changes as the locations change across the process domain. There are many valid parametric covariance functions, e.g., Exponential, Squared Exponential, and Matern; and Maximum Likelihood (ML) is the dominant method to estimate their parameters from computer experiments data (Santner et al. (2003)). However, the ML fitting procedure suffers from two main challenges: i) the negative loglikelihood is a nonconvex function of the covariance matrix; therefore, the covariance parameters may be poorly estimated, ii) the problem is computationally hard when the number of spatial locations $n$ is big. This is known as the big “n” problem in the literature.

In Davanloo et al. (2015), we proposed the Sparse Precision Selection (SPS) algorithm to deal with the first challenge with theoretical guarantees; and discussed a segmentation scheme on the training data to be able to solve big $n$ problems. A wide variety of computer experiments, including differential equation-based fluid dynamics and climate models, require approximation of a vector of responses obtained at each spatial or spatial-temporal location. In such instances, rather than considering each response independently, using the between-response covariance can significantly en-
hance the prediction performance. However, fitting multivariate GRFs not only suffers from the two challenges mentioned above; the parametrization of the matrix-valued covariance functions requires a higher-dimensional parameter vector which aggravates the difficulty of the GRF estimation problem further. The goal of this paper is to extend the theory of the SPS method (Davanloo et al. (2015)) to include the estimation of multivariate GRF models. The paper is organized as follows: Section 1.1 introduces the notation, and Section 2 provides some preliminary concepts related to the SPS method. In Section 3, the multivariate generalization of the SPS method and its theoretical guarantees are discussed. Section 4 includes numerical results. Finally, we summarize the main results in the paper and provide some future research directions in Section 5.

1.1. Notation. Throughout the paper, given $x \in \mathbb{R}^n$, $\|x\|$, $\|x\|_1$, $\|x\|_\infty$ denote the Euclidean, $\ell_1$, and $\ell_\infty$ norms, respectively. For $x \in \mathbb{R}^n$, $\text{diag}(x) \in S^n$ denotes a diagonal matrix with its diagonal equal to $x$. Given $X \in \mathbb{R}^{m \times n}$, we denote the vectorization of $X$ using $\text{vec}(X) \in \mathbb{R}^{np}$, obtained by stacking the columns of the matrix $X$ on top of one another. Moreover, let $r = \text{rank}(X)$, and $\sigma = [\sigma_i]_{i=1}^r \subset \mathbb{R}_{++}$ (positive orthant) denote the singular values of $X$; then, $\|X\|_F := \|\sigma\|$, $\|X\|_2 := \|\sigma\|_\infty$, and $\|X\|_* := \|\sigma\|_1$ denote the Frobenius, spectral, and nuclear norms of $X$, respectively. Given
$X, Y \in \mathbb{R}^{m \times n}, \langle X, Y \rangle := \text{Tr}(X^T Y)$ denotes the standard inner product.

Let $\mathcal{V}$ be a normed vector space with norm $\| \cdot \|_a$. For $\bar{x} \in \mathcal{V}$ and $r > 0$, $B_{\| \cdot \|_a}(\bar{x}, r) := \{ x \in \mathcal{V} : \| x - \bar{x} \|_a < r \}$ denotes the open ball centered at $\bar{x}$ with radius $r > 0$, and $\bar{B}_{\| \cdot \|_a}(\bar{x}, r)$ denotes its closure.

2. Preliminaries: the SPS method for a scalar GRF

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $y : \mathcal{X} \to \mathbb{R}$ be a Gaussian Random Field (GRF), where $y(x)$ denotes the value of the process at location $x \in \mathcal{X}$. Let $m(x) = \mathbb{E}(y(x))$ for $x \in \mathcal{X}$, and $c(x, x')$ be covariance function denoting the covariance between $y(x)$ and $y(x')$, i.e., $c(x, x') = \text{cov}(y(x), y(x'))$ for all $x, x' \in \mathcal{X}$.

Without loss of generality, we assume that the GRF has a constant mean equal to zero, i.e., $m(x) = 0$. Suppose the training data $\mathcal{D} = \{(x_i, y^{(r)}_i) : i = 1, \ldots, n, \ r = 1, \ldots, N\}$ contains $N$ realizations of the GRF at each of $n$ distinct locations in $\mathcal{D}^x := \{x_i\}_{i=1}^n \subset \mathcal{X}$. Let $y^{(r)} = [y^{(r)}_i]_{i=1}^n \in \mathbb{R}^n$ denote the vector of $r$-th realization values for locations in $\mathcal{D}^x$.

