ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2014)
76, *Part* 4, *pp.* 817–832

On semiparametric inference of geostatistical models via local Karhunen–Loève expansion

Tingjin Chu,

Renmin University of China, Beijing, People's Republic of China

Haonan Wang

Colorado State University, Fort Collins, USA

and Jun Zhu

University of Wisconsin, Madison, USA

[Received April 2012. Final revision July 2013]

Summary. We develop a semiparametric approach to geostatistical modelling and inference. In particular, we consider a geostatistical model with additive components, where the form of the covariance function of the spatial random error is not prespecified and thus is flexible. A novel, local Karhunen–Loève expansion is developed and a likelihood-based method is devised for estimating the model parameters and statistical inference. A simulation study demonstrates sound finite sample properties and a real data example is given for illustration. Finally, the theoretical properties of the estimates are explored and, in particular, consistency results are established.

Keywords: Geostatistics; Random field; Semiparametric methods; Spatial statistics

1. Introduction

Geostatistics are used in many scientific studies that involve analysis of spatially correlated data in a spatial domain (see, for example, Cressie (1993) and Stein (1999)). A geostatistical model, in its general form, is a random field for an attribute of interest such that the random field is a stochastic process over a continuous index within the spatial domain. On the basis of geostatistical data sampled at point locations, statistical inference about the geostatistical model can be drawn. The main purpose of this paper is to develop a novel semiparametric approach to spatial modelling and statistical inference that accounts for spatial dependence in a more robust manner and carries out the computation efficiently.

For a spatial linear model, Mardia and Marshall (1984) considered maximum likelihood estimates of the model parameters. The computational complexity of these maximum likelihood estimates for a sample of size N, however, is of the order of N^3, making the computation demanding for large N. To reduce such a computational burden, various methods based on approximations have been developed. One such method, covariance tapering, rescales the spatial correlation function by a weight function of the distance between two locations, effectively

Address for correspondence: Haonan Wang, Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA.
E-mail: wanghn@stat.colostate.edu

 1369–7412/14/76817

truncating the spatial correlation to zero when the distance exceeds a certain threshold. The resulting tapered covariance matrix as an approximation of the true covariance matrix is sparse and thus fast to compute at an appropriately chosen threshold (see, for example, Furrer *et al.* (2006), Kaufman *et al.* (2008), Du *et al.* (2009) and Chu *et al.* (2011)). Alternatively, Caragea and Smith (2007) partitioned the spatial domain into blocks and approximated the likelihood function by an estimating function that separates variability within blocks and between blocks. In a so-called small block case, the spatial processes in different small blocks are assumed to be independent, giving rise to a block diagonal covariance matrix that eases computation. Furthermore, the theoretical properties of the small block method were established under certain conditions. See also Vecchia (1988) and Stein *et al.* (2004).

The aforementioned methods, however, assume a parametric form for the spatial covariance function, on which the performance of statistical inference and the asymptotic results hinge. In contrast, semiparametric modelling offers an attractive alternative, as the spatial covariance function does not need to be prespecified. The corresponding approach tends to be more flexible and potentially more robust against model misspecification (see, for example, Im *et al.* (2007), Cressie and Johannesson (2008) and Zhang and Wang (2010)). For example, Cressie and Johannesson (2008) considered a flexible family of non-stationary spatial covariance functions and developed fixed rank kriging. Despite the added model flexibility, such methods primarily focus on spatial interpolation and there appears to be little or no theoretical backing. Thus, it is of interest to develop innovative semiparametric methods further in general and to explore their theoretical properties.

In this paper, we aim to develop a semiparametric approach to geostatistical modelling and inference. In particular, we consider a geostatistical model with additive components, namely a fixed mean possibly in the form of linear regression and Gaussian random errors. The spatial covariance function is left unspecified and, in particular, a novel local Karhunen–Loève expansion is developed to approximate the spatial random error. In addition, we devise a likelihood-based method for estimating the model parameters and drawing inference. The computational algorithm that is developed utilizes both Newton–Raphson iteration on a Stiefel manifold recently developed by Peng and Paul (2009) and the existing computational method for linear mixed models (see, for example, Pinheiro and Bates (2000)). Our approach applies to estimation of regression coefficients, selection of covariates and non-parametric estimation of the covariance function, as well as spatial interpolation. Although we approximate the likelihood function by employing a technique that is similar to the small block idea, our method does not assume a parametric form for the spatial covariance function like Caragea and Smith (2007). Finally, although more model flexibility is attained, it becomes substantially more challenging to establish the theoretical properties of semiparametric methods, which is an issue that is often not pursued in the existing literature. Here, we make an attempt to establish the consistency of likelihood-based estimates of regression coefficients and spatial covariance function.

The remainder of the paper is organized as follows. In Section 2, we describe a general geostatistical model and a local Karhunen–Loève expansion for the spatial random error. In Section 3, we develop a likelihood-based approximate method for parameter estimation and a modification of the estimation to increase the accuracy of the approximation. Methods for spatial interpolation and model selection are also developed. In Section 4, a simulation study is given to investigate the finite sample properties of the inference in comparison with several alternative approaches, followed by a real data example. We establish the consistency of the estimates in Section 5 and give a technical proof in Appendix A. Other technical details are given as on-line supplementary materials in a separate document.

2. Random-field model

Let R be a spatial domain of interest in $\mathbb{R}^d$, where $d \geqslant 1$ denotes the dimension of space. The following model for a random field $\{y(\mathbf{s}) : \mathbf{s} \in R\}$ is considered:

$$y(\mathbf{s}) = \mu(\mathbf{s}) + \varepsilon_1(\mathbf{s}) + \varepsilon_2(\mathbf{s}), \qquad \mathbf{s} \in R, \tag{1}$$

where $\mu(\mathbf{s})$ is an unknown mean function of location $\mathbf{s}$. Furthermore, the error $\varepsilon_1(\cdot)$ is assumed to be a stationary Gaussian process with mean 0 and a covariance function $\gamma(\mathbf{s} - \mathbf{s}')$, where $\mathbf{s}, \mathbf{s}' \in R$. The second error term $\varepsilon_2(\cdot)$ is assumed to be independently and identically distributed $N(0, \sigma^2)$ and independent of $\varepsilon_1(\cdot)$. i.e. the random field $y(\cdot)$ is decomposed into three additive components: a large-scale trend $\mu(\cdot)$, a small-scale spatial variation $\varepsilon_1(\cdot)$ and a measurement error $\varepsilon_2(\cdot)$; see Cressie (1993) for more details.

2.1. Local Karhunen–Loève expansion

Assume that the spatial domain R is compact and the error process $\varepsilon_1(\cdot)$ is square integrable over R. The Karhunen–Loève expansion of $\varepsilon_1(\mathbf{s})$ can be expressed as

$$\varepsilon_1(\mathbf{s}) = \sum_{j=1}^{\infty} \bar{\xi}_j \bar{\varphi}_j(\mathbf{s}), \qquad \mathbf{s} \in R,$$

where $\{\bar{\xi}_j : j = 1, 2, \ldots\}$ is a sequence of independent random variables and $\bar{\xi}_j \sim N(0, \bar{\lambda}_j)$, with variances $\bar{\lambda}_1 \geqslant \bar{\lambda}_2 \geqslant \ldots \geqslant 0$. Furthermore, $\{\bar{\varphi}_j(\cdot) : j = 1, 2, \ldots\}$ is a sequence of orthonormal eigenfunctions over R such that $\int_R \bar{\varphi}_j(\mathbf{s}) \bar{\varphi}_{j'}(\mathbf{s})\, \mathrm{d}\mathbf{s}$ equals 1 when $j = j'$ and 0 otherwise. For a review of the Karhunen–Loève expansion, see Adler and Taylor (2007) and the references therein.

For a random field, the application of the Karhunen–Loève expansion is limited. There is usually only one realization of the random field and, consequently, the variances λ_k may not be estimated consistently. To circumvent this issue, we introduce a notion of *local Karhunen–Loève expansion*.

First, we assume that the spatial domain R can be partitioned into K compact subdomains with identical shape, namely $R_1, \ldots, R_K$. Denote $R_k = R_1 + \mathbf{v}_k$, for a d-dimensional vector $\mathbf{v}_k$. Restricting the error $\varepsilon_1(\cdot)$ to each of the K subdomains gives rise to K error processes that are identically distributed, but not independent of each other, owing to the stationarity of the error process $\varepsilon_1(\cdot)$.

Next, we apply the Karhunen–Loève expansion to the error process within each subdomain. More specifically, we have

$$\varepsilon_1(\mathbf{s}) = \sum_{j=1}^{\infty} \xi_{j,k} \varphi_{j,k}(\mathbf{s}), \qquad \mathbf{s} \in R_k. \tag{2}$$

Here, for a fixed k, $\{\xi_{j,k}\}_{j=1}^{\infty}$ is a sequence of independent random variables such that $\xi_{j,k} \sim N(0, \lambda_j)$, with variances $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant 0$. For a given j, $\{\xi_{j,k}\}_{k=1}^{K}$ are identically, although not necessarily independently, distributed across all subdomains. Moreover, $\{\varphi_{j,k}(\cdot)\}_{j=1}^{\infty}$ is a sequence of orthonormal eigenfunctions on subdomain R_k. More importantly, as a direct consequence of stationarity of $\varepsilon_1(\cdot)$ over the domain R, for any given j and any $\mathbf{s} \in R_k$, $\varphi_{j,k}(\mathbf{s}) = \varphi_{j,1}(\mathbf{s} - \mathbf{v}_k)$, i.e. the orthonormal eigenfunctions $\varphi_{j,k}(\mathbf{s})$ are the same across all K subdomains up to a constant shift.

