Statistica Sinica Preprint No: SS-2017-0091.R1	
Title	Generalized Sparse Precision Matrix Selection for Fitting Multivariate Gaussian Random Fields to Large Data Sets
Manuscript ID	SS-2017-0091.R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0091
Complete List of Authors	S. Davanloo Tajbakhsh N.S. Aybat and E. del Castillo
Corresponding Author	E. del Castillo
E-mail	exd13@psu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

Statistica Sinica

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

S. Davanloo Tajbakhsh[1], N.S. Aybat[2], and E. del Castillo[2]

[1] *The Ohio State University and* [2] *The Pennsylvania State University*

Abstract: We present a new method for estimating multivariate, second-order stationary Gaussian Random Field (GRF) models based on the Sparse Precision matrix Selection (SPS) algorithm, proposed by Davanloo et al. (2015) for estimating scalar GRF models. Theoretical convergence rates for the estimated between-response covariance matrix and for the estimated parameters of the underlying spatial correlation function are established. Numerical tests using simulations and datasets 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.

1. Introduction

Gaussian Random Field (GRF) models are very popular in Machine Learning, e.g., (Rasmussen et al. (2006)); they also are widely used in Geostatistics, with applications in meteorology to model satellite data for fore-

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

casting or to solve inverse problems to tune weather models (Cressie et al. (2011)), and to model outputs of expensive-to-evaluate deterministic Finite Element Method (FEM) computer codes, e.g.,(Santner et al. (2003)). More recently, there have been applications of GRF to model stochastic simulations, e.g., queuing or inventory control models (Ankenman et al. (2010); Kleijnen (2010)), and to model free-form surfaces of manufactured products from noisy measurements for inspection or quality control purposes (Del Castillo et al. (2015)).

In a GRF model, a key role is played by the covariance or kernel 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, or Matern; and Maximum Likelihood (ML) is the dominant method to estimate their parameters from data (Santner et al. (2003)). The ML fitting procedure suffers from two main challenges: the negative loglikelihood is a *nonconvex* function of the covariance matrix, the problem is computationally hard when the number of spatial locations n is large. This is known as the *"big-n"* problem in the literature. Along with some other approximation methods, there is an important class that approximates the Gaussian likelihood using different forms of conditional

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

independence assumptions which reduces the computational complexity significantly, e.g., Snelson and Ghahramani (2006); Pourhabib et al. (2014) and references therein.

In (Davanloo et al. (2015)) we proposed a Sparse Precision Selection (SPS) algorithm for *univariate* processes to deal with the first challenge by providing theoretical guarantees on the SPS parameter estimates, and presented a *segmentation* scheme on the training data to be able to solve big-n problems. Given the nature of SPS, the segmentation does not result in discontinuities in the predicted process. In contrast, localized regression methods also rely on segmentation to reduce the computational cost; but may suffer from discontinuities on the predicted surface at the boundaries of the segments. In this paper, we present a Generalized SPS (GSPS) method for fitting a *multivariate* GRF process that deals with the two aforementioned challenges when there are possibly cross-correlated multiple responses that occur at each spatial location.

Compared to SPS (and also to GSPS), the likelihood approximation type GRF methods, e.g., Snelson and Ghahramani (2006), have the advantage of computational efficiency; but carry no guarantees on the quality of the parameter estimates as only an approximation to the likelihood function is optimized (compared to MLE, this is a small dimensional problem, but

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

still non-convex). SPS has theoretical error bound guarantees on hyper-parameter estimates (this is also the case for GSPS, see Theorem 4 below) – these bounds also imply error guarantees on prediction quality through the mean of the predictive distribution.

There is a wide variety of applications that require the approximation of a vector of *correlated* responses obtained at each spatial or spatial-temporal location. Climate models are classic geostatistical examples where environmental variables such as atmospheric CO_2 concentration, ocean heat uptake and global surface temperature are jointly modeled (Urban et al. (2010)). Lin (2008) uses a Multivariate GRF model to map spatial variations of five different heavy metals in soil. This is an application close to that of Kriging in mining engineering where the spatial occurrence of two metals may be cross-correlated, e.g., silver and lead. Multivariate GRFs are also popular in multi-task learning (Bonilla et a. (2008)), an area of machine learning where multiple related tasks need to be learned so that simultaneously learning them can be better than learning them in isolation without transfer of information between the tasks. The joint modeling of spatial responses is also useful in metrology when conducting *multi-fidelity analysis* (Forrester et al. (2008)), where an expensive, high fidelity spatial response needs to be predicted from predominantly low fidelity responses, which are

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

inexpensive – see also (Boyle et al. (2004)). Likewise, multivariate GRFs have been used to reconstruct 3-dimensional free-form surfaces of manufactured products through modeling each of the 3 coordinates of a measured point as a parametric surface response (Del Castillo et al. (2015)). Wang and Chen (2015) model the response surface of a catalytic oxidation process with two highly correlated response variables; Castellanos et al. (2015) estimate low dimensional spatio-temporal patterns of finger motion in repeated reach-to-grasp movements; Bhat et al. (2010) study a multi-output GRF for computer model calibration with multivariate spatial data to infer parameters in a climate model. In many of such applications multiple realizations of the GRF are sensed/measured over time ($N > 1$) over a fixed set of locations. GRF applications with $N > 1$ commonly arise in practice, including those in "metamodeling" of stochastic simulations for modeling an expensive-to-evaluate queuing or inventory control model, in modeling product surfaces for inspection or quality control purposes, and in models for which we observe a spatial process over time at the same locations for a system known to be static with respect to time.

Rather than considering each response independently, using the *between-response* covariance can significantly enhance the prediction performance. As mentioned by Cressie (2015), the principle of exploiting co-variation

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

to improve mean-squared prediction error goes back to Kolmogorov and Wiener in the first half of the Twentieth century. It is well-known that the minimum-mean-square-error predictor of a single response component of a multivariate GRF involves the between-response covariances of all responses (Santner et al. (2003)), a result that lies at the basis of the so-called Co-Kriging technique in Geostatistics (Cressie (2015)).

In this paper, we adopt a separable cross-covariance structure – see (3.2) – which has been an assumption made in the literature since at least Mardia and Goodall (1993), who proposed separability to model multivariate spatio-temporal data. Bhat et al. (2010) used separable cross-covariance for computer model calibration. See also (Gelfand et al. (2004); Banerjee et al. (2014); Gelfand and Banerjee (2010)) and (Genton and Kleiber (2015)). Li et al. (2008) proposed a technique to test the separability assumption for a multivariate random process. Gelfand and Banerjee (2010) mention an additional use of a separable covariance structure, as an example of this type of application, Banarjee and Gelfand employed such separable models (Banerjee and Gelfand (2002); Banerjee et al. (2014)) to analyze the relationship between shrub density and dew duration for a dataset consisting of 1129 locations in a west-facing watershed in the Negev desert in Israel.

