Krigings over space and time based on latent low-dimensional structures

Da Huang¹, Qiwei Yao²,∗ & Rongmao Zhang³

¹Department of Statistics, Fudan University, Shanghai 200433, China;
²Department of Statistics, London School of Economics, London WC²A 2AE, UK;
³School of Mathematics, Zhejiang University, Hangzhou 310058, China

Email: dahuang@fudan.edu.cn, q.yao@lse.ac.uk, rmzhang@zju.edu.cn

Received March 9, 2019; accepted September 15, 2019; published online April 23, 2020

Abstract  We propose a new nonparametric approach to represent the linear dependence structure of a spatio-temporal process in terms of latent common factors. Though it is formally similar to the existing reduced rank approximation methods, the fundamental difference is that the low-dimensional structure is completely unknown in our setting, which is learned from the data collected irregularly over space but regularly in time. Furthermore, a graph Laplacian is incorporated in the learning in order to take the advantage of the continuity over space, and a new aggregation method via randomly partitioning space is introduced to improve the efficiency. We do not impose any stationarity conditions over space either, as the learning is facilitated by the stationarity in time. Krigings over space and time are carried out based on the learned low-dimensional structure, which is scalable to the cases when the data are taken over a large number of locations and/or over a long time period. Asymptotic properties of the proposed methods are established. An illustration with both simulated and real data sets is also reported.

Keywords  aggregation via random partitioning, common factors, eigenanalysis, graph Laplacian, nugget effect, spatio-temporal processes

MSC(2020)  62M10, 62M30, 62H11

Citation: Huang D, Yao Q W, Zhang R M. Krigings over space and time based on latent low-dimensional structures. Sci China Math, 2021, 64: 823–848, https://doi.org/10.1007/s11425-019-1606-2

1 Introduction

Kriging, referring to the best spatial linear prediction, is named by Matheron after the South African mining engineer Daniel Krige. The key step in kriging is to identify and to estimate the covariance structure. The early applications of kriging are typically based on some parametric models for spatial covariance functions (see [10, Subsection 4.1] and the references therein). However, fitting those parametric covariance models to large spatial or spatio-temporal datasets is conceptually indefensible (see [16]). It also poses serious computational challenges. For example, a spatial kriging with observations from p locations involves inverting a p × p covariance matrix, which typically requires O(p³) operations with O(p²) memory. One attractive approach to overcome the computational burden is to introduce reduced
rank approximations for the underlying processes. Methods in this category include [18] using kernel convolutions, [8, 20, 33] using low rank basis functions (see also [10, Subsection 7.1.3]), [1, 12] using predictive processes, and [31] using thin-plate splines. However, as pointed out by [30], the reduced rank approximations often fail to capture the small-scale correlation structure accurately. An alternative approach is to seek sparse approximations for covariance functions (for example, Furrer et al. [13] and Kaufman et al. [22] proposed tapering methods by setting the covariances to 0 between any two locations with the distances beyond a threshold). Obviously these approaches overlook the correlations among the locations which are distantly apart from each other. Combining together both ideas of reducing rank and the tapering, Sang and Huang [28] and Zhang et al. [35] proposed a so-called full scale approximation method for large spatial and spatio-temporal datasets. Apart from the parametric strategies discussed above, Choi et al. [7] suggested a nonparametric way to build up spatio-temporal covariance.

In this paper, we propose a new nonparametric approach to represent the linear dependence structure of a spatio-temporal process. Different from all methods stated above, we impose neither distributional assumptions on the underlying process nor parametric forms on its covariance function. Under the setting that the observations are taken irregularly over space but regularly in time, we recover the linear dependent structure based on a latent factor representation. No stationarity conditions are imposed over space, though the stationary in time is assumed. The proposed latent factor model has a reduced rank representation in form. However, neither the factor process nor the loadings are assumed to be known in advance. This is a marked difference from the aforementioned reduced rank approximation methods. The motivation of our approach is to learn the linear dynamic structure across both space and time directly from data with little subjective assumptions. It captures the dependence across the locations over all distances automatically.

The latent factors and the corresponding loadings are estimated via eigenanalysis. However, it differs from the eigenanalysis for estimating latent factors for multiple time series (see [23], and the references therein) in at least three aspects. Firstly, we extract the information from the dependence among different locations instead of in time: the whole observations are randomly divided into two sets according to their locations, the estimation boils down to the eigen-decomposition related to the spatial covariance matrix of two data sets. One advantage of this approach is that it is free from the impact of the ‘nugget effect’ in the sense that we do not need to estimate the variances of, for example, measurement errors in order to recover the latent dependence structure. Secondly, we propose a new method of aggregation via repeatedly randomly partitioning the observations over space to improve the original estimation. This also overcomes the arbitrariness in dividing data in the partitioning. The proposed aggregation method is in the spirit of the Bagging of [4], though randomly partitioning instead of bootstrapping is used in our approach. Thirdly, we incorporate a graph Laplacian (see [17, p. 545]) into the eigenanalysis to take the advantage of the continuity over space, leading to further improvement in both estimation and kriging.

The number of latent factors is typically small or at least much smaller than the number of locations on which the data are recorded. Consequently the krigings can be performed via only inverting matrices of the size equal the number of factors. This is particularly appealing when dealing with large datasets. Though the eigenanalysis for estimating the latent factor structure requires $O(p^3)$ operation, the divide-and-conquer strategy makes our approach easily scalable to large datasets (see Subsection 3.3 below).

It is worth pointing out that our approach is designed for analyzing spatio-temporal data or pure spatial data but with repeated observations. Attribute to the advancement of information technology, large amount of data are collected routinely over space and time nowadays. The surge of the development of statistical methods and theories for modeling and forecasting spatio-temporal processes includes, among others, [5, 15, 19, 21, 25, 29, 35, 37] and [32] (see also the monograph [10]). In addition to the methods based on the low-dimensional covariance structures, the dynamic approach which, typically, specifies the standard Gaussian autoregressive model of order 1 (i.e., AR(1)), coupled with Markov chain Monte Carlo (MCMC) computation has gained popularity in modeling large spatio-temporal data. Cressie et al. [9] assumed a Gaussian AR(1) model for a low-dimensional latent process and developed a full scale Kalman filter in the context of spatio-temporal modeling (see also [10, Chapter 7]).

The rest of the paper is organized as follows. We specify the latent factor structure for a spatio-temporal
process in Section 2. The newly proposed estimation methods are presented in Section 3. The kriging over space and time is introduced in Section 4, in which we also state how to handle missing values. The asymptotic results for the proposed estimation and kriging methods are developed in Section 5. An illustration with both simulated and real data sets is reported in Section 6. Technical proofs are relegated to Appendix A.

2 Models

2.1 Setting

Consider the spatio-temporal process

\[ y_t(s) = z_t(s)' \beta(s) + \xi_t(s) + \varepsilon_t(s), \quad t = 0, \pm 1, \pm 2, \ldots, \quad s \in S \subset \mathbb{R}^2, \]

(2.1)

where \( z_t(s) \) is an \( m \times 1 \) observable covariant vector, \( \beta(s) \) is an unknown parameter vector, \( \varepsilon_t(s) \) is unobservable and represents the so-called nugget effect (over space) in the sense that

\[ E\{\varepsilon_t(s)\} = 0, \quad \text{Var}\{\varepsilon_t(s)\} = \sigma(s)^2, \quad \text{Cov}\{\varepsilon_t(u), \varepsilon_{t+2}(v)\} = 0, \quad \forall (t_1, u) \neq (t_2, v), \]

(2.2)

\( \xi_t(s) \) is a latent spatio-temporal process satisfying the conditions

\[ E\{\xi_t(s)\} = 0, \quad \text{Cov}\{\xi_t(u), \xi_{t+2}(v)\} = \Sigma_{t_1-t_2}(u, v). \]

(2.3)

Consequently, \( y_t(s) - z_t(s)' \beta(s) \) is (weakly) stationary in time \( t \), \( E\{y_t(s) - z_t(s)' \beta(s)\} = 0 \), and

\[ \text{Cov}\{y_{t_1}(u) - z_{t_1}(u)' \beta(u), y_{t_2}(v) - z_{t_2}(v)' \beta(v)\} = \Sigma_{|t_1-t_2|}(u, v) + \sigma(u)^2 I\{(t_1, u) = (t_2, v)\}. \]

(2.4)

Furthermore, we assume that \( \Sigma_t(u, v) \) is continuous in \( u \) and \( v \).

Model (2.1) does not impose any stationarity condition over space. However, it requires that \( y_t(\cdot) - z_t(\cdot)' \beta(\cdot) \) is second order stationary in time \( t \), which enables the learning of the dependence across different locations and periods. In practice the data often show some trends and seasonal patterns in time. The existing detrend and deseasonality methods in time series analysis can be applied to make data stationary in time.

2.2 A finite-dimensional representation for \( \xi_t(s) \)

Let \( L_2(S) \) be the Hilbert space consisting of all the square integrable functions defined on \( S \) equipped with the inner product

\[ \langle f, g \rangle = \int_S f(s)g(s)ds, \quad f, g \in L_2(S). \]

(2.5)

We assume that the latent process \( \xi_t(s) \) admits a finite-dimensional structure

\[ \xi_t(s) = \sum_{j=1}^d a_j(s)x_{ij}, \]

(2.6)

where \( a_1(\cdot), \ldots, a_d(\cdot) \) are deterministic and linear independent functions (i.e., none of them can be written as a linear combination of the others) in the Hilbert space \( L_2(S) \), and \( x_{t1}, \ldots, x_{td} \) are \( d \) latent time series. Obviously \( a_1(\cdot), \ldots, a_d(\cdot) \) (as well as \( x_{t1}, \ldots, x_{td} \)) are not uniquely defined by (2.6), as they can be replaced by any of their non-degenerate linear transformations. There is no loss of generality in assuming that \( a_1(\cdot), \ldots, a_d(\cdot) \) are orthonormal in the sense that

\[ \langle a_i, a_j \rangle = I(i = j), \]

(2.7)

as any set of linear independent functions in a Hilbert space can be standardized to this effect. Let \( x_t = (x_{t1}, \ldots, x_{td})' \). It follows from (2.3) that \( x_t \) is a \( d \)-variant stationary time series with mean \( 0 \), and

\[ \Sigma_0(u, v) = \text{Cov}\{\xi_t(u), \xi_t(v)\} = \sum_{i,j=1}^d a_i(u)a_j(v)\sigma_{ij}, \]

(2.8)
where \( \sigma_{ij} \) is the \((i,j)\)-th element of \( \text{Var}(x_t) \). Let
\[
\Sigma_0 \circ f(s) = \int_S \Sigma_0(s, u) f(u) du, \quad f \in L_2(S). \tag{2.9}
\]
Then \( \Sigma_0 \) is a non-negative definite operator defined on \( L_2(S) \) (see [2, Appendix A] for some basic facts on the operators in the Hilbert spaces). It follows from Mercer’s theorem (see [27]) that \( \Sigma_0 \) admits the spectral decomposition
\[
\Sigma_0(u, v) = \sum_{j=1}^{d} \lambda_j \varphi_j(u) \varphi_j(v), \tag{2.10}
\]
where \( \lambda_1 \geq \cdots \geq \lambda_d > 0 \) are the \( d \) positive eigenvalues of \( \Sigma_0(u, v) \), and \( \varphi_1, \ldots, \varphi_d \in L_2(S) \) are the corresponding eigenfunctions, i.e.,
\[
\Sigma_0 \circ \varphi_j(s) = \int_S \Sigma_0(s, u) \varphi_j(u) du = \lambda_j \varphi_j(s) \tag{2.11}
\]
(see Proposition 2.1 below).

**Proposition 2.1.** Let \( \text{rank}(\text{var}(x_t)) = d \). Then the following assertions hold:
(i) \( \Sigma_0 \) defined in (2.9) has exactly \( d \) positive eigenvalues.
(ii) The \( d \) corresponding orthonormal eigenfunctions can be expressed as
\[
\varphi_i(s) = \sum_{j=1}^{d} \gamma_{ij} a_j(s), \quad i = 1, \ldots, d,
\]
where \( \gamma_i \equiv (\gamma_{i1}, \ldots, \gamma_{id})' \), \( i = 1, \ldots, d \), are \( d \) orthonormal eigenvectors of matrix \( \text{Var}(x_t) \).

The above proposition shows that the finite-dimensional structure (2.6) can be identified via the covariance functions of \( \xi_t(s) \), though the representation of (2.6) itself is not unique. Note that the linear space spanned by the eigenfunctions \( \varphi_1(\cdot), \ldots, \varphi_d(\cdot) \) is called the kernel reproducing Hilbert space (KRHS) by \( \Sigma_0(\cdot, \cdot) \), meanwhile \{\( a_1(\cdot) \)\} and \{\( \varphi_1(\cdot) \)\} are two orthonormal basis of this KRHS. Furthermore, any orthonormal basis of this KRHS can be taken as \( a_1(\cdot), \ldots, a_d(\cdot) \). In Section 3 below, the estimation for \( a_1(\cdot), \ldots, a_d(\cdot) \) will be constructed in this spirit.

### 3 Estimation
Let \( \{y_t(s_i), z_t(s_i)\}, \quad i = 1, \ldots, p, \quad t = 1, \ldots, n \} \) be the available observations over space and time, where \( S_d = \{s_1, \ldots, s_p\} \subset S \) are typically irregularly spaced. The total number of observations is \( n \cdot p \).