In equation (2), the equivalence is defined in the L^2-sense. In practice, however, we approximate the error process $\varepsilon_1(\cdot)$ expressed in an infinite series by a finite sum, say J terms, i.e. we let $\varepsilon_1(\mathbf{s}) \approx \Sigma_{j=1}^{J} \xi_{j,k} \varphi_{j,k}(\mathbf{s}) = \Sigma_{j=1}^{J} \xi_{j,k} \varphi_{j,1}(\mathbf{s} - \mathbf{v}_k)$, for $\mathbf{s} \in R_k$.

2.2. Approximation of the eigenfunctions

An essential component of implementing the local Karhunen–Loève expansion is the computation of the eigenfunctions $\varphi_{j,k}(\cdot)$, which ideally can be obtained by solving integral equations, i.e. λ_k and $\varphi_{j,k}(\cdot)$ can be found by solving $\int_{R_k}\gamma(\mathbf{s}-\mathbf{s}')\varphi_{j,k}(\mathbf{s}')\,\mathrm{d}\mathbf{s}'=\lambda_j\varphi_{j,k}(\mathbf{s})$, for $j=1,2,\ldots$ and $\mathbf{s},\mathbf{s}'\in R_k$. In general, however, such eigenfunctions cannot be expressed explicitly except for certain special cases. Here, we propose to approximate these eigenfunctions by a set of known orthonormal basis functions.

Let $\boldsymbol{\phi}_1(\mathbf{s})=(\phi_{1,1}(\mathbf{s}),\ldots,\phi_{M,1}(\mathbf{s}))^{\mathrm{T}}$ be an M-dimensional vector of orthonormal basis functions on R_1. We propose to approximate the eigenfunctions $\boldsymbol{\varphi}_1(\mathbf{s})=(\varphi_{1,1}(\mathbf{s}),\ldots,\varphi_{J,1}(\mathbf{s}))^{\mathrm{T}}$ from the family $\{\mathbf{B}^{\mathrm{T}}\boldsymbol{\phi}_1(\mathbf{s}):\mathbf{B}^{\mathrm{T}}\mathbf{B}=\mathbf{I}_J\}$, where $\mathbf{B}$ is an $M\times J$ coefficient matrix and $\mathbf{I}_J$ is a $J\times J$ identity matrix. Suppose that an element from this approximating family, say $\mathbf{B}^{*\mathrm{T}}\boldsymbol{\phi}_1(\mathbf{s})$, provides an adequate approximation of $\boldsymbol{\varphi}_1(\mathbf{s})$. Then, on the subdomain R_k, the eigenfunctions $\boldsymbol{\varphi}_k(\mathbf{s})=(\varphi_{1,k}(\mathbf{s}),\ldots,\varphi_{J,k}(\mathbf{s}))^{\mathrm{T}}$ can be well approximated by $\mathbf{B}^{*\mathrm{T}}\boldsymbol{\phi}_k(\mathbf{s})$, where $\boldsymbol{\phi}_k(\mathbf{s})=\boldsymbol{\phi}_1(\mathbf{s}-\mathbf{v}_k)$.

Combining the truncated local Karhunen–Loève expansion and the eigenfunction approximation, we have

$$y(\mathbf{s})\approx\mu(\mathbf{s})+\boldsymbol{\phi}_k(\mathbf{s})^{\mathrm{T}}\mathbf{B}\boldsymbol{\xi}_k+\varepsilon_2(\mathbf{s}),\qquad\mathbf{s}\in R_k,\tag{3}$$

where $\boldsymbol{\xi}_k=(\xi_{1,k},\ldots,\xi_{J,k})^{\mathrm{T}}$ is a J-dimensional vector of random variables such that $\boldsymbol{\xi}_k\sim N(\mathbf{0},\boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda}=\mathrm{var}(\boldsymbol{\xi}_k)=\mathrm{diag}(\lambda_1,\ldots,\lambda_J)$, i.e. the error process $\varepsilon_1(\cdot)$ is approximated by a sum of independently distributed Gaussian random variables, which has substantial computational advantages, as we shall demonstrate later. In the special case of $d=1$, approximation (3) is a functional response model in functional data analysis (Yao *et al.*, 2005; Yao and Lee, 2006).

For the choice of basis functions, we consider the orthonormalized cubic B-spline basis for $d=1$ and orthonormalized radial basis function for $d\geqslant 2$ (Buhmann, 2003). In particular, the radial basis function is defined as $g(c\|\mathbf{s}-\boldsymbol{\kappa}\|)$, where g is a prespecified continuous function, $\mathbf{s}\in R$, $\boldsymbol{\kappa}$ is a knot point, $\|\cdot\|$ denotes the Euclidean distance and $c>0$ is a constant. Commonly used choices for g include $g(h)=h^2\log(h)$, which leads to thin plate splines, and $g(h)=\exp(-h^2)$, which results in Gaussian radial splines. In practice, the vector of basis functions can be orthonormalized.

3. Statistical inference

3.1. Constrained likelihood-based estimation

Henceforth, we shall restrict our attention to the case of model (1) with a linear trend, i.e. $\mu(\mathbf{s})=\mathbf{x}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta}$, where $\mathbf{x}(\mathbf{s})=(x_1(\mathbf{s}),\ldots,x_p(\mathbf{s}))^{\mathrm{T}}$ is a p-dimensional vector of covariates at location $\mathbf{s}$ and $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^{\mathrm{T}}$ is a p-dimensional vector of regression coefficients. Under this setting, model (3) resembles a linear mixed model but is subject to constraints due to the orthonormality of the basis functions. Consequently, standard statistical methods for estimating the parameters of a linear mixed model are not directly applicable. Peng and Paul (2009) considered a similar problem for functional data and implemented a manifold version of the Newton–Raphson method to optimize a likelihood-based criterion with such constraints. In addition, Paul and Peng (2009) established the consistency of the resulting estimates under the assumption that there are independent replicates per subject, which generally does not hold for geostatistical data. Here, we develop a new estimation procedure as follows.

Suppose that there are N sampling locations in the spatial domain R. Let $\{\mathbf{s}_{k,i}:i=1,\ldots,n_k\}$ denote the sampling locations in subdomain R_k and, thus, $\Sigma_{k=1}^{K}n_k=N$. Let $\mathbf{X}_k=(\mathbf{x}(\mathbf{s}_{k,1}),\ldots,\mathbf{x}(\mathbf{s}_{k,n_k}))^{\mathrm{T}}$ denote an $n_k\times p$ design matrix of the covariates and $\boldsymbol{\Phi}_k=(\boldsymbol{\phi}_k(\mathbf{s}_{k,1}),\ldots,\boldsymbol{\phi}_k(\mathbf{s}_{k,n_k}))^{\mathrm{T}}$ denote an $n_k\times M$ matrix of $\boldsymbol{\phi}_k(\cdot)$ evaluated at the sampling locations in the subdomain R_k.

Moreover, let $\mathbf{y}_k = (y(\mathbf{s}_{k,1}), \ldots, y(\mathbf{s}_{k,n_k}))^{\mathrm{T}}$ denote an n_k-dimensional vector of responses in R_k and $\mathbf{y} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_K^{\mathrm{T}})^{\mathrm{T}}$ denote an N-dimensional vector of responses in R, such that $\boldsymbol{\Sigma}_{0k} = \mathrm{var}(\mathbf{y}_k)$ is the true covariance matrix of $\mathbf{y}_k$ and $\boldsymbol{\Sigma}_0 = \mathrm{var}(\mathbf{y})$ is the true covariance matrix of $\mathbf{y}$.

On the basis of model (3), the corresponding approximating covariance matrix of $\boldsymbol{\Sigma}_{0k}$ is $\boldsymbol{\Sigma}_k = \boldsymbol{\Phi}_k^{\mathrm{T}} \mathbf{B} \boldsymbol{\Lambda} \mathbf{B}^{\mathrm{T}} \boldsymbol{\Phi}_k + \sigma^2 \mathbf{I}_{n_k}$. By ignoring the dependence among $\mathbf{y}_k$ in different subdomains, $\boldsymbol{\Sigma}_0$ can be approximated by a block diagonal matrix $\boldsymbol{\Sigma}_{\mathrm{KL}} = \mathrm{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$. Consequently, up to an additive constant, the negative log-likelihood function can be approximated as

$$L_K(\boldsymbol{\beta}, \sigma^2, \mathbf{B}, \boldsymbol{\Lambda}) = (2K)^{-1} \sum_{k=1}^{K} \{(\mathbf{y}_k - \mathbf{X}_k \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_k - \mathbf{X}_k \boldsymbol{\beta}) + \log |\boldsymbol{\Sigma}_k|\}. \tag{4}$$

Note that equation (4) provides a better approximation to the true negative log-likelihood function as the correlation of the observations between subdomains becomes weaker; see Section 5 for further discussion. Let $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\mathbf{B}}, \hat{\boldsymbol{\Lambda}})$ denote the estimates that are obtained from minimizing equation (4).

To carry out the minimization of equation (4), the following iterative algorithm is devised. First, for a given $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Lambda})$, we minimize equation (4) with respect to $\mathbf{B}$ subject to the constraint $\mathbf{B}^{\mathrm{T}} \mathbf{B} = \mathbf{I}_J$. Here, we implement a Newton–Raphson-type algorithm on a Stiefel manifold, which utilizes the intrinsic Riemannian geometric structure of such a manifold (Peng and Paul, 2009). Next, given $\mathbf{B}$, we minimize equation (4) with respect to $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Lambda}$. In the second step, for a fixed $\mathbf{B}$, equation (4) is the log-likelihood function of a linear mixed model and, thus, its minimization is straightforward (see, for example, Pinheiro and Bates (2000)).