For simplicity in estimation, the covariance function, $c(x, x')$ is typically assumed to belong to some parametric family $\{c(x, x'; \theta, \nu) : \theta \in \Theta, \nu \geq 0\}$ and $c(x, x', \theta) := \nu \rho(x, x', \theta)$, where $\rho(x, x', \theta)$ is a parametric correlation function where $\theta$ and $\nu$ denote the spatial correlation and variance parameters, respectively, and $\Theta \subset \mathbb{R}^q$ is a set that contains the true spatial correlation parameters – see e.g. Cressie (2015). Let $\theta^*$ and $\nu^*$ denote the unknown
true parameters of the process. Given a set of locations $D^x = \{x_i\}_{i=1}^n$, let $C(\theta, \nu) \in S^n_{++}$ be such that its $(i, j)^{th}$ element is $c(x_i, x_j; \theta, \nu)$ – throughout, $S^n_{++}$ and $S^n_+$ denote the set of $n$-by-$n$ symmetric, positive definite and positive semidefinite matrices, respectively. Let $C^* = C(\theta^*, \nu^*)$ denote the true covariance matrix corresponding to locations in $D^x = \{x_i\}_{i=1}^n$, and $P^* = (C^*)^{-1}$ denote the true precision matrix. In Davanloo et al. (2015), we proposed a two-stage method (SPS) to estimate the unknown process parameters $\theta^*$ and $\nu^*$. As described by Davanloo et al. (2015), the method is partially motivated from two empirical observations. First, the precision (inverse covariance) matrix of a GRF can be interpreted as a compressible signal due to the power-law decay observed when its entries are sorted by their magnitudes; hence, it can be well-approximated by a sparse matrix – compare it with the covariance matrix in Figure 1.

![Figure 1](image_url)

**Figure 1:** Decaying behavior of elements of the Precision and Covariance matrices for GRFs. The largest 1000 off-diagonal elements of the precision and covariance matrices (scaled by their maximums) plotted in descending order. The underlying GRF was evaluated over 100 randomly selected points in $X = \{x \in \mathbb{R}^2 : -50 \leq x \leq 50\}$ for three covariance functions with range and variance parameters equal to 10, and 1, respectively.
Secondly, we also observed that the precision matrix entries decay to zero in absolute value, as shown in Figure 2, as the corresponding pairwise distances between locations increase.

Let $a^*$ and $b^*$ be given constants such that $0 \leq a^* \leq \sigma_{\min}(P^*) \leq \sigma_{\max}(P^*) \leq b^* \leq \infty$. In the first stage of the SPS algorithm, we proposed to solve the following convex loglikelihood problem penalized with a weighted $\ell_1$-norm to estimate the true precision matrix corresponding to the given data locations $\mathcal{D}^x$:

$$
\hat{P} := \arg\min\{\langle S, P \rangle - \log \det(P) + \alpha \langle G, |P| \rangle : a^* I \preceq P \preceq b^* I\}, \quad (2.1)
$$

where $S = \frac{1}{N} \sum_{r=1}^{N} y^{(r)} y^{(r)\top} \in \mathcal{S}_+^n$ is the sample covariance matrix. The weight matrix $G \in \mathcal{S}^n$ is chosen as the matrix of pairwise distances:

$$
G_{ij} = \|x_i - x_j\|, \quad \text{if } i \neq j, \quad G_{ii} = \min\{\|x_i - x_j\| : j \in \mathcal{I} \setminus \{i\}\}, \quad (2.2)
$$

for all $(i, j) \in \mathcal{I} \times \mathcal{I}$, where $\mathcal{I} = \{1, 2, \ldots, n\}$ and $|.|$ is the elementwise absolute value operator. The sparsity structure of the estimated precision
matrix $\hat{P}$ encodes the conditional independence structure of a Gaussian Markov Random Filed (GMRF) approximation to the GRF. Using the Alternating Direction Method of Multipliers (ADMM, Boyd et al. (2011)) algorithm (2.1) can be solved efficiently.

In the second stage of the SPS method, we proposed to solve a least-square problem (2.3) to estimate the unknown parameters $\theta^*$ and $\nu^*$:

$$ (\hat{\theta}, \hat{\nu}) = \arg\min_{\theta \in \Theta, \nu \geq 0} \|C(\theta, \nu) - \hat{P}^{-1}\|_F^2. \quad (2.3) $$

In Davanloo et al. (2015), we showed how to solve each optimization problem, and also established theoretical convergence rate of the SPS estimator.

