Krigings Over Space and Time Based on Latent Low-Dimensional Structures*

Da Huang† Qiwei Yao‡ Rongmao Zhang⋆

†School of Management, Fudan University, Shanghai, 200433, China

‡Department of Statistics, London School of Economics, London, WC2A 2AE, U.K.

⋆School of Mathematics, Zhejiang University, Hangzhou, 310058, China

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

Abstract

We propose a new approach to represent nonparametrically 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 (Section 7.1.3 of Cressie and Wikle, 2011), 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 over time. 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. Their performance is further improved by a newly proposed aggregation method via randomly partitioning the observations accordingly to their locations. A low-dimensional correlation structure also makes the krigings 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. Illustration with both simulated and real data sets is also reported.

KEY WORDS: Aggregation via random partitioning; Blessing of dimensionality; Common factors; Covariance function; Eigenanalysis; Nugget effect; Spatio-temporal processes; Strong factors.

*Partially supported by National Statistical Research Project of China 2015LY77 and NSFC grants 11571080, 11571081, 71531006 (DH), by EPSRC grant EP/L01226X/1 (QY), and by NSFC grants 11371318 (RZ).

1 Introduction

Kriging, referring to the spatial best linear prediction, is named by Matheron after 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 Section 4.1 of Cressie and Wikle (2011) and references within. However fitting those parametric covariance models to large spatial or spatio-temporal datasets is conceptually indefensible (Hall, Fisher and Hoffmann, 1994). It also poses serious computational challenges. For example, a spatial kriging with observations from p locations involves inverting a $p \times p$ covariance matrix, which typically requires $O(p^3)$ operations with $O(p^2)$ memory. One attractive approach to overcome the computational burden is to introduce reduced rank approximations for the underlying processes. Methods in this category include Higdon (2002) using kernel convolutions, Wikle and Cressie (1999), Kammann and Wand (2003) and Cressie and Johannesson (2008) using low rank basis functions (see also Section 7.1.3 of Cressie and Wikle, 2011), and Banerjee *et al.* (2008) and Finley *et al.* (2009) using predictive processes. However as pointed out by Stein (2008), the reduced rank approximations often fail to capture small-scale correlation structure accurately. An alternative approach is to seek sparse approximations for covariance functions, see, e.g., Gneiting (2002) using compactly supported covariance functions, and Kaufman, Schervish and Nychka (2008) proposing a tempering method by setting the covariances to 0 between any two locations with the distances beyond a threshold. Obviously these approaches miss the correlations among the locations which are distantly apart from each other. Combining together both the ideas of reducing rank and the tempering, Sang and Huang (2012) and Zhang, Sang and Huang (2015) proposed a so-called full scale approximation method for large spatial and spatio-temporal datasets.

In this paper we propose a new nonparametric approach to represent the linear dependence structure of a spatio-temporal process. Different from all the methods stated above, we impose neither any distributional assumptions on the underlying process nor any 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. The key assumption is that the underlying process is stationary in time, though it can be non-stationary over space. Formally our latent factor model is a reduced rank representation. However both the factor process and the factor loadings are completely unknown. This is a marked difference from the aforementioned reduced rank approximation methods. The motivation for our approach is to learn the linear dynamic structure across both space and time directly from data with little subjective input. Therefore it captures the dependence across the locations over all distances automatically.

The latent factors and the corresponding loadings are estimated via an eigenanalysis of non-negative definite matrix, similar to the factor modelling for multiple time series of Lam, Yao and Bathia (2011) and Lam and Yao (2012). However we extract the information from the dependence across different locations instead of over time: the whole observations are divided into two sets randomly according to their locations, the estimation boils down to the singular value decomposition (SVD) of 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. To overcome the arbitrariness of the partition of the data, a new aggregation via randomly partitioning is proposed, which improves the original estimation. Kriging predictions over space in time are constructed based on the recovered latent factor structure.

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 to the number of factors. This is particularly appealing when dealing with large datasets. However the SVD for estimating the latent factor structure requires $O(p^3)$ operation. Nevertheless the nonparametric nature of our approach makes it particularly easy to make the method scalable to large datasets. See Section 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. With the advancement of information technol-

ogy, large amount of data are collected routinely over space and time nowadays. The surge of the development of statistical methods and theory for modelling and forecasting spatio-temporal processes includes, among others, Smith, Kolenikov and Cox (2003), Jun and Stein (2007), Li, Genton and Sherman (2007), Katzfuss and Cressie (2011), Castruccio and Stein (2013), Guinness and Stein (2013), Zhu, Fan and Kong (2014), Zhang, Sang and Huang (2015). See also the monograph Cressie and Wilkle (2011).

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 spelt out in Section 3. The kriging over space and time is presented in Section 4. The asymptotic results for the proposed estimation and kriging methods are developed in Section 5. We highlight the new results in the form of theorems. The results similar to the existing theorems (under different settings) are stated as propositions. Illustration with both simulated and real data is reported in Section 6. Additional remarks are given in Section 7. Technical proofs are relegated to the Appendix in a supplementary file.

2 Models

2.1 Setting

Consider spatio-temporal process

$$y_t(\mathbf{s}) = \mathbf{z}_t(\mathbf{s})'\boldsymbol{\beta}(\mathbf{s}) + \xi_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}), \quad t = 0, \pm 1, \pm 2, \cdots, \ \mathbf{s} \in \mathcal{S} \subset \mathcal{R}^2, \tag{2.1}$$

where $\mathbf{z}_t(\mathbf{s})$ is an $m \times 1$ observable covariant vector, $\boldsymbol{\beta}(\mathbf{s})$ is a unknown parameter vector, $\varepsilon_t(\mathbf{s})$ is unobservable and represents the so-called nugget effect (in space) in the sense that

$$E\{\varepsilon_t(\mathbf{s})\} = 0, \quad \mathrm{Var}\{\varepsilon_t(\mathbf{s})\} = \sigma(\mathbf{s})^2, \quad \mathrm{Cov}\{\varepsilon_{t_1}(\mathbf{u}), \varepsilon_{t_2}(\mathbf{v})\} = 0 \ \ \forall\, (t_1, \mathbf{u}) \neq (t_2, \mathbf{v}), \tag{2.2}$$

$\xi_t(\mathbf{s})$ is a latent spatio-temporal process satisfying the conditions

$$E\{\xi_t(\mathbf{s})\} = 0, \qquad \mathrm{Cov}\{\xi_{t_1}(\mathbf{u}), \xi_{t_2}(\mathbf{v})\} = \Sigma_{|t_1 - t_2|}(\mathbf{u}, \mathbf{v}). \tag{2.3}$$

Consequently, $y_t(\mathbf{s}) - \mathbf{z}_t(\mathbf{s})'\boldsymbol{\beta}(\mathbf{s})$ is (weakly) stationary in time t, $E\{y_t(\mathbf{s}) - \mathbf{z}_t(\mathbf{s})'\boldsymbol{\beta}(\mathbf{s})\} = 0$, and

$$\mathrm{Cov}\{y_{t_1}(\mathbf{u}) - \mathbf{z}_{t_1}(\mathbf{u})'\boldsymbol{\beta}(\mathbf{u}),\ y_{t_2}(\mathbf{v}) - \mathbf{z}_{t_2}(\mathbf{v})'\boldsymbol{\beta}(\mathbf{v})\} = \Sigma_{|t_1-t_2|}(\mathbf{u},\mathbf{v}) + \sigma(\mathbf{u})^2 I\{(t_1,\mathbf{u}) = (t_2,\mathbf{v})\}. \quad (2.4)$$

Finally, we assume that $\Sigma_t(\mathbf{u},\mathbf{v})$ is continuous in $\mathbf{u}$ and $\mathbf{v}$. Note that model (2.1) does not impose any stationarity conditions over space, though it requires that $y_t(\cdot) - \mathbf{z}_t(\cdot)'\boldsymbol{\beta}(\cdot)$ is second order stationary in time t.

2.2 A finite dimensional representation for $\xi_t(\mathbf{s})$

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

$$\langle f, g\rangle = \int_{\mathcal{S}} f(\mathbf{s})g(\mathbf{s})d\mathbf{s}, \qquad f, g \in L_2(\mathcal{S}). \quad (2.5)$$

We assume that the latent process $\xi_t(\mathbf{s})$ admits a finite-dimensional structure:

$$\xi_t(\mathbf{s}) = \sum_{j=1}^{d} a_j(\mathbf{s})x_{tj}, \quad (2.6)$$

where $a_1(\cdot), \cdots, a_d(\cdot)$ 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 $L_2(\mathcal{S})$, and $x_{t1}, \cdots, x_{td}$ are d random variables. Obviously $a_1(\cdot), \cdots, a_d(\cdot)$ (as well as $x_{t1}, \cdots, x_{td}$) 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 $a_1(\cdot), \cdots, a_d(\cdot)$ are orthonormal in the sense that

$$\langle a_i, a_j\rangle = I(i = j), \quad (2.7)$$

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

$$\Sigma_0(\mathbf{u},\mathbf{v}) = \mathrm{Cov}\{\xi_t(\mathbf{u}), \xi_t(\mathbf{v})\} = \sum_{i,j=1}^{d} a_i(\mathbf{u})a_j(\mathbf{v})\sigma_{ij}, \quad (2.8)$$

where σ_{ij} is the (i,j)-th element of $\mathrm{Var}(\mathbf{x}_t)$. Let

$$\Sigma_0 \circ f(\mathbf{s}) = \int_{\mathcal{S}} \Sigma_0(\mathbf{s},\mathbf{u})f(\mathbf{u})d\mathbf{u}, \qquad f \in L_2(\mathcal{S}). \quad (2.9)$$

Then Σ_0 is a non-negative definite operator defined in $L_2(\mathcal{S})$. See Appendix A of Bathia *et al.* (2010) for some basic facts on the operators in Hilbert spaces. It follows from Mercer's theorem (Mercer 1909) that Σ_0 admits the spectral decomposition

$$\Sigma_0(\mathbf{u}, \mathbf{v}) = \sum_{j=1}^{d} \lambda_j \varphi_j(\mathbf{u}) \varphi_j(\mathbf{v}), \tag{2.10}$$

where $\lambda_1 \geq \cdots \geq \lambda_d > 0$ are the d positive eigenvalues of $\Sigma_0(\mathbf{u}, \mathbf{v})$, and $\varphi_1, \cdots, \varphi_d \in L_2(\mathcal{S})$ are the corresponding eigenfunctions, i.e.

$$\Sigma_0 \circ \varphi_j(\mathbf{s}) = \int_{\mathcal{S}} \Sigma_0(\mathbf{s}, \mathbf{u}) \varphi_j(\mathbf{u}) d\mathbf{u} = \lambda_j \varphi_j(\mathbf{s}). \tag{2.11}$$

See Proposition 1 below.

Proposition 1 *Let rank*$(\mathrm{Var}(\mathbf{x}_t)) = d$. *Then the following assertions hold.*

(i) Σ_0 *defined in (2.9) has exactly* d *positive eigenvalues.*

(ii) The d *corresponding orthonormal eigenfunctions can be expressed as*

$$\varphi_i(\mathbf{s}) = \sum_{j=1}^{d} \gamma_{ij} a_j(\mathbf{s}), \qquad i = 1, \cdots, d,$$

where $\boldsymbol{\gamma}_i \equiv (\gamma_{i1}, \cdots, \gamma_{id})'$, $i = 1, \cdots, d$, *are* d *orthonormal eigenvectors of matrix* $\mathrm{Var}(\mathbf{x}_t)$.