An alternative method to minimize equation (4) is to treat the random vector $\boldsymbol{\xi}_k$ as missing data (James *et al.*, 2000). First, an EM algorithm is used to obtain the estimates for $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = \mathrm{var}(\mathbf{B}\boldsymbol{\xi}_k)$. Next, estimates for $(\mathbf{B}, \boldsymbol{\Lambda})$ are obtained through eigendecomposition $\hat{\boldsymbol{\Gamma}} = \hat{\mathbf{B}} \hat{\boldsymbol{\Lambda}} \hat{\mathbf{B}}^{\mathrm{T}}$, and, therefore, the orthogonality of eigenfunctions is ensured. A numerical comparison has been conducted, and both algorithms yield similar results. See section D of the on-line supplemental materials.

3.2. Pretapered and tapered estimates

The block diagonal matrix $\boldsymbol{\Sigma}_{\mathrm{KL}}$ approximates the true covariance matrix by ignoring the dependence of observations between subdomains. Such an approximation has great computational advantages; however, our numerical studies suggest that the amount of error can be large. Here, for a stationary isotropic random field, we propose an approach to recover some of the loss of accuracy that is caused by the approximation of the true covariance matrix.

Under the isotropic assumption, the covariance function of $\varepsilon_1(\mathbf{s})$, $\gamma(\mathbf{s} - \mathbf{s}')$, can be written as a function of the distance between $\mathbf{s}$ and $\mathbf{s}'$, $\gamma(\|\mathbf{s} - \mathbf{s}'\|)$. First, an estimated covariance function $\hat{\gamma}_1(h)$ for a given spatial lag h is obtained by taking an average of the covariance function estimates from minimizing equation (4) over the lag distances that are within a small neighbourhood Δh of h,

$$\hat{\gamma}_1(h) = N_h^{-1} \sum_{k=1}^{K} \sum_{1 \leqslant i \leqslant j \leqslant n_k} \hat{\gamma}(h_{k,ij})\, I\{h_{k,ij} \in [h - \Delta, h + \Delta h]\}, \qquad h \in [0, D],$$

where $h_{k,ij} = \|\mathbf{s}_{k,i} - \mathbf{s}_{k,j}\|$ for $\mathbf{s}_{k,i}, \mathbf{s}_{k,j} \in R_k$, $\hat{\gamma}(h_{ij})$ is the estimated covariance of $\varepsilon(\mathbf{s}_{k,i})$ and $\varepsilon(\mathbf{s}_{k,j})$, which is the (i, j)th entry of $\hat{\boldsymbol{\Sigma}}_k = \boldsymbol{\Phi}_k^{\mathrm{T}} \hat{\mathbf{B}} \hat{\boldsymbol{\Lambda}} \hat{\mathbf{B}}^{\mathrm{T}} \boldsymbol{\Phi}_k + \hat{\sigma}^2 \mathbf{I}_{n_k}$, $I\{\cdot\}$ is an indicator function and $N_h = \Sigma_{k=1}^{K} \Sigma_{1 \leqslant i \leqslant j \leqslant n_k} I\{h_{k,ij} \in [h - \Delta, h + \Delta h]\}$. Here, D is the largest spatial lag at which $\hat{\gamma}_1(h)$ is estimated, and it is usually slightly smaller than the diameter of subdomains.

Since $\hat{\gamma}_1(h)$ is not guaranteed to be positive definite, we implement a further transformation that was proposed by Hall and Patil (1994). In particular, let

$$\psi(\theta) = \int \exp(\mathrm{i}\theta h)\,\hat{\gamma}_1(h)\,\mathrm{d}h,$$

where $\theta \in \mathbb{R}$, and transform $\hat{\gamma}_1(h)$ to

$$\hat{\gamma}_2(h) = (2\pi)^{-1} \int_{\mathbb{R}} \cos(\theta h)\,\hat{\psi}(\theta)\,\mathrm{d}\theta, \tag{5}$$

where $\hat{\psi} = \max\{\psi, 0\}$. The resulting $\hat{\gamma}_2(h)$ will be referred to as a *pretapered estimate* of the covariance function. Although it is positive definite on $[0, D]$, it may not be continuous nor positive definite on $[0, \infty]$. Thus, we further adopt a tapering function $W(h, \omega)$, which is an isotropic auto-correlation function when $h \leqslant \omega$ and 0 when $h > \omega$ for a given threshold distance ω. Compactly supported correlation functions are often used as the tapering functions, such as $W(h, \omega) = (1 - h/\omega)\, I\{h \leqslant \omega\}$, where $I\{h \leqslant \omega\}$ is an indicator function (Wendland, 1995). A tapered estimate for the covariance function $\hat{\gamma}_3(h)$ can be obtained by $\hat{\gamma}_3(h) = \hat{\gamma}_2(h)\, W(h, \omega)$, which is a positive definite covariance function over $[0, \infty]$. The corresponding estimated covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathrm{T}} = [\hat{\gamma}_3(d_{ii'})]_{i,i'=1}^{N}$ is also positive definite, where $d_{ii'}$ is the distance between two sampling locations $\mathbf{s}_i$ and $\mathbf{s}_{i'}$. Using $\hat{\boldsymbol{\Sigma}}_{\mathrm{T}}$, we then update the estimates of $\boldsymbol{\beta}$ and σ^2, which are denoted by $\hat{\boldsymbol{\beta}}_{\mathrm{T}}$ and $\hat{\sigma}^2_{\mathrm{T}}$. Since $\hat{\boldsymbol{\Sigma}}_{\mathrm{T}}$ is a sparse matrix, the computation is fast even for large sample sizes. Moreover, our numerical examples show that the inverse of the resulting covariance matrix, $\hat{\boldsymbol{\Sigma}}_{\mathrm{T}}^{-1}$, is closer to the inverse of the true covariance matrix, $\boldsymbol{\Sigma}_0^{-1}$, when compared with the inverse of the previous covariance matrix estimate, $\hat{\boldsymbol{\Sigma}}_{\mathrm{KL}}^{-1}$.

To estimate the standard deviation of $\hat{\boldsymbol{\beta}}_{\mathrm{T}}$, a direct plug-in method is used to obtain the standard error $\mathrm{se}(\hat{\boldsymbol{\beta}}_{\mathrm{T}}) = \mathrm{diag}\{\hat{\sigma}^2_{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{T}}\mathbf{X})^{-1}\}^{1/2}$. The $100(1-\alpha)\%$ confidence interval for β_i is $[\hat{\beta}_{i,\mathrm{T}} + z(\alpha/2)\,\mathrm{se}(\hat{\beta}_{i,\mathrm{T}}), \hat{\beta}_{i,\mathrm{T}} + z(1-\alpha/2)\,\mathrm{se}(\hat{\beta}_{i,\mathrm{T}})]$, where $\hat{\beta}_{i,\mathrm{T}}$ and $\mathrm{se}(\hat{\beta}_{i,\mathrm{T}})$ are the coefficient estimate and standard error of β_i, which is the ith component of $\hat{\boldsymbol{\beta}}_{\mathrm{T}}$ and $\mathrm{se}(\hat{\boldsymbol{\beta}}_{\mathrm{T}})$ respectively. Here, $z(\cdot)$ is the quantile of the standard normal distribution. Alternatively, resampling techniques, such as the jackknife, can be used to obtain the standard error of $\hat{\boldsymbol{\beta}}_{\mathrm{T}}$. In particular, for pretapered and tapered covariance function estimates, pointwise confidence intervals can be constructed by using a jackknife method.

The tapered estimates $(\hat{\boldsymbol{\beta}}_{\mathrm{T}}, \hat{\sigma}^2_{\mathrm{T}}, \hat{\boldsymbol{\Sigma}}_{\mathrm{T}})$ which were developed in Section 3.2 can be applied to spatial prediction (Cressie, 1993) and variable selection (Wang and Zhu, 2009; Chu *et al.*, 2011). For details, see section A of the supplemental materials.

4. Numerical examples

4.1. Simulation study

We now investigate the finite sample properties of our proposed method by using local Karhunen–Loève expansion denoted as method KL. Four different scenarios are considered, which are combinations of two dimensions ($d = 1$ or $d = 2$) and two true covariance functions (exponential or not). For comparison, we consider three competing methods. The first alternative, ALT_1, is ordinary least squares that ignores spatial dependence. The second alternative, ALT_2, assumes a parametric covariance function and applies maximum likelihood for parameter estimation (Mardia and Marshall, 1984). Third, and last, we consider a 'small block' method, ALT_3, that was proposed by Caragea and Smith (2007). In ALT_2 and ALT_3, we assume that the error term follows an exponential covariance function regardless of the true underlying covariance structure.

4.1.1. $d = 1$, exponential covariance

Let the spatial domain be $R = [0, L]$ with $L = 30$, 60, 90. For a fixed sampling density 10, the corresponding sample size N is 300, 600 and 900 respectively. The linear regression model has seven covariates with regression coefficients $\beta = (4, 3, 2, 1, 0, 0, 0)^{\mathrm{T}}$. The covariates are generated from standard normal distributions with a cross-covariate correlation of 0.5. In addition, we standardize each covariate to have sample mean 0 and sample variance 1, and the response to have a sample mean 0. Consequently, there is no intercept in this model. For spatial dependence, we generate the error $\varepsilon_1(s)$ at the sampling location s from a zero-mean stationary and isotropic Gaussian process with an exponential covariance function $\gamma(h) = \sigma_1^2 \exp(-h/c_r)$, where σ_1^2 is a variance component and c_r is a range parameter. In addition, the measurement errors $\varepsilon_2(s)$ are independently generated from $N(0, \sigma_2^2)$. Let $\sigma_1^2 = 16$, $\sigma_2^2 = 4$ and $c_r = 2$. For methods KL and ALT_2, the subdomains are set to be intervals of equal length 6.