SPS is therefore based on a Gaussian Markov Random Field (GMRF) approximation to the GRF. A GRF process on a lattice is a GRF under the conditional independence assumption, i.e., when 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 $X$ for a GRF is uncountable; hence, in general GMRF models cannot represent GRFs exactly. Lindgren et al. (2011) recently established that the Matern GRFs are Markovian; in particular, they are Markovian when the smoothing parameter $\nu$ is such that $\nu - d/2 \in \mathbb{Z}_+$, where $d$ is the dimension of the input space – see Lindgren et al. (2011) and Fulgstad et
al. (2015) 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, the first stage of SPS lets the data determine the near-conditional independence pattern between variables through the precision matrix estimated via a weighted $\ell_1$-regularization. Furthermore, this first stage helps to “zoom into” the area where the true covariance parameters are located; hence, it helps not to get trapped in local optimum solutions in stage 2 of the method.

3. Multivariate GRF Models

From now on, let $y(\mathbf{x}) \in \mathbb{R}^p$ be the response vector at $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ of a multivariate Gaussian Random Field (GRF) $y : \mathcal{X} \to \mathbb{R}^p$ with zero mean and a cross-covariance function $c(\mathbf{x}, \mathbf{x}') = \text{cov} \left( y(\mathbf{x}), y(\mathbf{x}') \right) \in \mathbb{S}_{++}^p$. The cross-covariance function is a crucial object in multivariate GRF models which should converge to a symmetric and positive-definite matrix as $\|\mathbf{x} - \mathbf{x}'\| \to 0$. Similarly to the univariate case, the process is second-order stationarity if $c(\ldots)$ depends on $\mathbf{x}$ and $\mathbf{x}'$ only through $\mathbf{x} - \mathbf{x}'$, and it is isotropic if $c(\ldots)$ depends on $\mathbf{x}$ and $\mathbf{x}'$ only through $\|\mathbf{x} - \mathbf{x}'\|$.

The parametric structure of the cross-covariance matrix should be such that the resulting cross-covariance matrix is a positive-definite matrix. Gelfand
et al. (2004) and Banerjee et al. (2014) review some methods to construct a valid cross-covariance function. In these methods, parameter estimation involves solving nonconvex optimization problems.

In this study, we assume a separable cross-covariance function belonging to a parametric family, and propose a two-stage procedure for estimating the unknown parameters. The separable model assumes that the cross-covariance function is a multiplication of a spatial correlation function and a positive-definite between-response covariance matrix:

\[ c(x, x') = \rho(x, x') \Gamma^* \in \mathbb{S}_+^p, \quad (3.1) \]

where \( \rho : \mathcal{X} \times \mathcal{X} \to [0, 1] \) is the spatial correlation function, and \( \Gamma^* \in \mathbb{S}_+^p \) is the between-response covariance matrix. Furthermore, let \( y = [y(x_1)^T, \ldots, y(x_n)^T]^T \in \mathbb{R}^{np} \) denote the process values in long vector form corresponding to locations in \( D^x := \{x_i\}_{i=1}^n \subset \mathcal{X} \). Given the cross-covariance function (3.1), and the set of locations \( D^x \), \( y \) follows a multivariate Gaussian distribution with zero mean and covariance matrix equal to

\[ C^* = R^* \otimes \Gamma^*, \quad (3.2) \]

where \( R^* \in \mathbb{S}_{++}^n \) is the spatial correlation matrix such that \( R^*_{ij} = \rho(x_i, x_j) \) for \( i, j \in I := \{1, \ldots, n\} \), and \( \otimes \) denotes the Kronecker product. Hence,

\[ y \sim \mathcal{N}(0, C^*). \quad (3.3) \]
Let $\mathcal{D} = \{(x_i, y_i^{(r)}) : i \in \mathcal{I}, r = 1, \ldots, N\}$ be the training data set that contains $N$ realizations of the process over $n$ distinct locations $\mathcal{D}^x \subset \mathcal{X}$, i.e., for each $r \in \{1, \ldots, N\}$, $y^{(r)} = [y_i^{(r)}]_{i \in \mathcal{I}} \in \mathbb{R}^{np}$ is an independent realization of $y = [y(x_i)]_{i \in \mathcal{I}}$. Hence, $\{y^{(r)}\}_{r=1}^N$ are i.i.d. according to (3.3).