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

3 Estimation

Let $\{(y_t(\mathbf{s}_i), \mathbf{z}_t(\mathbf{s}_i)),\ i = 1, \cdots, p,\ t = 1, \cdots, n\}$ be the available observations over space and time, where $\mathcal{S}_o \equiv \{\mathbf{s}_1, \cdots, \mathbf{s}_p\} \subset \mathcal{S}$ are typically irregularly spaced. The total number of observations is $n \cdot p$.

3.1 Estimation for finite dimensional representations of $\xi_t(\mathbf{s})$

To simplify the notation, we first consider a special case $\boldsymbol{\beta}(\mathbf{s}) \equiv 0$ in (2.1) in Sections 3.1 & 3.2. Section 3.4 below considers the least squares regression estimation for $\boldsymbol{\beta}(\mathbf{s})$. Then the procedures describe in Sections 3.1 & 3.2 still apply if $\{y_t(\mathbf{s}_i)\}$ are replaced by the residuals from the regression estimation.

Now the observations are taken from the process

$$y_t(\mathbf{s}) = \xi_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}) = \sum_{j=1}^{d} a_j(\mathbf{s}) x_{tj} + \varepsilon_t(\mathbf{s}). \tag{3.1}$$

To exclude nugget effect in our estimation, we divide p locations $\mathbf{s}_1, \cdots, \mathbf{s}_p$ into two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ with, respectively, p_1 and p_2 elements, and $p_1 + p_2 = p$. Let $\mathbf{y}_{t,i}$ be a vector consisting of $y_t(\mathbf{s})$ with $\mathbf{s} \in \mathcal{S}_i$, $i = 1, 2$. Then $\mathbf{y}_{t,1}, \mathbf{y}_{t,2}$ are two vectors with lengths p_1 and p_2 respectively. Denoted by $\boldsymbol{\xi}_{t,1}$, $\boldsymbol{\xi}_{t,2}$ the corresponding vectors consisting of $\xi_t(\cdot)$. It follows from (3.1) that

$$\mathbf{y}_{t,1} = \boldsymbol{\xi}_{t,1} + \boldsymbol{\varepsilon}_{t,1} = \mathbf{A}_1 \mathbf{x}_t + \boldsymbol{\varepsilon}_{t,1}, \qquad \mathbf{y}_{t,2} = \boldsymbol{\xi}_{t,2} + \boldsymbol{\varepsilon}_{t,2} = \mathbf{A}_2 \mathbf{x}_t + \boldsymbol{\varepsilon}_{t,2}, \tag{3.2}$$

where $\mathbf{A}_i$ is a $p_i \times d$ matrix, its rows consist of the coefficients $a_j(\cdot)$ on the RHS of (3.1), and $\boldsymbol{\varepsilon}_{t,i}$ consists of $\varepsilon_t(\mathbf{s})$ with $\mathbf{s} \in \mathcal{S}_i$. There is no loss of generality in assuming $\mathbf{A}_1'\mathbf{A}_1 = \mathbf{I}_d$. This can be achieved by performing an orthogonal-triangular (QR) decomposition $\mathbf{A}_1 = \boldsymbol{\Gamma}\mathbf{R}$, and replacing $(\mathbf{A}_1, \mathbf{x}_t)$ by $(\boldsymbol{\Gamma}, \mathbf{R}\mathbf{x}_t)$ in the first equation in (3.2). Note $\mathcal{M}(\mathbf{A}_1) = \mathcal{M}(\boldsymbol{\Gamma})$, where $\mathcal{M}(\mathbf{A})$ denotes the linear space spanned by the columns of matrix $\mathbf{A}$. Thus $\mathcal{M}(\mathbf{A}_1)$ does not change from imposing the condition $\mathbf{A}_1'\mathbf{A}_1 = \mathbf{I}_d$. Similar we may also assume $\mathbf{A}_2'\mathbf{A}_2 = \mathbf{I}_d$, which however implies that $\mathbf{x}_t$ in the second equation in (3.2) will be different from that in the first equation. Hence we may re-write (3.2) as

$$\mathbf{y}_{t,1} = \mathbf{A}_1 \mathbf{x}_t + \boldsymbol{\varepsilon}_{t,1}, \qquad \mathbf{y}_{t,2} = \mathbf{A}_2 \mathbf{x}_t^\star + \boldsymbol{\varepsilon}_{t,2}, \tag{3.3}$$

where $\mathbf{A}_1'\mathbf{A}_1 = \mathbf{A}_2'\mathbf{A}_2 = \mathbf{I}_d$, $\mathbf{x}_t^\star = \mathbf{Q}\mathbf{x}_t$, and $\mathbf{Q}$ is an invertible $d \times d$ matrix. Note that $(\mathbf{A}_1, \mathbf{x}_t)$ and $(\mathbf{A}_2, \mathbf{x}_t^\star)$ are still not uniquely defined in (3.3), as they can be replaced, respectively, by $(\mathbf{A}_1\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_1'\mathbf{x}_t)$ and $(\mathbf{A}_2\boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_2'\mathbf{x}_t^\star)$ for any $d \times d$ orthogonal matrices $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$. However $\mathcal{M}(\mathbf{A}_1)$ and $\mathcal{M}(\mathbf{A}_2)$ are uniquely defined by (3.3).

Since $\mathbf{y}_{t,1}$ and $\mathbf{y}_{t,2}$ have no common elements, it follows from (3.1) and (2.2) that

$$\mathbf{\Sigma} \equiv \mathrm{Cov}(\mathbf{y}_{t,1}, \mathbf{y}_{t,2}) = \mathbf{A}_1 \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t^\star)\mathbf{A}_2'. \tag{3.4}$$

Note that $\mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t^\star) = \mathrm{Var}(\mathbf{x}_t)\mathbf{Q}$. When $p \gg d$, it is reasonable to assume that $\mathrm{rank}(\mathbf{\Sigma}) = \mathrm{rank}\{\mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t^\star)\} = \mathrm{rank}\{\mathrm{Var}(\mathbf{x}_t)\} = d$. Let

$$\mathbf{\Sigma}\mathbf{\Sigma}' = \mathbf{A}_1 \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t^\star)\mathrm{Cov}(\mathbf{x}_t^\star, \mathbf{x}_t)\mathbf{A}_1', \qquad \mathbf{\Sigma}'\mathbf{\Sigma} = \mathbf{A}_2 \mathrm{Cov}(\mathbf{x}_t^\star, \mathbf{x}_t)\mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t^\star)\mathbf{A}_2'. \tag{3.5}$$

Then these two matrices share the same d positive eigenvalues, and $\mathbf{\Sigma}\mathbf{\Sigma}'\mathbf{b} = 0$ for any vector $\mathbf{b}$ perpendicular to $\mathcal{M}(\mathbf{A}_1)$. Therefore, the d orthonormal eigenvectors of matrix $\mathbf{\Sigma}\mathbf{\Sigma}'$ corresponding to its d positive eigenvalues can be taken as the columns of $\mathbf{A}_1$. Similarly the d orthonormal eigenvectors of matrix $\mathbf{\Sigma}'\mathbf{\Sigma}$ corresponding to its d positive eigenvalues can be taken as the columns of $\mathbf{A}_2$. We construct the estimators for $\mathbf{A}_1$, $\mathbf{A}_2$ based on this observation.

Let $\widehat{\mathbf{\Sigma}}$ be the sample covariance of $\mathbf{y}_{t,1}$ and $\mathbf{y}_{t,2}$, i.e.

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{y}_{t,1} - \bar{\mathbf{y}}_1)(\mathbf{y}_{t,2} - \bar{\mathbf{y}}_2)', \tag{3.6}$$

where $\bar{\mathbf{y}}_i = n^{-1}\sum_t \mathbf{y}_{t,i}$. Let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots$ be the eigenvalues of $\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}'$, $\widehat{\gamma}_{1,1}, \widehat{\gamma}_{2,1}, \cdots$ and $\widehat{\gamma}_{1,2}, \widehat{\gamma}_{2,2}, \cdots$ be, respectively, the corresponding orthonormal eigenvectors of $\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}'$ and $\widehat{\mathbf{\Sigma}}'\widehat{\mathbf{\Sigma}}$. Then the estimators for $\mathbf{A}_1$ and $\mathbf{A}_2$ are defined as

$$\widehat{\mathbf{A}}_1 = (\widehat{\gamma}_{1,1}, \cdots, \widehat{\gamma}_{d,1}), \qquad \widehat{\mathbf{A}}_2 = (\widehat{\gamma}_{1,2}, \cdots, \widehat{\gamma}_{d,2}). \tag{3.7}$$

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

$$\widehat{\mathbf{x}}_t = \widehat{\mathbf{A}}_1'\mathbf{y}_{t,1}, \qquad \widehat{\mathbf{x}}_t^\star = \widehat{\mathbf{A}}_2'\mathbf{y}_{t,2}. \tag{3.8}$$

Consequently,

$$\widehat{\boldsymbol{\xi}}_{t,1} = \widehat{\mathbf{A}}_1\widehat{\mathbf{x}}_t = \widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1'\mathbf{y}_{t,1}, \qquad \widehat{\boldsymbol{\xi}}_{t,2} = \widehat{\mathbf{A}}_2\widehat{\mathbf{x}}_t^\star = \widehat{\mathbf{A}}_2\widehat{\mathbf{A}}_2'\mathbf{y}_{t,2}. \tag{3.9}$$

See also (3.2).

In practice, we also need to estimate d. We adopt the ratio estimation method of Lam and Yao (2012), i.e. define the estimator as

$$\widehat{d} = \max_{1 \leq j < p_*} \widehat{\lambda}_j / \widehat{\lambda}_{j+1}, \tag{3.10}$$

where $p_* \ll \min(p_1, p_2)$ is a prespecified integer (*e.g.* $p_* = \min(p_1, p_2)/2$).