Fitting *multivariate* GRFs can be difficult due to the two challenges

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

mentioned above. In particular, the parametrization of the matrix-valued covariance functions requires a higher-dimensional parameter vector that aggravates the difficulty of GRF estimation (Banerjee et al. (2014); Cressie et al. (2011)). The goal of this paper is to extend the theory of the univariate SPS method (Davanloo et al. (2015)) to include the hyper-parameter estimation of *multivariate* GRF models for which the error bounds on the approximation quality can be established. 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, GSPS, the multivariate generalization of the SPS method is described and compared with other methods for fitting multivariate GRF, and theoretical guarantees of the GSPS estimates are discussed. Section 4 includes numerical results. 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, ℓ_1, and ℓ_∞ norms, respectively. For $x \in \mathbb{R}^n$, $\mathrm{diag}(x) \in \mathbb{S}^n$ 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 $\mathbf{vec}(X) \in \mathbb{R}^{np}$, obtained by stacking the columns of the matrix X on top of one another. Let $r = \mathbf{rank}(X)$, and $\sigma = [\sigma_i]_{i=1}^r \subset \mathbb{R}^r_{++}$ (positive orthant) denote the singular values of

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

X; then, $\|X\|_F := \|\sigma\|$, $\|X\|_2 := \|\sigma\|_\infty$, and $\|X\|_* := \|\sigma\|_1$ denote the Frobenius, spectral, and nuclear norms of X, respectively. Given $X, Y \in \mathbb{R}^{m\times n}$, $\langle X, Y\rangle := \mathbf{Tr}(X^\top Y)$ denotes the standard inner product. Let $\mathcal{V}$ be a normed vector space with norm $\|.\|_a$. For $\bar{x} \in \mathcal{V}$ and $r > 0$, $\mathcal{B}_{\|.\|_a}(\bar{x}, r) := \{x \in \mathcal{V} : \ \|x - \bar{x}\|_a < r\}$ denotes the open ball centered at $\bar{x}$ with radius $r > 0$, and $\bar{\mathcal{B}}_{\|.\|_a}(\bar{x}, r)$ 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 GRF, where $y(\mathbf{x})$ denotes the value of the process at location $\mathbf{x} \in \mathcal{X}$. Let $m(\mathbf{x}) = \mathbb{E}(y(\mathbf{x}))$ for $\mathbf{x} \in \mathcal{X}$, and $c(\mathbf{x}, \mathbf{x}')$ be the spatial covariance function denoting the covariance between $y(\mathbf{x})$ and $y(\mathbf{x}')$. Without loss of generality, we assume that the GRF has a mean equal to zero. Suppose 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 $\mathcal{D}^x := \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$. 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 $\mathcal{D}^x$.

For simplicity in estimation, the covariance function, $c(\mathbf{x}, \mathbf{x}')$, is typically assumed to belong to some parametric family $\{c(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}, \nu) : \ \boldsymbol{\theta} \in \Theta, \ \nu \geq 0\}$ and $c(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}) := \nu\rho(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta})$, where $\rho(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta})$ is a parametric correlation function where $\boldsymbol{\theta}$ and ν denote the *spatial correlation* and *variance* parameters, respectively, and $\Theta \subset \mathbb{R}^q$ is a set that contains the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

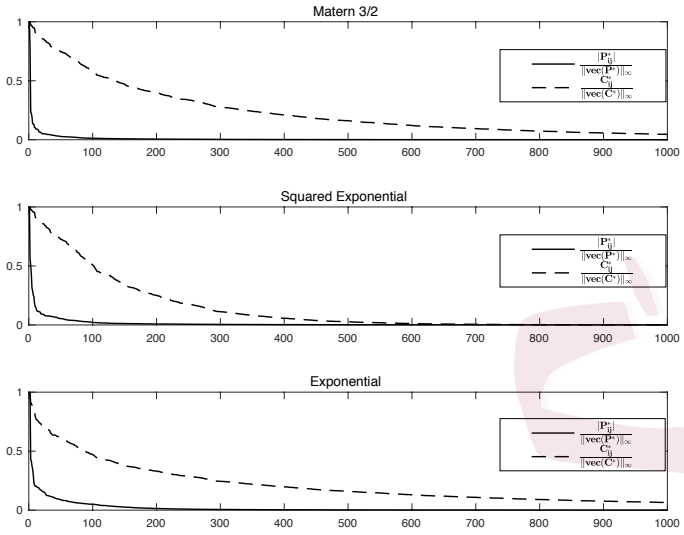

Figure 1: Decaying behavior of elements of the Precision and Covariance matrices for GRFs. The largest 1000 off-diagonal elements of the precision and covariance matrices (scaled by their maximums) plotted in descending order. The underlying GRF was evaluated over 100 randomly selected points in $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^2 : -50 \leq \mathbf{x} \leq 50\}$ for three covariance functions with range and variance parameters equal to 10, and 1, respectively.

true spatial correlation parameters – see e.g. Cressie (2015). Let $\boldsymbol{\theta}^*$ and ν^* denote the unknown *true* parameters of the process. Given a set of locations $\mathcal{D}^x = \{\mathbf{x}_i\}_{i=1}^n$, let $C(\boldsymbol{\theta}, \nu) \in \mathbb{S}_{++}^n$ be such that its $(i,j)^{th}$ element is $c(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}, \nu)$ – throughout, $\mathbb{S}_{++}^n$ and $\mathbb{S}_{+}^n$ denote the set of n-by-n symmetric, positive definite and positive semidefinite matrices, respectively.

Let $C^* = C(\boldsymbol{\theta}^*, \nu^*)$ denote the true covariance matrix corresponding to locations in $\mathcal{D}^x = \{\mathbf{x}_i\}_{i=1}^n$, 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 $\boldsymbol{\theta}^*$ and ν^*. The method is motivated by the results in numerical linear algebra which demonstrate that if the elements of a matrix show a decay property, then the elements

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

of its inverse also show a similar behavior – see Benzi (2016) and Jaffard (1990). The latter author considers the two decay classes defined as follows:

Definition 1. 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 $|A_{ij}| \leq K_{\gamma'} \exp\left(-\gamma' d(\mathbf{x}_i, \mathbf{x}_j)\right)$ for all $1 \leq i, j \leq n$. Moreover, A belongs to the class $\mathcal{Q}_\gamma$ for some $\gamma > 1$ if there exists a constant K such that $|A_{ij}| \leq K \left(1 + d(\mathbf{x}_i, \mathbf{x}_j)\right)^{-\gamma}$ for all $1 \leq i, j \leq n$.

Theorem 1. *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 (inverse covariance) matrix of a GRF makes it a *compressible signal* (Candes (2006)); hence, one can argue that it can be well-approximated by a sparse matrix – compare it with the covariance matrix depicted in Figure 1. 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.

Let a^* and b^* be given constants, $0 \leq a^* \leq \sigma_{\min}(P^*) \leq \sigma_{\max}(P^*) \leq b^* \leq \infty$. In the first stage of the SPS algorithm, we solve a convex loglikelihood

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

problem penalized with a weighted ℓ_1-norm to estimate the true precision matrix corresponding to the given data locations $\mathcal{D}^x$:

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

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

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

for all $(i, j) \in \mathcal{I} \times \mathcal{I}$, where $\mathcal{I} = \{1, 2, ..., n\}$ 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 Field (GMRF) approximation to the GRF. Using ADMM, the Alternating Direction Method of Multipliers, see Boyd et al. (2011), (2.1) can be solved efficiently. Indeed, since $-\log\det(.)$ is strongly convex and has a Lipschitz continuous gradient for $0 < a^* \leq b^* < \infty$, an ADMM iterate sequence converges to the optimal solution with a *linear* rate (Deng and Yin (2015)).