As in the univariate case, suppose the correlation function belongs to a parametric family $\{\rho(x, x'; \theta) : \theta \in \Theta\}$, where $\Theta$ is a closed convex set containing the true parameter vector, $\theta^*$, of the correlation function $\rho$. Given $\mathcal{D}^x = \{x_i\}_{i \in \mathcal{I}}$, define $R^* := R(\theta^*)$, where $R(\theta) \in \mathbb{S}_+^n$ is such that

$$R(\theta) = [r_{ij}(\theta)]_{i,j \in \mathcal{I}}, \quad r_{ij}(\theta) = \rho(x_i, x_j; \theta) \quad \forall i, j \in \mathcal{I}. \quad (3.4)$$

The sample covariance matrix $S \in \mathbb{S}_+^{np}$ is calculated as $S = \frac{1}{N} \sum_{r=1}^N y^{(r)} y^{(r)\top}$. Furthermore, let $G \in \mathbb{S}_+^n$ such that $G_{ij} > 0$ for all $i, j \in \mathcal{I}$; in particular, we fix $G$ as in (2.2) based on inter-distances. Let $P^* = (C^*)^{-1}$ be the true precision matrix corresponding to locations in $\mathcal{D}^x$, and let $a^*$ and $b^*$ be some given constants such that $0 \leq a^* \leq \sigma_{\min}(P^*) \leq \sigma_{\max}(P^*) \leq b^* \leq \infty$. To estimate $P^*$, we propose to solve the following convex program:

$$\hat{P} = \arg\min_{a^* I \preceq P \preceq b^* I} \langle S, P \rangle - \log \det(P) + \alpha \langle G \otimes (1_p 1_p^\top), |P| \rangle, \quad (3.5)$$

where $|.|$ is the element-wise absolute value operator, and $1_p \in \mathbb{R}^p$ denotes the vector of all ones. This objective penalizes the elements of the precision matrix with weights proportional to the distance between their locations.
Problem (3.5) can be solved efficiently using the ADMM implementation proposed in Davanloo et al. (2015). Indeed, for $0 < a^* \leq b^* < \infty$, the function $-\log \det(\cdot)$ is strongly convex and has a Lipschitz continuous gradient; therefore, the ADMM iterate sequence converges to the optimal solution with a linear rate – see Deng and Yin (2015).

Let $\hat{C} := \hat{P}^{-1}$, and for all $(i, j) \in \mathcal{I} \times \mathcal{I}$ define block matrices $S^{ij} \in \mathbb{S}^p$, $\hat{C}^{ij} \in \mathbb{S}^p$ and $\Sigma^{ij} \in \mathbb{S}^p$ such that $S = [S^{ij}]$, $\hat{C} = [\hat{C}^{ij}]$ and $C^* = [\Sigma^{ij}]$, i.e., $S^{ij} \in \mathbb{S}^p$, $\hat{C}^{ij} \in \mathbb{S}^p$ and $\Sigma^{ij} \in \mathbb{S}^p$ are the sample, estimated and true covariance matrices between the locations $x_i$ and $x_j$. The following establishes a probability bound for the estimation error $\hat{P} - P^*$.

**Theorem 1.** Let $\{y^{(r)}\}^N_{r=1} \subset \mathbb{R}^{nq}$ be independent realizations of a GRF with zero-mean and stationary covariance function $c(x, x'; \theta^*)$ observed over $n$ distinct locations $\{x_i\}_{i \in \mathcal{I}}$ with $\mathcal{I} := \{1, \ldots, n\}$; furthermore, let $C^* = R(\theta^*) \otimes \Gamma^*$ be the true covariance matrix, and $P^* := C^{*-1}$ be the corresponding true precision matrix, where $R(\theta)$ is defined in (3.4). Finally, let $\hat{P}$ be the SPS estimator computed as in (3.5) 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 $M > 0$, $N \geq N_0 := \lceil 2[(M + 2) \ln(np) + \ln 4] \rceil$, and $b^* \geq \sigma_{\max}(P^*)$,

$$Pr\left(\|\hat{P} - P^*\|_F \leq 2b^*2p(n + \|G\|_F)\alpha\right) \geq 1 - (np)^{-M}, \quad (3.6)$$

for all $\alpha$ such that $40 \max_{i=1,\ldots,p} (\Gamma^*_i)^{\frac{N_0}{N}} \leq \alpha \leq 40 \max_{i=1,\ldots,p} (\Gamma^*_i)$. 