#### 3.1 Estimation for finite-dimensional representations of \( \xi_t(s) \)
To simplify the notation, we first consider a special case \( \beta(s) \equiv 0 \) in (2.1) in Subsections 3.1 and 3.2. Subsection 3.4 below considers the least squares regression estimation for \( \beta(s) \). Then the procedures described in Subsections 3.1 and 3.2 still apply if \( \{y_t(s_i)\} \) are replaced by the residuals from the regression estimation.

Now under (2.6),
\[
y_t(s) = \xi_t(s) + \varepsilon_t(s) = \sum_{j=1}^{d} a_j(s)x_{tj} + \varepsilon_t(s). \tag{3.1}
\]
To exclude nugget effect in our estimation, we divide \( p \) locations \( s_1, \ldots, s_p \) into two disjoint sets \( S_1 \) and \( S_2 \) with, respectively, \( p_1 \) and \( p_2 \) elements, and \( p_1 + p_2 = p \). Let \( y_{t,i} \) be a vector consisting of \( y_t(s) \) with \( s \in S_i, \quad i = 1, 2 \). Then \( y_{t,1} \) and \( y_{t,2} \) are two vectors with lengths \( p_1 \) and \( p_2 \), respectively. Denote by \( \xi_{t,1} \) and \( \xi_{t,2} \) the corresponding vectors consisting of \( \xi_t(\cdot) \). It follows from (3.1) that
\[
y_{t,1} = \xi_{t,1} + \varepsilon_{t,1} = A_1x_t + \varepsilon_{t,1}, \quad y_{t,2} = \xi_{t,2} + \varepsilon_{t,2} = A_2x_t + \varepsilon_{t,2}, \tag{3.2}
\]
where $A_1$ is a $p_1 \times d$ matrix, its rows consist of the coefficients $a_j(\cdot)$ on the right-hand side of (3.1), and $\varepsilon_{t,i}$ consists of $\varepsilon_t(s)$ with $s \in S_i$. There is no loss of generality in assuming $A_1 A_1 = I_d$. This can be achieved by performing an orthogonal-triangular (QR) decomposition $A_1 = GR$, and replacing $(A_1, x_t)$ by $(\Gamma, Rx_t)$ in the first equation in (3.2). Note $\mathcal{M}(A_1) = \mathcal{M}(\Gamma)$, where $\mathcal{M}(A)$ denotes the linear space spanned by the columns of matrix $A$. Thus $\mathcal{M}(A_1)$ does not change from imposing the condition $A_1 A_1 = I_d$. Similarly, we may also assume $A_2 A_2 = I_d$, which however implies that $x_t$ in the second equation in (3.2) is unlikely to be the same as that in the first equation. Hence we may re-write (3.2) as
\begin{equation}
y_{t,1} = A_1 x_t + \varepsilon_{t,1}, \quad y_{t,2} = A_2 x_t^* + \varepsilon_{t,2},
\end{equation}
(3.3)
where $A_1^2 A_1 = A_2^2 A_2 = I_d$, $x_t^* = Q x_t$, and $Q$ is an invertible $d \times d$ matrix. Note that $(A_1, x_t)$ and $(A_2, x_t^*)$ are still not uniquely defined in (3.3), as they can be replaced, respectively, by $(A_1 \Gamma_1, \Gamma_1' x_t)$ and $(A_2 \Gamma_2, \Gamma_2' x_t)$ for any $d \times d$ orthogonal matrices $\Gamma_1$ and $\Gamma_2$. However, $\mathcal{M}(A_1)$ and $\mathcal{M}(A_2)$ are uniquely defined by (3.3).

Since $y_{t,1}$ and $y_{t,2}$ have no common elements, it follows from (3.1) and (2.2) that
\begin{equation}
\Sigma \equiv \text{Cov}(y_{t,1}, y_{t,2}) = A_1 \text{Cov}(x_t, x_t^*) A_2.'
\end{equation}
(3.4)
Note that Cov($x_t, x_t^*$) = Var($x_t$)$Q$. When $p \gg d$, it is reasonable to assume that $\text{rank}(\Sigma) = \text{rank}\{\text{Cov}(x_t, x_t^*)\} = \text{rank}\{\text{Var}(x_t)\} = d$.

Let $\Sigma^\prime$ be the sample covariance of $y_{t,1}$ and $y_{t,2}$, i.e.,
\begin{equation}
\hat{\Sigma} = \frac{1}{n} \sum_{t=1}^{n} (y_{t,1} - \bar{y}_1)(y_{t,2} - \bar{y}_2)',
\end{equation}
(3.6)
where $\bar{y}_t = n^{-1} \sum_t y_{t,i}$. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots$ be the eigenvalues of $\hat{\Sigma} \Sigma'$. A natural estimator of $d$ is defined as
\begin{equation}
\hat{d} = \max_{1 \leq j < p_*} \frac{\hat{\lambda}_j}{\lambda_{j+1}},
\end{equation}
(3.7)
where $p_* \ll \min(p_1, p_2)$ is a prespecified integer (e.g., $p_* = \min(p_1, p_2)/2$). This estimation method is based on the fact that $\lambda_j/\lambda_{j+1}$ are positive and finite constants for $j = 1, \ldots, d - 1$, and $\lambda_d/\lambda_{d+1} = \infty$. However, $\lambda_j/\lambda_{j+1}$ is asymptotically ‘0/0’ for $j = d + 1, \ldots, p - 1$. In practice, we mitigate this difficulty by comparing the ratios for $j < p_* \ll \min(p_1, p_2)$. Asymptotic properties of the ratio estimators under different settings have been established in, e.g., [6, 23, 36]. The finite sample performance of the ratio estimators are also reported in those papers.

Consequently, the $\hat{d}$ orthonormal eigenvectors of $\hat{\Sigma} \Sigma'$ (or $\hat{\Sigma}' \hat{\Sigma}$), corresponding to the eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{d}}$, can be taken as the estimated columns of $A_1$ (or $A_2$). However, such an estimator ignores the fact that $\varepsilon_t(\cdot)$ is continuous over the set $S_i$, which should be taken into account to improve the estimation. To achieve this, denote by $s_{1,i}, \ldots, s_{p_1,i}$, the $p_1$ locations in $S_i$ arranged according to the order such that the $j$-th component of $y_{t,1}$ is the observation taken at the location $s_{j,i}$. We define a graph Laplacian $L = G - W$, where $W = (w_{ij})$ is a weight matrix with $w_{ii} = 0$ and, e.g., $w_{ij} = 1/(1 + \|s_{i} - s_{j}\|)$ (\|\cdot\| denotes the Euclidean norm) for $i \neq j$, and $G = (g_{ij})$ with $g_{ii} = \sum_j w_{ij}$ and $g_{ij} = 0$ for all $i \neq j$. Then it holds that for any column vector $a = (a_1, \ldots, a_p)'$,
\begin{equation}
a' L a = \sum_{i=1}^{p} g_{ii} a_i^2 - \sum_{i,j=1}^{p} w_{ij} a_i a_j = \frac{1}{2} \sum_{i,j=1}^{p} w_{ij} (a_i - a_j)^2 \quad (3.8)
\end{equation}
(see [17, p.545]). By requiring $a'La \leq c_0$ for some small positive constant $c_0$, the components of $a$ at nearby locations will be close to each other. Hence the columns of $A_1$ are obtained by solving the following optimization problem:

$$\hat{\gamma}_i = \arg\max_{\gamma} \gamma' \Sigma \Sigma' \gamma \quad \text{subject to} \quad \|\gamma\| = 1 \quad \text{and} \quad \gamma' L \gamma \leq c_0,$$

and for $j = 2, \ldots, \hat{d}$,

$$\hat{\gamma}_j = \arg\max_{\gamma} \gamma' \Sigma \Sigma' \gamma \quad \text{subject to} \quad \|\gamma\| = 1, \quad \gamma' \hat{\gamma}_i = 0 \quad \text{for} \quad 1 \leq i < j, \quad \text{and} \quad \gamma' L \gamma \leq c_0.$$ 

Note the optimization problem above is a quadratically constrained quadratic program (QCQP). We may transform it to its ‘dual problem’ via the Lagrangian dual function. Then the constrained maximization can be recast as an eigen-problem for the symmetric (but necessarily non-negative definite) matrix $\Sigma \Sigma' - \tau L$ stated below, where $\tau > 0$ controls the penalty according to $L$:

$\text{Find the orthonormal eigenvectors } \hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}} \text{ of } \Sigma \Sigma' - \tau L \text{ corresponding to its } \hat{d} \text{ largest eigenvalues.}$

Please refer to [3] or other optimization textbooks for details.

Denote the resulting estimator for the loading matrix $A_1$ by

$$\hat{A}_1 = (\hat{\gamma}_1, \ldots, \hat{\gamma}_{\hat{d}}).$$

(3.8)

The estimator for $A_2$, denoted by $\hat{A}_2$, is constructed in the same manner.

By (3.2), the estimators for the two different representations of the latent processes are defined as

$$\hat{x}_i = \hat{A}_1' y_{i,1}, \quad \hat{x}_i^* = \hat{A}_2' y_{i,2}. \quad (3.9)$$

Consequently,

$$\hat{\xi}_{i,1} = \hat{A}_1 \hat{x}_i = \hat{A}_1 \hat{A}_1' y_{i,1}, \quad \hat{\xi}_{i,2} = \hat{A}_2 \hat{x}_i^* = \hat{A}_2 \hat{A}_2' y_{i,2} \quad (3.10)$$

(see also (3.2)).

**Remark 3.1.** (i) The assumption that the matrix $\Sigma$ in (3.4) has rank $d$ implies that all the latent factors are spatially correlated (see (3.1)). In the unlikely scenarios that some latent factors are only serially correlated but spatially uncorrelated, we should include autocovariance matrices in the estimation [23]. To this end, let

$$\hat{\Sigma}_i(k) = \frac{1}{n} \sum_{t=1}^{n-k} (y_{t+k,i} - \bar{y}_i)(y_{t,i} - \bar{y}_i)',$$

$$\hat{\Sigma}_{12}(k) = \frac{1}{n} \sum_{t=\max\{1, -k\}}^{\min\{n-k, n\}} (y_{t+1,1} - \bar{y}_1)(y_{t,2} - \bar{y}_2).$$

Assume $p_1 = p_2$ for simplicity. Put

$$M_1 = \hat{\Sigma} \hat{\Sigma}' + \sum_{j=1}^{k_0} (\hat{\Sigma}_1(j) \hat{\Sigma}_1(j)' + \hat{\Sigma}_{12}(j) \hat{\Sigma}_{12}(j)' + \hat{\Sigma}_{12}(-j) \hat{\Sigma}_{12}(-j)'),$$

$$M_2 = \hat{\Sigma}' \hat{\Sigma} + \sum_{j=1}^{k_0} (\hat{\Sigma}_2(j) \hat{\Sigma}_2(j)' + \hat{\Sigma}_{12}(j)' \hat{\Sigma}_{12}(j) + \hat{\Sigma}_{12}(-j)' \hat{\Sigma}_{12}(-j)),$$

where $\hat{\Sigma}$ is defined in (3.6), and $k_0 \geq 1$ is an integer. Then we replace $\hat{\Sigma} \hat{\Sigma}'$ by $M_1$ in (3.8), and replace $\hat{\Sigma}' \hat{\Sigma}$ by $M_2$ for computing $\hat{A}_2$. Empirical evidences in modeling high-dimensional time series indicate that the estimation is not sensitive to the choice of $k_0$, and small values of $k_0$ such as 1 to 5 are sufficient for most applications (see [6,23,24]). Since using $M_1$ and $M_2$ does not add anything fundamentally new, we proceed with the simple version only.
(ii) The proposed procedure encapsulates all the dependence across space and time into \( d \) latent factors. Those latent factors, specified objectively by sample covariances (and autocovariances) of the data, capture all the linear correlations parsimoniously. The real data example in Subsection 6.2 below, and also those not shown in this paper, indicate that the estimated \( d \) is often small.

**Remark 3.2.** The two-subsample procedure’s function is indeed beyond excluding the nugget effect. In fact, it is the first but essential step to improve our analysis. In most related projects, the main task is to separate the signal and noise, i.e., remove nugget effects without losing useful information. Hence, our random partitioning strategy provides an effective way to separate the signal and noise, i.e., remove nugget effects without losing useful information. Hence, as we can see, it eases the extraction of quantities of interest and improves estimation of the remaining parameters.

**Remark 3.3.** It is easy to see that \( \Sigma \Sigma' \) is non-negative definite because for any compatible vector \( \gamma \), we have

\[
\gamma' \Sigma \Sigma' \gamma = (\Sigma' \gamma)'(\Sigma' \gamma) \geq 0.
\]

Hence, all eigenvalues of \( \Sigma \Sigma' \) are either zero or positive. Moreover, \( \text{rank}(\Sigma \Sigma') = d \) because \( \text{rank}(\Sigma) = d \). So \( \text{rank}(\Sigma \Sigma') \) has \( d \) positive eigenvalues. Let \( U_1 \) be the matrix whose columns are the eigenvectors with respect to the positive eigenvalues of \( \Sigma \Sigma' \). Naturally any vector \( a_1 \) orthogonal to \( \mathcal{M}(U_1) \) will result in \( \Sigma \Sigma' a_1 = 0 \). On the other hand, because

\[
\Sigma \Sigma' = A_1 \text{Cov}(x_t, x_t') \text{Cov}(x_t', x_t) A_1',
\]

any vector \( u_1 \) perpendicular to \( \mathcal{M}(A_1) \) will result in \( \Sigma \Sigma' u_1 = 0 \). Moreover, given both \( U_1 \) and \( A_1 \) being \((p_1 \times d)\)-dimensional matrices, we have \( \mathcal{M}(A_1) = \mathcal{M}(U_1) \). Hence, we may find a \( d \times d \) orthonormal matrix \( H_1 \) such that \( A_1 = U_1 H_1 \). Now, we have

\[
\Sigma \Sigma' = U_1 \text{Cov}(H_1 x_t, x_t') \text{Cov}(x_t', H_1 x_t) U_1'.
\]

As we noted above, \((A_1, x_t)\) is not unique. So we may simply let \( A_1^* = U_1 \) and \( x^* = H_1 x \).