Remark 1 *(i) The d-dimensional structure (2.6) is reflected by the fact that the matrices defined in (3.5) share the same d non-zero eigenvalues. Thus assumption (2.6) can be checked from data. When p is fixed,* $\widehat{\lambda}_j$ *is a* $\sqrt{n}$*-consistent estimator for its non-zero true value for* $1 \le j \le d$*, and* $\widehat{\lambda}_j = O_P(1/n)$ *for* $d < j \le \min(p_1, p_2)$*; see Proposition 2 in Section 5.1 below. This super-fast convergence rate for the zero-eigenvalues is extremely beneficial in identifying the dimension d. Similar results for diverging p are presented in Proposition 3.*

(ii) The estimator $\widehat{d}$ *is obtained by comparing the ratios of the successive eigenvalues directly, based on the observation that* λ_j/λ_{j+1} *are positive and finite constants for* $j = 1, \cdots, d-1$*, and* $\lambda_d/\lambda_{d+1} = \infty$*. However the ratio is asymptotically '0/0' for* $j = d+1, \cdots, p$*. 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. Lam and Yao (2012), Chang* et al. *(2015), and Zhang* et al. *(2015). The (fine) finite sample performance of the ratio estimators are also reported in those papers.*

(iii) In the above estimation, we have not taken into account the fact that $\xi_t(\cdot)$ *may exhibit certain degree of continuity over the set* $\mathcal{S}$*. To this effect, we may require the eigenvectors of* $\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}'$ *(or* $\widehat{\mathbf{\Sigma}}'\widehat{\mathbf{\Sigma}}$*) satisfy the constraints*

$$\widehat{\mathbf{a}}_j' \mathbf{L} \widehat{\mathbf{a}}_j \le c_0, \qquad j = 1, \cdots, p \tag{3.11}$$

where $c_0 > 0$ *is a constant, and* $\mathbf{L} \equiv \mathbf{G} - \mathbf{W}$ *is a graph Laplacian, i.e.* $\mathbf{W} = (w_{ij})$ *is a weight matrix with* $w_{ii} = 0$ *and, e.g.* $w_{ij} = 1/(1 + \|\mathbf{s}_i - \mathbf{s}_j\|)$ *(*$\|\cdot\|$ *denotes the Euclidean norm) for* $i \ne j$*, and* $\mathbf{G} = (g_{ij})$ *with* $g_{ii} = \sum_j w_{ij}$ *and* $g_{ij} = 0$ *for all* $i \ne j$*. See, e.g., Hastie, Tibshirani and Friedman (2009, pp.545). Then it holds that for any* $\mathbf{x} = (x_1, \cdots, x_p)'$*,*

$$\mathbf{x}'\mathbf{L}\mathbf{x} = \sum_{i=1}^{p} g_{ii} x_i^2 - \sum_{i,j=1}^{p} w_{ij} x_i x_j = \frac{1}{2} \sum_{i,j=1}^{p} w_{ij} (x_i - x_j)^2.$$

Hence (3.11) ensures that the loadings for the nearby sites are similar. The constrained eigen-problem can be recast as the eigenanalysis for matrix $\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}' - \tau\mathbf{L}$*, where* $\tau > 0$ *controls the penalty according to* $\mathbf{L}$*.*

3.2 Aggregating via random partitioning

The estimation for the latent variable $\xi_t(\cdot)$ depends on partitioning $\mathcal{S}_o = \{\mathbf{s}_1, \cdots, \mathbf{s}_p\}$ into two non-overlapping sets $\mathcal{S}_1$ and $\mathcal{S}_2$; see (3.9). Since the estimation procedure presented in Section 3.1 puts $\mathcal{S}_1$ and $\mathcal{S}_2$ on equal footing, we set $p_1 = [p/2]$. By randomly dividing $\mathcal{S}_o$ into $\mathcal{S}_1$ and $\mathcal{S}_2$ with the sizes p_1 and p_2 respectively, the estimates for $\boldsymbol{\xi}_{t,1}$ and $\boldsymbol{\xi}_{t,2}$ are obtained as in (3.9). We repeat this randomization J times, where $J \geq 1$ is a large integer, leading to the J pairs of the estimates $(\widehat{\boldsymbol{\xi}}^j_{t,1}, \widehat{\boldsymbol{\xi}}^j_{t,2})$ for $j = 1, \cdots, J$. The aggregating estimator over the randomized partitions is

$$\widetilde{\xi}_t(\mathbf{s}_i) = \frac{1}{J}\sum_{j=1}^{J} \widehat{\xi}^j_t(\mathbf{s}_i), \qquad j = 1, \cdots, p, \tag{3.12}$$

where $\widehat{\xi}^j_t(\mathbf{s}_i)$ is a component of either $\widehat{\boldsymbol{\xi}}^j_{t,1}$ or $\widehat{\boldsymbol{\xi}}^j_{t,2}$, depending on $\mathbf{s}_i \in \mathcal{S}_1$ or $\mathcal{S}_2$ in the j-th randomized partition of $\mathcal{S}_o$. Similar to the Bagging method, the choice of J is not critical. In our numerical experiments, we set $J = 100$.

Theorem 1 *For $\widehat{\xi}_t(\mathbf{s}_j)$ defined in (3.9) and $\widetilde{\xi}_t(\mathbf{s}_j)$ defined in (3.12), the following two assertions hold.*

(i) For $k = 1, \cdots, n$ and $\ell = 1, \cdots, p$,

$$E\Big(\big\{\widetilde{\xi}_k(\mathbf{s}_\ell) - y_k(\mathbf{s}_\ell)\big\}^2 \Big| \{y_t(\mathbf{s}_i)\}\Big) \leq E\Big(\big\{\widehat{\xi}_k(\mathbf{s}_\ell) - y_k(\mathbf{s}_\ell)\big\}^2 \Big| \{y_t(\mathbf{s}_i)\}\Big), \tag{3.13}$$

and

$$\mathrm{E}\Big(\frac{1}{np}\sum_{t=1}^{n}\sum_{j=1}^{p} \big\{\widetilde{\xi}_t(\mathbf{s}_j) - \xi_t(\mathbf{s}_j)\big\}^2 \Big| \{\xi_t(\mathbf{s}_i),\, y_t(\mathbf{s}_i)\}\Big) \leq \mathrm{E}\Big(\frac{1}{np}\sum_{t=1}^{n}\sum_{j=1}^{p} \big\{\widehat{\xi}_t(\mathbf{s}_j) - \xi_t(\mathbf{s}_j)\big\}^2 \Big| \{\xi_t(\mathbf{s}_i),\, y_t(\mathbf{s}_i)\}\Big). \tag{3.14}$$

(ii) Let the condition in Lemma 3 in the Appendix in the supplementary file hold. As $p^\delta n^{-1/2} \to 0$, it holds that

$$\mathrm{E}\Big(\frac{1}{np}\sum_{t=1}^{n}\sum_{j=1}^{p} \big\{\widehat{\xi}_t(\mathbf{s}_j) - \xi_t(\mathbf{s}_j)\big\}^2 \Big| \{\xi_t(\mathbf{s}_i),\, y_t(\mathbf{s}_i)\}\Big) = O_p(p^\delta/n + p^{(\delta-1)/2}n^{-/2} + p^{-1}).$$

Further, if $p^{1+\delta}/n \to 0$, it holds in probability that

$$\mathrm{E}\Big(\frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{p} \big\{\widehat{\xi}_t(\mathbf{s}_j) - \xi_t(\mathbf{s}_j)\big\}^2 \Big| \{\xi_t(\mathbf{s}_i),\, y_t(\mathbf{s}_i)\}\Big) = \mathrm{E}\Big[\boldsymbol{\varepsilon}'_{t,1}\mathbf{A}_1\mathbf{A}'_1\boldsymbol{\varepsilon}_{t,1} + \boldsymbol{\varepsilon}'_{t,2}\mathbf{A}_2\mathbf{A}'_2\boldsymbol{\varepsilon}_{t,2}\Big].$$

The inequality in Theorem 1(i) is in the same spirit as Breiman's inequality for Bagging; see (4.2) in Breiman (1996). Note that all the conditional expectations in Theorem 1 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 $p_0 \equiv p!/(p_1!p_2!)$ different partitions, each being taken with probability $1/p_0$. Denote by $\widehat{\xi}_k^{(1)}(\cdot), \cdots, \widehat{\xi}_k^{(p_0)}(\cdot)$ the resulting p_0 estimates as in (3.9). Then

$$E\Big(\big\{\widetilde{\xi}_k(\mathbf{s}_\ell) - y_k(\mathbf{s}_\ell)\big\}^2 \Big| \{y_t(\mathbf{s}_i)\}\Big) \;=\; E\Big(\big\{\frac{1}{J}\sum_{l=1}^{J}(\widehat{\xi}_k^{l}(\mathbf{s}_\ell) - y_k(\mathbf{s}_\ell))\big\}^2 \Big| \{y_t(\mathbf{s}_i)\}\Big)$$

$$\leq E\Big(\frac{1}{J}\sum_{l=1}^{J}\big\{(\widehat{\xi}_k^{l}(\mathbf{s}_\ell) - y_k(\mathbf{s}_\ell))\big\}^2 \Big| \{y_t(\mathbf{s}_i)\}\Big)$$

$$= \frac{1}{p_0}\sum_{j=1}^{p_0}\big\{\widehat{\xi}_k^{(j)}(\mathbf{s}_\ell) - y_k(\mathbf{s}_\ell)\big\}^2 = E\Big(\big\{\widehat{\xi}_k(\mathbf{s}_\ell) - y_k(\mathbf{s}_\ell)\big\}^2 \Big| \{y_t(\mathbf{s}_i)\}\Big).$$

This completes the proof for (3.13). Note that (3.14) can be established in the same manner. The proof for the second part of the theorem is given in the Appendix in the supplementary file.

3.3 Scalable to large datasets

The estimators $\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{A}}_2$ in (3.7) were obtained from the singular value decomposition (SVD) for the $p_1 \times p_2$ matrix $\widehat{\mathbf{\Sigma}}$ in (3.6), 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 divided $\mathcal{S}_0$ into p/q sets $\mathcal{S}_1^*, \cdots, \mathcal{S}_q^*$, and each $\mathcal{S}_i^*$ contains q locations, where q is an integer such that the SVD can be performed comfortably with the available computing capacity. We estimate $\xi_t(\cdot)$ at the q locations in $\mathcal{S}_i^*$ for each of $i = 1, \cdots, p/q$ 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 $\widehat{\xi}_t(\mathbf{s})$ for $\mathbf{s} \in \mathcal{S}_i^*$ as in (3.9).

(iii) Repeat (i) and (ii) above J times, aggregate the estimates as in (3.12).

Alternatively, we can randomly choose $2q$ locations from $\mathcal{S}_0$ to perform the estimation (3.9). Repeating the estimation a large number (say, greater than $Jp/(2q)$) of times, we then aggregate

the estimates at each location as in (3.12). 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 $\mathbf{z}_t(\cdot)$ in (2.1), the regression coefficient vector $\boldsymbol{\beta}(\cdot)$ can be estimated by the least squares method. To this end, let

$$\mathbf{y}(\mathbf{s}_i) = (y_1(\mathbf{s}_i), \cdots, y_n(\mathbf{s}_i))', \qquad \mathbf{Z}(\mathbf{s}_i) = (\mathbf{z}_1(\mathbf{s}_i), \cdots, \mathbf{z}_n(\mathbf{s}_i))'. \tag{3.15}$$

It follows from (2.1) that

$$\mathbf{y}(\mathbf{s}_i) = \mathbf{Z}(\mathbf{s}_i)\boldsymbol{\beta}(\mathbf{s}_i) + \mathbf{e}(\mathbf{s}_i),$$

where $\mathbf{e}(\mathbf{s}_i) = (\xi_1(\mathbf{s}_i) + \varepsilon_1(\mathbf{s}_i), \cdots, \xi_n(\mathbf{s}_i) + \varepsilon_n(\mathbf{s}_i))'$. Thus the least squares estimator for $\boldsymbol{\beta}(\mathbf{s}_i)$ is defined as

$$\widehat{\boldsymbol{\beta}}(\mathbf{s}_i) = \{\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i)\}^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{y}(\mathbf{s}_i), \qquad i = 1, \cdots, p. \tag{3.16}$$

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

However in the presence of the endogeneity in the sense $\mathrm{Cov}(\mathbf{z}_t(\mathbf{s}),\, \xi_t(\mathbf{s})) \neq 0$, the regression estimator $\widehat{\boldsymbol{\beta}}(\mathbf{s}_i)$ in (3.16) is practically the estimator for

$$\boldsymbol{\beta}(\mathbf{s}_i)^\star \equiv \boldsymbol{\beta}(\mathbf{s}_i) + \mathrm{Var}(\mathbf{z}_t(\mathbf{s}_i))^{-1}\mathrm{Cov}(\mathbf{z}_t(\mathbf{s}_i),\, \xi_t(\mathbf{s}_i))$$

instead, as (2.1) can be written as $y_t(\mathbf{s}) = \mathbf{z}_t(\mathbf{s})'\boldsymbol{\beta}(\mathbf{s})^\star + \xi_t(\mathbf{s})^\star + \varepsilon_t(\mathbf{s})$, where

$$\xi_t(\mathbf{s})^\star = \xi_t(\mathbf{s}) - \mathbf{z}_t(\mathbf{s})'\mathrm{Var}(\mathbf{z}_t(\mathbf{s}_i))^{-1}\mathrm{Cov}(\mathbf{z}_t(\mathbf{s}_i), \xi_t(\mathbf{s}_i)).$$

It is easy to see that $\mathrm{Cov}(\mathbf{z}_t(\mathbf{s}),\ \xi_t(\mathbf{s})^\star) = 0$. Hence $\widehat{\boldsymbol{\beta}}(\mathbf{s}_i)$ is a consistent estimator for $\boldsymbol{\beta}(\mathbf{s}_i)^\star$. 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 $\xi_t(\mathbf{s})^\star$ instead.