Proof. See the supplementary materials.

Given that $C^* = R^* \otimes \Gamma^*$, and the diagonal elements of the spatial correlation matrix $R^*$ are equal to one, we have $\Sigma^{ii} = \Gamma^*$. Therefore, we propose to estimate the between-response covariance matrix $\Gamma^*$ by taking the average of the $p \times p$ matrices along the diagonal of $\hat{C}$, i.e.,

$$
\hat{\Gamma} := \frac{1}{n} \sum_{i=1}^{n} \hat{C}^{ii} \in S^n_{++},
$$

(3.7)

Note that (3.5) implies that $\hat{P} \in S^{np}_{++}$; hence, $\hat{C} \in S^{np}_{++}$ as well. Therefore, all its block-diagonal elements are positive definite, i.e., $\hat{\Sigma}^{ii} \in S^n_{++}$ for $i = 1, ..., n$. Since $\hat{\Gamma}$ is a convex combination of $\hat{\Sigma}^{ii} \in S^n_{++}$, $i = 1, ..., n$ and the cone of positive definite matrices is a convex set, we also have $\hat{\Gamma} \in S^n_{++}$.

A probability bound in the estimation error of the covariance matrices is shown in the following theorem.

**Theorem 2.** Given $M > 0$, $N \geq N_0 := \lceil 2 \left( (M + 2) \ln(np) + \ln 4 \right) \rceil$, and $a^*, b^*$ such that $0 < a^* \leq \sigma_{\min}(P^*) \leq \sigma_{\max}(P^*) \leq b^* < \infty$, let $\hat{P}$ be the SPS estimator as in (3.5). Then $\hat{\Gamma}$, defined in (3.7), and $\hat{C} = \hat{P}^{-1}$ satisfy

$$
Pr\left( \max\{\|\hat{C} - C^*\|_2, \|\hat{\Gamma} - \Gamma^*\|_2\} \leq 2 \left( \frac{b^*}{a^*} \right)^2 p(n + \|G\|_F) \alpha \right) \geq 1 - (np)^{-M},
$$

(3.8)

for all $\alpha$ such that $40 \max_{i=1,...,p} (\Gamma^*_{ii}) \sqrt{\frac{N}{N_0}} \leq \alpha \leq 40 \max_{i=1,...,p} (\Gamma^*_{ii})$. 


Proof. From (3.8), we have
\[ \|\hat{C} - C^*\|_2 \leq \frac{1}{a^*} \|\hat{P} - P^*\|_2 \leq \frac{1}{a^*} \|\hat{P} - P^*\|_F \leq 2 \left( \frac{b^*}{a^*} \right)^2 p(n + \|G\|_F)\alpha, \]
where the first inequality follows from the Lipschitz continuity of $P \mapsto P^{-1}$ on the domain $P \succeq 0$ with respect to the spectral norm $\|\cdot\|_2$. Hence, given that $\Gamma^* = \Sigma^{ii}$ for all $i \in \mathcal{I}$, we have $\|\hat{C}^{ii} - \Gamma^*\|_2 \leq \left( \frac{b^*}{a^*} \right)^2 p(n + \|G\|_F)\alpha$ for all $i \in \mathcal{I}$. Therefore, from convexity of $X \mapsto \|X - \Gamma^*\|_2$, it follows that
\[ \|\hat{\Gamma} - \Gamma^*\|_2 \leq \frac{1}{n} \sum_{i \in \mathcal{I}} \|\hat{C}^{ii} - \Gamma^*\|_2 \leq 2 \left( \frac{b^*}{a^*} \right)^2 p(n + \|G\|_F)\alpha. \]

Given $\mathcal{D}^x = \{x_i\}_{i \in \mathcal{I}} \subset \mathcal{X}$, define $R : \mathbb{R}^q \to \mathbb{S}^n$ over $\Theta \subset \mathbb{R}^q$ as in (3.4), i.e., $R(\theta) = [r_{ij}(\theta)]_{i,j \in \mathcal{I}} \in \mathbb{S}^n$ and $r_{ij}(\theta) = \rho(x_i, x_j; \theta)$ for all $(i, j) \in \mathcal{I} \times \mathcal{I}$. To estimate the true parameter vector of the spatial correlation function, $\theta^*$, we propose to solve
\[ \hat{\theta} \in \arg\min_{\theta \in \Theta} \frac{1}{2} \sum_{i,j \in \mathcal{I}} \|r_{ij}(\theta)\hat{\Gamma} - \hat{C}^{ij}\|_F^2. \]  
(3.9)