### 3.2 Aggregating via random partitioning

The estimation for the latent variable \( \xi_t(\cdot) \) depends on partitioning \( S_o = \{s_1, \ldots, s_p\} \) into two non-overlapping sets \( S_1 \) and \( S_2 \) (see (3.10)). Since the estimation procedure presented in Subsection 3.1 puts \( S_1 \) and \( S_2 \) on equal footing, we set \( p_1 = \lfloor p/2 \rfloor \) and \( p_2 = p - p_1 \). By randomly dividing \( S_o \) into \( S_1 \) and \( S_2 \) with the sizes \( p_1 \) and \( p_2 \) respectively, the estimates for \( \xi_{t,1} \) and \( \xi_{t,2} \) are obtained as in (3.10). We repeat this randomization \( J \) times, where \( J \geq 1 \) is a large integer, leading to the \( J \) pairs of the estimates \((\hat{\xi}_{t,1}, \hat{\xi}_{t,2})\) for \( j = 1, \ldots, J \). The aggregating estimator over the randomized partitions is

\[
\bar{\xi}_t(s_i) = \frac{1}{J} \sum_{j=1}^{J} \xi_{t,j}(s_i), \quad j = 1, \ldots, p, \quad (3.11)
\]

where \( \xi_{t,j}(s_i) \) is a component of either \( \bar{\xi}_{t,1} \) or \( \bar{\xi}_{t,2} \), depending on \( s_i \in S_1 \) or \( S_2 \) in the \( j \)-th randomized partition of \( S_o \). Similar to the Bagging method of [4], the choice of \( J \) is not critical. In our numerical experiments, we set \( J = 100 \).

**Theorem 3.4.** For \( k = 1, \ldots, n \) and \( \ell = 1, \ldots, p \),

\[
E(\{\xi_k(s_{i\ell}) - y_k(s_{i\ell})\}^2 \mid \{y_{i\ell}(s_i)\}) \leq E(\{\bar{\xi}_k(s_{i\ell}) - y_k(s_{i\ell})\}^2 \mid \{y_{i\ell}(s_i)\}), \quad (3.12)
\]
and

\[
E \left( \frac{1}{np} \sum_{t=1}^{n} \sum_{j=1}^{p} (\xi_t(s_j) - \xi_t(s_j)) \right)^2 \mid \{\xi_t(s_i), y_t(s_i)\} \\
\leq E \left( \frac{1}{np} \sum_{t=1}^{n} \sum_{j=1}^{p} (\xi_t(s_j) - \xi_t(s_j)) \right)^2 \mid \{\xi_t(s_i), y_t(s_i)\}\right).
\] (3.13)

Theorem 3.4 is in the same spirit as Breiman’s inequality for Bagging (see [4, (4.2)]). Note that all the conditional expectations in Theorem 3.4 above are taken with respect to the random partitioning of the location set \( \mathcal{S}_o \) into \( \mathcal{S}_1 \) and \( \mathcal{S}_2 \). There are in total \( p_0 = p!/p_1!p_2! \) different partitions, each being taken with probability \( 1/p_0 \). Denote by \( \hat{\xi}_k^{(1)}(\cdot), \ldots, \hat{\xi}_k^{(p_0)}(\cdot) \) the resulting \( p_0 \) estimates as in (3.10). Then

\[
E(\hat{\xi}_k(s_l) - y_k(s_l))^2 \mid \{y_k(s_i)\} = E \left( \left( \frac{1}{J} \sum_{l=1}^{J} (\hat{\xi}_k(s_l) - y_k(s_l)) \right)^2 \mid \{y_k(s_i)\} \right) \\
\leq E \left( \frac{1}{J} \sum_{l=1}^{J} ((\hat{\xi}_k(s_l) - y_k(s_l))^2 \mid \{y_k(s_i)\} \right) \\
= \frac{1}{p_0} \sum_{j=1}^{p_0} (\hat{\xi}_k^{(j)}(s_l) - y_k(s_l))^2 \\
= E((\hat{\xi}_k(s_l) - y_k(s_l))^2 \mid \{y_k(s_i)\}).
\]

This completes the proof for (3.12). Note that (3.13) can be established in the same manner.

### 3.3 Scalability to large datasets

The estimator \( \hat{A}_1 \) in (3.8) was obtained from an eigenanalysis which requires \( O(p_1 p_2^2) \) operations. This is computational challenging when \( p \) is large. However, our approach can be easily adapted to large \( p \), which is in the spirit of ‘divide and conquer’.

We randomly divide \( \mathcal{S}_o \) into \( p/q \) sets \( \mathcal{S}_1, \ldots, \mathcal{S}_q^* \), and each \( \mathcal{S}_i^* \) contains \( q \) locations, where \( q \) is an integer such that the eigenanalysis for \( q \times q \) matrices can be performed comfortably with the available computing capacity. We estimate \( \xi_i(\cdot) \) at the \( q \) locations in \( \mathcal{S}_i^* \) for each \( i = 1, \ldots, p/q \) separately using the aggregation algorithm below.

(i) Randomly select \( q \) locations from \( \mathcal{S}_0 - \mathcal{S}_i^* \).

(ii) Combine the data on the locations in \( \mathcal{S}_i^* \) and the locations selected in (i). By treating the combined data as the whole sample, calculate \( \hat{\xi}_i(s) \) for \( s \in \mathcal{S}_i^* \) as in (3.10).

(iii) Repeat (i) and (ii) above \( J \) times, and aggregate the estimates as in (3.11).

Note through the ‘divide and conquer’ strategy, the computational cost may be reduced to \( O(pq^2) \), which is much less than the standard cost of \( O(p^3) \).

Alternatively, we can randomly choose \( 2q \) locations from \( \mathcal{S}_0 \) to perform the estimation (3.10). Repeating the estimation a large number (say, greater than \( Jp/(2q) \)) of times, we then aggregate the estimates at each location as in (3.11). This is a computationally more efficient approach with the drawback that the number of the estimates obtained at each location is not directly under control.

### 3.4 Regression estimation

In the presence of observable covariant \( z_i(\cdot) \) in (2.1), the regression coefficient vector \( \beta(\cdot) \) can be estimated by the least squares method. To this end, let

\[
y(s_i) = (y_1(s_i), \ldots, y_n(s_i))^\prime, \quad Z(s_i) = (z_1(s_i), \ldots, z_n(s_i))^\prime.
\] (3.14)

It follows from (2.1) that

\[
y(s_i) = Z(s_i)\beta(s_i) + e(s_i),
\]
Thus the least squares estimator for $\beta(s_i)$ is defined as

$$\hat{\beta}(s_i) = \{Z(s_i)'Z(s_i)\}^{-1}Z(s_i)'y(s_i), \quad i = 1, \ldots, p.$$  \hspace{1cm} (3.15)

Then by replacing the original data $y_t(s_i)$ by the regression residuals $y_t(s_i) - \hat{z}_t(s_i)'\hat{\beta}(s_i)$, we proceed to estimate the finite-dimensional structure of $\xi_t(\cdot)$ as described in Subsection 3.1 above.

However, in the presence of the endogeneity in the sense $\text{Cov}(z_t(s), \xi_t(s)) \neq 0$, the regression estimator $\hat{\beta}(s_i)$ in (3.15) is practically an estimator for

$$\beta(s_i)^* \equiv \beta(s_i) + \text{Var}(z_t(s_i))^{-1}\text{Cov}(z_t(s_i), \xi_t(s_i))$$

instead, as (2.1) can be written as $y_t(s) = z_t(s)'\beta(s)^* + \xi_t(s)^* + \epsilon_t(s)$, where

$$\xi_t(s)^* = \xi_t(s) - z_t(s)'\text{Var}(z_t(s_i))^{-1}\text{Cov}(z_t(s_i), \xi_t(s_i)).$$

It is easy to see that

$$\text{Cov}(z_t(s), \xi_t(s)^*) = 0.$$ 

Hence $\hat{\beta}(s_i)$ is a consistent estimator for $\beta(s_i)^*$. Furthermore, the estimation based on the residuals described above is still valid though the finite-dimensional structure (2.6) is now imposed upon the latent process $\xi_t(s)^*$ instead.

## 4 Kriging

We firstly state a general lemma on linear prediction which shows explicitly the terms required in order to carry out kriging for spatio-temporal process $y_t(s)$.

**Lemma 4.1.** For any random vectors $\zeta$ and $\eta$ with $E(||\zeta||^2 + ||\eta||^2) < \infty$, the best linear predictor for $\zeta$ based on $\eta$ is defined as $\tilde{\zeta} = \alpha_0 + B_0\eta$, where

$$(\alpha_0, B_0) = \arg\inf_{\alpha, B} E\{||\zeta - \alpha - B\eta||^2\}.$$ 

In fact,

$$B_0 = \text{Cov}(\zeta, \eta)\{\text{Var}(\eta)\}^{-1}, \quad \alpha_0 = E\zeta - B_0E\eta.$$ 

Furthermore,

$$E\{(\tilde{\zeta} - \zeta)(\tilde{\zeta} - \zeta)'\} = \text{Var}(\zeta) - \text{Cov}(\zeta, \eta)\{\text{Var}(\eta)\}^{-1}\text{Cov}(\eta, \zeta).$$  \hspace{1cm} (4.1)

With the above lemma, we can predict any value $y_t(s)$. With two scenarios considered below, we illustrate how to calculate inverses of large covariance matrices by taking advantages from the finite-dimensional structure (2.6): all matrices to be inverted are of the sizes $d \times d$ only, regardless of the size of $p$. Technically we repeatedly use the following formulas for the inverses of partitioned matrices.

**Lemma 4.2.** For an invertible block-partitioned matrix

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix},$$

it holds that

$$H^{-1} = \begin{pmatrix} H_{11}^{-1} + H_{11}^{-1}H_{12}(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1}H_{21}H_{11}^{-1} & -H_{11}^{-1}H_{12}(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1} \\ -(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1}H_{21}H_{11}^{-1} & (H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1} \end{pmatrix}.$$  \hspace{1cm} (4.2)

provided $H_{11}^{-1}$ exists.
Formula (4.2) can be proved by checking $H^{-1}H = I$ directly. Note by the famous Shermann-Woodbury-Morrison formula,

$$ (H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1} = H_{22}^{-1} + H_{22}^{-1}H_{21}(H_{11} - H_{12}H_{22}^{-1}H_{21})^{-1}H_{12}H_{22}^{-1} \tag{4.3} $$

provided both $H_{11}^{-1}$ and $H_{22}^{-1}$ exist.

4.1 Kriging over space

The goal is to predict the unobserved value $y_t(s_0)$ for some $s_0 \in S$, $1 \leq t \leq n$, and $s_0 \neq s_j$ for $1 \leq j \leq p$, based on the observations $y_t = (y_t', y_t')'$ only, where $y_{t,1}$ and $y_{t,2}$ are defined as in (3.2). We introduce two predictors below. We always use the notation $K_h(\cdot) = h^{-1}K(\cdot/h)$, where $K(\cdot)$ denotes a kernel function, $h > 0$ is a bandwidth, and $K$ and $h$ may be different at different places.

To simplify the notation, we assume $\beta(s) \equiv 0$ in (2.1). As indicated in Subsection 3.4, this effectively implies to replace the observations $y_t(s_j)$ by the regression residuals. For kriging, we also need to estimate $\beta(s_0)$ based on $\beta(s_j)$, $j = 1, \ldots, p$, given in (3.15). It can be achieved by, for example, using the kernel smoothing

$$ \hat{\beta}(s_0) = \sum_{j=1}^{p} \hat{\beta}(s_j)K_h(s_j - s_0) / \sum_{j=1}^{p} K_h(s_j - s_0), \tag{4.4} $$

where $K(\cdot)$ is a density function defined on $\mathbb{R}^2$, $h > 0$ is a bandwidth. Furthermore, a local linear smoothing can be applied to improve the accuracy of the estimation (see, e.g., [11, Chapter 3]). By the standard argument it can be shown (see the supplementary document) that

$$ |\hat{\beta}(s_0) - \beta(s_0)| = O_p(h^2 + n^{-1/2}), $$

provided that the conditions in Theorem 5.3 in Subsection 5.2 below hold. Note that if $\beta(s_j), j = 1, \ldots, p$, were all known, the above error rate reduces to $O_p(h^2)$, as $\beta(\cdot)$ is deterministic and continuous (see Condition 4 in Subsection 5.2 below). The term of order $n^{-1/2}$ reflects the errors in estimation for $\beta(s_j)$.

In the rest of Section 4, we adhere with the assumption $\beta(s) \equiv 0$.