4 Kriging

First we state a general lemma on linear prediction which shows explicitly the terms required in order to carry out kriging for spatio-temporal process $y_t(\mathbf{s})$.

Lemma 1 *For any random vectors $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ with $E(\|\boldsymbol{\zeta}\|^2 + \|\boldsymbol{\eta}\|^2) < \infty$, the best linear predictor for $\boldsymbol{\zeta}$ based on $\boldsymbol{\eta}$ is defined as $\widehat{\boldsymbol{\zeta}} = \boldsymbol{\alpha}_0 + \mathbf{B}_0\boldsymbol{\eta}$, where*

$$(\boldsymbol{\alpha}_0, \mathbf{B}_0) = \arg\inf_{\boldsymbol{\alpha},\mathbf{B}} E\big\{\|\boldsymbol{\zeta} - \boldsymbol{\alpha} - \mathbf{B}\boldsymbol{\eta}\|^2\big\}.$$

In fact,

$$\mathbf{B}_0 = \mathrm{Cov}(\boldsymbol{\zeta}, \boldsymbol{\eta})\{\mathrm{Var}(\boldsymbol{\eta})\}^{-1}, \qquad \boldsymbol{\alpha}_0 = E\boldsymbol{\zeta} - \mathbf{B}_0 E\boldsymbol{\eta}.$$

Furthermore,

$$E\{(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})'\} = \mathrm{Var}(\boldsymbol{\zeta}) - \mathrm{Cov}(\boldsymbol{\zeta}, \boldsymbol{\eta})\{\mathrm{Var}(\boldsymbol{\eta})\}^{-1}\mathrm{Cov}(\boldsymbol{\eta}, \boldsymbol{\zeta}). \tag{4.1}$$

With the above lemma, we can predict any value $y_t(\mathbf{s})$. 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 the size of p. Technically we repeatedly use the following formulas for the inverses of partitioned matrices.

Lemma 2 *For an invertible block-partitioned matrix $\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{pmatrix}$, it holds that*

$$\mathbf{H}^{-1} = \begin{pmatrix} \mathbf{H}_{11}^{-1} + \mathbf{H}_{11}^{-1}\mathbf{H}_{12}(\mathbf{H}_{22} - \mathbf{H}_{21}\mathbf{H}_{11}^{-1}\mathbf{H}_{12})^{-1}\mathbf{H}_{21}\mathbf{H}_{11}^{-1} & -\mathbf{H}_{11}^{-1}\mathbf{H}_{12}(\mathbf{H}_{22} - \mathbf{H}_{21}\mathbf{H}_{11}^{-1}\mathbf{H}_{12})^{-1} \\ -(\mathbf{H}_{22} - \mathbf{H}_{21}\mathbf{H}_{11}^{-1}\mathbf{H}_{12})^{-1}\mathbf{H}_{21}\mathbf{H}_{11}^{-1} & (\mathbf{H}_{22} - \mathbf{H}_{21}\mathbf{H}_{11}^{-1}\mathbf{H}_{12})^{-1} \end{pmatrix} \tag{4.2}$$

provided $\mathbf{H}_{11}^{-1}$ exists. Furthermore,

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

provided both $\mathbf{H}_{11}^{-1}$ and $\mathbf{H}_{22}^{-1}$ exist.

Formula (4.2) can be proved by checking $\mathbf{H}^{-1}\mathbf{H} = \mathbf{I}$ directly, while (4.3) follows from (4.2) by comparing the (1,1) and (2,2) blocks on the RHS of (4.2).

4.1 Kriging over space

The goal is to predict the unobserved value $y_t(\mathbf{s}_0)$ for some $\mathbf{s}_0 \in \mathcal{S}$, $1 \le t \le n$, and $\mathbf{s}_0 \neq \mathbf{s}_j$ for $1 \le j \le p$, based on the observations $\mathbf{y}_t \equiv (\mathbf{y}'_{t,1}, \mathbf{y}'_{t,2})'$ only, where $\mathbf{y}_{t,1}, \mathbf{y}_{t,2}$ 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 $\boldsymbol{\beta}(\mathbf{s}) \equiv 0$ in (2.1). As indicated in Section 3.4, this effectively implies to replace the observations $y_t(\mathbf{s}_j)$ by the regression residuals. For kriging, we also need to estimate $\boldsymbol{\beta}(\mathbf{s}_0)$ based on $\widehat{\boldsymbol{\beta}}(\mathbf{s}_j)$, $j = 1, \cdots, p$, given in (3.16). It can be achieved by, for example, using the kernel smoothing:

$$\widehat{\boldsymbol{\beta}}(\mathbf{s}_0) = \sum_{j=1}^{p} \widehat{\boldsymbol{\beta}}(\mathbf{s}_j) K_h(\mathbf{s}_j - \mathbf{s}_0) \Big/ \sum_{j=1}^{p} K_h(\mathbf{s}_j - \mathbf{s}_0), \tag{4.4}$$

where $K(\cdot)$ is a density function defined on $\mathcal{R}^2$, $h > 0$ is a bandwidth. Furthermore, a local linear smoothing can be applied to improve the accuracy of the estimation; see, e.g. Chapter 3 of Fan and Gijbels (1996). By the standard argument it can be shown that

$$|\widehat{\boldsymbol{\beta}}(\mathbf{s}_0) - \boldsymbol{\beta}(\mathbf{s}_0)| = O_p(h^2 + n^{-1/2}),$$

provided that the conditions in Theorem 2 in Section 5.2 below hold. Note that if $\boldsymbol{\beta}(\mathbf{s}_j)$, $j = 1, \cdots, p$, were all known, the above error rate reduces to $O_p(h^2)$, as $\boldsymbol{\beta}(\cdot)$ is deterministic and continuous. See Condition 4 in Section 5.2 below. The $n^{-1/2}$ reflects the errors in estimation for $\boldsymbol{\beta}(\mathbf{s}_j)$. In the rest of Section 4, we adhere with the assumption $\boldsymbol{\beta}(\mathbf{s}) \equiv 0$.

It follows from Lemma 1 that the best linear predictor for $y_t(\mathbf{s}_0)$ based on $\mathbf{y}_t$ is

$$\widehat{y}_t(\mathbf{s}_0) = \mathrm{Cov}(y_t(\mathbf{s}_0), \mathbf{y}_t)\mathrm{Var}(\mathbf{y}_t)^{-1}\mathbf{y}_t. \tag{4.5}$$

It follows from (4.1) that

$$E[\{\widehat{y}_t(\mathbf{s}_0) - y_t(\mathbf{s}_0)\}^2] = \mathrm{Var}\{y_t(\mathbf{s}_0)\} - \mathrm{Cov}(y_t(\mathbf{s}_0), \mathbf{y}_t)\mathrm{Var}(\mathbf{y}_t)^{-1}\mathrm{Cov}(\mathbf{y}_t, y_t(\mathbf{s}_0))$$

$$= \sigma(\mathbf{s}_0)^2 + \mathrm{Var}\{\xi_t(\mathbf{s}_0)\} - \mathrm{Cov}(\xi_t(\mathbf{s}_0), \boldsymbol{\xi}_t)\{\mathrm{Var}(\boldsymbol{\xi}_t) + \mathbf{D}\}^{-1}\mathrm{Cov}(\boldsymbol{\xi}_t, \xi_t(\mathbf{s}_0)), \tag{4.6}$$

where $\mathbf{D} = \mathrm{Var}(\boldsymbol{\varepsilon}_t)$ is a diagonal matrix, $\boldsymbol{\varepsilon}_t = (\boldsymbol{\varepsilon}'_{t,1}, \boldsymbol{\varepsilon}'_{t,2})'$ and $\boldsymbol{\xi}_t = (\boldsymbol{\xi}'_{t,1}, \boldsymbol{\xi}'_{t,2})'$. See (3.2).

To apply predictor $\widehat{y}_t(\mathbf{s}_0)$ in (4.5) in practice, we need to estimate both $\mathrm{Cov}(y_t(\mathbf{s}_0), \mathbf{y}_t)$ and $\mathrm{Var}(\mathbf{y}_t)$. Since $\mathrm{Cov}(y_t(\mathbf{s}_0), \mathbf{y}_t) = \mathrm{Cov}(\xi_t(\mathbf{s}_0), \mathbf{y}_t)$, it can be estimated by

$$c(\mathbf{s}_0) = \frac{1}{n}\sum_{t=1}^{n}(\widehat{\xi}_t(\mathbf{s}_0) - \bar{\xi}(\mathbf{s}_0))(\mathbf{y}_t - \bar{\mathbf{y}}_t),$$

where $\widehat{\xi}_t(\mathbf{s}_0)$ is a kernel estimator for $\xi_t(\mathbf{s}_0)$ defined as

$$\widehat{\xi}_t(\mathbf{s}_0) = \sum_{j=1}^{p}\widehat{\xi}_t(\mathbf{s}_j)K_h(\mathbf{s}_j - \mathbf{s}_0)\Big/\sum_{j=1}^{p}K_h(\mathbf{s}_j - \mathbf{s}_0) \tag{4.7}$$

with $\widehat{\xi}_t(\mathbf{s}_1), \cdots \widehat{\xi}_t(\mathbf{s}_p)$ defined in (3.9) (see also (4.4) above), and $\bar{\xi}(\mathbf{s}_0) = n^{-1}\sum_t \widehat{\xi}_t(\mathbf{s}_0)$. Thus a realistic predictor for $y_t(\mathbf{s}_0)$ is

$$\widehat{y}^r_t(\mathbf{s}_0) = c(\mathbf{s}_0)\widehat{\boldsymbol{\Sigma}}_y^{-1}\mathbf{y}_t, \tag{4.8}$$

where $\widehat{\boldsymbol{\Sigma}}_y = n^{-1}\sum_{t=1}^{n}(\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})'$ is the sample variance of $\mathbf{y}_t$. Nevertheless it turns out that

$$\widehat{y}^r_t(\mathbf{s}_0) = \widehat{\xi}_t(\mathbf{s}_0). \tag{4.9}$$

To show this, let $w_j = K_h(\mathbf{s}_j - \mathbf{s}_0)\big/\sum_{j=1}^{p}K_h(\mathbf{s}_j - \mathbf{s}_0)$. It follows from (3.9) that

$$\begin{aligned}
\widehat{y}^r_t(\mathbf{s}_0) &= (w_1, \cdots, w_p)\Big[\frac{1}{n}\sum_{t=1}^{n}(\widehat{\boldsymbol{\xi}}_t - \bar{\boldsymbol{\xi}})(\mathbf{y}_t - \bar{\mathbf{y}})'\Big]\widehat{\Sigma}_y^{-1}\mathbf{y}_t \\
&= (w_1, \cdots, w_p)\begin{pmatrix} \widehat{\mathbf{A}}_1\widehat{\mathbf{A}}'_1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{A}}_2\widehat{\mathbf{A}}'_2 \end{pmatrix}\Big[\frac{1}{n}\sum_{t=1}^{n}(\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})'\Big]\widehat{\Sigma}_y^{-1}\mathbf{y}_t \\
&= (w_1, \cdots, w_p)\begin{pmatrix} \widehat{\mathbf{A}}_1\widehat{\mathbf{A}}'_1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{A}}_2\widehat{\mathbf{A}}'_2 \end{pmatrix}\mathbf{y}_t = (w_1, \cdots, w_p)\begin{pmatrix} \widehat{\boldsymbol{\xi}}_{t,1} \\ \widehat{\boldsymbol{\xi}}_{t,2} \end{pmatrix} = \widehat{\xi}_t(\mathbf{s}_0).
\end{aligned}$$