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

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

SPS is therefore based on a Gaussian Markov Random Field (GMRF) approximation to the GRF. In general, GMRF models cannot represent GRFs *exactly*. Lindgren et al. (2011) established that the Matern GRFs are Markovian; in particular, they are Markovian 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) 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 ℓ_1-regularization. This first stage helps to "zoom into" the area where the true covariance parameters are located; hence, it helps to not get trapped in local optimum solutions in the *second stage* 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 $c(\mathbf{x}, \mathbf{x}') = \operatorname{cov}(y(\mathbf{x}), y(\mathbf{x}')) \in \mathbb{S}^p_{++}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

The cross-covariance function is a crucial object in multivariate GRF models which should converge to a symmetric and positive-definite matrix as $\|\mathbf{x} - \mathbf{x}'\| \to 0$. Similar to the univariate case, the process is *second-order stationarity* if $c(.,.)$ depends on $\mathbf{x}$ and $\mathbf{x}'$ only through $\mathbf{x} - \mathbf{x}'$, and it is *isotropic* if $c(.,.)$ 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.

Here 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 (see Gelfand and Banerjee (2010); Gelfand et al. (2004), and the references therein):

$$c(\mathbf{x}, \mathbf{x}') = \rho(\mathbf{x}, \mathbf{x}')\ \Gamma^* \in \mathbb{S}_+^p, \tag{3.1}$$

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. Let $\mathbf{y} = [y(\mathbf{x}_1)^\top, ..., y(\mathbf{x}_n)^\top]^\top \in \mathbb{R}^{np}$ denote the process values in long vector form corresponding to locations

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

in $\mathcal{D}^x := \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$. Given the cross-covariance function (3.1), and the set of locations $\mathcal{D}^x$, $\mathbf{y}$ follows a multivariate Gaussian distribution with zero mean and covariance matrix equal to

$$C^* = R^* \otimes \Gamma^*, \tag{3.2}$$

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

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, C^*). \tag{3.3}$$

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

Suppose the correlation function belongs to a parametric family $\{\rho(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, where Θ is a *closed convex* set containing the true parameter vector, $\boldsymbol{\theta}^*$, of the correlation function ρ. Given $\mathcal{D}^x = \{\mathbf{x}_i\}_{i \in \mathcal{I}}$, let $R^* := R(\boldsymbol{\theta}^*)$, where $R(\boldsymbol{\theta}) \in \mathbb{S}^n_{++}$ is such that

$$R(\boldsymbol{\theta}) = [r_{ij}(\boldsymbol{\theta})]_{i,j \in \mathcal{I}}, \quad r_{ij}(\boldsymbol{\theta}) = \rho(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) \quad \forall i, j \in \mathcal{I}. \tag{3.4}$$

For a GRF model with known parameters, the *best linear unbiased* prediction at a new location $\mathbf{x}_0$ is given by the mean of the conditional

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

distribution $p(\mathbf{y}(\mathbf{x}_0)|\{\mathbf{y}^{(r)}\}_{r=1}^N, \mathcal{D}^x)$,

$$\hat{\mathbf{y}}(\mathbf{x}_0) = (\mathbf{r}(\mathbf{x}_0; \boldsymbol{\theta}^*)^\top \otimes \Gamma^*)(R(\boldsymbol{\theta}^*) \otimes \Gamma^*)^{-1} \sum_{r=1}^N \mathbf{y}^{(r)}/N, \tag{3.5}$$

where $\mathbf{r}(\mathbf{x}_0; \boldsymbol{\theta}^*) \in \mathbb{R}^n$ contains the spatial correlation between the new point $\mathbf{x}_0$ and n observed data points – see (Santner et al., 2003). This prediction equation is a *continuous* function of the parameters $\boldsymbol{\theta}^*$ and Γ^*; hence, *biased* estimation of the parameters will translate to poor prediction performance. Then, too, (3.5) shows the importance of considering the between-response covariance matrix Γ^*.

The sample covariance matrix $S \in \mathbb{S}_+^{np}$ is $S = \frac{1}{N}\sum_{r=1}^N \mathbf{y}^{(r)}\mathbf{y}^{(r)\top}$. Let $G \in \mathbb{S}^n$ be such that $G_{ij} > 0$ for all $i, j \in \mathcal{I}$. Here, 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 convex program

$$\hat{P} = \underset{a^*I \preceq P \preceq b^*I}{\operatorname{argmin}} \ \langle S, P\rangle - \log\det(P) + \alpha \left\langle G \otimes (\mathbf{1}_p\mathbf{1}_p^\top), |P| \right\rangle, \tag{3.6}$$

where $|.|$ is the element-wise absolute value operator, and $\mathbf{1}_p \in \mathbb{R}^p$ 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 (3.6) can be solved efficiently using the ADMM implementation

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

proposed in Davanloo et al. (2015). Indeed, for $0 < a^* \leq b^* < \infty$, the function $-\log\det(.)$ 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 $S^{ij} \in \mathbb{S}^p$, $\hat{C}^{ij} \in \mathbb{S}^p$ and $\Sigma^{ij} \in \mathbb{S}^p$ are the sample, estimated and true covariance matrices between the locations $\mathbf{x}_i$ and $\mathbf{x}_j$. We have a probability bound for the estimation error $\hat{P} - P^*$.

Theorem 2. *Let $\{\mathbf{y}^{(r)}\}_{r=1}^N \subset \mathbb{R}^{nq}$ 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 $\mathcal{I} := \{1, ..., n\}$; let $C^* = R(\boldsymbol{\theta}^*) \otimes \Gamma^*$ be the true covariance matrix, and $P^* := C^{*-1}$ be the corresponding true precision matrix, where $R(\boldsymbol{\theta})$ is defined in* (3.4)*. If $\hat{P}$ is the GSPS estimator computed as in* (3.6) *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 $M > 0$, $N \geq N_0 := \lceil 2\big[(M+2)\ln(np) + \ln 4\big]\rceil$, and $b^* \geq \sigma_{\max}(P^*)$,*

$$Pr\Big(\|\hat{P} - P^*\|_F \leq 2{b^*}^2 p(n + \|G\|_F)\alpha\Big) \geq 1 - (np)^{-M}, \tag{3.7}$$

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

Proof. See the supplementary materials. □

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

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

$$\hat{\Gamma} := \frac{1}{n} \sum_{i=1}^{n} \hat{C}^{ii} \in \mathbb{S}_{++}^{n}. \tag{3.8}$$

Now (3.6) implies that $\hat{P} \in \mathbb{S}_{++}^{np}$, hence, $\hat{C} \in \mathbb{S}_{++}^{np}$ as well. Therefore, all its block-diagonal elements are positive definite, $\hat{\Sigma}^{ii} \in \mathbb{S}_{++}^{n}$ for $i = 1, ..., n$. Since $\hat{\Gamma}$ is a convex combination of $\hat{\Sigma}^{ii} \in \mathbb{S}_{+}^{n},\ i = 1, ..., n$ and the cone of positive definite matrices is a convex set, we also have $\hat{\Gamma} \in \mathbb{S}_{++}^{n}$.