It follows from Lemma 4.1 that the best linear predictor for $\hat{y}_t(s_0)$ based on $y_t$ is

$$ \hat{y}_t(s_0) = \text{Cov}(y_t(s_0), y_t)\text{Var}(y_t)^{-1}y_t. \tag{4.5} $$

It follows from (4.1) that

$$ E[\tilde{y}_t(s_0) - y_t(s_0)]^2 = \text{Var}(y_t(s_0)) - \text{Cov}(y_t(s_0), y_t)\text{Var}(y_t)^{-1} \text{Cov}(y_t, y_t(s_0)) = \sigma(s_0)^2 + \text{Var}[\xi_t(s_0)] - \text{Cov}(\xi_t(s_0), \xi_t)[\text{Var}(\xi_t) + D]^{-1}\text{Cov}(\xi_t, \xi_t(s_0)), \tag{4.6} $$

where $D = \text{Var}(\varepsilon_t)$ is a diagonal matrix, $\varepsilon_t = (\varepsilon_t', \varepsilon_t')'$ and $\xi_t = (\xi_t', \xi_t')'$ (see (3.2)).

To apply predictor $\hat{y}_t(s_0)$ in (4.5) in practice, we need to estimate both $\text{Cov}(y_t(s_0), y_t)$ and $\text{Var}(y_t)$. Since $\text{Cov}(y_t(s_0), y_t) = \text{Cov}(\xi_t(s_0), y_t)$, it can be estimated by

$$ c(s_0) = \frac{1}{n} \sum_{k=1}^{n} (\hat{\xi}_t(s_0) - \hat{\xi}(s_0))(y_k - \bar{y}), $$

where $\hat{\xi}_t(s_0)$ is a kernel estimator for $\xi_t(s_0)$ defined as

$$ \hat{\xi}_t(s_0) = \sum_{j=1}^{p} \hat{\xi}_t(s_j)K_h(s_j - s_0) / \sum_{j=1}^{p} K_h(s_j - s_0), \tag{4.7} $$

with $\hat{\xi}_t(s_1), \ldots, \hat{\xi}_t(s_p)$ defined in (3.10) (see also (4.4) above), and

$$ \hat{\xi}(s_0) = n^{-1} \sum_t \hat{\xi}_t(s_0). $$
Thus a realistic predictor for $y_t(s_0)$ is

$$\hat{y}_t'(s_0) = c(s_0)\hat{\Sigma}^{-1}_y y_t,$$

where

$$\hat{\Sigma}_y = n^{-1} \sum_{k=1}^{n} (y_k - \bar{y})(y_k - \bar{y})'$$

is the sample variance of $y_t$. Nevertheless it turns out that

$$\hat{y}_t'(s_0) = \hat{\xi}_t(s_0).$$

(4.9)

To show this, let

$$w_j = \frac{K_h(s_j - s_0)}{\sum_{j=1}^{p} K_h(s_j - s_0)}.$$

It follows from (3.10) that

$$\hat{y}_t'(s_0) = (w_1, \ldots, w_p) \left[ \frac{1}{n} \sum_{k=1}^{n} (\hat{\xi}_k - \bar{\xi})(y_k - \bar{y})' \right] \hat{\Sigma}_y^{-1} y_t$$

$$= (w_1, \ldots, w_p) \left( \begin{array}{cc} \hat{A}_1\hat{A}_1' & 0 \\ 0 & \hat{A}_2\hat{A}_2' \end{array} \right) \left[ \frac{1}{n} \sum_{k=1}^{n} (y_k - \bar{y})(y_k - \bar{y})' \right] \hat{\Sigma}_y^{-1} y_t$$

$$= (w_1, \ldots, w_p) \left( \begin{array}{cc} \hat{A}_1\hat{A}_1' & 0 \\ 0 & \hat{A}_2\hat{A}_2' \end{array} \right) y_t = (w_1, \ldots, w_p) \left( \begin{array}{c} \hat{\xi}_{t,1} \\ \hat{\xi}_{t,2} \end{array} \right) = \hat{\xi}_t(s_0).$$

It is worth pointing out that expression (4.8) involves inverting $p \times p$ matrix $\hat{\Sigma}_y$, which is difficult when $p$ is large, while (4.9) paves the way for computing the predictor $\hat{y}_t'(s_0)$ without the need to compute $\hat{\Sigma}_y^{-1}$ directly.

By Theorem 3.4, a better predictor than $\hat{y}_t'(s_0)$ in (4.9) is

$$\tilde{y}_t'(s_0) \equiv \tilde{\xi}_t(s_0) = \frac{\sum_{j=1}^{p} \tilde{\xi}(s_j)K_h(s_j - s_0)}{\sum_{j=1}^{p} K_h(s_j - s_0)},$$

(4.10)

where $\tilde{\xi}_t(s_j)$ is defined in Equation (3.11).

Both $\hat{y}_t'(s_0)$ and $\tilde{y}_t'(s_0)$ are the approximate linear estimators for $\xi_t(s_0)$ based on $\xi_t(s_1), \ldots, \xi_t(s_p)$. Note that $y_t(s_0) = \xi_t(s_0) + \varepsilon_t(s_0)$, and the nugget effect term $\varepsilon_t(s_0)$ is unpredictable. The best (unrealistic) predictor for $y_t(s_0)$ is $\xi_t(s_0)$. It is indeed recommended to predict $\xi_t(s_0)$ instead of $y_t(s_0)$ directly (see also [10, pp. 136–137]).

**Remark 4.3.**  
(i) The realistic kriging estimators $\hat{y}_t'(s_0)$ and $\tilde{y}_t'(s_0)$ actually make the full use of all the available data, since they were induced from (4.5). Note that the ideal (and unrealistic) predictor for $y_t(s_0)$ is $\sum_{1 \leq j \leq d} a_j(s_0) x_j$, and $\hat{x}_{tj}$ and $\tilde{x}_{tj}$ are the estimators for $x_{tj}$ based on all the available data from time 1 to $n$. It follows from (4.7)–(4.9) that $\tilde{y}_t'(s_0)$ is a realistic optimal predictor for $y_t(s_0)$ based on $\{ \hat{x}_{tj}, j = 1, \ldots, d \}$.

(ii) When the number of observations in the vicinity of $s_0$ is small, the kernel based predictor (4.7) may perform poorly. One alternative is to impose a parametric spatial covariance function and to perform the kriging based on the parametric model (see [10, Subsections 4.1.1 and 6.1]). How to identify an appropriate parametric model using the nonparametric analysis presented in this paper deserves a dedicated study.

**4.2  Kriging in time**

**4.2.1  Prediction methods**

The goal now is to predict the future values $y_{n+j}(s_1), \ldots, y_{n+j}(s_p)$, for some $j \geq 1$, based on $y_n, \ldots, y_{n-j_0}$, where $0 \leq j_0 < n$ is a prescribed integer. When $j_0 = n - 1$, we use all the available data to predict the
future values. Since \( \varepsilon_{t+j}(\cdot) \) is unpredictable, a more effective approach is to predict

\[
x_{n+j} = (x_{n+j,1}, \ldots, x_{n+j,d})'
\]

based on \( x_n, \ldots, x_{n-j_0} \), as the ideal predictor for \( y_{n+j}(s_i) \) is \( \xi_{n+j}(s_i) \) (see (3.1)).

Since our procedure to recover the latent process \( x_t \) requires to split \( y_t \) into two subvectors \( y_{t,1} \) and \( y_{t,2} \), leading to two different configurations \( x_t \) and \( x_t' \) in (3.3), we apply the prediction procedure in Subsection 4.2.2 below to each of \( x_t \) and \( x_t^* \). Then the predictors for \( y_{n+j,1} \) and \( y_{n+j,2} \) are defined as

\[
y_{n,1}(j) = A_1 x_n(j), \quad y_{n,2}(j) = A_2 x_n^*(j),
\]

where \( x_n(j) \) is the predictor for \( x_{n+j} \), and \( x_n^*(j) \) is the predictor for \( x_{n+j}^* \). In practice, \( A_i, x_t \) and \( x_t^* \) are replaced by their estimators defined in (3.8) and (3.9).

The predictors defined above depend on a single partition \( S_n = S_1 \cup S_2 \). By repeatedly randomly partitioning \( S_n \) for \( J \) times, we may obtain the aggregated predicted values for \( y_{n+j}(s_i) \) in the same manner as in (3.11).

Since \( \xi_t(s_1), \ldots, \xi_t(s_p) \) are correlated with each other, we should not model \( \xi_t \) at each location separately. Instead modeling the factor process \( x_t \) catches the temporal dynamics much more parsimoniously.

An alternative approach, not pursued here, would be to build a dynamic model for \( x_t \), leading to the model-based forecasts. For example, [9] adopted the Gaussian AR(1) specification for the latent process and facilitated the forecasting by a Kalman filter.

### 4.2.2 Predicting \( x_{n+j} \) and \( x_{n+j}^* \)

We only state the method for predicting \( x_{n+j} \). It can be applied to predicting \( x_{n+j}^* \) exactly in the same manner.

Let \( X' = (x'_1, \ldots, x'_{n-j_0}) \),

\[
W_k \equiv \text{Var} \left( \begin{array}{c}
x_t' \\
x_{t-1}' \\
\vdots \\
x_{t-k}'
\end{array} \right) = 
\begin{pmatrix}
\Sigma_x(0) & \Sigma_x(1) & \cdots & \Sigma_x(k) \\
\Sigma_x(1)' & \Sigma_x(0) & \cdots & \Sigma_x(k-1) \\
\vdots & \vdots & \ddots & \vdots \\
\Sigma_x(k)' & \Sigma_x(k-1)' & \cdots & \Sigma_x(0)
\end{pmatrix}, \quad k \geq 0,
\]

(4.12)

\[
R_{j_0} \equiv \left( \Sigma_x(j), \Sigma_x(j+1), \ldots, \Sigma_x(j+j_0) \right),
\]

where \( \Sigma_x(k) = \text{Cov}(x_{t+k}, x_t) \). By Lemma 4.1, the best linear predictor for \( x_{n+j} \) is

\[
x_{n}(j) = R_{j_0} W_{j_0}^{-1} X.
\]

The key is to be able to calculate the inverse of \((j_0+1)d \times (j_0+1)d\) matrix \( W_{j_0} \). This can be done by calculating \( W_0^{-1}, W_1^{-1}, \ldots \) recursively based on

\[
W_{k+1}^{-1} = 
\begin{pmatrix}
W_k^{-1} + W_k^{-1} U_k V_k W_k^{-1} - W_k^{-1} U_k V_k \\
-V_k U_k' W_k^{-1} & V_k
\end{pmatrix},
\]

(4.13)

where

\[
U_k' = (\Sigma_x(k+1)', \ldots, \Sigma_x(1)'), \quad V_k = (\Sigma_x(0) - U_k' W_k^{-1} U_k)^{-1}
\]

(see (4.2)). Note that only \( d \times d \) inverse matrices are involved in this recursion.

In practice we replace \( \Sigma_x(k) \) in \( R_{j_0} \) and \( W_{j_0} \) by \( \hat{\Sigma}_x(k) = \hat{A}_1 \hat{\Sigma}_{y,1}(k) \hat{A}_1 \), and replace \( X \) by

\[
\hat{X} = (y_{t,1}' \hat{A}_1, \ldots, y_{t-k,1}' \hat{A}_1)',
\]

where

\[
\hat{\Sigma}_{y,1}(k) = \frac{1}{n} \sum_{t=1}^{n-k} (y_{t+k,1} - \hat{y}_1)(y_{t,1} - \hat{y}_1)', \quad \hat{y}_1 = \frac{1}{n} \sum_{t=1}^{n} y_{t,1}.
\]
The resulting predictor for \(x_{n+j}\) is denoted by \(\hat{x}_n(j)\).

We may define \(\hat{x}^*_n(j)\) in the same manner as \(\hat{x}_n(j)\) with \((y_{t,1}, \hat{A}_1)\) replaced by \((y_{t,2}, \hat{A}_2)\). Consequently, the practical feasible predictor for \(y_{n+j}\) is defined in two similar formulas

\[
\hat{y}_{n,1}(j) = \hat{A}_1 \hat{x}_n(j), \quad \hat{y}_{n,2}(j) = \hat{A}_2 \hat{x}^*_n(j)
\]  

(4.14).

### 4.3 Handling missing values

It is not uncommon that a large data set contains some missing values. We assume that the number of missing values is small in the sense that the number of the available observations at each given time \(t\) is of the order \(p\), and the number of the available observations at each location \(s_i\) is of the order \(n\). We outline below how to apply the proposed method when some observations are missing.

First for \(\Sigma \equiv (\sigma_{ij})\) defined in (3.4), we may estimate each \(\sigma_{ij}\) separately using all the available pairs \((y^1_{t,i}, y^2_{t,j})\) with \(1 \leq i \leq n\), where \(y^\ell_{t,i}\) denotes the \(i\)-th element of \(y_{t,\ell}\), \(\ell = 1, 2\). With the estimated \(\Sigma\), we may derive the estimators \(\hat{A}_1\) and \(\hat{A}_2\) as in (3.8).

For the simplicity in notation, suppose that \(y_1(s_1)\) is missing. Let \(y^\ell_t\) denote all the available observations at time \(t = 1\). By Lemma 4.1, the kriging predictor for \(y_1(s_1)\) is

\[
\tilde{\xi}_1(s_1) = \text{Cov}(y_1(s_1), y^\ell_t) \{\text{Var}(y^\ell_t)\}^{-1} y^\ell_t.
\]  

(4.15)

We may estimate \(\text{Cov}(y_1(s_1), y^\ell_t)\) and \(\text{Var}(y^\ell_t)\) in the same manner as that for estimating \(\Sigma\) described above. Replacing all the missing values with their kriging estimates, we may proceed the estimation for \(\tilde{\xi}_1(s_j)\) and \(\tilde{\xi}_1(s_j)\) as in Subsections 3.1 and 3.2.