The objective function of (3.9) can be written in a more compact form as the parametric function below, with parameters $\Gamma \in \mathbb{S}^p$ and $C \in \mathbb{S}^{np}$:
\[ f(\theta; \Gamma, C) := \frac{1}{2} \|R(\theta) \otimes \Gamma - C\|_F^2. \]  
(3.10)

Let $\theta = [\theta_1, \ldots, \theta_q]^{\top}$, and $R'_k : \mathbb{R}^q \to \mathbb{S}^n$ such that $R'_k(\theta) = [\frac{\partial}{\partial \theta_k} r_{ij}(\theta)]_{i,j \in \mathcal{I}}$ for $k = 1, \ldots, q$. Similarly, $R''_{k\ell} : \mathbb{R}^q \to \mathbb{S}^n$ such that $R''_{k\ell}(\theta) = [\frac{\partial^2}{\partial \theta_k \partial \theta_\ell} r_{ij}(\theta)]_{i,j \in \mathcal{I}}$. 

for \(1 \leq k, \ell \leq q\). Let \(Z(\theta; \Gamma, C) := R(\theta) \otimes \Gamma - C\); hence, \(f(\theta; \Gamma, C) = \|Z(\theta; \Gamma, C)\|_F^2/2\); and define \(Z'_k(\theta; \Gamma) := R'_k(\theta) \otimes \Gamma\) for \(k = 1, \ldots, q\).

**Lemma 1.** Suppose \(\rho(x, x'; \theta)\) is twice continuously differentiable in \(\theta\) over \(\Theta\) for all \(x, x' \in X\), then there exists \(\gamma^* > 0\) such that

\[
\nabla_\theta^2 f(\theta^*; \Gamma^*, C^*) \succeq \gamma^* I
\]

if and only if \(\{\text{vec}(R'_k(\theta^*))\}_{k=1}^q \subset \mathbb{R}^{n^2}\) are linearly independent.

**Proof.** Clearly, \(\nabla_\theta f(\theta; \Gamma, C) = [\langle Z'_1(\theta; \Gamma, C), Z(\theta; \Gamma, C) \rangle, \ldots, \langle Z'_q(\theta; \Gamma, C), Z(\theta; \Gamma, C) \rangle]^T\).

Hence, it can be shown that for \(1 \leq k \leq q\)

\[
\frac{\partial}{\partial \theta_k} f(\theta; \Gamma, C) = \|\Gamma\|_F^2 \langle R'_k(\theta), R(\theta) \rangle - \langle C, R'_k(\theta) \otimes \Gamma \rangle,
\]

and from the product rule for derivatives, it follows that for \(1 \leq k, \ell \leq q\)

\[
\frac{\partial^2}{\partial \theta_k \partial \theta_\ell} f(\theta; \Gamma, C) = \|\Gamma\|_F^2 \langle R'_k(\theta), R'_\ell(\theta) \rangle + \langle R''_{k\ell}(\theta) \otimes \Gamma, R(\theta) \otimes \Gamma - C \rangle.
\]

Thus, since \(C^* = r(\theta^*) \otimes \Gamma^*\), we have

\[
\frac{\partial^2}{\partial \theta_k \partial \theta_\ell} f(\theta; \Gamma^*, C^*) = \|\Gamma^*\|_F^2 \langle R'_k(\theta^*), R'_\ell(\theta^*) \rangle.
\]

Therefore, \(\nabla_\theta^2 f(\theta^*; \Gamma^*, C^*) = \|\Gamma^*\|_F^2 J(\theta^*)^T J(\theta^*)\), where \(J(\theta) \in \mathbb{R}^{n^2 \times q}\) such that \(J(\theta) := [\text{vec}(R'_1(\theta)) \ldots \text{vec}(R'_q(\theta))].\) Hence, if \(\{\text{vec}(R'_k(\theta^*))\}_{k=1}^q \subset \mathbb{R}^{n^2}\) are linearly independent, then there exists \(\gamma^* > 0\) such that \(\nabla_\theta^2 f(\theta^*; \Gamma^*, C^*) \succeq \gamma^* I.\)

**Remark.** We would like to comment on the linear independence condition stated in Lemma 1. As an example, consider the anisotropic exponential
correlation function \( \rho(x, x', \theta) = \exp \left( - (x - x')^\top \text{diag}(\theta)(x - x') \right) \), where \( q = d \), and \( \Theta = \mathbb{R}^d_+ \). Let \( \mathcal{X} = [-\beta, \beta]^d \) for some \( \beta > 0 \), and suppose \( \{x_i\}_{i \in I} \) is a set of independent identically distributed uniform random samples inside \( \mathcal{X} \). Then it can be easily shown that for the anisotropic exponential correlation function, the condition in Lemma 1 holds with probability 1, i.e., \( \{\text{vec}(R_k'(\theta^*))\}_{k=1}^d \) are linearly independent w.p. 1.