Theorem 3. *Given $M > 0$, $N \geq N_0 := \lceil 2\big[(M+2)\ln(np) + \ln 4\big]\rceil$, 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* (3.6)*. Then $\hat{\Gamma}$, defined in* (3.8)*, and $\hat{C} = \hat{P}^{-1}$ satisfy*

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

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

Proof. From (3.7), we have

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

where the first inequality follows from the Lipschitz continuity of $P \mapsto P^{-1}$ on the domain $P \succeq a^*\mathbf{I}$ with respect to the spectral norm $\|.\|_2$. Hence, given

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

that $\Gamma^* = \Sigma^{ii}$ for all $i \in \mathcal{I}$, we have $\|\hat{C}^{ii} - \Gamma^*\|_2 \leq 2\left(\frac{b^*}{a^*}\right)^2 p(n + \|G\|_F)\alpha$ for all $i \in \mathcal{I}$. From convexity of $X \mapsto \|X - \Gamma^*\|_2$, it follows that

$$\|\hat{\Gamma} - \Gamma^*\|_2 \leq \sum_{i\in\mathcal{I}} \frac{1}{n}\|\hat{C}^{ii} - \Gamma^*\|_2 \leq 2\left(\frac{b^*}{a^*}\right)^2 p(n + \|G\|_F)\alpha.$$

□

Remark. For Theorems 2 and 3 to hold, α should be in the interval $40 \max_{i=1,\ldots,p}(\Gamma^*_{ii})\sqrt{\frac{N_0}{N}} \leq \alpha \leq 40 \max_{i=1,\ldots,p}(\Gamma^*_{ii})$; for $N \geq N_0$ this interval is non-empty. The trade-off here is such that smaller α makes the estimation error bounds inside the probabilities tighter – hence, desirable; at the same time, smaller α makes the estimated precision matrix less sparse and requires more memory to store a denser estimated precision matrix. Although the upper-bound on α is fixed, one can play with the lower bound; in particular, one can make it smaller by requiring more realizations N.

Given $\mathcal{D}^x = \{\mathbf{x}_i\}_{i\in\mathcal{I}} \subset \mathcal{X}$, define $R : \mathbb{R}^q \to \mathbb{S}^n$ over $\Theta \subset \mathbb{R}^q$ as in (3.4). To estimate the true parameter vector of the spatial correlation function, $\boldsymbol{\theta}^*$, we propose to solve

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \frac{1}{2}\sum_{i,j\in\mathcal{I}} \|r_{ij}(\boldsymbol{\theta})\hat{\Gamma} - \hat{C}^{ij}\|_F^2. \tag{3.9}$$

The objective function of (3.9) can be written in a more compact form as a parametric function with parameters $\Gamma \in \mathbb{S}^p$ and $C \in \mathbb{S}^{np}$,

$$f(\boldsymbol{\theta}; \Gamma, C) := \frac{1}{2}\|R(\boldsymbol{\theta}) \otimes \Gamma - C\|_F^2. \tag{3.10}$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

for $k = 1, \ldots, q$. Take $R''_{k\ell} : \mathbb{R}^q \to \mathbb{S}^n$ such that $R''_{k\ell}(\boldsymbol{\theta}) = [\frac{\partial^2}{\partial\theta_k \partial\theta_\ell} r_{ij}(\boldsymbol{\theta})]_{i,j\in\mathcal{I}}$ for $1 \leq k, \ell \leq q$. Let $Z(\boldsymbol{\theta}; \Gamma, C) := R(\boldsymbol{\theta}) \otimes \Gamma - C$; hence, $f(\boldsymbol{\theta}; \Gamma, C) = \|Z(\boldsymbol{\theta}; \Gamma, C)\|_F^2/2$, and let $Z'_k(\boldsymbol{\theta}; \Gamma) := R'_k(\boldsymbol{\theta}) \otimes \Gamma$ for $k = 1, \ldots, q$.

Lemma 1. *Suppose $\rho(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$ over Θ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, then there exists $\gamma^* > 0$ such that $\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^*; \Gamma^*, C^*) \succeq \gamma^* \mathbf{I}$ if and only if $\{\mathbf{vec}(R'_k(\boldsymbol{\theta}^*))\}_{k=1}^q \subset \mathbb{R}^{n^2}$ are linearly independent.*

Proof. Clearly, $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \Gamma, C) = \left[\langle Z'_1(\boldsymbol{\theta}; \Gamma), Z(\boldsymbol{\theta}; \Gamma, C)\rangle, \ldots, \left\langle Z'_q(\boldsymbol{\theta}; \Gamma), Z(\boldsymbol{\theta}; \Gamma, C)\right\rangle\right]^\top$. Hence, it can be shown that for $1 \leq k \leq q$,

$$\frac{\partial}{\partial\theta_k} f(\boldsymbol{\theta}; \Gamma, C) = \|\Gamma\|_F^2 \langle R'_k(\boldsymbol{\theta}), R(\boldsymbol{\theta})\rangle - \langle C, R'_k(\boldsymbol{\theta}) \otimes \Gamma\rangle \tag{3.11}$$

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

$$\frac{\partial^2}{\partial\theta_k \partial\theta_\ell} f(\boldsymbol{\theta}; \Gamma, C) = \|\Gamma\|_F^2 \langle R'_k(\boldsymbol{\theta}), R'_\ell(\boldsymbol{\theta})\rangle + \left\langle R''_{k\ell}(\boldsymbol{\theta}) \otimes \Gamma, R(\boldsymbol{\theta}) \otimes \Gamma - C\right\rangle. \tag{3.12}$$

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

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

Therefore, $\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^*; \Gamma^*, C^*) = \|\Gamma^*\|_F^2 \; J(\boldsymbol{\theta}^*)^\top J(\boldsymbol{\theta}^*)$, where $J(\boldsymbol{\theta}) \in \mathbb{R}^{n^2 \times q}$ is such that $J(\boldsymbol{\theta}) := [\mathbf{vec}(R'_1(\boldsymbol{\theta})) \ldots \mathbf{vec}(R'_q(\boldsymbol{\theta}))]$. Hence, there exists $\gamma^* > 0$ such that $\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^*; \Gamma^*, C^*) \succeq \gamma^* I$ when $\{\mathbf{vec}(R'_k(\boldsymbol{\theta}^*))\}_{k=1}^q \subset \mathbb{R}^{n^2}$ are linearly independent. □

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

Remark. As to the linear independence condition stated in Lemma 1, consider the *anisotropic exponential correlation* function $\rho(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}) = \exp\big(-(\mathbf{x}-\mathbf{x}')^\top \operatorname{diag}(\boldsymbol{\theta})(\mathbf{x}-\mathbf{x}')\big)$, where $q = d$, and $\Theta = \mathbb{R}^d_+$. 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 easily shown that for the anisotropic exponential correlation function, the condition in Lemma 1 holds with probability 1.

The next result builds on Lemma 1, and shows the convergence of the GSPS estimator as the number of samples per location, N, increases.