For each sample size N, we simulate 100 data sets and, for each data set, we estimate β by using our KL method as well as the three alternatives. The mean and standard deviation of the resulting estimates are reported in Table A of the on-line supplemental materials. The results show that ALT_2 performs the best, as expected. As the sample size increases, the performances of all four estimates improve in terms of smaller biases and variances. Moreover, by accounting for spatial dependence, parameter estimation using methods ALT_2, ALT_3 and KL all outperform

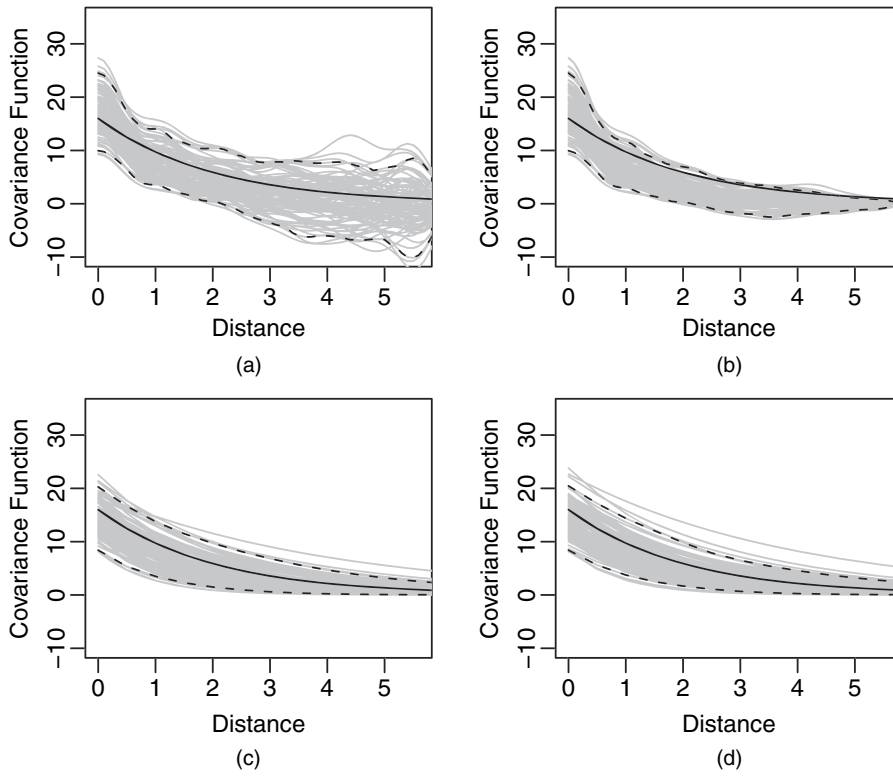

Fig. 1. Estimated covariance functions and 95% pointwise simulation intervals using our proposed method with both (a) pretapered estimates and (b) tapered estimates, (c) maximum likelihood ALT_2 and (d) a small block method ALT_3: ———, true exponential covariance function; ▭, estimated covariance function from each simulated data set; – – –, pointwise simulation intervals

ALT_1. The estimates from ALT_3 and KL tend to those of ALT_2, which suggests that the effect of the covariance matrix approximation becomes smaller as the sample size increases.

The estimated covariance functions along with 95% pointwise simulation intervals by method KL are illustrated in Fig. 1. In particular, the estimates are quite close. Here, 95% pointwise simulation intervals are defined as $[\{\hat{\gamma}_{(97)}(h)+\hat{\gamma}_{(98)}(h)\}/2, \{\hat{\gamma}_{(2)}(h)+\hat{\gamma}_{(3)}(h)\}/2]$ for $h \in [0, D]$, where $\hat{\gamma}_{(i)}(h)$ is the ith largest value of $\{\hat{\gamma}_{[i]}(h) : i = 1, \ldots, 100\}$, and $\hat{\gamma}_{[i]}(h)$ is the estimate for $\gamma(h)$ from the ith simulated data set. For the pretapered estimate $\hat{\gamma}_2(h)$, the true covariance function falls well within the 95% pointwise simulation intervals. However, when the spatial lag increases, the pointwise simulation intervals do not narrow as in methods ALT_2 and ALT_3, owing to a non-parametric form of the error process. For the tapered estimate of covariance function $\hat{\gamma}_3(h)$, the true covariance functions fall within 95% pointwise simulation intervals except when the spatial lag becomes close to 6, owing to tapering beyond distance 6.

In Fig. 2, the computing time for all the three methods KL, ALT_2 and ALT_3 is reported. It can be seen that, for relatively small sample sizes, the three methods take up about the same amount of time. As the sample size increases, however, the computing time for ALT_2 increases dramatically compared with both KL and ALT_3 whose computing time is similar. This large difference in computing time is expected, as ALT_2 involves large matrix inversion, but underscores the usefulness of our KL method. Similar observations are made in the other three scenarios.

To evaluate the performance of spatial prediction, we define a mean-squared prediction error (MSPE) as $n^{-1}\Sigma_{i=1}^{n}\{\tilde{y}(\mathbf{s}_{0i}) - y(\mathbf{s}_{0i})\}^2$, where $\mathbf{s}_{01}, \ldots, \mathbf{s}_{0n}$ are n unsampled locations in R, $y(\mathbf{s}_{0i})$ is the true value and $\tilde{y}(\mathbf{s}_{0i})$ is the predicted value at location $\mathbf{s}_{0i}$, for $i = 1, \ldots, n$. In the simulation, for each sample size N, an additional 10% observations are generated at new locations to form a test set. The sample mean and standard deviation of the MSPE values from 100 simulations are

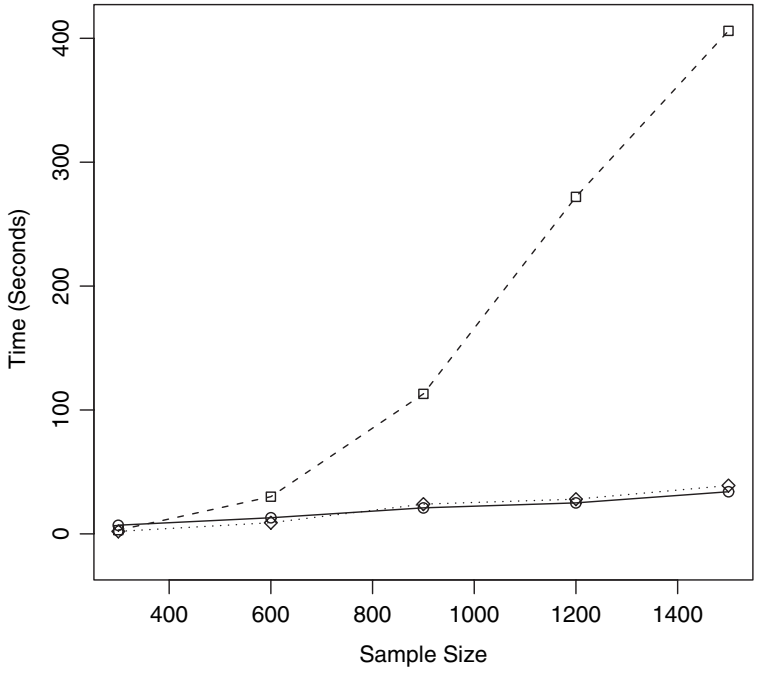

Fig. 2. Computing times for the KL (○), ALT_2 (□) and ALT_3 (◊) methods *versus* various choices of sample size

Table 1. Simulation results from scenarios 1 (left-hand panel) and 2 (right-hand panel): sample mean and standard deviation SD of the MSPE under methods KL, ALT_1, ALT_2 and ALT_3 for sample size $N = 300, 600, 900$

N	*Method*	*Exponential covariance*		*Misspecified covariance*	
		mean(MSPE)	*SD(MSPE)*	*mean(MSPE)*	*SD(MSPE)*
300	KL	5.83	1.52	4.88	1.30
	ALT_1	17.84	6.32	18.97	6.43
	ALT_2	5.69	1.49	4.96	1.29
	ALT_3	5.72	1.49	4.99	1.27
600	KL	5.70	1.20	4.81	0.98
	ALT_1	18.92	4.36	19.81	3.97
	ALT_2	5.62	1.17	4.93	0.99
	ALT_3	5.68	1.19	4.99	1.00
900	KL	5.61	0.84	4.76	0.72
	ALT_1	19.21	4.12	19.92	3.46
	ALT_2	5.51	0.81	4.90	0.74
	ALT_3	5.56	0.82	4.95	0.75