### 5 Asymptotic properties

In this section, we investigate the asymptotic properties of the proposed methods. For any matrix \(M\), let \(\|M\|_{\min} = \sqrt{\lambda_{\min}(MM^\prime)}\) and \(\|M\| = \sqrt{\lambda_{\max}(MM^\prime)}\), where \(\lambda_{\min}\) and \(\lambda_{\max}\) denote, respectively, the minimum and the maximum eigenvalues. When \(M\) is a vector, \(\|M\|\) reduces to its Euclidean norm.

#### 5.1 On latent finite-dimensional structures

We state in this subsection some asymptotic results on the estimation of the factor loading spaces \(\mathcal{M}(A_1)\) and \(\mathcal{M}(A_2)\). They pave the way to establish the properties for the kriging estimation presented in Subsection 5.2 below. Proposition 5.1 below is similar to those in [23] and [6] but with the extra features due to the graph Laplacian incorporated in order to perturb the continuity over space. Nevertheless its proof is similar and, therefore, is omitted.

For any two \(k \times d\) orthogonal matrices \(B_1\) and \(B_2\) with \(B_1^\prime B_1 = B_2^\prime B_2 = I_d\), we measure the distance between the two linear spaces \(\mathcal{M}(B_1)\) and \(\mathcal{M}(B_2)\) by

\[
D(\mathcal{M}(B_1), \mathcal{M}(B_2)) = \sqrt{1 - \frac{1}{d} \text{tr}(B_1 B_1^\prime B_2 B_2^\prime)}.
\]  

(5.1)

It can be shown that \(D(\mathcal{M}(B_1), \mathcal{M}(B_2)) \in [0, 1]\), being 0 if and only if \(\mathcal{M}(B_1) = \mathcal{M}(B_2)\), and 1 if and only if \(\mathcal{M}(B_1)\) and \(\mathcal{M}(B_2)\) are orthogonal. We introduce some regularity conditions first. Put

\[
y_t = (y_t(s_1), \ldots, y_t(s_p))^\prime, \quad Z_t = (z_t(s_1), \ldots, z_t(s_p)).
\]

**Condition 1.** \(\{(y_t, Z_t), t = 0, \pm 1, \pm 2, \ldots\}\) is a strictly stationary and \(\alpha\)-mixing process with

\[
\max_{1 \leq i \leq p} \{\mathbb{E}|y_t(s_i)|^\gamma + \mathbb{E}\|z_t(s_i)\|^{\gamma}\} < \infty
\]
for some $\gamma > \max\{\beta, 4\}$, $\beta > 2$ and the $\alpha$-mixing coefficients $\alpha_m$ satisfying the condition

$$\alpha_m = O(m^{-\theta}) \quad \text{for some} \quad \theta > \gamma\beta/(\gamma - \beta). \quad (5.2)$$

Furthermore, $\min_{1 \leq i \leq p} \lambda_{\min}(\text{Var}(z_i(s_i))) > c_0$ for some positive constant $c_0$.

**Condition 2.** Let $\Sigma_2 = \text{Cov}(x_i, x_i^*)$, where $x_i$ and $x_i^*$ are defined in (3.3). There exists a constant $\delta \in [0, 1]$ for which $\|\Sigma_2\|_{\min} \asymp \|\Sigma_2\|^{1-\delta}$.

The constant $\delta$ in Condition 2 reflects the strength of factors. Intuitively a strong factor is linked with most components of $y_{i,1}$ and $y_{i,2}$, implying that the corresponding coefficients in $A_1$ or $A_2$ are non-zero. Therefore it is relatively easy to recover those strong factors from the observations. Unfortunately the mathematical definition of the factor strength is tangled with the standardization condition $A_1 A_1' = A_2 A_2' = I_d$ (see [23, Remark 1(i)] and [24, Lemma 1]). To simplify the presentation, Condition 2 assumes that all the factors in (3.3) are of the same strength which is measured by a constant $\delta \in [0, 1]$; $\delta = 0$ indicates that the strength of the factors is at its strongest, and $\delta = 1$ corresponds to the weakest factors.

**Proposition 5.1.** Let Conditions 1 and 2 hold, and $p^3 n^{-1/2} + pn^{-\beta/2} + p^{2\beta-2} \|L\| \to 0$ as $n \to \infty$.

Then

(i) $|\lambda_i - \lambda_l| = O_p(p^{2-\delta} n^{-1/2} + \tau \|L\|)$ for $1 \leq i \leq d$;

(ii) $|\tau_j| = O_p(p^{2n-1} + \tau \|L\|)$ for $d < i \leq p$; and

(iii) $D(M(A), M(A)) = O_p(p^2 n^{-1/2} + p^{2\beta-2} \tau \|L\|)$ ($i = 1, 2$), provided that $d$ is known.

**Remark 5.2.** (i) Proposition 5.1 indicates that stronger factors result in a better estimation for the factor loading spaces, and, consequently, a better recovery of the factor process. This is due to the fact that $\lambda_d - \lambda_{d+1}$ increases as $\delta$ decreases, where $\lambda_i$ denotes the $i$-th largest eigenvalue of $\Sigma \Sigma'$, and $\Sigma$ is defined in (3.4). Especially with the strongest factors (i.e., $\delta = 0$), $D(M(\hat{A}), M(A))$ attains the standard error rate $n^{-1/2} + p^{-\beta/2} \tau \|L\|$. This phenomenon is coined as ‘blessing of dimensionality’ as in [23].

(ii) Proposition 5.1(iii) can be made adaptive to the unknown $d$ (see [2, Remark 5]). See also [6, Theorem 2.4] on how to make $d$ defined in (3.7) be a consistent estimator for $d$.

(iii) The condition $p^{2\beta-2} \|L\| \to 0$ in Proposition 5.1 controls the perturbation between $\hat{A}$ and $A$, which is implied by either $p^{2\beta-1} \tau \to 0$ (as $\|L\| \leq p$) or $\|L\| \leq C$ and $p^{2\beta-2} \tau \to 0$. By the perturbation theory (see [14, Theorem 8.1.10]), the bound $\|\hat{A} - A\|$ depends on $\|\Sigma \Sigma' + \tau L - \Sigma \Sigma'\|$, which is bounded from above by $\|\Sigma \Sigma' - \Sigma \Sigma'\| + \tau \|L\|$. This leads to the upper bound of Proposition 5.1(iii).

### 5.2 On kriging

We now consider the asymptotic properties for the kriging methods proposed in Section 4. To simplify the presentation, we always assume that $d$ is known. We introduce some regularity conditions first.

**Condition 3.** The kernel $K(\cdot)$ is a symmetric density function on $\mathbb{R}^2$ with a bounded support.

**Condition 4.** In (3.1), $\beta(\cdot)$ and $\alpha_j(\cdot)/\|\alpha(s_i)\|$, $j = 1, \ldots, d$, are twice continuously differentiable and bounded functions on $S$, where $\alpha(s_0) = (a_1(s_0), \ldots, a_d(s_0))$.

**Condition 5.** There exists a positive and continuously differentiable sampling intensity $f(s)$ on $S$ such that as $p \to \infty$,

$$\frac{1}{p} \sum_{s \in S} I(s \in A) = \int_A f(s) ds(1 + o(1)) \quad (5.3)$$

holds for any measurable set $A \subseteq S$.

Theorem 5.3 below presents the asymptotic properties of the two spatial kriging methods in (4.9) and (4.10). Since

$$E[(\hat{y}_t(s_0) - y_t(s_0))^2] = E[(\hat{y}_t(s_0) - \xi_t(s_0))^2] + \text{Var}(\xi_t(s_0)),$$

it is more relevant to measure the difference between a predictor and $\xi_t(s_0)$ directly.
Theorem 5.3. Let the bandwidth $h \to 0$, $ph \to \infty$ and $p^\delta n^{-1/2} + p\beta/2 + p^{2\delta - 2}\|L\| \to 0$ as $n \to \infty$. It holds under Conditions 1–5 that
\[
\max\{\|\hat{y}_t(s_0) - \xi_t(s_0)\|, \|\hat{y}_t(s_0) - \xi_t(s_0)\|\} = O_p(h^2 + p^\delta (nh)^{-1/2} + (ph)^{-1/2} + p^{2\delta - 2}h^{-1/2}\|L\|).
\]

Theorem 5.4 below considers the convergence rates for the kriging predictions in time. Recall $\hat{y}_{n,1}(j)$, $\hat{y}_{n,2}(j)$, $\bar{x}_n(j)$ and $\bar{x}_n^*(j)$ as defined in (4.14).

Theorem 5.4. Let Conditions 1 and 2 hold. As $n, p \to \infty$ and $p^\delta n^{-1/2} + p^{2\delta - 2}\|L\| \to 0$,
\begin{enumerate}[(a)]  
\item $p^{-1/2}\|\bar{x}_n(j) - x_n(j)\| = O_p(p^\delta n^{-1/2} + p^{2\delta - 2}\|L\| + p^{-1/2})$, $p^{-1/2}\|\bar{x}_n^*(j) - x_n^*(j)\| = O_p(p^\delta n^{-1/2} + p^{2\delta - 2}\|L\| + p^{-1/2})$;  
\end{enumerate}  and
\begin{enumerate}[(b)]  
\item $p^{-1/2}\|\hat{y}_{n,i}(j) - y_{n,i}(j)\| = O_p(p^\delta n^{-1/2} + p^{2\delta - 2}\|L\| + p^{-1/2})$ for $i = 1, 2$.  
\end{enumerate}

Theorems 5.3 and 5.4 indicate that strong factors result in better predictions (see also Remark 5.2(i) above).

6 Numerical properties

We illustrate the finite sample properties of the proposed methods via both simulated and real data.

6.1 Simulation

For simplicity, we let $s_1, \ldots, s_p$ be drawn randomly from the uniform distribution on $[-1,1]^2$ and $y_t(s_i)$ be generated from (3.1) in which $d = 3$, $\varepsilon_t(s)$ are independent and standard normal, and
\[
\begin{align*}
    a_1(s) &= s_1/2, \quad a_2(s) = s_2/2, \quad a_3(s) = (s_1^2 + s_2^2)/2, \\
    x_{i1} &= -0.8x_{i-1,1} + \epsilon_{i1}, \quad x_{i2} = \epsilon_{i2} - 0.5\epsilon_{i-1,2}, \quad x_{i3} = -0.6x_{i-1,3} + \epsilon_{i3} + 0.3\epsilon_{i-1,3}.
\end{align*}
\]

In the above expressions, $\epsilon_{i\ell}$ are independent and standard normal. The signal-noise ratio, which is defined as
\[
\frac{\int_{s\in[-1,1]^2} \sqrt{\text{VAR}(\xi_t(s))} \text{d}s}{\int_{s\in[-1,1]^2} \sqrt{\text{VAR}(\varepsilon_t(s))} \text{d}s},
\]
is about 0.72.

With $n = 80$, $n = 160$ or $n = 320$, and $p = 50$, $p = 100$ or $p = 200$, we draw 200 samples from each setting. With each sample, we calculate $d$ as in (3.7), and the factor loadings $\hat{A}_1$ and $\hat{A}_2$ as in (3.8). For the latter, we choose the tuning parameter $\tau$ over 101 grid points between 0 and 10 by a five-fold cross-validation: we randomly divide $s_1, \ldots, s_n$ into 5 groups of the same size. Each time we use the data at the locations in four groups for estimation, and predict the values at the locations in the other group by spatial kriging (4.9). We use Gaussian kernel in (4.7) with bandwidth $h$ selected by the leave-one-out cross validation method.

As the estimated value $\hat{d}$ may not always be equal to $d$, and $A_1$ and $A_2$ are not half-orthogonal matrices in the model specified above, we extend the distance measure for two linear spaces (5.1) as follows:
\[
D(M(\hat{A}_i), M(A_i)) = \left(1 - \frac{1}{\max(d, \hat{d})} \text{tr}\{\hat{A}_i \hat{A}_i' (A_i'A_i)^{-1} A_i'A_i^{-1}\}\right)^{1/2}.
\]
It can be shown that $D(M(\hat{A}_i), M(A_i)) \in [0,1]$, being 0 if and only if $M(\hat{A}_i) = M(A_i)$, and 1 if and only if $M(\hat{A}_i)$ and $M(A_i)$ are orthogonal. It reduces to (5.1) when $\hat{d} = d$ and $A_i'A_iI = I_d$.

Figure 1 depicts the boxplots of the average distance
\[
\frac{1}{2}(D(M(\hat{A}_1), M(A_1)) + D(M(\hat{A}_2), M(A_2)))
\]
over 200 replications under different settings. As expected, the errors in estimating $M(A_1)$ and $M(A_2)$ decrease as $n$ increases. Perhaps more interesting is the phenomenon that the estimation errors do not
increase as the number of locations \( p \) increases. Note that the three factors specified in the above model are all strong factors. According to Proposition 5.1(iii), \( D(\mathcal{M}(\hat{A}_1),\mathcal{M}(\hat{A}_2)) = O_p(n^{-1/2} + \|L\|/p^2) \) when \( \delta = 0 \) (see also Remark 5.2(i)). Figure 1 also shows that the estimation errors with \( p = 50 \) are significantly greater than those with \( p = 100 \) and \( p = 200 \). This is due to greater errors in estimating \( d \) with smaller \( p \) (see Table 1 below). Note that Proposition 5.1(iii) assumes \( d \) to be known.