It is worth pointing out that expression (4.8) involves inverting $p \times p$ matrix $\widehat{\boldsymbol{\Sigma}}_y$, which is difficult when p is large, while (4.9) paves the way for computing the predictor $\widehat{y}^r_t(\mathbf{s}_0)$ without the need to compute $\widehat{\boldsymbol{\Sigma}}_y^{-1}$ directly; see (4.8).

By Theorem 1, a better predictor than $\widehat{y}_t^r(\mathbf{s}_0)$ in (4.9) is

$$\widetilde{y}_t^r(\mathbf{s}_0) \equiv \widetilde{\xi}_t(\mathbf{s}_0) = \sum_{j=1}^{p} \widetilde{\xi}_t(\mathbf{s}_j) K_h(\mathbf{s}_j - \mathbf{s}_0) \Big/ \sum_{j=1}^{p} K_h(\mathbf{s}_j - \mathbf{s}_0), \tag{4.10}$$

where $\widetilde{\xi}_t(\mathbf{s}_j)$ is defined in (3.12).

Both $\widehat{y}_t^r(\mathbf{s}_0)$ and $\widetilde{y}_t^r(\mathbf{s}_0)$ are the approximate linear estimators for the $\xi_t(\mathbf{s}_0)$ based on $\xi_t(\mathbf{s}_1), \cdots, \xi_t(\mathbf{s}_p)$. Note that $y_t(\mathbf{s}_0) = \xi_t(\mathbf{s}_0) + \varepsilon_t(\mathbf{s}_0)$, and the nugget effect term $\varepsilon_t(\mathbf{s}_0)$ is unpredictable. The best (unrealistic) predictor for $y_t(\mathbf{s}_0)$ is $\xi_t(\mathbf{s}_0)$. It is indeed recommended to predict $\xi_t(\mathbf{s}_0)$ instead of $y_t(\mathbf{s}_0)$ directly. See also pp.136-137 of Cressie and Wikle (2011).

4.2 Kriging in time

4.2.1 Prediction methods

The goal now is to predict the future values $y_{n+j}(\mathbf{s}_1), \cdots, y_{n+j}(\mathbf{s}_p)$, for some $j \geq 1$, based on $\mathbf{y}_n, \cdots, \mathbf{y}_{n-j_0}$, 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 $\mathbf{x}_{n+j} = (x_{n+j,1}, \cdots, x_{n+j,d})'$ based on $\mathbf{x}_n, \cdots, \mathbf{x}_{n-j_0}$, as the ideal predictor for $y_{n+j}(\mathbf{s}_i)$ is $\xi_{n+j}(\mathbf{s}_i)$; see (3.1).

Since our procedure to recover the latent process $\mathbf{x}_t$ requires to split $\mathbf{y}_t$ into two subvectors $\mathbf{y}_{t,1}$, $\mathbf{y}_{t,2}$, leading to two different configurations $\mathbf{x}_t$ and $\mathbf{x}_t^\star$ in (3.3), we will apply the prediction procedure in Section 4.2.2 below to each of $\mathbf{x}_t$ and $\mathbf{x}_t^\star$. Then the predictors for $\mathbf{y}_{n+j,1}$ and $\mathbf{y}_{n+j,2}$ are defined as

$$\mathbf{y}_{n,1}(j) = \mathbf{A}_1 \mathbf{x}_n(j), \qquad \mathbf{y}_{n,2}(j) = \mathbf{A}_2 \mathbf{x}_n^\star(j), \tag{4.11}$$

where $\mathbf{x}_n(j)$ is the predictor for $\mathbf{x}_{n+j}$, and $\mathbf{x}_n^\star(j)$ is the predictor for $\mathbf{x}_{n+j}^\star$. In practice, $\mathbf{A}_i, \mathbf{x}_t, \mathbf{x}_t^\star$ are replaced by their estimators defined in (3.7) and (3.8).

The predictors defined above depend on a single partition $\mathcal{S}_o = \mathcal{S}_1 \cup \mathcal{S}_2$. By repeating random partition of $\mathcal{S}_o$ J times, we may obtain the aggregated predicted values for $y_{n+j}(\mathbf{s}_i)$ in the same manner as in (3.12).

Since $\xi_t(\mathbf{s}_1), \cdots, \xi_t(\mathbf{s}_p)$ are correlated with each other, we should not model ξ_t at each location

separately. Instead modeling the factor process $\mathbf{x}_t$ catches the temporal dynamics much more parsimoniously.

4.2.2 Predicting $\mathbf{x}_{n+j}$ and $\mathbf{x}^\star_{n+j}$

We only state the method for predicting $\mathbf{x}_{n+j}$. It can be applied to predicting $\mathbf{x}^\star_{n+j}$ exactly in the same manner.

Let $\mathbf{X}' = (\mathbf{x}'_n, \cdots, \mathbf{x}'_{n-j_0})$,

$$\mathbf{W}_k \equiv \mathrm{Var}\begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \\ \vdots \\ \mathbf{x}_{t-k} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_x(0) & \boldsymbol{\Sigma}_x(1) & \cdots & \boldsymbol{\Sigma}_x(k) \\ \boldsymbol{\Sigma}_x(1)' & \boldsymbol{\Sigma}_x(0) & \cdots & \boldsymbol{\Sigma}_x(k-1) \\ & \cdots & \cdots & \\ \boldsymbol{\Sigma}_x(k)' & \boldsymbol{\Sigma}_x(k-1)' & \cdots & \boldsymbol{\Sigma}_x(0) \end{pmatrix}, \quad k \geq 0, \tag{4.12}$$

$$\mathbf{R}_{j_0} \equiv (\boldsymbol{\Sigma}_x(j), \boldsymbol{\Sigma}_x(j+1). \cdots, \boldsymbol{\Sigma}_x(j+j_0)),$$

where $\boldsymbol{\Sigma}_x(k) = \mathrm{Cov}(\mathbf{x}_{t+k}, \mathbf{x}_t)$. By Lemma 1, the best linear predictor for $\mathbf{x}_{n+j}$ is

$$\mathbf{x}_n(j) = \mathbf{R}_{j_0}\mathbf{W}_{j_0}^{-1}\mathbf{X}.$$

The key is to be able to calculate the inverse of $(j_0+1)d \times (j_0+1)d$ matrix $\mathbf{W}_{j_0}$. This can be done by calculating $\mathbf{W}_0^{-1}, \mathbf{W}_1^{-1}, \cdots$ recursively based on

$$\mathbf{W}_{k+1}^{-1} = \begin{pmatrix} \mathbf{W}_k^{-1} + \mathbf{W}_k^{-1}\mathbf{U}_k\mathbf{V}_k\mathbf{U}_k'\mathbf{W}_k^{-1} & -\mathbf{W}_k^{-1}\mathbf{U}_k\mathbf{V}_k \\ -\mathbf{V}_k\mathbf{U}_k'\mathbf{W}_k^{-1} & \mathbf{V}_k \end{pmatrix}, \tag{4.13}$$

where

$$\mathbf{U}_k' = (\boldsymbol{\Sigma}_x(k+1)', \cdots, \boldsymbol{\Sigma}_x(1)'), \qquad \mathbf{V}_k = (\boldsymbol{\Sigma}_x(0) - \mathbf{U}_k'\mathbf{W}_k^{-1}\mathbf{U}_k)^{-1}.$$

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

In practice we replace $\boldsymbol{\Sigma}_x(k)$ in $\mathbf{R}_{j_0}$ and $\mathbf{W}_{j_0}$ by $\widehat{\boldsymbol{\Sigma}}_x(k) = \widehat{\mathbf{A}}_1'\widehat{\boldsymbol{\Sigma}}_{y,1}(k)\widehat{\mathbf{A}}_1$, and replace $\mathbf{X}$ by

$$\widehat{\mathbf{X}} = (\mathbf{y}'_{t,1}\widehat{\mathbf{A}}_1, \cdots, \mathbf{y}'_{t-k,1}\widehat{\mathbf{A}}_1)',$$

where

$$\widehat{\boldsymbol{\Sigma}}_{y,1}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(\mathbf{y}_{t+k,1} - \bar{\mathbf{y}}_1)(\mathbf{y}_{t,1} - \bar{\mathbf{y}}_1)', \qquad \bar{\mathbf{y}}_1 = \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t,1}.$$

The resulting predictor for $\mathbf{x}_{n+j}$ is denoted by $\widehat{\mathbf{x}}_n(j)$.

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

$$\widehat{y}_{n,1}(j) = \widehat{\mathbf{A}}_1 \widehat{\mathbf{x}}_n(j), \qquad \widehat{y}_{n,2}(j) = \widehat{\mathbf{A}}_2 \widehat{\mathbf{x}}^{\star}_n(j), \tag{4.14}$$

see (4.11).

5 Asymptotic properties

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

5.1 On finite-dimensional representation of $\xi_t(\mathbf{s})$

We state in this subsection some asymptotic results on the estimation of the factor loading spaces $\mathcal{M}(\mathbf{A}_1)$ and $\mathcal{M}(\mathbf{A}_2)$. They paves the way to establish the properties for the kriging estimation presented in Section 5.2 below. The results presented in this section are similar to those in Lam and Yao (2012) and Chang, Guo and Yao (2015), though our setting is different in the sense that the estimation is based on the correlations across the space (instead of across time). Therefore we omit their proofs.