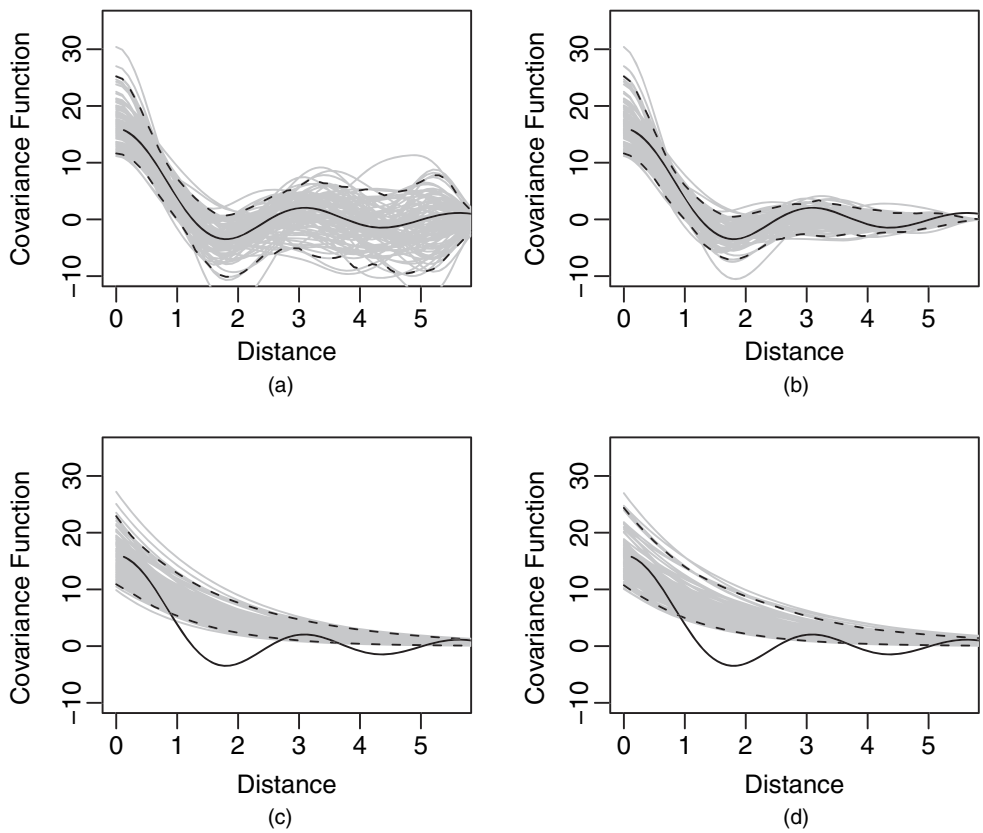

Fig. 3. Estimated covariance functions and 95% pointwise simulation intervals using our proposed method with both (a) pretapered estimates and (b) tapered estimates, (c) maximum likelihood ALT_2 and (d) a small block method ALT_3: ———, true sinusoidal covariance function; ═══, estimated covariance function from each simulated data set; – – –, pointwise simulation intervals

Table 2. Precipitation data: regression coefficient estimates and standard errors under methods KL, ALT_1, ALT_2 and ALT_3

Term	*Estimates and standard errors for the following methods:*							
	KL	*SE*	ALT_1	*SE*	ALT_2	*SE*	ALT_3	*SE*
Elevation	0.281	0.058	0.221	0.047	0.305	0.055	0.235	0.052
Slope	0.020	0.031	0.074	0.041	0.158	0.026	0.027	0.029
Aspect	0.000	0.027	0.051	0.034	−0.004	0.022	0.005	0.025
B1M	0.196	0.184	0.142	0.214	0.214	0.157	0.254	0.170
B2M	0.036	0.074	0.069	0.093	0.058	0.064	0.017	0.068
B3M	0.037	0.131	0.059	0.160	0.017	0.109	−0.015	0.112
B4M	−0.400	0.214	−0.472	0.242	−0.043	0.183	−0.381	0.199
B5M	0.090	0.105	0.155	0.137	0.043	0.089	0.115	0.098
B6M	−0.190	0.135	−0.357	0.166	−0.162	0.116	−0.212	0.124
B7M	0.158	0.116	0.241	0.150	0.172	0.098	0.121	0.110

reported in Table 1. Our KL method performs similarly to methods ALT_2 and ALT_3, whereas ALT_1 gives rather poor prediction. As for variable selection, KL, ALT_2 and ALT_3 perform similarly and satisfactorily, and all are slightly better than ALT_1, as shown in Table A of the on-line supplemental materials.

4.1.2. $d = 1$, misspecified covariance

Here the set-up is the same as scenario 1 except for the spatial dependence structure. Specifically, the error process $\varepsilon_1(\mathbf{s})$ follows a sinusoidal covariance function: $\gamma(h) = \sigma_1^2 \sin(h/c_r)c_r/h$. Moreover, the measurement error terms $\varepsilon_2(\mathbf{s})$ are independently generated from $N(0, \sigma_2^2)$. Let $\sigma_1^2 = 16$, $\sigma_2^2 = 4$ and $c_r = 0.4$. The results are reported in Table B of the on-line supplemental materials.

Again, as the sample size increases, the estimation of all four methods improves. In addition, methods ALT_2, ALT_3 and KL perform better than ALT_1 for both parameter estimation and prediction. This suggests that it is important to consider spatial dependence, even if the spatial covariance function is misspecified. However, our method KL outperforms ALT_2 and ALT_3 by providing a more robust estimate of the covariance function, as illustrated in Fig. 3. Unlike scenario 1, the estimated covariance functions from methods ALT_2 and ALT_3 are not close to the true underlying function, owing to the model misspecification when applying maximum likelihood. In contrast, the performance of KL is satisfactory. Similar results regarding spatial prediction and variable selection are attained as scenario 1.

When $d = 2$, similar conclusions can be drawn regarding parameter estimation, spatial prediction and variable selection. For brevity, those results are included in section B of the on-line supplemental materials. In addition, from a practical viewpoint, there are several tuning parameters in our proposed method. Here, we propose an empirical rule. See section C of the on-line supplemental materials for more detailed discussion.

4.2. Data example

The data set consists of January precipitation (inches per 24-hour period) on the log-scale from 259 weather stations in Colorado (Reich and Davis, 2008; Chu *et al.*, 2011). There are 10

covariates of interest, namely elevation, slope, aspect and seven spectral bands from moderate resolution imaging spectroradiometer satellite imagery (B1M–B7M). To investigate the relationship between precipitation and these covariates, we first fit a spatial linear model with an exponential covariance function via ordinary least squares, maximum likelihood and the small block method. The parameter estimates and their standard errors in Table 2 suggest that the regression coefficients for elevation, B1M, B4M, B6M and B7M are possibly significant. We then fit the data by using our proposed method and the results are similar to the three alternative methods, although the maximum likelihood estimate and the small block method appear to have slightly smaller standard errors.

5. Theoretical aspect

In this section, we shall establish the consistency of estimates in Section 3.1. Recall that R denotes the compact domain of interest and N is the number of sampling locations in R. Also, λ_j is the jth-largest eigenvalue in the local Karhunen–Loève expansion and n_k is the number of sampling locations in subdomain R_k. Let $\boldsymbol{\beta}_0$ denote the true parameters. We assume that σ^2 is known with $\sigma^2 = 1$, without loss of generality.

Let $\lambda_i(\mathbf{A}_1)$ denote the ith-largest eigenvalue of a square matrix $\mathbf{A}_1$. Furthermore, for an $n \times m$ matrix $\mathbf{A}_2 = (a_{ij})_{i,j=1}^{n,m}$, the Frobenius norm and L_2-norm are defined as $\|\mathbf{A}_2\|_{\mathrm{F}} = (\Sigma_{i=1}^{n} \Sigma_{j=1}^{m} a_{ij}^2)^{1/2}$ and $\|\mathbf{A}_2\|_2 = \max\{\lambda_i(\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2)^{1/2} : i = 1, \ldots, m\}$ respectively. Moreover, if $\mathbf{A}_3$ is an $n \times l$ matrix, it holds that $\|\mathbf{A}_2\mathbf{A}_3\|_{\mathrm{F}} \leqslant \|\mathbf{A}_2\|_{\mathrm{F}}\|\mathbf{A}_3\|_{\mathrm{F}}$, $\|\mathbf{A}_2\mathbf{A}_3\|_{\mathrm{F}} \leqslant \|\mathbf{A}_2\|_2\|\mathbf{A}_3\|_{\mathrm{F}}$ and $\|\mathbf{A}_2\mathbf{A}_3\|_2 \leqslant \|\mathbf{A}_2\|_2\|\mathbf{A}_3\|_2$.

We assume the following regularity conditions.

Assumption 1. There exist $0 < c_1, c_2\, c_3 < \infty$, such that

(a) $c_1 \geqslant \lambda_1 > \ldots > \lambda_J > \lambda_{J+1}$,
(b) $\max_{1 \leqslant j \leqslant J}(\lambda_j - \lambda_{j+1})^{-1} \leqslant c_2$ and
(c) $\lambda_1(\boldsymbol{\Sigma}_0) \leqslant c_3$.

Assumption 2. The eigenfunctions $\{\varphi_{j,1}(\cdot)\}_{j=1}^{J}$ are four times continuously differentiable and satisfy $\max_{1 \leqslant j \leqslant J} \|\varphi_{j,1}^{(4)}(\cdot)\|_\infty \leqslant C_0$ for some $0 < C_0 < \infty$.

Assumption 3. For n_k, $\underline{n} \leqslant n_k \leqslant \bar{n}$, where $\underline{n} \geqslant 4$, $\bar{n}/\underline{n} = O(1)$ and $\bar{n} = \mathcal{O}(K^{\kappa})$ for some $\kappa \geqslant 0$.

Assumption 4. There exist $(\mathbf{B}^*, \boldsymbol{\Lambda}^*)$, such that, $\delta_K = \max_{1 \leqslant k \leqslant K} n_k^{-1}\|\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^*\|_{\mathrm{F}}$ and $\bar{n}\delta_K = \mathcal{O}[\{M \log(K)/K\}^{1/2}]$, where $\boldsymbol{\Sigma}_k^* = \boldsymbol{\Phi}_k^{\mathrm{T}}\mathbf{B}^*\boldsymbol{\Lambda}^*\mathbf{B}^{*\mathrm{T}}\boldsymbol{\Phi}_k + \sigma^2\mathbf{I}_{n_k}$ and $\mathbf{B}^{*\mathrm{T}}\mathbf{B}^* = \mathbf{I}_r$.