Figure 2 presents the boxplots of

\[
\text{MSE}(\hat{\xi}_t) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (\bar{\xi}_t(s_j) - \xi_t(s_j))^2, \quad \text{MSE}(\bar{\xi}_t) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (\bar{\xi}_t(s_j) - \xi_t(s_j))^2, \quad (6.1)
\]

where \( \bar{\xi}_t(s_j) \) and \( \bar{\xi}_t(s_j) \) are defined, respectively, in (3.10) and (3.11). We set \( J = 100 \) for the aggregation estimates \( \xi_t(s_j) \). As shown by Theorem 3.4, \( \bar{\xi}_t(s_j) \) always provides more accurate estimate for \( \xi_t(s_j) \) than \( \hat{\xi}_t(s_j) \). Furthermore, the MSE decreases when either \( n \) or \( p \) increases.

Note that estimating \( \hat{A}_1 \) and \( \hat{A}_2 \) with \( \tau > 0 \) makes use the continuity of the loading functions \( a_i(\cdot) \). Table 2 lists the means and the standard errors, over 200 replications, of MSE\((\hat{\xi}_t)\) with \( \hat{\xi}_t(s_j) \) calculated using either \( \tau \) selected by the five-fold cross-validation (i.e., \( \tau > 0 \)) or \( \tau = 0 \). The improvement from using the continuity is more pronounced when \( n \) and \( p \) are small.

To illustrate the kriging performance, with each sample we also draw additional 50 ‘post-sample’ data points at the locations randomly drawn from \( U[-1,1]^2 \). For each \( t = 1, \ldots, n \), we calculate the spatial kriging estimate \( \hat{y}_t^k(\cdot) \) in (4.9) at each of the 50 post-sample locations. The mean squared predictive error is computed as

\[
\text{MSPE}(\hat{y}_t^k) = \frac{1}{50n} \sum_{t=1}^{n} \sum_{s_0 \in \mathcal{S}^*} (\hat{y}_t^k(s_0) - y_t(s_0))^2, \quad (6.2)
\]

where \( \mathcal{S}^* \) is the set consisting of the 50 post-sample locations. Similarly, we repeat this exercise for \( \hat{y}_t^k(\cdot) \) in (4.10). To compare with other low-rank methods, we apply the fixed rank kriging (FRK) method to interpolate \( \hat{y}_t^\text{FRK}(\cdot) \) for the same data sets. The computation is implemented via the FRK package of R (please refer to [34, Chapter 4] for details). To check the performance of the kriging in time, we also generate two post-sample surfaces at times \( n+1 \) and \( n+2 \) for each sample. The mean of squared predictive error (MSPE) is calculated as follows:

\[
\text{MSPE}(\hat{y}_{n+\ell}) = \frac{1}{p} \sum_{j=1}^{p} (\hat{y}_{n+\ell}(s_j) - y_{n+\ell}(s_j))^2, \quad \ell = 1, 2. \quad (6.3)
\]

Since the FRK package only provides one period forecast, the column with respect to \( \hat{y}_{n+2}^{\text{FRK}}(\cdot) \) is left blank. We repeat the above exercise for the aggregation estimator \( \hat{y}_{n+\ell}^{\text{Ag}} \) with \( J = 100 \).
Table 1  Means of $\hat{d}$, and means and standard errors (in parentheses) of MSPE for kriging over space and in time

| n   | p   | $\hat{d}$ | MSPE($\hat{d}_{FRK}$) | MSPE($\hat{d}_t$) | MSPE($\hat{d}_{t+1}$) | MSPE($\hat{y}_{t+1}$) | MSPE($\hat{y}_{t+2}$) | MSPE($\hat{y}_{t+2}$) | MSPE($\hat{y}_{t+2}$) |
|-----|-----|-----------|------------------------|--------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| 80  | 50  | 2.240     | 1.275 (0.055)          | 1.176 (0.100)      | 1.156 (0.089)          | 1.612 (0.620)          | 1.583 (0.564)          | 1.510 (0.521)          | 1.612 (0.620)          |
| 160 | 50  | 2.725     | 1.275 (0.043)          | 1.121 (0.059)      | 1.108 (0.036)          | 1.617 (0.535)          | 1.510 (0.419)          | 1.454 (0.409)          | 1.634 (0.643)          |
| 320 | 50  | 2.935     | 1.274 (0.039)          | 1.107 (0.046)      | 1.094 (0.041)          | 1.571 (0.569)          | 1.417 (0.443)          | 1.373 (0.418)          | 1.489 (0.531)          |
| 80  | 100 | 2.725     | 1.254 (0.047)          | 1.075 (0.066)      | 1.071 (0.064)          | 1.639 (0.626)          | 1.524 (0.457)          | 1.473 (0.452)          | 1.787 (0.801)          |
| 160 | 100 | 2.975     | 1.256 (0.043)          | 1.054 (0.024)      | 1.049 (0.020)          | 1.588 (0.605)          | 1.407 (0.360)          | 1.377 (0.347)          | 1.609 (0.549)          |
| 320 | 100 | 3.000     | 1.251 (0.037)          | 1.047 (0.016)      | 1.042 (0.015)          | 1.604 (0.627)          | 1.455 (0.477)          | 1.434 (0.475)          | 1.647 (0.555)          |
| 80  | 200 | 2.905     | 1.259 (0.049)          | 1.038 (0.047)      | 1.036 (0.046)          | 1.538 (0.461)          | 1.550 (0.479)          | 1.513 (0.463)          | 1.774 (0.740)          |
| 160 | 200 | 3.000     | 1.250 (0.043)          | 1.026 (0.019)      | 1.024 (0.019)          | 1.626 (0.657)          | 1.414 (0.358)          | 1.396 (0.357)          | 1.649 (0.536)          |
| 320 | 200 | 3.000     | 1.251 (0.036)          | 1.024 (0.011)      | 1.022 (0.011)          | 1.557 (0.528)          | 1.345 (0.328)          | 1.332 (0.326)          | 1.538 (0.487)          |
Figure 2  (Color online) Boxplot of $\text{MSE}(\hat{\xi})$ (red) and $\text{MSE}(\hat{\xi})$ (blue) in a simulation with 200 replications

Table 2  Means and standard errors (in parentheses) of $\text{MSE}(\hat{\xi})$ with $\hat{\xi}(s_i)$ calculated using either $\tau > 0$ selected by five-fold cross-validation or $\tau = 0$

| $n$  | $p$  | $\tau > 0$         | $\tau = 0$         |
|------|------|---------------------|---------------------|
| 80   | 50   | 0.0874 (0.0310)     | 0.1084 (0.0380)     |
| 160  | 50   | 0.0658 (0.0162)     | 0.0778 (0.0230)     |
| 320  | 50   | 0.0583 (0.0079)     | 0.0662 (0.0137)     |
| 80   | 100  | 0.0224 (0.0127)     | 0.0268 (0.0159)     |
| 160  | 100  | 0.0154 (0.0020)     | 0.0162 (0.0034)     |
| 320  | 100  | 0.0144 (0.0010)     | 0.0149 (0.0011)     |
| 80   | 200  | 0.0054 (0.0048)     | 0.0059 (0.0057)     |
| 160  | 200  | 0.0039 (0.0002)     | 0.0039 (0.0003)     |
| 320  | 200  | 0.0037 (0.0002)     | 0.0037 (0.0002)     |

The means and the standard errors of the MSPE in the 200 replications for each settings are listed in Table 1. It is clear that the proposed kriging methods are superior to the FRK method for all settings. For our methods, in general MSPE decreases as $n$ increases. For the kriging over space, MSPE also decreases as $p$ increases. See also Theorem 5.3, noting $\delta = 0$ when all the factors are strong. MSPEs of the kriging over space are smaller than those of the kriging in time. This is understandable from comparing Theorems 5.3 and 5.4. The aggregated kriging always outperforms the non-aggregate counterparts. Last but not least, the ratio estimator (3.7) for $d$ works well for reasonably large $n$ and $p$. 
6.2 Real data analysis

We illustrate the proposed methods with the monthly temperature records (in Celsius) at the 176 monitoring stations in China from January 1970 to December 2000. All series are of the length \( n = 372 \). For each series, we remove the annually seasonal component by subtracting the average temperature of the same months. The distance among the stations are calculated as the great circle distance based on their longitudes and latitudes.

For kriging over space, we randomly select \( p = 126 \) stations for estimation, and predict the values at the other 50 stations. The mean squared predictive error for the non-aggregation estimates (4.8) are calculated as follows:

\[
\text{MSPE}(\hat{y}') = \frac{1}{50 \times 372} \sum_{i=1}^{372} \sum_{s_0 \in S^*} \left( \hat{y}_i(s_0) - y_i(s_0) \right)^2.
\]

We also apply the aggregation (with \( J = 100 \)) estimator \( \bar{y}_i(\cdot) \) in (4.10) to improve the kriging accuracy. To avoid the sampling bias in selecting stations, we replicate this exercise 100 times via randomly dividing the 176 stations into two sets of sizes 126 and 50. The estimated \( d \)-values are equal to 1 in the 98 replications, and are 2 in the two other replications. The means of MSPE over the 100 replications for \( \hat{y}^\text{FRK} \), \( \hat{y}' \) and \( \bar{y}_i(\cdot) \) are 1.1819, 1.0990 and 1.0931, with the corresponding standard errors 0.1026, 0.0862, and 0.0807, respectively. For our methods, in the training step, the average MSPE of cross-validation are 0.2069 with optimal \( \tau \), where \( \tau > 0 \), and 0.2173 with \( \tau \) being equal to zero. Among all 100 replications, the optimal \( \tau \)'s are always greater than zero.

For kriging in time, we consider one-step-ahead and two-step-ahead post-sample prediction (with \( j_0 = 6 \)) for all the 176 locations in each of the last 24 months in the data set. The corresponding mean squared predictive error at each step is defined as

\[
\text{MSPE}(\bar{y}_{n+\ell}) = \frac{1}{176} \sum_{j=1}^{176} \left( \bar{y}_{n+\ell}(s_j) - y_{n+\ell}(s_j) \right)^2, \quad \ell = 1, 2.
\]

We also apply the aggregation estimator \( \bar{y}_{n+\ell}(\cdot) \) with \( J = 100 \). The means and standard errors of MSPE(\( \bar{y}_{n+\ell} \)) over the last 24 months are 4.1895 and 5.8431. Their counterparts with respect to MSPE(\( \hat{y}^\text{FRK} \)) are 2.0151 and 1.4212 for \( \ell = 1 \), while 2.1943 and 1.5778 for \( \ell = 2 \). Lastly, the means and standard errors of MSPE(\( \bar{y}_{n+\ell} \)) are 2.0136 and 1.4194 for \( \ell = 1 \), and 2.1942 and 1.5759 for \( \ell = 2 \). As we expected, the one-step-ahead prediction is more accurate than the two-step-ahead prediction.

Overall the kriging over space is more accurate than those in time. The aggregation via random partitioning of locations improves the prediction, though the improvement is not substantial in this example.

Acknowledgements The first author was supported by National Statistical Research Project of China (Grant No. 2015LY77) and National Natural Science Foundation of China (Grant Nos. 11571080, 11571081, 71531006 and 71672042). The second author was supported by Engineering and Physical Sciences Research Council (Grant No. EP/L01226X/1). The third author was supported by National Natural Science Foundation of China (Grant Nos. 11371318 and 11771390), Zhejiang Province Natural Science Foundation (Grant No. R16A010001) and the Fundamental Research Funds for the Central Universities. The authors thank Professor Noel Cressie for helpful comments and suggestions.

References
1 Banerjee S, Gelfand A, Finley A O, et al. Gaussian predictive process models for large spatial data sets. J R Stat Soc Ser B Stat Methodol, 2008, 70: 825–848
2 Bathia N, Yao Q, Ziegelmann F. Identifying the finite dimensionality of curve time series. Ann Statist, 2010, 38: 3352–3386
3 Boyd S, Vandenberghe L. Convex Optimization. Cambridge: Cambridge University Press, 2004
4 Breiman L. Bagging predictors. Machine Learning, 1996, 24: 123–140
5 Castruccio S, Stein M L. Global space-time models for climate ensembles. Ann Appl Stat, 2013, 7: 1593–1611
In this appendix, we present the proofs of main theorems, lemmas and other propositions.

Proof of Proposition 2.1. The first part of the proposition can be proved in the same manner as [2, Proposition 1], which is omitted. To prove the second part, it follows from (2.9) and (2.8) that any
Thus $\lambda_i\varphi_i(s) = \sum_j \gamma_{ij}a_j(s)$. Now it follows from (2.11) and (2.8) that

$$
\Sigma_0 \circ \varphi_i(s) = \sum_{k,l,j} \sigma_{kl} \varphi_i(s)\langle a_k, a_j \rangle = \sum_{k,l} \sum_j \sigma_{kl} \gamma_{ij} a_k(s) = \sum_k \lambda_i \gamma_{ik} a_k(s) = \lambda_i \varphi_i(s).
$$

Since $a_1, \ldots, a_d$ are orthonormal, it holds that

$$
\sum_j \sigma_{kj} \gamma_{ij} = \lambda_i \gamma_{ik}, \quad k = 1, \ldots, d. \tag{A.1}
$$

As $\sigma_{kj}$ is the $(k,j)$-th element of matrix $\text{Var}(x_\ell)$, (A.1) is equivalent to $\text{Var}(x_\ell)\gamma_i = \lambda_i \gamma_i$, i.e., $\gamma_i$ is an eigenvector of $\text{Var}(x_\ell)$ corresponding to the eigenvalue $\lambda_i, i = 1, \ldots, d$. Furthermore,

$$
I(i = k) = \langle \phi_i, \phi_k \rangle = \sum_j \gamma_{ij} \gamma_{kj} = \gamma_i^2.
$$

Thus $\gamma_1, \ldots, \gamma_d$ are orthogonal. \hfill \Box

To prove Theorem 3.4(ii), we firstly introduce Lemma A.1 below. For the simplicity in presentation, we assume that the $d$ positive eigenvalues of $\Sigma \Sigma'$, defined in (3.5), are distinct from each other. Then both $A_1$ and $A_2$ are uniquely defined if we line up each of the two sets of the $d$ orthonormal eigenvectors (i.e., the columns of $A_1$ and $A_2$) in the descending order of their corresponding eigenvalues, and we require that the first non-zero element of each of those eigenvector to be positive (see the discussion below (3.5) above).