The next result builds on Lemma 1, and it shows the convergence of the SPS estimator as the number of samples per location, \( N \), increases.

**Theorem 3.** Suppose \( \theta^* \in \text{int} \Theta \), and \( \rho(x, x'; \theta) \) is twice continuously differentiable in \( \theta \) over \( \Theta \) for all \( x, x' \in \mathcal{X} \). Suppose \( \{\text{vec}(R_k'(\theta^*))\}_{k=1}^q \subset \mathbb{R}^{n^2} \) are linearly independent. For any given \( M > 0 \) and \( N \geq N_0 := [2(M + 2) \ln(np) + \ln 16] \), let \( \hat{\theta}^{(N)} \) be the SPS estimator of \( \theta^* \), i.e., \( \hat{\theta} = \text{argmin}_{\theta \in \Theta} f(\theta; \hat{\Gamma}, \hat{C}) \), and \( \hat{\Gamma} \) be computed as in (3.7). Then for any sufficiently small \( \epsilon > 0 \), there exists \( N \geq N_0 \) satisfying \( N = O(N_0/\epsilon^2) \) such that setting \( \alpha = 40 \max_{i=1,\ldots,p} \sqrt{\frac{N_0}{N}} \) in (3.5) implies \( \|\hat{\theta}^{(N)} - \theta^*\| \leq \epsilon \) and \( \|\hat{\Gamma} - \Gamma^*\| = O(\epsilon) \) with probability at least \( 1 - (np)^{-M} \); moreover, the STAGE-II function \( f(\cdot; \hat{\Gamma}, \hat{C}) \) is strongly convex around the estimator \( \hat{\theta} \).

**Proof.** See the supplementary material. \( \square \)

A summary of the proposed algorithm for fitting multivariate GRFs models is provided in Algorithm 1.
Algorithm 1 SPS algorithm to fit multivariate GRFs

**Input:** $\mathcal{D} = \{(x_i, y_i^{(r)})\}_{i=1}^n \subset \mathcal{X} \times \mathbb{R}^p$, $i \in \mathcal{I}$, $r = 1, \ldots, N$

/* Compute the sample covariance and distance matrices*/

$y^{(r)} \leftarrow [y(x_1)^T, \ldots, y(x_n)^T]^T \in \mathbb{R}^{np}$, $r = 1, \ldots, N$

$S \leftarrow \frac{1}{N} \sum_{r=1}^N y^{(r)} y^{(r)\top}$

$G_{ij} \leftarrow \|x_i - x_j\|_2$, if $i \neq j$, $G_{ii} \leftarrow \min\{\|x_i - x_j\|_2 : j \in \mathcal{I} \setminus \{i\}\}$

/* Compute the precision matrix and its inverse */

$\hat{P} \leftarrow \arg\min\{\langle S, P \rangle - \log \det(P) + \alpha \langle G \otimes (1_q 1_q^T), |P| \rangle : a^* \mathbf{I} \preceq P \preceq b^* \mathbf{I}\}$

$\hat{C} \leftarrow \hat{P}^{-1}$

/* Compute the between response covariance matrix */

$\hat{\Gamma} \leftarrow \frac{1}{n} \sum_{i \in \mathcal{I}} \hat{C}_{ii}$

/* Compute the spatial correlation parameter vector */

$\hat{\theta} \leftarrow \arg\min_{\theta \in \Theta} \frac{1}{2} \sum_{i,j \in \mathcal{I}} \|\rho_{ij}(\theta) \hat{\Gamma} - \hat{C}_{ij}\|_F^2$