Theorem 4. *Suppose $\boldsymbol{\theta}^* \in \mathbf{int}\,\boldsymbol{\Theta}$, and $\rho(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$ over Θ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Suppose $\{\mathbf{vec}(R'_k(\boldsymbol{\theta}^*))\}_{k=1}^q \subset \mathbb{R}^{n^2}$ are linearly independent. For any given $M > 0$ and $N \geq N_0 := \lceil 2(M+2)\ln(np) + \ln 16\rceil$, let $\hat{\boldsymbol{\theta}}^{(N)}$ be the GSPS estimator of $\boldsymbol{\theta}^*$, $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta}; \hat{\Gamma}, \hat{C})$, and let $\hat{\Gamma}$ be computed as in (3.8). Then, for any sufficiently small $\epsilon > 0$, there exists $N \geq N_0$ satisfying $N = \mathcal{O}(N_0/\epsilon^2)$ such that setting $\alpha = 40\max_{i=1,\ldots,p}(\Gamma^*_{ii})\sqrt{\frac{N_0}{N}}$ in (3.6) implies $\|\hat{\boldsymbol{\theta}}^{(N)} - \boldsymbol{\theta}^*\| \leq \epsilon$ and $\|\hat{\Gamma} - \Gamma^*\| = \mathcal{O}(\epsilon)$ with probability at least $1-(np)^{-M}$; moreover, the STAGE-II function $f(\cdot; \hat{\Gamma}, \hat{C})$ is strongly convex around the estimator $\hat{\boldsymbol{\theta}}$.*

Proof. See the supplementary material. □

Remark. In Theorem 4, α is explicitly set equal to the lower bound,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

i.e., $\alpha = 40 \max_{i=1,\ldots,p} (\Gamma^*_{ii}) \sqrt{\frac{N_0}{N}} = 40 \max_{i=1,\ldots,p} (\Gamma^*_{ii}) \sqrt{\frac{\left\lceil 2\left\lceil (M+2)\ln(np)+\ln 4 \right\rceil \right\rceil}{N}}$. Here M controls the probability bound; hence, the only unknown is $\max_{i=1,\ldots,p} (\Gamma^*_{ii})$ – we implicitly assume that this quantity can be estimated empirically or we have a prior knowledge about it. Theorem 4 also guides us in how to select α. One $\|\hat{\boldsymbol{\theta}}^{(N)} - \boldsymbol{\theta}^*\| \leq \epsilon$ and $\|\hat{\Gamma} - \Gamma^*\| = \mathcal{O}(\epsilon)$ whenever $N = \mathcal{O}(N_0/\epsilon^2)$; which implies a choice of $\alpha = \mathcal{O}(\epsilon)$. In the simulations provided in Section 4, α is set equal to $c\sqrt{\log(np)/N}$, where c is chosen 10^{-2} after some preliminary cross-validation studies.

A summary of the proposed algorithm for fitting multivariate GRF models follows.

Algorithm 1 GSPS algorithm to fit multivariate GRFs

input: $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i^{(r)})\}_{i=1}^n \subset \mathcal{X} \times \mathbb{R}^p,\ i \in \mathcal{I},\ r = 1, ..., N\}$
/* Compute the sample covariance and distance matrices*/
$\mathbf{y}^{(r)} \leftarrow [\mathbf{y}(\mathbf{x}_1)^T, ..., \mathbf{y}(\mathbf{x}_n)^T]^T \in \mathbb{R}^{np},\ r = 1, ..., N$
$S \leftarrow \frac{1}{N} \sum_{r=1}^N \mathbf{y}^{(r)} \mathbf{y}^{(r)\top}$
$G_{ij} \leftarrow \|\mathbf{x}_i - \mathbf{x}_j\|_2, \quad \text{if } i \neq j,\ G_{ii} \leftarrow \min\{\|\mathbf{x}_i - \mathbf{x}_j\|_2 :\ j \in \mathcal{I} \setminus \{i\}$
/* Compute the precision matrix and its inverse */
$\hat{P} \leftarrow \text{argmin}\{\langle S, P\rangle - \log\det(P) + \alpha \left\langle G \otimes (\mathbf{1}_q \mathbf{1}_q^T), |P| \right\rangle :\ a^*\mathbf{I} \preceq P \preceq b^*\mathbf{I}\}$
$\hat{C} \leftarrow \hat{P}^{-1}$
/* 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*/
$\hat{\boldsymbol{\theta}} \leftarrow \text{argmin}_{\boldsymbol{\theta} \in \Theta} \frac{1}{2} \sum_{i,j \in \mathcal{I}} \|\rho_{ij}(\boldsymbol{\theta})\, \hat{\Gamma} - \hat{C}^{ij}\|_F^2$
return: $\hat{\Gamma}$ and $\hat{\boldsymbol{\theta}}$

3.1. Connection to SPS. The main difference between the SPS method and GSPS is how $\hat{\Gamma}$, the estimator for Γ^*, is computed (when $p = 1$, $\Gamma^* \in \mathbb{R}_{++}$ corresponds to the variance parameter $\nu^* > 0$ in SPS), and

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

this difference in the way Γ^* is estimated has significant implications on the numerical stability of solving STAGE-II problem, and on the proof technique to show consistency of the hyperparameter estimate as the number of process realizations, N, increases.

The STAGE-II problem proposed in (3.9), $\min_{\boldsymbol{\theta}\in\mathbb{R}^d_+} \frac{1}{2}\|R(\boldsymbol{\theta})\otimes\hat{\Gamma}-\hat{C}\|_F^2$ where $\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^n \hat{C}_{ii}$, behaves well (although it is non-convex in general), and Theorem 4 shows that the STAGE-II objective is strongly convex around a neighborhood of the estimator. In all our numerical tests, standard nonlinear optimization techniques were able to compute a point close to the global minimizer very efficiently.

As to using GSPS to fit a *multivariate* GRF as opposed to using SPS to fit p independent *univariate* GRFs to p responses, the simulations and real-data examples in Sections 4.2 and 4.3 show that the proposed GSPS method performs significantly better than modeling each response independently.

3.2. Computational Complexity. The computational bottleneck of GSPS method is the singular value decompositions (SVD) that arises when solving the STAGE-I problem using the ADMM algorithm. The per-iteration complexity is $\mathcal{O}((np)^3)$. However, the STAGE-I problem is strongly convex; and ADMM has a linear rate (Deng and Yin (2015)). Therefore, an ϵ-optimal solution can be computed within $\mathcal{O}(\log(1/\epsilon))$ iterations of ADMM,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

thus, the overall complexity to solve STAGE-I is $\mathcal{O}((np)^3 \log(1/\epsilon))$. Likelihood approximation methods do not have such iteration complexity results due to the nonconvexity of the approximate likelihood problem, even though they have cheaper per-iteration-complexity. In case of an isotropic process, the STAGE-II problem in (3.9) is one dimensional and it can simply be solved by using bisection. If the process is anisotropic, then (3.9) is nonconvex in general. That said, this problem is low dimensional due to $d \ll n$; hence, standard nonlinear optimization techniques can compute a local minimizer very efficiently – we also show that STAGE-II objective is strongly convex around a neighborhood of the estimator. In all our numerical tests, STAGE-II problem was solved in much shorter time than the STAGE-I problem; hence, it does not affect the overall complexity significantly. In the code, we use golden-section search for isotropic processes, and Knitro's nonconvex solver to solve (3.9) for general anisotropic processes.

To eliminate $\mathcal{O}((np)^3)$ complexity due to an SVD computation per ADMM iteration and due to computing $\hat{C}$, we used a segmentation scheme. We partition the data to K segments, each one composed of $\approx n/K$ points chosen uniformly at random among n locations, and assuming conditional independence between blocks. In Davanloo et al. (2015), we discussed two blocking/segmentation schemes: Spatial Segmentation (SS) and Ran-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

dom Selection (RS). Solving the STAGE-I problem with blocking schemes assumes a conditional independence assumption between blocks. In SS scheme such conditional independence assumption is potentially violated for points along the common boundary between two blocks. The RS scheme, however, works numerically better for "big-n" scenarios. We believe that with the RS scheme the infill asymptotics make the blocks conditionally independent to a reasonable degree. Using such blocking schemes, the bottleneck complexity reduces to $\mathcal{O}((np/K)^3)$ by solving STAGE-I problem for each block; hence, solving STAGE-I and computing $\hat{C}$, which we assume to be block diagonal, requires a total complexity of $\mathcal{O}(\log(1/\epsilon)\ (np)^3/K^2)$ and this bottleneck complexity can be controlled by properly choosing K.