Using the same notation as in (3.8), we denote by $\hat{A}_1^{(j)}$ and $\hat{A}_2^{(j)}$ the estimated factor loading matrices in (3.8) with the $j$-th partition, by $\Sigma^{(j)}$ the covariance matrix in (3.4), and by $x_t^{(j)}$, $x_t^{*^{(j)}}$ the estimated latent factors in (3.9), $j = 1, \ldots, p_0$, where $p_0 = [p_1/p_2!]$. Assume that the $d$ positive eigenvalues of $\Sigma^{(j)}(\Sigma^{(j)})'$ are distinct. Then $A_1^{(j)}$ and $A_2^{(j)}$ can be uniquely defined as above. Now we are ready to state the lemma.

**Lemma A.1.** Let Condition 1 hold. Let the $d$ positive eigenvalues of $\Sigma^{(j)}(\Sigma^{(j)})'$ be distinct, and Condition 2 hold for $x_t^{(j)}$ and $x_t^{*^{(j)}}$ for all $j = 1, \ldots, p_0$. Then as $p^\delta n^{-1/2} + p^{2\delta - 2}\|L\| = o(1)$, it holds that

$$
\max_{1 \leq j \leq p_0} \{ \| \hat{A}_1^{(j)} - A_1^{(j)} \| + \| \hat{A}_2^{(j)} - A_2^{(j)} \| \} = O_p(p^\delta n^{-1/2} + p^{2\delta - 2}\|L\|).
$$

**Proof.** Since $\max_{1 \leq j \leq p_0} \| \hat{A}_2^{(j)} - A_2^{(j)} \|$ can be shown similarly to $\max_{1 \leq j \leq p_0} \| \hat{A}_1^{(j)} - A_1^{(j)} \|$, we only prove $\max_{1 \leq j \leq p_0} \| \hat{A}_1^{(j)} - A_1^{(j)} \|$ here. Note that for any $1 \leq j \leq p_0$,

$$
\| \hat{\Sigma}^{(j)}(\hat{\Sigma}^{(j)})' - \tau L - \Sigma^{(j)}(\Sigma^{(j)})' \| \leq \| \hat{\Sigma}^{(j)} - \Sigma^{(j)} \|^2 + 2\| \Sigma^{(j)} \| \| \hat{\Sigma}^{(j)} - \Sigma^{(j)} \| + \| \Sigma^{(j)} - \Sigma^{(j)} \| + \tau \| L \|. \tag{A.2}
$$

Since $x_t^{(j)}$ satisfies Condition 2, it follows that $\| \Sigma^{(j)} \| = O(p^{1-\delta})$ (see [24]). On the other hand, by the mixing condition of $\{y_t\}$, we have

$$
\sup_j \| \hat{\Sigma}^{(j)} - \Sigma^{(j)} \|^2 = \sup_j \left\| \frac{1}{n} \sum_{t=1}^n \left( \langle y_t^{(j)}, y_t^{(j)} \rangle - \langle \hat{y}_t^{(j)}, \hat{y}_t^{(j)} \rangle - \text{Cov}(y_t^{(j)}, y_t^{(j)}) \right) \right\|^2
\leq \sum_{i=1}^p \sum_{j=1}^p \left\{ \frac{1}{n} \sum_{t=1}^n [y_t(s_i) - \langle \hat{y}_t, \hat{y}_t \rangle][y_t(s_j) - \langle \hat{y}_t, \hat{y}_t \rangle] - \text{Cov}[y_t(s_i), y_t(s_j)] \right\}^2
= O_p(p^2/n).
$$

Thus, by (A.2),

$$
\sup_j \| \hat{\Sigma}^{(j)}(\hat{\Sigma}^{(j)})' - \tau L - \Sigma^{(j)}(\Sigma^{(j)})' \| = O_p(p^{2-\delta} n^{-1/2} + \tau \| L \|). \tag{A.3}
$$
By (A.3) and a similar argument to [24, Theorem 1], we can show that
\[
\max_{1 \leq j \leq p_0} \| \tilde{A}_1^{(j)} - A_1^{(j)} \| = O_p(p^\delta n^{-1/2} + p^{2\delta-2} \|L\|)
\]
and complete the proof of Lemma A.1.

**Proof of Theorem 3.4(ii).** Note that
\[
\mathbb{E}\left[ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left( \hat{\xi}_i(s_i) - \hat{\xi}_i(s_i) \right) \right] \mathbb{E}\left[ \left( \xi_i(s_i), y_i(s_i) \right) \right] \\
= \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left[ A_1^{(j)} x_i^{(j)} - A_1^{(j)} x_i^{(j)} \right] \left[ \tilde{A}_1^{(j)} x_i^{(j)} - A_1^{(j)} x_i^{(j)} \right] \\
+ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left[ \tilde{A}_1^{(j)} x_i^{(j)} - A_1^{(j)} x_i^{(j)} \right] \left[ \tilde{A}_1^{(j)} x_i^{(j)} - A_1^{(j)} x_i^{(j)} \right] \\
= \Sigma_1 + \Sigma_2.
\]

By Lemma A.1, we have
\[
\Sigma_1 = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left[ \tilde{A}_1^{(j)} (A_1^{(j)})' - A_1^{(j)} (A_1^{(j)})' \right] A_1^{(j)} x_i^{(j)} + \tilde{A}_1^{(j)} (A_1^{(j)})' \varepsilon_i^{(j)} \\
\times \left[ \tilde{A}_1^{(j)} (A_1^{(j)})' - A_1^{(j)} (A_1^{(j)})' \right] A_1^{(j)} x_i^{(j)} + \tilde{A}_1^{(j)} (A_1^{(j)})' \varepsilon_i^{(j)} \\
= \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left[ A_1^{(j)} x_i^{(j)} - A_1^{(j)} x_i^{(j)} \right] \left[ \tilde{A}_1^{(j)} (A_1^{(j)})' - A_1^{(j)} (A_1^{(j)})' \right] A_1^{(j)} x_i^{(j)} \\
+ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left[ \tilde{A}_1^{(j)} x_i^{(j)} - A_1^{(j)} x_i^{(j)} \right] A_1^{(j)} x_i^{(j)} \\
+ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left[ \tilde{A}_1^{(j)} x_i^{(j)} - A_1^{(j)} x_i^{(j)} \right] A_1^{(j)} x_i^{(j)} \\
= O_p(p^\delta/n + p^{(\delta-1)/2} n^{-\gamma/2} + p^{\delta-2} \|L\|) + \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left( \varepsilon_i^{(j)} \right)^2.
\]

Since
\[
\mathbb{E}\left( \frac{1}{p_0} \sum_{j=1}^{p_0} \left( \varepsilon_i^{(j)} \right)^2 \right) \leq \frac{1}{p_0} \sum_{j=1}^{p_0} \mathbb{E}[(\varepsilon_i^{(j)})^2] \leq \mathbb{E}[\varepsilon_i^{(j)}]^2 < \infty,
\]

it follows from Markov’s inequality that
\[
\frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left( \varepsilon_i^{(j)} \right)^2 \leq \frac{1}{p_0} \sum_{j=1}^{p_0} \mathbb{E}[(\varepsilon_i^{(j)})^2] \rightarrow P \mathbb{E}[(\varepsilon_i^{(j)})^2] \varepsilon_i^{(j)}.
\]

Thus, by (A.4), we have the following two conclusions:

(i) When \( n \rightarrow \infty \), \( \Sigma_1 = O_p(p^\delta/n + p^{(\delta-1)/2} n^{-\gamma/2} + p^{\delta-2} \|L\| + p^{-1}) \).

(ii) When \( p^\delta + n + p^{\delta-1} \|L\| \rightarrow 0 \), \( p \Sigma_1 \rightarrow P \mathbb{E}[(\varepsilon_i^{(j)})^2] \varepsilon_i^{(j)} \).

Similarly, the above properties hold also for \( \Sigma_2 \). Hence,
\[
\mathbb{E}\left[ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left( \xi_i(s_i) - \xi_i(s_i) \right) \right] \mathbb{E}\left[ \left( \xi_i(s_i), y_i(s_i) \right) \right] = O_p(p^\delta/n + p^{(\delta-1)/2} n^{-\gamma/2} + p^{\delta-2} \|L\| + p^{-1}).
\]
Furthermore, when $p^{1+\delta}/n + p^{\delta-1}\|L\| \to 0$,

$$E\left[\frac{1}{n} \sum_{t=1}^{n} \sum_{i=1}^{p} \left(\xi_t(s_i) - \xi_i(s_i)\right)^2 \left(\xi_t(s_i), y_i(s_i)\right)^T\right] = E\left[\xi_t^{(1)} A_1^{(1)} (A_1^{(1)})^T + (\xi_t^{(2)}) A_2^{(1)} (A_2^{(1)})^T\right]$$

in probability.

\[\]  

\[\text{Lemma A.2.} \quad \text{Let Condition 1 hold and } pn^{-\beta/2} \to 0. \text{ Then} \]

$$\lim_{n \to \infty} P\left\{ \min_{1 \leq j \leq p} \lambda_{\min}[n^{-1}Z(s_i)^T Z(s_i)] \geq c_0/2 \right\} = 1.$$  

\[\text{Proof.} \]

Let $z_t^j(s_i), j = 1, \ldots, m$ be the components of $z_t(s_i)$. Since $\{z_t(s_i)\}$ is a stationary $\alpha$-mixing process satisfying Condition 1, by [26, Lemma 12.22], we have that for any $1 \leq j, k \leq m$,

$$E\left[\frac{1}{n} \sum_{i=1}^{n} (z_t^j(s_i) z_t^k(s_i) - E[z_t^j(s_i) z_t^k(s_i)])\right] = O(n^{-\beta/2}). \quad (A.6)$$

Since $m$ is finite, it follows that

$$E\|n^{-1}Z(s_i)^T Z(s_i) - \operatorname{Var}(z_t(s_i))\|_F^2 = O(n^{-\beta/2}), \quad (A.7)$$

where $\| \cdot \|_F$ denotes the Frobenius norm. Now suppose that $\min_{1 \leq i \leq p} \lambda_{\min}[n^{-1}Z(s_i)^T Z(s_i)] < c_0/2$. Since $\min_{1 \leq i \leq p} \lambda_{\min}[\operatorname{Var}(z_t(s_i))] > c_0$, and

$$\operatorname{Var}(z_t(s_i)) = [\operatorname{Var}(z_t(s_i)) - n^{-1}Z(s_i)^T Z(s_i)] + n^{-1}Z(s_i)^T Z(s_i),$$

it holds that

$$\max_{1 \leq i \leq p} \|n^{-1}Z(s_i)^T Z(s_i) - \operatorname{Var}(z_t(s_i))\|_F \geq c_0/2. \quad (A.8)$$

However, by (A.7), it follows that

$$\begin{align*}
P\left\{ \max_{1 \leq i \leq p} \|n^{-1}Z(s_i)^T Z(s_i) - \operatorname{Var}(z_t(s_i))\|_F \geq c_0/2 \right\} \\
\leq \sum_{i=1}^{p} (c_0/2)^{-\beta} E\|n^{-1}z_t(s_i)^T z_t(s_i) - \operatorname{Var}(z_t(s_i))\|_F^2 \\
= O_p(pn^{-\beta/2}) = o(1). \quad (A.9)
\end{align*}$$

This implies that

$$P\left\{ \min_{1 \leq i \leq p} \lambda_{\min}[n^{-1}z_t(s_i)^T z_t(s_i)] < c_0/2 \right\} = o(1)$$

and completes the proof of Lemma A.2.

\[\text{Proof for the convergence rate of } \tilde{\beta}(s_0). \]

Let

$$e_i(s) = y_i(s) - z_t(s)^T \beta(s)$$

and

$$w_i = \frac{K_h(s_i - s_0)}{\sum_{i=1}^{p} K_h(s_i - s_0)}.$$  

Then $e(s) = (e_1(s), \ldots, e_n(s))^T$ and

$$\tilde{\beta}(s_0) = \sum_{i=1}^{p} [Z(s_i)^T Z(s_i)]^{-1} [Z(s_i)^T e(s_i)] w_i + \sum_{i=1}^{p} \beta(s_i) w_i \equiv I_1 + I_2.$$  

For any twice differentiable function $g(s) = g(s_1, s_2), s = (s_1, s_2) \in \mathbb{R}^2$, define $g_1(s) = \partial g(s)/\partial s_1$, $g_2(s) = \partial g(s)/\partial s_2$, $g_{11}(s) = \partial^2 g(s)/\partial s_1^2$ and $g_{22}(s) = \partial^2 g(s)/\partial s_2^2$. Under Conditions 3 and 4 and Taylor's expansion, it can be shown that as $p \to \infty$,

$$I_2 - \beta(s_0) = \sum_{i=1}^{p} (\beta(s_i) - \beta(s_0)) w_i.$$
\[
\begin{align*}
&= \frac{h^2}{f(s_0)} \left[ \beta_1(s_0)f_1(s_0) + \frac{1}{2} f(s_0) \beta_1(s_0) \right] \int_R \int_R x^2 K(x,y) \, dx \, dy \\
&\quad + \frac{h^2}{f(s_0)} \left[ \beta_2(s_0)f_2(s_0) + \frac{1}{2} f(s_0) \beta_2(s_0) \right] \int_R \int_R y^2 K(x,y) \, dx \, dy + o(h^2). \\
&\quad (A.10)
\end{align*}
\]

As for \( I_1 \), by H{"o}lder’s inequality, it follows that
\[
\|I_1\| \leq \left( \max_{1 \leq i \leq p} \|Z(s_i)' Z(s_i)/n\|_F \right) \sum_{i=1}^{p} \|Z(s_i)' e(s_i)/n\| w_i. \quad (A.11)
\]

By Lemma A.2, we have \( \max_{1 \leq i \leq p} \lambda_{\max}\{(n^{-1} Z(s_i)' Z(s_i))^{-1}\} \leq 2/c_0 \) holds in probability. Since the dimension of \( z_i(s) \) is fixed, it follows that
\[
\max_{1 \leq i \leq p} \|Z(s_i)' Z(s_i)/n\|_F \leq c_1 \quad (A.12)
\]

holds in probability for some positive constant \( c_1 \). On the other hand, it is easy to get that
\[
\max_{1 \leq i \leq p} E\|n^{-1} Z(s_i)' e(s_i)\| = O(n^{-1/2}),
\]

and hence,
\[
E\left[ \sum_{i=1}^{p} \|n^{-1} Z(s_i)' e(s_i)\| w_i \right] = O(n^{-1/2}). \quad (A.13)
\]

It follows from (A.11)–(A.13) that
\[
\|I_1\| = O_p(n^{-1/2}). \quad (A.14)
\]

Thus, by (A.10) and (A.13), we have \( |\hat{\beta}(s_0) - \beta(s_0)| = O_p(h^2 + n^{-1/2}) \).