For any two $k \times d$ orthogonal matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ with $\mathbf{B}_1'\mathbf{B}_1 = \mathbf{B}_2'\mathbf{B}_2 = \mathbf{I}_d$, we measure the distance between the two linear spaces $\mathcal{M}(\mathbf{B}_1)$ and $\mathcal{M}(\mathbf{B}_2)$ by

$$D(\mathcal{M}(\mathbf{B}_1), \mathcal{M}(\mathbf{B}_2)) = \sqrt{1 - \frac{1}{d}\mathrm{tr}(\mathbf{B}_1\mathbf{B}_1'\mathbf{B}_2\mathbf{B}_2')}. \tag{5.1}$$

It can be shown that $D(\mathcal{M}(\mathbf{B}_1), \mathcal{M}(\mathbf{B}_2)) \in [0, 1]$, being 0 if and only if $\mathcal{M}(\mathbf{B}_1) = \mathcal{M}(\mathbf{B}_2)$, and 1 if and only if $\mathcal{M}(\mathbf{B}_1)$ and $\mathcal{M}(\mathbf{B}_2)$ are orthogonal. We consider two asymptotic modes: (i) p is fixed while $n \to \infty$, and (ii) $p \to \infty$ at a slower rate while $n \to \infty$. We introduce some regularity

conditions first. Put

$$\mathbf{y}_t = (y_t(\mathbf{s}_1), \cdots, y_t(\mathbf{s}_p))', \qquad \mathbf{Z}_t = (\mathbf{z}_t(\mathbf{s}_1), \cdots, \mathbf{z}_t(\mathbf{s}_p)).$$

Condition 1. $\{(\mathbf{y}_t, \mathbf{Z}_t),\, t = 0, \pm 1, \pm 2, \cdots\}$ is a strictly stationary and α-mixing process with $\max_{1\le i\le p}[\mathrm{E}|y_t(\mathbf{s}_i)|^\gamma + \mathrm{E}||\mathbf{z}_t(\mathbf{s}_i)||^\gamma] < \infty$ for some $\gamma > \max\{\beta, 4\}$, $\beta > 2$ and the α-mixing coefficients α_m satisfying the condition

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

Further, $\min_{1\le i\le p} \lambda_{\min}(\mathrm{Var}(\mathbf{z}_t(\mathbf{s}_i))) > c_0$ for some positive constant c_0.

Proposition 2 *Let p be a fixed constant and Condition 1 hold. Then as $n \to \infty$,*

(i) $|\widehat{\lambda}_i - \lambda_i| = O_p(n^{-1/2})$ *for* $1 \le i \le d$,

(ii) $|\widehat{\lambda}_i| = O_p(n^{-1})$ *for* $d < i \le p$, *and*

(iii) $D(\mathcal{M}(\widehat{\mathbf{A}}_i), \mathcal{M}(\mathbf{A}_i)) = O_p(n^{-1/2})$ $(i = 1, 2)$, *provided that d is known.*

The asymptotic results in (iii) of the above proposition can be made adaptive to unknown d, requiring a consistent estimator for d; see Remark 5 of Bathia, Yao and Ziegelmann (2010). Similar to Theorem 2.4 of Chang, Guo and Yao (2015), the ratio estimator $\widehat{d}$ in (3.10) can be modified to ensure the consistency by avoiding the technical difficulties in handling asymptotic '0/0'. See also Remark 1 (ii) in Section 3.1 above.

To handle the high-dimensional settings with $p = o(n^c)$ for some $c > 0$, we need to quantify the strength of latent factors (i.e. the components of $\mathbf{x}_t$ and $\mathbf{x}_t^\star$) in (3.3). See Lam, Yao and Bathia (2011) and Lam and Yao (2012). Intuitively a strong factor is linked with most components of $\mathbf{y}_{t,1}$ or $\mathbf{y}_{t,2}$, implying that the corresponding coefficients in $\mathbf{A}_1$ or $\mathbf{A}_2$ are non-zero. Therefore it is relatively easy to recover those strong factors from the observations. Unfortunately a formal mathematical definition of the factor strength is tangled with the standardization condition $\mathbf{A}_1'\mathbf{A}_1 = \mathbf{A}_2'\mathbf{A}_2 = \mathbf{I}_d$. To simplify the presentation, we assume that all the factors in (3.3) are of the same strength which is measured by a constant $\delta \in [0, 1]$ in Condition 2 below: $\delta = 0$ indicates

that the strength of the factors is at its strongest, and $\delta = 1$ corresponds to the weakest factors. See Remark 1(i) of Lam and Yao (2012) and Lemma 1 of Lam, Yao and Bathia (2011) on how the factor strength is represented by this δ.

Condition 2. Let $\mathbf{\Sigma}_x = \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t^\star)$, where $\mathbf{x}_t$ and $\mathbf{x}_t^\star$ are defined in (3.3). There exists a constant $\delta \in [0, 1]$ for which $||\mathbf{\Sigma}_x||_{\min} \asymp ||\mathbf{\Sigma}_x|| \asymp p^{1-\delta}$.

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

(i) $|\widehat{\lambda}_i - \lambda_i| = O_p(p^{2-\delta} n^{-1/2})$ *for* $1 \le i \le d$,

(ii) $|\widehat{\lambda}_i| = O_p(p^2 n^{-1})$ *for* $d < i \le p$, *and*

(iii) $D(\mathcal{M}(\widehat{\mathbf{A}}_i), \mathcal{M}(\mathbf{A}_i)) = O_p(p^\delta n^{-1/2})$ *(*$i = 1, 2$*), provided that d is known.*

Remark 2 *Proposition 3 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 δ decreases, where λ_i denotes the i-th largest eigenvalue of $\mathbf{\Sigma}\mathbf{\Sigma}'$, where $\mathbf{\Sigma}$ is defined in (3.4). Especially with the strongest factors (i.e. $\delta = 0$), $D(\mathcal{M}(\widehat{\mathbf{A}}_i), \mathcal{M}(\mathbf{A}_i))$ attains the standard convergence rate $\sqrt{n}$, in spite of diverging p. This phenomenon is coined as 'blessing of dimensionality' as in Lam and Yao (2012).*

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 $\mathcal{R}^2$ with a bounded support.

Condition 4. In (3.1) $\boldsymbol{\beta}(\cdot)$ and $a_j(\cdot)/||\mathbf{a}(\mathbf{s}_0)||$, $j = 1, \cdots, d$, are twice continuously differentiable and bounded functions on $\mathcal{S}$, where $\mathbf{a}(\mathbf{s}_0) = (a_1(\mathbf{s}_0), \cdots, a_d(\mathbf{s}_0))$.

Condition 5. There exists a positive and continuously differentiable sampling intensity $f(s)$ on $\mathcal{S}$ such that for any measurable set $A \subset \mathcal{S}$,

$$\frac{1}{p}\sum_{\mathbf{s}\in\mathcal{S}} I(\mathbf{s}\in A) \to \int_A f(\mathbf{s})\, d\mathbf{s}, \quad \text{as } p\to\infty.$$

Theorem 2 below presents the asymptotic properties of the two spatial kriging methods in (4.9) and (4.10). Since

$$E[\{\widehat{y}_t^r(\mathbf{s}_0) - y_t(\mathbf{s}_0)\}^2] = E[\{\widehat{y}_t^r(\mathbf{s}_0) - \xi_t(\mathbf{s}_0)\}^2] + \mathrm{Var}(\varepsilon_t(\mathbf{s}_0)),$$

it is more relevant to measure the difference between a predictor and $\xi_t(\mathbf{s}_0)$ directly. Hence Theorem 2 below.

Theorem 2 *Let bandwidth $h\to 0$ and $ph\to\infty$ as $n\to\infty$. It holds under Conditions 1–5 that*

$$\max\{|\widehat{y}_t^r(\mathbf{s}_0) - \xi_t(\mathbf{s}_0)|, |\widetilde{y}_t^r(\mathbf{s}_0) - \xi_t(\mathbf{s}_0)|\} = O_p\{h^2 + p^\delta (nh)^{-1/2} + (ph)^{-1/2}\}.$$

Remark 3 *Theorem 2 indicates that stronger factors result in better prediction for $\xi_t(\mathbf{s}_0)$, and, therefore, better kriging prediction.*

Theorem 3 below considers the convergence rates for the kriging predictions in time. Recall $\widehat{\mathbf{y}}_{n,1}(j)$, $\widehat{\mathbf{y}}_{n,2}(j)$, $\widehat{\mathbf{x}}_n(j)$ and $\widehat{\mathbf{x}}_n^\star(j)$ as defined in (4.14).

Theorem 3 *Let Conditions 1 and 2 hold. As $n,p\to\infty$ and $p^\delta n^{-1/2}\to 0$,*

(a) $p^{-\frac{1}{2}}||\widehat{\mathbf{x}}_n(j) - \mathbf{x}_n(j)|| = O_p(p^{\frac{\delta}{2}}n^{-\frac{1}{2}} + p^{-\frac{1}{2}})$, $p^{-\frac{1}{2}}||\widehat{\mathbf{x}}_n^\star(j) - \mathbf{x}_n^\star(j)|| = O_p(p^{\frac{\delta}{2}}n^{-\frac{1}{2}} + p^{-\frac{1}{2}})$, *and*

(b) $p^{-\frac{1}{2}}||\widehat{\mathbf{y}}_{n,i}(j) - \mathbf{y}_{n,i}(j)|| = O_p(p^{\frac{\delta}{2}}n^{-\frac{1}{2}} + p^{-\frac{1}{2}})$ *for $i=1,2$.*

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 $\mathbf{s}_1, \cdots, \mathbf{s}_p$ be drawn randomly from the uniform distribution on $[-1,1]^2$ and $y_t(\mathbf{s}_i)$ be generated from (3.1) in which $d = 3$, $\varepsilon_t(\mathbf{s})$ are independent and normal with mean 0 and the standard deviation $(1 + s_1^2 + s_2^2)/2$, and

$$a_1(\mathbf{s}) = (s_1 - s_2)/2, \qquad a_2(\mathbf{s}) = \cos\left\{(2(s_1^2 + s_2^2))^{1/2}\pi\right\}, \qquad a_3(\mathbf{s}) = 1.5 s_1 s_2,$$

$$x_{t1} = -0.8 x_{t-1,1} + e_{t1}, \qquad x_{t2} = e_{t2} - 0.9 e_{t-1,2}, \qquad x_{t3} = -0.8 x_{t-1,3} + e_{t3} + 0.3 e_{t-1,3}.$$

In the above expressions, e_{ti} are independent and standard normal. The signal-noise-ratio, which is defined as

$$\frac{\int_{\mathbf{s}\in[-1,1]^2} \sqrt{\mathrm{VAR}(\xi_t(\mathbf{s}))} d\mathbf{s}}{\int_{\mathbf{s}\in[-1,1]^2} \sqrt{\mathrm{VAR}(\varepsilon_t(\mathbf{s}))} d\mathbf{s}},$$

is about 1.70.