4. Numerical results

In this section, comprehensive simulation analyses are reported for the study of the performance of the proposed method. N realizations of a zero-mean p-variate GRF with anisotropic spatial correlation function were simulated in a square domain $\mathcal{X} = [0, 10]^d$ over n distinct points. The separable covariance function is the product of an anisotropic exponential spatial correlation function $\rho(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}^*) = \exp\big(-(\mathbf{x}-\mathbf{x}')^\top \operatorname{diag}(\boldsymbol{\theta}^*)(\mathbf{x}-\mathbf{x}')\big)$ and a p-variate between-response covariance matrix $\Gamma^* \in \mathbb{S}^p_{++}$. The correlation function parameter vector $\boldsymbol{\theta}^*_\ell$ was sampled uniformly from the surface of a

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

hyper-sphere in $\mathbb{R}^d$ in the positive orthant for each replication $\ell \in \{1, ..., L\}$. The between-response covariance matrix was $\Gamma_\ell^* = A^\top A$ for $A \in \mathbb{R}^{w \times p}$ such that $w > p$, where the elements of A were sampled independently from $\mathcal{N}(0, 1)$ per replication. To solve the STAGE-I problem, the sparsity parameter α in (2.1) was $c\sqrt{\log(np)/N}$ for some constant c. After some preliminary cross-validation studies, we set c equal to 10^{-2}. In our code, we used golden-section search for isotropic processes which requires a univariate optimization in STAGE-II, and we used Knitro's nonconvex solver to solve (3.9) for general anisotropic processes.

4.1. Parameter estimate consistency

To compare the quality of the GSPS parameter estimate with the Maximum Likelihood Estimate (MLE), for ten replicates, we simulated N independent realizations of GRF, described above under different scenarios, and the mean of $\{\|\hat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}^*\|\}_{\ell=1}^{10}$ and $\{\|\hat{\Gamma}_\ell - \Gamma^*\|_F\}_{\ell=1}^{10}$ are reported.

To deal with the nonconcavity of the likelihood, the MLEs were calculated from ten random initial solutions and the best final solutions are reported. To solve the problem in (3.6) for scenarios with $np > 2000$, we used the *Random Selection* (RS) blocking scheme as described in Davanloo et al. (2015). Tables 1 and 2 show the results for p-variate GRF models with $p = 2$ and $p = 5$, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

			N=1		N=10		N=40		
d	n	Method	$\|\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}^*\|_2$	$\|\hat{\Gamma}_\ell - \Gamma^*\|_F$	$\|\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}^*\|_2$	$\|\hat{\Gamma}_\ell - \Gamma^*\|_F$	$\|\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}^*\|_2$	$\|\hat{\Gamma}_\ell - \Gamma^*\|_F$	Time (sec)
2	100	GSPS	0.43	0.89	0.34	0.66	0.21	0.53	14.9
		MLE	0.38	0.78	0.36	0.70	0.26	0.61	21.3
	500	GSPS	0.39	0.81	0.29	0.60	0.13	0.43	312.3
		MLE	0.37	0.83	0.32	0.62	0.19	0.50	496.1
	1000	GSPS	0.33	0.73	0.23	0.57	0.08	0.34	2342.5
		MLE	0.32	0.74	0.28	0.58	0.11	0.40	3216.5
5	100	GSPS	0.49	0.96	0.38	0.71	0.26	0.56	18.9
		MLE	0.46	0.93	0.42	0.71	0.36	0.61	36.5
	500	GSPS	0.44	0.88	0.33	0.69	0.29	0.53	527.4
		MLE	0.46	0.89	0.38	0.67	0.34	0.59	1023.4
	1000	GSPS	0.40	0.81	0.30	0.62	0.29	0.50	2987.3
		MLE	0.43	0.92	0.35	0.66	0.34	0.56	6120.8
10	100	GSPS	0.55	1.05	0.39	0.82	0.35	0.58	29.1
		MLE	0.57	1.02	0.56	0.89	0.53	0.69	75.2
	500	GSPS	0.47	0.99	0.35	0.73	0.31	0.49	613.8
		MLE	0.54	1.00	0.53	0.81	0.50	0.58	4125.6
	1000	GSPS	0.41	0.89	0.31	0.71	0.29	0.43	4920.5
		MLE	0.51	0.97	0.49	0.76	0.47	0.50	7543.3

Table 1: Comparison of GSPS vs. MLE for p=2 response variables

			N=1		N=10		N=40		
d	n	Method	$\|\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}^*\|_2$	$\|\hat{\Gamma}_\ell - \Gamma^*\|_F$	$\|\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}^*\|_2$	$\|\hat{\Gamma}_\ell - \Gamma^*\|_F$	$\|\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}^*\|_2$	$\|\hat{\Gamma}_\ell - \Gamma^*\|_F$	Time (sec)
2	100	GSPS	0.66	1.43	0.38	0.91	0.30	0.76	17.2
		MLE	0.62	1.40	0.57	1.30	0.41	1.28	26.3
	500	GSPS	0.58	1.35	0.35	0.87	0.27	0.73	363.4
		MLE	0.57	1.32	0.51	1.24	0.39	1.15	512.5
	1000	GSPS	0.49	1.24	0.31	0.82	0.24	0.70	2835.4
		MLE	0.49	1.22	0.42	1.19	0.33	1.10	3913.7
5	100	GSPS	0.73	1.49	0.50	0.92	0.39	0.79	25.6
		MLE	0.71	1.47	0.62	1.36	0.49	1.35	53.1
	500	GSPS	0.60	1.41	0.44	1.00	0.36	0.75	665.6
		MLE	0.64	1.43	0.54	1.26	0.44	1.24	1424.3
	1000	GSPS	0.54	1.32	0.39	1.06	0.31	0.74	3783.6
		MLE	0.63	1.36	0.47	1.20	0.38	1.17	7346.7
10	100	GSPS	0.77	1.57	0.59	0.98	0.52	0.85	45.3
		MLE	0.79	1.60	0.67	1.39	0.61	1.43	87.2
	500	GSPS	0.65	1.47	0.54	1.03	0.46	0.81	717.6
		MLE	0.74	1.56	0.60	1.31	0.52	1.37	4994.3
	1000	GSPS	0.59	1.39	0.49	1.08	0.42	0.75	6001.3
		MLE	0.66	1.48	0.53	1.27	0.45	1.29	8223.1

Table 2: Comparison of GSPS vs. MLE for p=5 response variables

For fixed n, the parameter estimation error increases with the dimension of the input space d, which is reasonable due to higher number of parameters

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

in the anisotropic correlation function. Furthermore, the errors increase with p, the number of responses. As expected, increasing the point density n helps in improving the estimation of the parameters, a result in accordance to the expected effect of infill asymptotics.

Overall, the GSPS method results in better parameter estimates compared to MLE with relative performance improvements becoming more obvious as p and d increase. Further, as the number of realizations N increases, GSPS performs consistently better than MLE. The robust performance of the proposed method is theoretically guaranteed for $N \geq N_0$ by Theorem 4.