Assumption 5. There exist constants $\rho_1, d_1, d_2, K_1 > 0$, such that, for $(\mathbf{B}, \boldsymbol{\Lambda}) \in \Theta\{(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*); \rho_1\}$, $d_1\underline{n}^2 a_K^2 < (1/K)\Sigma_{k=1}^{K}\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_{\mathrm{F}}^2 < d_2\bar{n}^2 a_K^2$ for all $K \geqslant K_1$, where $\Theta\{(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*); \rho\}$ is the neighbourhood centred at $(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*)$ with radius ρ in $\mathbb{R}^p \times \mathcal{S}_{M,J} \times \mathbb{R}^J$, and $\mathcal{S}_{M,J}$ is a *Stiefel manifold*. For more details, see Appendix A.

Assumption 6. For $(\varphi_{1,k}^*(\cdot), \ldots, \varphi_{J,k}^*(\cdot)) = \mathbf{B}^{*\mathrm{T}}(\phi_{1,k}(\cdot), \ldots, \phi_{M,k}(\cdot))$,

$$\max_{1 \leqslant j \leqslant J} \|\varphi_{j,k}(\cdot) - \varphi_{j,k}^*(\cdot)\| \leqslant c_{\phi,3} M^{-4} \max_{1 \leqslant j \leqslant J} \|\varphi_{j,k}^{(4)}(\cdot)\|_\infty, \tag{6}$$

$$\|\boldsymbol{\Phi}_k\|^2 \leqslant \bar{n}c_{g,1} + c_{\phi,0}^{-1} d_\eta (M^{3/2} \log(K) \vee [M\{\bar{n} \log(K)\}^{1/2}]), \tag{7}$$

$$\|\boldsymbol{\Phi}_k\|^2 \leqslant c_{\phi,2}\bar{n}M, \tag{8}$$

where $c_{\phi,0}$, $c_{\phi,2}$, $c_{\phi,3}$ and $c_{g,1}$ are constants.

Assumption 7. For $(\mathbf{B}, \mathbf{\Lambda}) \in \Theta\{(\boldsymbol{\beta}_0, \mathbf{B}^*, \mathbf{\Lambda}^*); \rho_1\}$, let $\alpha_K(\cdot)$ be the α-mixing coefficient of random variable $\mathrm{tr}\{(\mathbf{\Sigma}_k^{-1} - \mathbf{\Sigma}_k^{*-1})(\mathbf{S}_k - \mathbf{\Sigma}_{0k})\}$. Then $\alpha_K(2) = o(K^{-(2+2\kappa)Mr-\eta-2})$, where $\mathbf{S}_k = (\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_0)(\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_0)^{\mathrm{T}}$ and ρ_1 is defined in assumption 4.

Assumption 8. As $K \to \infty$, $M^{-1}\{K/\log(K)\}^{1/9} = \mathcal{O}(1)$, $M = o[\{K/\log(K)^{1/2}\}]$ and $\bar{n}^4 M^2 \times \log(K) = o(K)$.

Assumption 9. There exist constants $C_1, C_2, N_1 > 0$, such that $C_1\mathbf{I}_N \leqslant \mathbf{X}^{\mathrm{T}}\mathbf{X}/N \leqslant C_2\mathbf{I}_N$, for all $N \geqslant N_1$.

Assumptions 1 and 2 are about the spatial covariance structure and the Karhunen–Loève expansion. Assumption 3 is a boundedness condition for the number of sampling locations in subdomains (Paul and Peng, 2009). Assumption 4 assumes that there are optimal parameters $(\mathbf{B}^*, \mathbf{\Lambda}^*)$ such that the difference of the true covariance matrix $\mathbf{\Sigma}_{0k}$ and optimal covariance matrix $\mathbf{\Sigma}_k^*$ in every subdomain tends to 0 uniformly, as the number of subdomains $K \to \infty$. Moreover, assumption 4 requires that λ_{J+1} decay sufficiently fast (e.g. the expansion (2) of the process $\varepsilon_1(\mathbf{s})$ has finite J terms). In this case, the existence of the optimal parameters can be shown by using the spline approximation theory, which is well established for the one-dimensional space. Assumption 5 is about the properties of the fixed sampling locations, whereas assumption 6 is about the properties for the spline basis and eigenfunctions. Assumption 7 assumes that correlations between subdomains decrease in the sense of increasing domain. Assumption 8 specifies the relationship between $\bar{n}$, M and K (Paul and Peng, 2009); see section F of the on-line supplemental materials for a detailed discussion. The assumption about the design matrix $\mathbf{X}$ is made in assumption 9.

The following theorem 1 establishes the consistency for the estimates of the regression coefficients $\boldsymbol{\beta}$ and the spatial covariance function parameters $\mathbf{B}$ and $\mathbf{\Lambda}$, using our proposed method.

Theorem 1. Suppose that assumptions 1–9 hold. Then there is a minimizer $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{B}}, \hat{\mathbf{\Lambda}})$ of equation (4), such that, for $a_K = \{\bar{n}^2 M \log(K)/K\}^{1/2}$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = \mathcal{O}_p(N^{-1/2}),$$
$$\|\hat{\mathbf{B}} - \mathbf{B}^*\|_{\mathrm{F}} = \mathcal{O}_p(a_K),$$
$$\|\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*\|_{\mathrm{F}} = \mathcal{O}_p(a_K).$$

Theorem 1 shows that, for the regression coefficient vector, there is a local minimizer $\hat{\boldsymbol{\beta}}$ converging to the true parameter $\boldsymbol{\beta}_0$ at the rate of $N^{1/2}$. For the spatial covariance function parameters $\mathbf{B}$ and $\mathbf{\Lambda}$, the convergence is also achievable but at a slower rate of a_K. However, the convergence rate is not as slow as it might appear, especially for $\mathbf{B}$, since both $\mathbf{B}$ and $\mathbf{\Lambda}$ are converging in the Frobenius norm, i.e., for $\mathbf{B}$, the convergence is for the square sum of MJ parameters. It is also worth mentioning that the resulting estimates $(\hat{\mathbf{B}}, \hat{\mathbf{\Lambda}})$ are the primary building blocks to obtain the more accurate pretapered and tapered estimates for covariance functions in Section 3.2.

Next, we establish the convergence rate for covariance function estimates $\hat{\gamma}(\cdot)$ from equation (4). For simplicity, we suppress k in $\xi_{j,k}, \xi_{j,k}^*, \hat{\xi}_{j,k}, \lambda_{j,k}, \lambda_{j,k}^*, \varphi_{j,k}(\cdot), \varphi_{j,k}^*(\cdot), \hat{\varphi}_{j,k}(\cdot)$ and $\phi_{j,k}(\cdot)$. Let $\varepsilon_1(\mathbf{s}) = \Sigma_{j=1}^{\infty}\xi_j\varphi_j(\mathbf{s})$, $\varepsilon_1(\mathbf{s})^* = \Sigma_{j=1}^{J}\xi_j^*\varphi_j^*(\mathbf{s})$, and $\hat{\varepsilon}_1(\mathbf{s}) = \Sigma_{j=1}^{J}\hat{\xi}_j\hat{\varphi}_j(\mathbf{s})$, with the corresponding covariance functions $\gamma(\mathbf{s}, \mathbf{s}')$, $\gamma^*(\mathbf{s}, \mathbf{s}')$ and $\hat{\gamma}(\mathbf{s}, \mathbf{s}')$ respectively.

Theorem 2. Suppose that assumptions 1–9 hold. Then we have $\|\gamma(\mathbf{s}, \mathbf{s}') - \hat{\gamma}(\mathbf{s}, \mathbf{s}')\|_\infty = \|\gamma(\mathbf{s}, \mathbf{s}') - \gamma^*(\mathbf{s}, \mathbf{s}')\|_\infty + \mathcal{O}_p(M^{1/2}J^{1/2}a_K)$, $\|\mathbf{s} - \mathbf{s}'\| \leqslant D$.

$\|\gamma(\mathbf{s}, \mathbf{s}') - \gamma^*(\mathbf{s}, \mathbf{s}')\|_\infty$ is the truncated part of the spatial random process. If we assume that $\|\gamma(\mathbf{s}, \mathbf{s}') - \gamma^*(\mathbf{s}, \mathbf{s}')\|_\infty = \mathcal{O}_p(J^{-\alpha})$, with $\alpha > 0$, then $\|\gamma(\mathbf{s}, \mathbf{s}') - \hat{\gamma}(\mathbf{s}, \mathbf{s}')\|_\infty = \mathcal{O}_p(\max\{J^{-\alpha},$

$M^{1/2}J^{1/2}a_K\}$). By assumption 8, we can further express the convergence rates in theorem 1 and theorem 2 in terms of the total number of observations N.

Acknowledgements

The authors thank the Joint Editor, the Associate Editor and the referees for their helpful comments. Funding has been provided for this research from US Department of Agriculture Co-operative State Research, Education and Extension Service Hatch and McIntire–Stennis projects. The research of Tingjin Chu was supported by the National Natural Science Foundation of China (grant 11301536). The research of Haonan Wang was partially supported by National Science Foundation grants DMS-0854903 and DMS-1106975.

Appendix A: Neighbourhood in the parameter space

To maximize the approximated log-likelihood function in equation (4), a major challenge is the orthogonal constraint of the matrix parameter $\mathbf{B}$; i.e. $\mathbf{B}$ is taken from the set $\mathcal{S}_{M,J} = \{\mathbf{A} \in \mathbb{R}^{M\times J} : \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_J\}$, which is the well-known *Stiefel manifold*. Each matrix $\mathbf{B}$ can be considered as a point on the manifold $\mathcal{S}_{M,J}$.