Setting $n = 60, 120$ and 240, $p = 50, 100, 200$ and 400, we draw 200 samples from each setting. With each sample, we calculate $\widehat{d}$ as in (3.10), and the factor loadings $\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{A}}_2$ as in (3.7). Figure 1 depicts the boxplots of the average distance

$$\frac{1}{2}\{D(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{A}_1)) + D(\mathcal{M}(\widehat{\mathbf{A}}_2), \mathcal{M}(\mathbf{A}_2))\}$$

over 200 samples for each setting. Since the estimated value $\widehat{d}$ is not always equal to d, and $\mathbf{A}_1$, $\mathbf{A}_2$ are not half-orthogonal matrices in the model below, we extend the distance measure for two linear spaces (5.1) to the following:

$$D(\mathcal{M}(\widehat{\mathbf{A}}_i), \mathcal{M}(\mathbf{A}_i)) = \left(1 - \frac{1}{\max(d, \widehat{d})} \mathrm{tr}\{\widehat{\mathbf{A}}_i \widehat{\mathbf{A}}_i' \mathbf{A}_i (\mathbf{A}_i'\mathbf{A}_i)^{-1} \mathbf{A}_i'\}\right)^{1/2}.$$

It can be shown that $D(\mathcal{M}(\widehat{\mathbf{A}}_i), \mathcal{M}(\mathbf{A}_i)) \in [0,1]$, being 0 if and only if $\mathcal{M}(\widehat{\mathbf{A}}_i) = \mathcal{M}(\mathbf{A}_i)$, and 1 if and only if $\mathcal{M}(\widehat{\mathbf{A}}_i)$ and $\mathcal{M}(\mathbf{A}_i)$ are orthogonal. It reduces to (5.1) when $\widehat{d} = d$ and $\mathbf{A}_i'\mathbf{A}_i = \mathbf{I}_d$. Figure 1 presents the boxplots of the above distance measure over 200 replications under different settings. As expected, the errors in estimating $\mathcal{M}(\mathbf{A}_1)$ and $\mathcal{M}(\mathbf{A}_2)$ 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

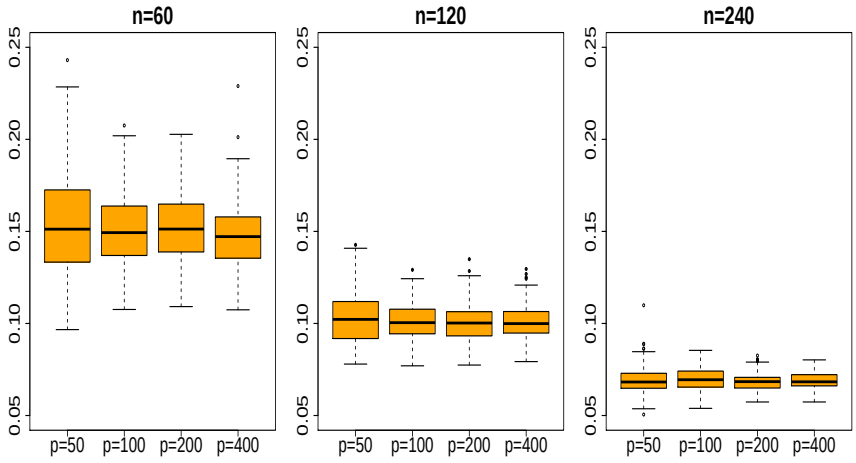

Figure 1: Boxplot of $\frac{1}{2}\{D(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{A}_1)) + D(\mathcal{M}(\widehat{\mathbf{A}}_2), \mathcal{M}(\mathbf{A}_2))\}$ from a simulation with 200 replications.

all strong factors. According to Proposition 3(iii), $D(\mathcal{M}(\widehat{\mathbf{A}}_i), \mathcal{M}(\mathbf{A}_i)) = O_p(n^{-1/2})$ when $\delta = 0$. See also Remark 2. The ratio estimator (3.10) for d works very well. Among all the settings, we observe the occurrence of the event $\{\widehat{d} \neq d\}$ only when $n = 60$ and $p = 50$ with the relative frequency smaller than 5%.

For each sample, we also evaluate $\widehat{\xi}_t(\mathbf{s}_j)$ as in (3.9). Specifically, we adopt the Gaussian kernel and the bandwidth selected by the leave-one-out cross-validation method in minimizing the mean squared error. We repeat this exercise also for the aggregation estimator $\widetilde{\xi}_t(\cdot)$ defined in (3.12) with $J = 100$. The boxplots of

$$\text{MSE}(\widehat{\xi}) = \frac{1}{np}\sum_{t=1}^{n}\sum_{j=1}^{p}\{\widehat{\xi}_t(\mathbf{s}_j) - \xi_t(\mathbf{s}_j)\}^2,$$

and $\text{MSE}(\widetilde{\xi})$, defined in the same manner, are displayed in Figure 2. For each combination of n and p, we draw the box plots of $\text{MSE}(\widehat{\xi})$ and $\text{MSE}(\widetilde{\xi})$ side by side. As indicated by Theorem 1, $\widetilde{\xi}_t(\mathbf{s}_j)$ provides much more accurate estimate for $\xi_t(\mathbf{s}_j)$ than $\widehat{\xi}_t(\mathbf{s}_j)$. Furthermore the MSE decreases when either n or p increases.

To illustrate the kriging performance, for each sample we also generate data at 50 'post-sample' locations drawn randomly from $U[-1, 1]^2$. For each $t = 1, \cdots, n$, we calculate the spatial kriging estimate $\widehat{y}_t^r(\cdot)$ in (4.9) at each of the 50 post-sample locations. The mean squared predictive error

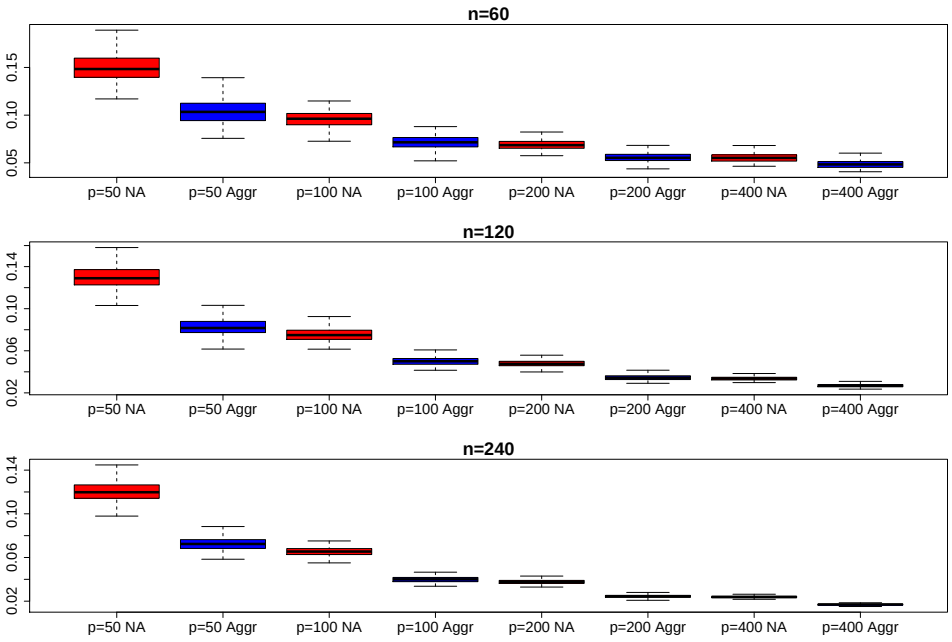

Figure 2: Boxplot of MSE($\widehat{\xi}$) (red) and MSE($\widetilde{\xi}$) (blue) in a simulation with 200 replications.

is computed as

$$\text{MSPE}(\widehat{y}^r) = \frac{1}{50n}\sum_{t=1}^{n}\sum_{\mathbf{s}_0\in\mathcal{S}^*}\{\widehat{y}_t^r(\mathbf{s}_0) - \xi_t(\mathbf{s}_0)\}^2,$$

where $\mathcal{S}^*$ is the set consisting of the 50 post-sample locations. Similarly, we repeat this exercise for $\widetilde{y}_t^r(\cdot)$ in (4.10). 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. Setting $j_0 = 3$, we compute both the one-step-ahead and two-step-ahead predictions at time n. The mean of square predictive error (MSPE) is calculated as follows.

$$\text{MSPE}(\widehat{\text{y}}_{\text{n}+\ell}^{\text{r}}) = \frac{1}{\text{p}}\sum_{\text{j}=1}^{\text{p}}\{\widehat{\text{y}}_{\text{n}+\ell}^{\text{r}}(\mathbf{s}_{\text{j}}) - \xi_{\text{n}+\ell}(\mathbf{s}_{\text{j}})\}^2, \qquad \ell = 1, 2.$$

We repeat the above exercise for the aggregation estimator $\tilde{y}_{n+\ell}^r$ with $J = 100$.

The means and standard errors of the MSPEs in the 200 replications for each settings are listed in Table 1. In general MSPE decreases as n increases. For the kriging over space, MSPE also decreases as p increases. See also Theorem 2, noting $\delta = 0$ when all the factors are strong (Lam and Yao, 2012). MSPEs of the kriging over space are smaller than those of the kriging in time. This is understandable from comparing Theorem 2 and Theorem 3. Last but not least, the

Table 1: Means and standard errors (in parentheses) of the mean squared predictive errors (MSPE)

		Kriging in Space		Kriging in Time			
n	p	MSPE($\widehat{y}_t^r$)	MSPE($\widetilde{y}_t^r$)	MSPE($\widehat{y}_{t+1}^r$)	MSPE($\widetilde{y}_{t+1}^r$)	MSPE($\widehat{y}_{t+2}^r$)	MSPE($\widetilde{y}_{t+2}^r$)
60	50	0.387(0.184)	0.369(0.177)	1.650(1.536)	1.555(1.489)	2.468(2.090)	2.371(2.046)
120	50	0.362(0.120)	0.345(0.120)	1.241(1.021)	1.168(1.027)	1.751(1.495)	1.692(1.500)
240	50	0.355(0.123)	0.337(0.123)	1.104(0.985)	1.061(0.970)	1.793(1.498)	1.755(1.472)
60	100	0.180(0.065)	0.174(0.065)	1.308(1.208)	1.249(1.217)	2.338(2.319)	2.280(2.317)
120	100	0.171(0.059)	0.163(0.058)	1.128(1.092)	1.090(1.096)	1.881(1.507)	1.857(1.516)
240	100	0.165(0.049)	0.156(0.048)	1.099(0.905)	1.068(0.898)	1.898(1.542)	1.878(1.533)
60	200	0.084(0.028)	0.081(0.028)	1.282(1.095)	1.244(1.082)	2.320(2.049)	2.265(2.036)
120	200	0.076(0.023)	0.072(0.022)	1.224(1.115)	1.207(1.119)	2.216(2.004)	2.206(2.019)
240	200	0.076(0.020)	0.072(0.020)	1.030(0.940)	1.016(0.941)	1.695(1.521)	1.688(1.517)
60	400	0.041(0.013)	0.040(0.013)	1.335(1.154)	1.314(1.151)	2.096(1.913)	2.071(1.911)
120	400	0.037(0.010)	0.036(0.010)	1.130(1.036)	1.117(1.033)	1.995(1.920)	1.983(1.913)
240	400	0.035(0.009)	0.033(0.009)	0.994(0.842)	0.985(0.842)	1.548(1.412)	1.542(1.410)

aggregated kriging methods always outperform their non-aggregate counterparts.

6.2 Real Data Analysis

We illustrate the proposed methods with a real data set which consists of the monthly temperature records (in Celsius) at the 176 monitoring stations in China in January 1970 – 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.

$$\text{MSPE}(\widehat{y}^r) = \frac{1}{50 \times 372} \sum_{t=1}^{372} \sum_{\mathbf{s}_0 \in \mathcal{S}^*} \left\{ \widehat{y}_t^r(\mathbf{s}_0) - y_t(\mathbf{s}_0) \right\}^2.$$

We also apply the aggregation (with $J = 200$) estimator $\widetilde{y}_t(\cdot)$ in (4.10) to improve the kriging accuracy. To avoid the sampling bias in site selection, we replicate this exercise 100 times via randomly selecting 126 sites for estimation each time. The mean and the standard errors over the 100 replications are 0.0463 and 0.2225 for MSPE($\widetilde{y}^r$), and 0.0461 and 0.2216 for MSPE($\widetilde{y}^r$). The gain from using the aggregation method $\widetilde{y}^r$ is not substantial for this example.