4.2. Prediction consistency

To evaluate prediction performance, we compared the GSPS method against multiple *univariate* SPS (mSPS) and against the Convolved Multiple output Gaussian Process (CMGP) method of Alvarez and Lawrence (2011). Given the size of the training data n, none of the approximations in Alvarez and Lawrence (2011) with induced points were used.

For ten replicates, we simulated N independent realizations of the GRF defined at the beginning of Section 4, under different scenarios, to learn the model parameters. We also simulated the p-variate response over a fixed set of $n_0 = 1000$ test locations per replicate. The mean of the conditional distribution $p(\mathbf{y}(\mathbf{x}_0)|\{\mathbf{y}^{(r)}\}_{r=1}^{N}, \mathcal{D}^x)$ was used to predict at these test locations.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

The mean of Mean Squared Prediction Error (MSPE) over ten replicates, p outputs, and n_0 test points are reported for $p = 2$ and $p = 5$ in Tables 3 and 4, respectively.

Table 3: MSPE comparison for $p = 2$ response variables

d	n	Method	$N = 1$	$N = 10$	$N = 40$
2	100	mSPS	7.02	2.68	2.08
		GSPS	6.71	2.12	1.44
		CMGP	6.40	2.39	1.61
	400	mSPS	6.76	2.22	1.87
		GSPS	5.53	1.89	0.91
		CMGP	5.16	2.04	1.33
5	100	mSPS	7.12	3.09	2.39
		GSPS	6.98	2.45	1.52
		CMGP	6.74	2.95	1.99
	400	mSPS	7.34	3.04	2.24
		GSPS	5.88	2.45	1.05
		CMGP	6.32	2.89	1.73
10	100	mSPS	7.83	4.15	3.23
		GSPS	7.11	3.34	2.02
		CMGP	6.97	3.67	2.39
	400	mSPS	7.65	3.53	2.65
		GSPS	6.13	2.96	1.22
		CMGP	6.63	3.32	2.28

Table 4: MSPE comparison for $p = 5$ response variables

d	n	Method	$N = 1$	$N = 10$	$N = 40$
2	100	mSPS	7.83	4.42	3.08
		GSPS	7.05	3.89	2.11
		CMGP	6.74	3.71	2.49
	400	mSPS	7.51	3.78	2.18
		GSPS	6.81	2.96	1.32
		CMGP	6.23	3.36	2.03
5	100	mSPS	8.54	5.30	3.32
		GSPS	7.19	4.43	2.01
		CMGP	7.10	4.97	2.86
	400	mSPS	8.22	4.15	2.63
		GSPS	7.00	3.10	1.45
		CMGP	7.45	4.04	2.65
10	100	mSPS	9.23	5.67	3.43
		GSPS	7.23	4.68	2.19
		CMGP	8.53	5.25	3.24
	400	mSPS	8.54	4.24	2.94
		GSPS	7.08	3.23	1.63
		CMGP	7.82	4.20	2.87

The mean of the Mean Squared Prediction Error (MSPE) comparison of multiple SPS (mSPS), Generalized SPS (GSPS) and Convolved Multiple Gaussian Process (CMGP) of Alvarez and Lawrence (2011) for p response variables

The prediction performance of the GSPS method is almost always better than that of the mSPS method. Learning the cross-covariance between different responses provides additional useful information that helps improve the prediction performance of the joint model, GSPS. Comparing GSPS vs. CMGP, we observe relatively better performance of CMGP over GSPS when $N = 1$ in a lower-dimensional input space, e.g., $(N, d) = (1, 2)$. As n, the number of locations, increases, the GSPS predictions become bet-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

ter than CMGP even if $N = 1$, e.g., for $(N, d) = (1, 5)$, GSPS does better than CMPG for $n = 400$. The prediction performance of GSPS improves significantly with increasing N, the number of realizations of the process. In $d = 10$ dimensional space, GSPS performs consistently better, even when $N = 1$ for both $p = 2$ and $p = 5$. However, CMGP with 50 inducing points is significantly faster than GSPS in the learning phase.

4.3. Performance on a real data set

We used a data set to compare the prediction performance of GSPS with the naive method of using multiple univariate SPS (mSPS) fits, and with the two approximation methods proposed in Alvarez and Lawrence (2011). The data set consists of n=9635 (x, y, z) measurements obtained by a laser scanner from a free-form surface of a manufactured product. Del Castillo et al. (2015) proposed modeling each coordinate, separately, as a function of the corresponding (u, v) surface coordinates (obtained using the ISOMAP algorithm by Tenenbaum et al. (2000)). These (u, v) coordinates are selected such that their pairwise Euclidean distance is equal to the pairwise geodesic distances between their corresponding (x, y, z) points along the surface. We modeled $(x(u, v), y(u, v), z(u, v))$ as a multivariate GRF using GSPS, and compared against fitting $p = 3$ independent *univariate* GRF using the SPS method (mSPS).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

Given the large size of the data set, n=9635, we used the *Random Selection* blocking scheme as described in Davanloo et al. (2015) for varying number of blocks; hence, there are different number of observations per block. Table 5 reports the MSPE and the corresponding standard errors (std. error) obtained from 10-fold cross validation.

Method	n/block	MSPE	std. error
mSPS	100	0.0932	0.0047
mSPS	500	0.0621	0.0021
mSPS	1000	0.0842	0.0013
GSPS	100	0.0525	0.0023
GSPS	500	0.0167	0.0012
GSPS	1000	0.0285	0.0019

Table 5: 10-fold cross validation to evaluate prediction performance of multiple SPS (mSPS) and GSPS for the metrology data set with n=9635 data points.

According to Table 5, the best predictions are obtained when the number of observations per block is 500. We also compared the GSPS method with 500 data points per block against the two approximation methods developed in Alvarez and Lawrence (2011), the Full Independent Training Conditional (FITC) method and the Partially Independent Training Conditional (PITC) method. These methods were run for different numbers of inducing points $K \in \{100, 500, 1000\}$ with their initial locations selected at random. The locations of the inducing points along with the hyper-parameters' values are found by maximizing the likelihood through a scaled conjugate gradient method proposed by Alvarez and Lawrence

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

(2011). Results are reported in Table 6.

Method	MSPE	std. error
mSPS (n/block=500)	0.0621	0.0021
GSPS (n/block=500)	0.0167	0.0012
FITC (K=100)	0.0551	0.0042
FITC (K=500)	0.0463	0.0011
FITC (K=1000)	0.0174	0.0010
PITC (K=100)	0.0698	0.0062
PITC (K=500)	0.0421	0.0021
PITC (K=1000)	0.0197	0.0007

Table 6: 10-fold cross validation to compare prediction performance of mSPS, GSPS vs. FITC and PITC methods by Alvarez and Lawrence (2011) for the metrology data set with n=9635 data points

The best prediction performance for both FITC and PITC approximations are obtained for the larger K values as this represents a better approximation of the underlying GRF. The GSPS method performed better than FITC and PITC for all K parameter choices. As expected, fitting p univariate GRF models (mSPS) performed worse than the multivariate methods.

5. Conclusions and future research

We presented a new two-stage estimation method to fit multivariate Gaussian Random Field (GRF) models with separable covariance functions. Theoretical convergence rates for the estimated between-response covariance matrix and the estimated correlation function parameter are established with respect to the number of process realizations. Numerical studies confirm the theoretical results.