Let $\mathcal{T}_{\mathbf{B}}$ be the tangent space of $\mathcal{S}_{M,J}$ at the point $\mathbf{B}$. In particular, any element in the tangent space, $\mathbf{U} \in \mathcal{T}_{\mathbf{B}}'$, can be expressed as $\mathbf{U} = \mathbf{B}\mathbf{A}_U + \mathbf{C}_U$, where $\mathbf{A}_U = -\mathbf{A}_U^{\mathrm{T}}$ and $\mathbf{B}^{\mathrm{T}}\mathbf{C}_U = \mathbf{0}$; see Edelman *et al.* (1998) for more details. On the manifold $\mathcal{S}_{M,J}$, the geodesic emanating from $\mathbf{B}$ along the direction $\mathbf{U}$ can be written as, for any $t \geqslant 0$, $\mathbf{G}_{\mathbf{B},\mathbf{U}}(t) = \mathbf{B}\mathbf{M}_{\mathbf{B},\mathbf{U}}(t) + \mathbf{Q}\mathbf{N}_{\mathbf{B},\mathbf{U}}(t)$, where

$$\begin{pmatrix} \mathbf{M}_{\mathrm{B},\mathrm{U}}(t) \\ \mathbf{N}_{\mathrm{B},\mathrm{U}}(t) \end{pmatrix} = \exp\left\{ t \begin{pmatrix} \mathbf{B}^{\mathrm{T}}\mathbf{U} & -\mathbf{R}^{\mathrm{T}} \\ \mathbf{R} & 0 \end{pmatrix} \right\} \begin{pmatrix} \mathbf{I}_r \\ \mathbf{0} \end{pmatrix}. \tag{9}$$

Here $\exp(\cdot)$ is the usual matrix exponential functional, and $\mathbf{QR}$ is the QR-decomposition of $(\mathbf{I}_M - \mathbf{B}\mathbf{B}^{\mathrm{T}})\mathbf{U}$. The function $\mathbf{G}_{\mathbf{B},\mathbf{U}}(t)$ is the exponential map on the manifold $\mathcal{S}_{M,J}$ at $\mathbf{B}$ along the direction $\mathbf{U}$, which essentially maps a tangent vector to a point on the manifold.

The geodesic, along with the exponential mapping, provides a useful way to define a neighbourhood on the manifold. For instance, we can define the neighbourhood as $\{\mathbf{G}_{\mathbf{B},\mathbf{U}}(t) : \text{for some sufficiently small } t \text{ and } \mathbf{U}\}$. The magnitudes of t and $\mathbf{U}$ will determine the size of the neighbourhood around $\mathbf{B}$. For convenience, we let $t = 1$, since $\mathbf{G}_{\mathbf{B},t\mathbf{U}}(t) = \mathbf{G}_{\mathbf{B},t\mathbf{U}}(1)$.

Finally, we define a *neighbourhood* in parameter space, centred at $(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*)$ and with size ρ, by

$$\Theta\{(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*); \rho\} = \left\{(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Lambda}) : \boldsymbol{\beta} = \exp(\mathbf{E})\boldsymbol{\beta}_0, \mathbf{B} = \mathbf{G}_{\mathbf{B}^*,\mathbf{U}}(1), \boldsymbol{\Lambda} = \exp(\mathbf{D})\boldsymbol{\Lambda}^*\right\}, \tag{10}$$

where $\mathbf{U} = \mathbf{B}^*\mathbf{A}_{\mathbf{U}} + \mathbf{C}_{\mathbf{U}}$, $\mathbf{A}_{\mathbf{U}} = -\mathbf{A}_{\mathbf{U}}^{\mathrm{T}}$ and $\mathbf{B}^{*\mathrm{T}}\mathbf{C}_{\mathbf{U}} = \mathbf{0}$, $\mathbf{D}$ is an $r \times r$ diagonal matrix, $\mathbf{E}$ is a $p \times p$ diagonal matrix and $\|\mathbf{A}_{\mathbf{U}}\|_{\mathrm{F}}^2 + \|\mathbf{C}_{\mathbf{U}}\|_{\mathrm{F}}^2 + \|\mathbf{D}\|_{\mathrm{F}}^2 + \|n^{1/2}a_K\mathbf{E}\|_{\mathrm{F}}^2 = \rho^2$. This neighbourhood extends the notion of restricted parameter space of $(\mathbf{B}, \boldsymbol{\Lambda})$ in Paul and Peng (2009), by incorporating the vector of regression coefficient $\boldsymbol{\beta}$. Paul and Peng (2009) also provided two important expansions, which will be used in our proof of theorem 1:

$$\mathbf{B}^{*\mathrm{T}}(\mathbf{G}_{\mathbf{B}^*,\mathbf{U}}(1) - \mathbf{B}^*) = \mathbf{B}^{*\mathrm{T}}\mathbf{U} + \mathcal{O}[\{\|\mathbf{B}^{*\mathrm{T}}\mathbf{U}\|_{\mathrm{F}} + \|(\mathbf{I}_M - \mathbf{B}^*\mathbf{B}^{*\mathrm{T}})\mathbf{U}\|_{\mathrm{F}}\}\|\mathbf{U}\|_{\mathrm{F}}], \tag{11}$$

$$(\mathbf{I}_M - \mathbf{B}^*\mathbf{B}^{*\mathrm{T}})\mathbf{G}_{\mathbf{B}^*,\mathbf{U}}(1) = (\mathbf{I}_M - \mathbf{B}^*\mathbf{B}^{*\mathrm{T}})\mathbf{U} + \mathcal{O}\{\|(\mathbf{I}_M - \mathbf{B}^*\mathbf{B}^{*\mathrm{T}})\mathbf{U}\|_{\mathrm{F}}\|\mathbf{U}\|_{\mathrm{F}}\} \tag{12}$$

as $\|\mathbf{U}\|_{\mathrm{F}} \to 0$.

Appendix B: Proof of theorem 1

It suffices to show that, given $\eta > 0$, for sufficiently large K, there is a constant c_η, such that

$$P\left\{\inf_{(\boldsymbol{\beta},\mathbf{B},\boldsymbol{\Lambda}) \in \Theta(c_\eta a_K)} L_K(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Lambda}) > L_K(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*)\right\} \geqslant 1 - \mathcal{O}(K^{-\eta}),$$

where $\Theta(c_\eta a_K) \equiv \Theta\{(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*); c_\eta a_K\}$ is the neighbourhood defined in Appendix A. Note that $L_K(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*) = L_K(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Lambda}) + \{L_K(\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*)\}$. We shall quantify $L_K(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Lambda})$ and $L_K(\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*)$, which are denoted by I and II respectively.

First, we show that

$$P\left\{\inf_{(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Lambda}) \in \Theta(c_\eta a_K)} L_K(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Lambda}) > 0\right\} \geqslant 1 - \mathcal{O}(K^{-\eta}). \tag{13}$$

It can be shown that, for any $(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Lambda}) \in \Theta(c_\eta a_K)$,

$$\begin{aligned}
2K\{L_K(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Lambda})\} &= \sum_{k=1}^{K} \{(\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}) - (\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_0)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_0)\} \\
&= \sum_{k=1}^{K} \{2(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{X}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_0)\} + \sum_{k=1}^{K} \{(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{X}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \mathbf{X}_k (\boldsymbol{\beta}_0 - \boldsymbol{\beta})\} \\
&\equiv I_1 + I_2.
\end{aligned}$$

The term I_1 has a normal distribution with mean $\mathbf{0}$ and variance $4(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{KL}}^{-1} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_{\mathrm{KL}}^{-1} \mathbf{X} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})$. Moreover, by assumption (1) and the definition of $\Theta(c_\eta a_K)$, there is a constant $c_4 > 0$, such that

$$4(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{KL}}^{-1} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_{\mathrm{KL}}^{-1} \mathbf{X} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \leqslant c_4 (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}).$$

Together with assumption 9, we have

$$4(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{KL}}^{-1} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_{\mathrm{KL}}^{-1} \mathbf{X} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \leqslant c_4 C_2 \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_2^2,$$

which yields $I_1 = \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_2 \mathcal{O}_p(N^{1/2})$.

Next, we consider the term I_2. By assumptions 1 and 8 and the definition of $\Theta(c_\eta a_K)$, there is a constant $c_5 > 0$, such that $I_2 = (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{KL}}^{-1} \mathbf{X} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \geqslant c_5 (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})$. By assumption 9, we have $I_2 \geqslant c_5 C_1 N \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_2^2$. For a sufficiently large c_η, I_2 dominates I_1 and, thus, inequality (13) follows.

For term II, we follow arguments similar to those in Paul and Peng (2009) and establish its uniform bound as

$$P\left\{\inf_{(\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Lambda}) \in \Theta(c_\eta a_K)} L_K(\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*) > (c_\eta a_K)^2\right\} \geqslant 1 - \mathcal{O}(K^{-\eta}). \tag{14}$$

For any fixed $\boldsymbol{\beta}_0$, we can express term II as

$$\begin{aligned}
\mathrm{II} &= K^{-1} \sum_{k=1}^{K} V(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_k^*) + (2K)^{-1} \sum_{k=1}^{K} \mathrm{tr}\{(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1})(S_k - \boldsymbol{\Sigma}_{0k})\} \\
&\quad + (2K)^{-1} \sum_{k=1}^{K} \mathrm{tr}\{(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1})(\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^*)\} \equiv \mathrm{II}_1 + \mathrm{II}_2 + \mathrm{II}_3,
\end{aligned}$$

where $\mathbf{S}_k = (\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_0)(\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_0)^{\mathrm{T}}$ and

$$V(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_k^*) = \tfrac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}_k^{-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{-1/2}\} - \tfrac{1}{2}\log|\mathbf{I}_{n_k} + \boldsymbol{\Sigma}_k^{-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{-1/2}|.$$

In section E of the on-line supplemental materials, we bound the above three terms individually in lemmas 2–4. A key device is the Bernstein inequality of an array of weakly dependent random variables; see lemma 5.3 of Sun and Lahiri (2006). As a consequence, we have $P\{L_K(\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \mathbf{B}^*, \boldsymbol{\Lambda}^*) \leqslant (c_\eta a_K)^2\} = \mathcal{O}(K^{-(2+2\kappa)MJ-\eta})$, for each point $(\mathbf{B}, \boldsymbol{\Lambda}) \in \Theta_0(c_\eta a_K)$, where $\Theta_0(c_\eta a_K) \equiv \{(\mathbf{B}, \boldsymbol{\Lambda}) : (\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Lambda}) \in \Theta(c_\eta a_K)\}$.