Proof of Theorem 5.3. Let \( x_i^0 = x_i I(s_i \in S_1) + x_i^1 I(s_i \in S_2) \). Then
\[
\hat{\xi}_i(s_0) - \xi_i(s_0) = \sum_{i=1}^{p} (a_i(s_i) x_i^0 - a_i(s_i) x_i^0) w_i + \sum_{j=1}^{d} \sum_{i=1}^{p} (a_j(s_i) - a_j(s_0)) x_i j w_i
\]
\[
= J_1 + J_2. \quad (A.15)
\]

Similar to (A.10), we have
\[
\left| \sum_{i=1}^{p} (a_j(s_i) - a_j(s_0))/\|a(s_0)\| w_i \right| = O(h^2),
\]

which implies that
\[
|J_2| = O(h^2)(\|a(s_0)\|) \sum_{j=1}^{d} |x_j^0| = O(dh^2 \cdot \|a(s_0)\| \cdot \|x_i^0\|) = O_p(h^2), \quad (A.16)
\]

where we use the facts that \( \|x_i^0\| = O_p(p^{1/2}) \) and \( \|a(s_0)\| = O(p^{1/2}) \), which are followed by \( \lambda_{\min}\{E(x_i^0 x_i^0)\} \approx p^{1/2} \) and
\[
(\|a(s_0)\|^2) \lambda_{\min}\{E(x_i^0 x_i^0)\} \|a(s_0)\|^2 \leq (\|a(s_0)\|^2) E[(a_i(s_0)) x_i^0] x_i^0 = E[a_i(s_0) x_i^0] x_i^0 = E_y \|a_i(s_0)\| < \infty.
\]

By Proposition 5.1(iii) and the same arguments as in \([6, \text{Theorem 2.2}]\), we have that for \( i = 1, 2 \),
\[
p^{-1/2} \|A_i x_i^0 - A_i x_i^0\| = O_p(\|A_i - A_i\| + n^{-1/2} + p^{-1/2}) = O_p(n^{-1/2} p^\delta + p^{2\delta-2} \|L\| + p^{-1/2}),
\]
which combining with Hölder’s inequality implies that
\[
J_1 \leq \left\{ \sum_{i=1}^{p} |(\hat{a'}(s_i)\hat{x}_i^o - a'(s_i)x_i^o)|^2 \right\}^{1/2} \left( \sum_{i=1}^{p} u_i^2 \right)^{1/2}
\]
\[
\leq \left( \sum_{i=1}^{2} \|A_i\hat{x}_i^o - A_ix_i^o\| \right) \left( \sum_{i=1}^{p} u_i^2 \right)^{1/2}
\]
\[
= O_p \{p^{1/2}(n^{-1/2}p^\delta + p^{2\delta - 2} \|L\| + p^{-1/2})\} \cdot O((ph)^{-1/2})
\]
\[
= O_p \{p^{1/2}(nh)^{-1/2} + (ph)^{-1/2} + p^{2\delta - 2} h^{-1/2} \|L\|\}.
\]

(A.17)

Thus, (ii) follows from (A.16) and (A.17). Similarly, we can show that (A.17) holds also for $\tilde{\xi}_t(s_0)$. \hfill $\square$

**Proof of Theorem 5.4.** For simplicity, we only show the case with spatial points over $S_1$, i.e., $y_{t1} = A_1x_t + \varepsilon_{t,1}$. The case for points over $S_2$ can be shown similarly. Let

\[
\tilde{\Sigma}_c(k) = \frac{1}{n} \sum_{t=1}^{n-k} (\varepsilon_{t+k,1} - \bar{\varepsilon}_1)(\varepsilon_{t,1} - \bar{\varepsilon}_1)',
\]

\[
\tilde{\Sigma}_{xx}(k) = \frac{1}{n} \sum_{t=1}^{n-k} (x_{t+k} - \bar{x})(x_{t,1} - \bar{x})',
\]

\[
\tilde{\Sigma}_{ex}(k) = \frac{1}{n} \sum_{t=1}^{n-k} (\varepsilon_{t+k,1} - \bar{\varepsilon}_1)(x_t - \bar{x})'
\]

and

\[
\tilde{\Sigma}_{xe}(k) = \frac{1}{n} \sum_{t=1}^{n-k} (x_{t+k} - \bar{x})(\varepsilon_t - \bar{\varepsilon}_1)'.
\]

It follows that for any $k$,

\[
\tilde{\Sigma}_x(k) - \Sigma_x(k) = (\hat{A}'_1 - A'_1)A_1\tilde{\Sigma}_{xx}(k)A'_1 + \tilde{\Sigma}_{ex}(k)A'_1(\hat{A}_1 - A_1) + (\tilde{\Sigma}_{xe}(k) - \Sigma_x(k))
\]

\[
+ \hat{A}'_1\tilde{\Sigma}_e(k) + \tilde{\Sigma}_{xe}(k)A_1 + \hat{A}'_1\tilde{\Sigma}_{ex}(k)A_1 + \hat{A}'_1\tilde{\Sigma}_{ex}(k)A'_1\hat{A}_1
\]

\[
=: \sum_{j=1}^{6} L_j.
\]

By $\|\hat{A}_1 - A_1\| = O_p(n^{-1/2}p^\delta + p^{2\delta - 2}\|L\|)$, it follows that

\[
\|L_1\| + \|L_2\| = O(\|\hat{A}_1 - A_1\| \cdot \|\tilde{\Sigma}_{xx}(k)\|) = O_p(n^{-1/2}p + p^{\delta - 1}\|L\|).
\]

(A.18)

By [23, (A.1)], we have

\[
\|L_3\| \leq \|\tilde{\Sigma}_{xx}(k) - \Sigma_x(k)\|_F = O(d\|\tilde{\Sigma}_{xx}(k) - \Sigma_x(k)\|) = O(p^{1-\delta}n^{-1/2}).
\]

(A.19)

It is easy to get that

\[
\|\tilde{\Sigma}_{xx}(k)\| = O_p(p^{1-\delta/2}n^{-1/2}) = \|\tilde{\Sigma}_{xx}(k)\| \quad \text{and} \quad \|\tilde{\Sigma}_e(k)\| = O_p(ps^{-1/2})
\]

(see, for example, [24, Lemma 2]). Thus,

\[
\|L_4\| + \|L_5\| + \|L_6\| = O_p(ps^{-1/2}).
\]

(A.20)

Combining (A.18)–(A.20) yields that for any $0 \leq k \leq j_0$,

\[
\|\tilde{\Sigma}_x(k) - \Sigma_x(k)\| = O_p(ps^{-1/2} + p^{\delta - 1}\|L\|).
\]

(A.21)
Thus, by $p^{\delta n^{-1/2}} + p^{2\delta-2}\|L\| = o(1)$, we get $pn^{-1/2} + p^{\delta-1}\|L\| = o(p^{1-\delta})$ and in probability,
\begin{equation}
\|\hat{\Sigma}_x(k)\|_{\min} \asymp \|\Sigma_x(k)\|_{\min} \asymp p^{1-\delta} \asymp \|\Sigma_x(k)\| \asymp \|\hat{\Sigma}_x(k)\|.
\end{equation}
(A.22)

Since $j_0$ is fixed, from (A.21) it follows that
\begin{equation}
\|\hat{R}_{j_0} - R_{j_0}\| \asymp \|W_{j_0} - \hat{W}_{j_0}\| \asymp O_p(pn^{-1/2} + p^{\delta-1}\|L\|)
\end{equation}
(A.23)
and from (A.22) it follows that
\begin{equation}
\|R_{j_0}\| = O(p^{1-\delta}) \quad \text{and} \quad \|\hat{W}_{j_0}^{-1}\| \asymp \|W_{j_0}^{-1}\| \asymp O_p(p^{\delta-1}).
\end{equation}
(A.24)

Since
\begin{equation}
\|\hat{x}_t - x_t\| = \|\vspace{0.5em}(\hat{A}_1 - A_1)'Ax_t + (\hat{A}_1 - A_1)'\varepsilon_{t,1} + A_1'\varepsilon_{t,1}\| = O_p(p^{1/2+\delta_1}n^{-1/2} + p^{2\delta-3/2}\|L\| + 1)
\end{equation}
\begin{equation}
\text{and} \quad p^{\delta/2}(p^{\delta n^{-1/2}} + p^{2\delta-2}\|L\|) = o(1), \quad \text{it follows that} \quad \|\hat{X}\| = O_p(p^{\delta/2}). \quad \text{Note that}
\end{equation}
\begin{align*}
\hat{x}_{n+j} - x_n(j) &= \hat{R}_{j_0}\hat{W}_{j_0}^{-1}\hat{X} - R_{j_0}W_{j_0}^{-1}X \\
&= (\hat{R}_{j_0} - R_{j_0})\hat{W}_{j_0}^{-1}\hat{X} + R_{j_0}\hat{W}_{j_0}^{-1}(W_{j_0} - \hat{W}_{j_0})W_{j_0}^{-1}\hat{X} + R_{j_0}W_{j_0}^{-1}(\hat{X} - X).
\end{align*}

By (A.23) and (A.24), we have
\begin{equation}
\|(\hat{R}_{j_0} - R_{j_0})\hat{W}_{j_0}^{-1}\hat{X}\|^2 = O_p(\|\hat{R}_{j_0} - R_{j_0}\|^2 \cdot \|\hat{W}_{j_0}^{-1}\|^2 \cdot \|\hat{X}\|) = O_p(p^{1+\delta n^{-1}}).
\end{equation}
(A.25)

Similarly,
\begin{align*}
\|R_{j_0}\hat{W}_{j_0}^{-1}(W_{j_0} - \hat{W}_{j_0})W_{j_0}^{-1}\hat{X}\|^2 &= O_p(\|R_{j_0}\|^2 \cdot \|\hat{W}_{j_0}^{-1}\|^2 \cdot \|W_{j_0} - \hat{W}_{j_0}\|^2 \cdot \|W_{j_0}^{-1}\|^2 \cdot \|\hat{X}\|) \\
&= O_p(p^{1+\delta n^{-1}}).
\end{align*}
(A.26)

On the other hand, by (A.24) and $\|\hat{x}_t - x_t\| = O_p(p^{1/2+\delta_1}n^{-1/2} + p^{2\delta-3/2}\|L\| + 1)$, we have
\begin{equation}
\|R_{j_0}W_{j_0}^{-1}(\hat{X} - X)\| = O_p(p^{1/2+\delta_1}n^{-1/2} + p^{2\delta-3/2}\|L\| + 1).
\end{equation}
(A.27)

Thus,
\begin{equation}
\|\hat{x}_{n+j} - x_n(j)\| = O_p(p^{1/2+\delta_1}n^{-1/2} + p^{2\delta-3/2}\|L\| + 1)
\end{equation}
holds and Theorem 5.4(a) is proved.

As for Conclusion (b), by Conclusion (a) and Proposition 5.1(iii), we have
\begin{align*}
\|\hat{y}_{n+j} - y_n(j)\| &= \|\hat{A}\hat{x}_{n+j} - A x_n(j)\| \\
&\leq \|\hat{A}(A - A)x_n(j)\| + \|\hat{A}\hat{x}_{n+j} - x_n(j)\| \\
&= O_p(p^{\delta_1 n^{-1/2}}p^{1/2+\delta_1}n^{-1/2} + p^{1/2+\delta_1}n^{-1/2} + p^{2\delta-3/2}\|L\| + 1) \\
&= O_p(p^{1/2+\delta_1}n^{-1/2} + p^{2\delta-3/2}\|L\| + 1).
\end{align*}

This gives (b) as desired and completes the proof of Theorem 5.4. □