We considered separable covariance functions. Future research may

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

consider non-separable covariance functions, e.g., convolutions of covariance functions, or kernel convolutions. As another potential future work, we propose estimating the cross-covariance matrix $\hat{\Gamma}$ at the outset by solving $\hat{\Gamma} = \text{argmin}_{\Gamma}\{\|\Gamma - \frac{1}{n}\sum_{i=1}^{n} S^{ii}\|_F : \Gamma \succeq \epsilon \mathbf{I}\}$. Then we solve, as the new STAGE-I,

$$\hat{P}_\rho = \underset{P_\rho}{\text{argmin}} \left\langle S, P_\rho \otimes \hat{\Gamma}^{-1} \right\rangle - \log\det(P_\rho \otimes \hat{\Gamma}^{-1}) + \alpha \left\langle G \otimes (\mathbf{1}_p \mathbf{1}_p^\top), |P_\rho \otimes \hat{\Gamma}^{-1}| \right\rangle$$
$$\text{s.t.} \qquad a^* \lambda_{\max}(\hat{\Gamma}) I \preceq P_\rho \preceq b^* \lambda_{\min}(\hat{\Gamma}) I.$$

Here $\log\det(P_\rho \otimes \hat{\Gamma}^{-1}) = p \log\det(P_\rho) - n \log\det(\hat{\Gamma})$. Hence, there exists some $S_\rho,\ G_\rho \in \mathbb{S}^n$, which can be computed very efficiently, such that

$$\hat{P}_\rho = \underset{P_\rho}{\text{argmin}} \left\{ \langle S_\rho, P_\rho \rangle - p \log\det(P_\rho) + \alpha \langle G_\rho, |P_\rho| \rangle : \ a^* \lambda_{\max}(\hat{\Gamma}) I \preceq P_\rho \preceq b^* \lambda_{\min}(\hat{\Gamma}) I \right\}.$$

Such an approach would be much easier to solve in terms of computational complexity – the overall complexity is $\mathcal{O}(\log(1/\epsilon) n^3)$ for this STAGE-I problem. Further work could then be devoted to proving consistency of the resulting estimator and its rate could be compared to $\log(1/\epsilon^2)$ of GSPS.

Supplementary Materials

The proofs for Theorems 2 and 4 are provided in the online supplementary materials.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

References

Alvarez, M. A., and Lawrence, N. D. (2011). Computationally efficient convolved multiple output Gaussian processes. Journal of Machine Learning Research, 1459-1500.

Ankenman, B., Nelson, B. L., and Staum, J. (2010). Stochastic Kriging for Simulation Metamodeling. Operations Research, 58, 371-382.

Banerjee, S., and Gelfand, A. E. (2002). Prediction, interpolation and regression for spatially misaligned data. Sankhya: The Indian Journal of Statistics, Series A, 227-245.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). Hierarchical modeling and analysis for spatial data. Boca Raton, FL: CRC Press.

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

Bhat, K., Haran, M., and Goes, M. (2010). Computer model calibration with multivariate spatial output: A case study. Frontiers of Statistical Decision Making and Bayesian Analysis, 168-184.

Bonilla, E.V., Chai, K.M.A., and Williams, C.K.I. (2007). Multi-task Gaussian Process Prediction. NIPs. Vol. 20.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 3(1), 1–122.

Boyle, P. and Frean, M. R. (2004). Dependent Gaussian Processes. NIPS vol. 17, pp. 217-224.

Cands, E. J. (2006). Compressive sampling. Proceedings of the international congress of mathematicians. Vol. 3.

Castellanos, L., Vu, V. Q., Perel, S., Schwartz, A. B., and Kass, R. E. (2015). A multivariate gaussian process factor model for hand shape during reach-to-grasp movements. Statistica Sinica, 5-24.

Cressie, N., (2015). Statistics for spatial data, 2nd ed. John Wiley & Sons.

Cressie, N., and Wikle, K. (2011). Statistics for Spatio-Temporal Data. John Wiley & Sons.

Davanloo Tajbakhsh, S., Aybat, N. S., Del Castillo, E. (2015). Sparse Precision Matrix Selection for Fitting Gaussian Random Field Models to Large Data Sets. *arXiv:1405.5576*.

Del Castillo, E., Colosimo, B., and Davanloo Tajbakhsh, S. (2015). Geodesic Gaussian Processes for the Reconstruction of a 3D Free-Form Surface, Technometrics, 57:1, pp. 87–99.

Deng, W., and Yin, W. (2015). On the Global and Linear Convergence of the Generalized Alternating Direction Method of Multipliers. Journal of Scientific Computing, 66(3), 889-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

916.

Forrester, A., Sobester, A., and Keane, A. (2008). Engineering design via surrogate models. John Wiley & Sons.

Fulgstad, G. A., Lindgren, F., Simpson, D., and Rue, H. (2015). Exploring a new class of non-stationary spatial Gaussian random fields with varying anisotropy. Statistica Sinica, 25, 115-133.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. Test, 13(2), 263-312.

Gelfand, A. E., and Banerjee, S. (2010). Multivariate spatial process models. Handbook of Spatial Statistics, 495-515.

Genton, M. G., and Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. Statistical Science, 30(2), 147-163.

Jafard S. (1990) Proprietes des matrices bien localisees pres de leur digonale et quelques applications. In Annales de l'IHP Analyse non Lineaire, 7:461–476.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer Berlin Heidelberg.

Kleijnen, J.P.C. (2010). Design and analysis of simulation experiments. NY: Springer.

Li, B., Genton, M. G., and Sherman, M. (2008). Testing the covariance structure of multivariate random fields. Biometrika, 813-829.

Lin, Y.P., (2008). Multivariate geostatistical methods to identify and map spatial variations of soils heavy metals. Environmental Geology, 42, pp. 1-10.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4), 423–498.

Mardia, K. V., and Goodall, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. Multivariate environmental statistics, 6(76), 347-385.

Ok, E. A. (2007). Real Analysis with Economic Applications. Princeton University Press.

Pourhabib, A., Faming L., and Ding, Y. (2014). Bayesian site selection for fast Gaussian process regression. IIE Transactions 46.5: 543–555.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electronic Journal of Statistics, 5, 935-980.

Rasmussen, C.E., and Williams, C.K. (2006). Gaussian Processes for Machine Learning, MIT Press.

Rue, H. and Held, L. (2005). Gaussian Markov Random Fields: Theory and Applications. NY:

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0091

Chapman & Hall/CRC.

Santner, T. J., Williams, B. J., and Notz, W. I., (2003). The design and analysis of computer experiments. NY: Springer.

Snelson, E., and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. Advances in Neural Information Processing Systems. 18, 1257.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000), A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science, 290, 2319–2323.

Urban, N. M., and Keller, K. (2010), Probabilistic Hindcasts and Projections of the Coupled Climate, Carbon Cycle, and Atlantic Meridional Overturning Circulation Systems: A Bayesian Fusion of Century-Scale Observations With a Simple Model, Tellus A, 62, 737–750.

Wang, B., and Chen, T. (2015). Gaussian process regression with multiple response variables. Chemometrics and Intelligent Laboratory Systems, 142, 159-165.

Dept. of Integrated Systems Engineering, Columbus OH 43210 USA
E-mail: davanloo-tajbakhsh.1@osu.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??)