Furthermore, define a restricted neighbourhood in $\mathcal{S}_{M,J} \otimes \mathbb{R}^J$, centred at $(\mathbf{B}_1, \boldsymbol{\Lambda}_1)$ with size ω_K, as $\mathrm{Ne}(\mathbf{B}_1, \boldsymbol{\Lambda}_1; \omega_K) = \{(\mathbf{B}, \boldsymbol{\Lambda}) : \|\mathbf{B} - \mathbf{B}_1\|_{\mathrm{F}}^2 + \|\boldsymbol{\Lambda} - \boldsymbol{\Lambda}_1\|_{\mathrm{F}}^2 \leqslant \omega_K^2\}$. There is a finite set in $\mathcal{S}_{M,J} \otimes \mathbb{R}^J$, which is denoted by $\mathcal{C}[\omega_K]$, such that $\bigcup_{(\mathbf{B}_1, \boldsymbol{\Lambda}_1) \in \mathcal{C}[\omega_K]} \mathrm{Ne}(\mathbf{B}_1, \boldsymbol{\Lambda}_1; \omega_K) \supset \Theta_0(c_\eta a_K)$. In fact, by a standard construction of such neighbourhoods on a sphere in $\mathbb{R}^p$ ($p = MJ - J(J-1)/2$), there exists $\mathcal{C}[\omega_K]$, in which the number of elements is of order $\max\{1, (a_K \omega_K^{-1})^p\}$.

Thus, by assumption 3, for a sufficiently large K, taking $\omega_K = (\bar{n}^2 K)^{-1}$ yields

$$P\left\{\inf_{(\mathbf{B},\mathbf{\Lambda})\in\mathcal{C}[\omega_K]} L_K(\beta_0,\mathbf{B},\mathbf{\Lambda}) - L_K(\beta_0,\mathbf{B}^*,\mathbf{\Lambda}^*) > (c_\eta a_K)^2\right\} \geqslant 1-\mathcal{O}(K^{-\eta}).$$

Together with lemma 5 in the on-line supplemental materials, we have equation (14). Finally, combining expressions (13) and (14), theorem 1 follows.

Appendix C: Proof of theorem 2

By triangle inequality, we have

$$\|\gamma(\mathbf{s},\mathbf{s}') - \hat{\gamma}(\mathbf{s},\mathbf{s}')\|_\infty \leqslant \|\gamma(\mathbf{s},\mathbf{s}') - \gamma^*(\mathbf{s},\mathbf{s}')\|_\infty + \|\gamma^*(\mathbf{s},\mathbf{s}') - \hat{\gamma}(\mathbf{s},\mathbf{s}')\|_\infty.$$

Therefore, we need only to show that $\|\gamma^*(\mathbf{s},\mathbf{s}') - \hat{\gamma}(\mathbf{s},\mathbf{s}')\|_\infty = \mathcal{O}_p(M^{1/2}J^{1/2}a_K)$. Note that

$$\begin{aligned}\|\gamma^*(\mathbf{s},\mathbf{s}') - \hat{\gamma}(\mathbf{s},\mathbf{s}')\|_\infty &\leqslant \|\sum_{j=1}^J \lambda_j^*\{\varphi_j^*(\mathbf{s})\varphi_j^*(\mathbf{s}') - \hat{\varphi}_j(\mathbf{s})\hat{\varphi}_j(\mathbf{s}')\}\|_\infty + \|\sum_{j=1}^J(\lambda_j^* - \hat{\lambda}_j)\hat{\varphi}_j(\mathbf{s})\hat{\varphi}_j(\mathbf{s}')\|_\infty \\ &= \mathrm{It} + \mathrm{IIt}.\end{aligned}$$

By theorem 1 and the Cauchy–Schwarz inequality, we have

$$\sum_{j=1}^J \|\varphi_j^*(\mathbf{s}) - \hat{\varphi}_j(\mathbf{s})\|_\infty = \|\mathbf{B}^*\mathbf{\Phi} - \hat{\mathbf{B}}\mathbf{\Phi}\|_\infty \leqslant (MJ)^{1/2}\|\mathbf{B}^* - \hat{\mathbf{B}}\|_{\mathrm{F}} \max_{j=1,\ldots,M}\|\phi_j(\mathbf{s})\|_\infty = \mathcal{O}_p(M^{1/2}J^{1/2}a_K).$$

Moreover, applying the triangle inequality yields

$$\mathrm{It} \leqslant \|\sum_{j=1}^J \lambda_j^*\varphi_j^*(\mathbf{s})\{\varphi_j^*(\mathbf{s}') - \hat{\varphi}_j(\mathbf{s}')\}\|_\infty + \|\sum_{j=1}^J \lambda_j^*\hat{\varphi}_j(\mathbf{s}')\{\varphi_j^*(\mathbf{s}) - \hat{\varphi}_j(\mathbf{s})\}\|_\infty = \mathcal{O}_p(M^{1/2}J^{1/2}a_K).$$

For term IIt, we have

$$\mathrm{IIt} \leqslant \sum_{j=1}^J |\lambda_j^* - \hat{\lambda}_j|\|\hat{\varphi}_j(\mathbf{s})\|_\infty\|\hat{\varphi}_j(\mathbf{s}')\|_\infty \leqslant J^{1/2}\|\Lambda^* - \hat{\Lambda}\|_{\mathrm{F}}\|\hat{\varphi}_j(\mathbf{s})\|_\infty^2 = \mathcal{O}_p(a_K J^{1/2}).$$

References

Adler, R. and Taylor, J. (2007) *Random Fields and Geometry*. New York: Springer.

Buhmann, M. D. (2003) *Radial Basis Functions: Theory and Implementations*. Cambridge: Cambridge University Press.

Caragea, P. and Smith, R. (2007) Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *J. Multiv. Anal.*, **98**, 1417–1440.

Chu, T., Zhu, J. and Wang, H. (2011) Penalized maximum likelihood estimation and variable selection in geostatistics. *Ann. Statist.*, **39**, 2607–2625.

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

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc.* B, **70**, 209–226.

Du, J., Zhang, H. and Mandrekar, V. S. (2009) Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Ann. Statist.*, **37**, 3330–3361.

Edelman, A., Arias, T. A. and Smith, S. T. (1998) The geometry of algorithms with orthogonality constraints. *SIAM J. Matrx Anal. Appl.*, **20**, 303–353.

Furrer, R., Genton, M. G. and Nychka, D. (2006) Covariance tapering for interpolation of large spatial datasets. *J. Computnl Graph. Statist.*, **15**, 502–523.

Hall, P. and Patil, P. (1994) Properties of nonparametric estimators of autocovariance for stationary random fields. *Probab. Theor. Reltd Flds*, **99**, 399–423.

Im, H. K., Stein, M. L. and Zhu, Z. (2007) Semiparametric estimation of spectral density with irregular observations. *J. Am. Statist. Ass.*, **102**, 726–735.

James, G., Hastie, T. and Sugar, C. (2000) Principal component models for sparse functional data. *Biometrika*, **87**, 587–601.

Kaufman, C. G., Schervish, M. J. and Nychka, D. W. (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Am. Statist. Ass.*, **103**, 1545–1555.

Mardia, K. and Marshall, R. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **73**, 135–146.

Paul, D. and Peng, J. (2009) Consistency of restricted maximum likelihood estimators of principal components. *Ann. Statist.*, **37**, 1229–1271.
Peng, J. and Paul, D. (2009) A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *J. Computnl Graph. Statist.*, **18**, 995–1015.
Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-effects Models in S and S-PLUS*. New York: Springer.
Reich, R. and Davis, R. (2008) *Lecture Notes of Quantitative Spatial Analysis*. Fort Collins: Colorado State University.
Stein, M. (1999) *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.
Stein, M. L., Chi, Z. and Welty, L. J. (2004) Approximating likelihoods for large spatial data sets. *J. R. Statist. Soc.* B, **66**, 275–296.
Sun, S. and Lahiri, S. (2006) Bootstrapping the sample quantile of a weakly dependent sequence. *Sankhya* A, **68**, 130–166.
Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *J. R. Statist. Soc.* B, **50**, 297–312.
Wang, H. and Zhu, J. (2009) Variable selection in spatial regression via penalized least squares. *Can. J. Statist.*, **37**, 1–18.
Wendland, H. (1995) Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Computnl Math.*, **4**, 389–396.
Yao, F. and Lee, T. C. M. (2006) Penalized spline models for functional principal component analysis. *J. R. Statist. Soc.* B, **68**, 3–25.
Yao, F., Müller, H. and Wang, J. (2005) Functional data analysis for sparse longitudinal data. *J. Am. Statist. Ass.*, **100**, 577–590.
Zhang, H. and Wang, Y. (2010) Kriging and cross-validation for massive spatial data. *Environmetrics*, **21**, 290–304.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplemental materials for "On semiparametric inference of geostatistical models via local Karhunen–Loève expansion"'.