**Return:** $\hat{\Gamma}$ and $\hat{\theta}$

4. Numerical results

A simulation was performed over a square domain $\mathcal{X} = [0, 10]^d$ over $n$ spatially distinct points. $N$ realizations of a $p$-variate zero-mean anisotropic GRF with a separable covariance function were used to estimate the parameters of the model. The covariance function was set equal to the product of the exponential correlation function $\rho(x, x', \theta^*) = \exp\left(- (x - x')^\top \text{diag}(\theta^*) (x - x')\right)$ and the between-response covariance matrix $\Gamma^* \in \mathbb{S}^p_{++}$. $N$ was set equal to $N_0$ in Theorem 1 with $M$ set equal to 1. To compare, we also calculated the maximum likelihood estimate (MLE) of the parameters. To deal with the nonconcavity of the likelihood, the MLEs were calculated from 10 random initial solutions and the best final solutions were reported. Each simulation scenario is replicated 10 times and the mean (and standard deviation) of $\{\|\hat{\theta}_\ell - \theta^*\|\}_{\ell=1}^{10}$ and $\{\|\hat{\Gamma}_\ell - \Gamma^*\|_F\}_{\ell=1}^{10}$ are reported in
Table 1. To solve problem (3.5) in scenarios with \( np > 1000 \), we used the Random Selection (RS) segmentation scheme as described in Davanloo et al. (2015). As we can see from the results, the mean of \( \{\|\hat{\theta}_\ell - \theta^*\|\}_{\ell=1}^{10} \) is an increasing function of the input space dimension \( d \) which is reasonable due to higher number of parameters in the anisotropic correlation function. Furthermore, the mean of \( \{\|\hat{\Gamma}_\ell - \Gamma^*\|_F\}_{\ell=1}^{10} \) increases with \( p \), the number of responses. In most of the scenarios, \( n = 1000 \) (vector observations) result in better parameters estimates compared to the \( n = 100 \) case, a result in accordance to the expected effect of infill asymptotics. Overall, the Generalized SPS method results in better parameter estimates compared to MLE with relative performance improvements becoming more obvious as \( p \) and \( d \) increase.

### Table 1: Generalized SPS method vs. MLE. The numbers are the mean (standard deviation) of \( \{\|\hat{\theta}_\ell - \theta^*\|\}_{\ell=1}^{10} \) and \( \{\|\hat{\Gamma}_\ell - \Gamma^*\|_F\}_{\ell=1}^{10} \).

<table>
<thead>
<tr>
<th>( d )</th>
<th>( n )</th>
<th>Method</th>
<th>( p=2 )</th>
<th>( p=5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>100</td>
<td>GSPS</td>
<td>0.15 (0.03)</td>
<td>0.25 (0.11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLE</td>
<td>0.18 (0.05)</td>
<td>0.29 (0.18)</td>
</tr>
<tr>
<td>1000</td>
<td>0.10 (0.02)</td>
<td>0.22 (0.10)</td>
<td>0.90 (0.33)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.52 (0.23)</td>
<td>0.49 (0.20)</td>
<td>1.35 (0.63)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
<td>GSPS</td>
<td>0.36 (0.12)</td>
<td>0.42 (0.14)</td>
</tr>
<tr>
<td></td>
<td>0.50 (0.17)</td>
<td>0.48 (0.20)</td>
<td>1.26 (0.63)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.33 (0.12)</td>
<td>0.39 (0.12)</td>
<td>0.69 (0.16)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.49 (0.20)</td>
<td>0.46 (0.17)</td>
<td>1.07 (0.48)</td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusions and future research

A new two-stage estimation method is proposed to fit multivariate Gaussian Random Field (GRF) models with separable covariance functions. Theoret-
ical convergence rate for the estimated between-response covariance matrix and the correlation function parameter estimate are established. Numerical studies confirm the provided theoretical results. From the statistical perspective, the first stage provides a Gaussian Markov Random Field (GMRF) approximation to the underlying GRF without discretizing the input space or assuming a sparsity structure for the precision matrix. From an optimization perspective, the first stage helps to “zoom into” the region where the global optimal covariance parameters exist, facilitating the second stage least-squares optimization. In this research, we considered separable covariance functions. Future research may consider non-separable covariance functions, e.g., convolution of covariance functions, or kernel convolution.

References


REFERENCES


Dept. of Statistics, Blacksburg VA 24061, USA
E-mail: sdt144@vt.edu

Dept. of Industrial and Manufacturing Engineering, University Park PA 16802 USA
E-mail: nsa10@psu.edu

Dept. of Industrial and Manufacturing Engineering, University Park PA 16802 USA
E-mail: exd13@psu.edu

(Received March 2016; accepted ??? 20??)