For kriging in time, we consider one-step-ahead and two-step-ahead post-sample prediction (with $j_0 = 12$) 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

$$\text{MSPE}(\widehat{y}^r_{n+\ell}) = \frac{1}{176}\sum_{j=1}^{176}\left\{\widehat{y}^r_{n+\ell}(\mathbf{s}_j) - y_{n+\ell}(\mathbf{s}_j)\right\}^2, \qquad \ell = 1, 2.$$

We also apply the aggregation estimator $\widetilde{y}^r_{n+\ell}(\cdot)$ with $J = 200$. The means and the standard errors of MSPE($\widehat{y}^r_{n+\ell}$) over the last 24 months are 1.9107 and 1.5090 for $\ell = 1$, and 2.0181 and 1.4892 for $\ell = 2$. The means and the standard errors of MSPE($\widetilde{y}^r_{n+\ell}$) are 1.9093 and 1.5083 for $\ell = 1$, and 2.0167 and 1.4892 for $\ell = 2$. As we expected, the one-step-ahead prediction is more accurate than the two-step-ahead prediction.

It is clear that the kriging in space is much 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.

7 Final remarks

The fundamental reason for not imposing any distributional assumptions in model (2.1) is the stationarity in time imposed on the underlying process. It enables us to learn the dependence across different locations in a complete nonparametric manner. In practice the data often show some trends or seasonal pattern in time. The existing detrend and deseasonality methods in time series analysis can be applied to make data stationarity.

The assumption that matrix $\mathbf{\Sigma}$ 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, those factors cannot be recovered by the method presented in Section 3.1. They will be left in the residuals $\widehat{\varepsilon}_t(\mathbf{s}_j) \equiv y_t(\mathbf{s}_j) - \widehat{\xi}_t(\mathbf{s}_j)$. We can recover those factors from the residuals using the factor modelling method for multiple time series of Lam and Yao (2012).

References

Banerjee, S., Gelfand, A., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society*, **B**, **70**, 825-848.

Bathia, N., Yao, Q. and Ziegelmann, F. (2010). Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, **38**, 3352-3386.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123-140.

Castruccio, S. and Stein, M. L. (2013). Global space-time models for climate ensembles. *Annals of Applied Statistics*, **7**, 1593-1611.

Chang, J., Guo, B. and Yao, Q. (2015). High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. *Journal of Econometrics*, **189**, 297-312.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society*, **B**, **70**, 209-226.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

Finley, A., Sang, H., Banerjee, S. and Gelfand, A. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Analysis*, **53**, 2873-2884.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis*, **83**, 493-508.

Guinness, J. and Stein, M. L. (2013). Interpolation of nonstationary high frequency spatial-temporal temperature data. *Annals of Applied Statistics*, **7**, 1684-1708.

Hall, P., Fisher, N. I., and Hoffmann, B. (1994). On the Nonparametric Estimation of Covariance Functions. *The Annals of Statistics*, **22**, 2115-2134.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* (eds C. W. Anderson, V. Barnett, P. C. Chatwin and A. H. El-Shaarawi), pp. 37-54. London: Springer.

Jun, M. and Stein, M. L. (2007). An approach to producing space-time covariance functions on spheres. *Technometrics*, **49**, 468-479.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Applied Statistics*, **52**, 1-18.

Katzfuss, M. and Cressie, N. (2011). Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *Journal of Time Series Analysis*, **32**, 430-446.

Kaufman, C., Schervish, M. and Nychka, D. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, **103**, 1545-1555.

Lam, C. and Yao, Q. (2012). Factor modelling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, **40**, 694-726

Lam, C., Yao, Q. and Bathia, N. (2011). Estimation for latent factors for high-dimensional time series. *Biometrika*, **98**, 901-918.

Li, B., Genton, M. G. and Sherman, M. (2007). A nonparametric assessment of properties of space-time covariance functions. *Journal of the American Statistical Association*, **102**, 736-744.

Lin, Z. and Lu, C. (1996). *Limit Theory on Mixing Dependent Random Variables*. Kluwer Academic Publishers, New York.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, **209**, 415-446.

Sang, H. and Huang J. Z. (2012). A full-scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society*, **B**, **74**, 111-132.

Smith, R. L., Kolenikov, S. and Cox, L. H. (2003). Spatiotemporal modelling of $PM_{2.5}$ data with missing values. *Journal of Geophysical Research*, **108**, No.D24, DOI:10.1029/2002JD002914.

Stein, M. (2008). A modeling approach for large spatial data sets. *Journal of the Korean Statistical Society*, **37**, 3-10.

Wikle, C. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, **86**, 815-829.

Zhang, B., Sang, H., Huang, J. Z. (2015). Full-scale approximations of spatio-temporal covariance models for large datasets. *Statistica Sinica*, **25**, 99-114.

Zhang, R., Robinson, P. and Yao, Q. (2015). Identifying cointegration by eigenanalysis. *Available at* `arXiv:1505.00821`.

Zhu, H., Fan, J. and Kong, L. (2014). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, **109**, 1084-1098.

Supplementary document of "Krigings over space and time based on latent low-dimensional structures"

Appendix: Technical proofs

Proof of Proposition 1. The first part of the proposition can be proved in the same manner as Proposition 1 of Bathia *et al.* (2010), which is omitted. To prove the second part, it follows (2.9) and (2.8) that any eigenfunction of Σ_0 must be the linear combination of $a_1, \cdots, a_d$, i.e. $\varphi_i(\mathbf{s}) = \sum_j \gamma_{ij} a_j(\mathbf{s})$. Now it follows from (2.11) and (2.8) that

$$\Sigma_0 \circ \varphi_i(\mathbf{s}) = \sum_{k,\ell,j} \sigma_{k\ell}\gamma_{ij} a_k(\mathbf{s})\langle a_\ell, a_j\rangle = \sum_{k,j} \sigma_{kj}\gamma_{ij} a_k(\mathbf{s}) = \sum_k \lambda_i \gamma_{ik} a_k(\mathbf{s}) = \lambda_i \varphi_i(\mathbf{s}).$$

Since $a_1, \cdots, a_d$ are orthonormal, it must hold that

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

As σ_{kj} is the (k,j)-th element of matrix $\mathrm{Var}(\mathbf{x}_t)$, (A.1) is equivalent to $\mathrm{Var}(\mathbf{x}_t)\boldsymbol{\gamma}_i = \lambda_i \boldsymbol{\gamma}_i$, i.e. $\boldsymbol{\gamma}_i$ is an eigenvector of $\mathrm{Var}(\mathbf{x}_t)$ corresponding to the eigenvalue λ_i, $i = 1, \cdots, d$. Furthermore,

$$I(i=k) = \langle \phi_i, \phi_k\rangle = \sum_{j,\ell} \gamma_{ij}\gamma_{k\ell}\langle a_j, a_\ell\rangle = \sum_j \gamma_{ij}\gamma_{kj} = \boldsymbol{\gamma}_i'\boldsymbol{\gamma}_k.$$

Thus $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_d$ are orthogonal. ■

To prove Theorem 1(ii), we first introduce Lemma 3 below. For the simplicity in presentation, we assume that the d positive eigenvalues of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}'$, defined in (3.5), are distinct from each other. Then both $\mathbf{A}_1$ and $\mathbf{A}_2$ are uniquely defined if we line up each of the two sets of the d orthonormal eigenvectors (i.e. the columns of $\mathbf{A}_1$ and $\mathbf{A}_2$) in the descending order of their corresponding eigenvalues, and we require that the first non-zero element of each those eigenvector to be positive. See the discussion below (3.5) above.

Using the same notation as in (3.7), we denote by $\widehat{\mathbf{A}}_1^{(j)}$, $\widehat{\mathbf{A}}_2^{(j)}$ the estimated factor loading matrices in (3.7) with the j-th partition, by $\boldsymbol{\Sigma}^{(j)}$ the covariance matrix in (3.4), and by $\mathbf{x}_t^{(j)}$, $\mathbf{x}_t^{*(j)}$ the estimated latent factors in (3.8), $j = 1, \cdots, p_0 = p!/(p_1!p_2!)$. Assume that the d positive eigenvalues of $\boldsymbol{\Sigma}^{(j)}(\boldsymbol{\Sigma}^{(j)})'$ are distinct. Then $\widehat{\mathbf{A}}_1^{(j)}$ and $\widehat{\mathbf{A}}_2^{(j)}$ can be uniquely defined as above. Now we are ready to state the lemma.

Lemma 3 *Let Condition 1 hold. Let the d positive eigenvalues of* $\mathbf{\Sigma}^{(j)}(\mathbf{\Sigma}^{(j)})'$ *be distinct, and Condition 2 hold for* $\mathbf{x}_t^{(j)}$ *and* $\mathbf{x}_t^{*(j)}$ *for all* $j = 1, \cdots, p_0$. *Then as* $p^\delta n^{-1/2} = o(1)$, *it holds that*

$$\max_{1\le j\le p_0}\{||\widehat{\mathbf{A}}_1^{(j)} - \mathbf{A}_1^{(j)}|| + ||\widehat{\mathbf{A}}_2^{(j)} - \mathbf{A}_2^{(j)}||\} = O_P(p^\delta n^{-1/2}).$$

Proof. Since the proof of $\max_{1\le j\le p_0}||\widehat{\mathbf{A}}_1^{(j)} - \mathbf{A}_1^{(j)}||$ and $\max_{1\le j\le p_0}||\widehat{\mathbf{A}}_2^{(j)} - \mathbf{A}_2^{(j)}||$ are similar. We only give the proof for $\max_{1\le j\le p_0}||\widehat{\mathbf{A}}_1^{(j)} - \mathbf{A}_1^{(j)}||$. Note that for any $1 \le j \le p_0$,

$$||\widehat{\mathbf{\Sigma}}^{(j)}(\widehat{\mathbf{\Sigma}}^{(j)})' - \mathbf{\Sigma}^{(j)}(\mathbf{\Sigma}^{(j)})'|| \le ||\widehat{\mathbf{\Sigma}}^{(j)} - \mathbf{\Sigma}^{(j)}||^2 + 2||\mathbf{\Sigma}^{(j)}|| \times ||\widehat{\mathbf{\Sigma}}^{(j)} - \mathbf{\Sigma}^{(j)}||. \tag{A.2}$$

Since $\mathbf{x}_t^{(j)}$ satisfies Condition 2, we have $||\mathbf{\Sigma}^{(j)}|| = O(p^{1-\delta})$, see Lam, Yao and Bathia (2011). On the other hand, by the mixing condition of $\{\mathbf{y}_t\}$, it follows that

$$\begin{aligned}
\sup_j||\widehat{\mathbf{\Sigma}}^{(j)} - \mathbf{\Sigma}^{(j)}||^2 &= \sup_j||\frac{1}{n}\sum_{t=1}^n\{(\mathbf{y}_{t,1}^{(j)} - \bar{\mathbf{y}}_1^{(j)})(\mathbf{y}_{t,2}^{(j)} - \bar{\mathbf{y}}_2^{(j)})' - \mathrm{Cov}(\mathbf{y}_{t,1}^{(j)}, \mathbf{y}_{t,2}^{(j)})\}||^2 \\
&\le \sum_{i=1}^p\sum_{j=1}^p\Big\{\frac{1}{n}\sum_{t=1}^n[y_t(\mathbf{s}_i) - \bar{y}(\mathbf{s}_i)][y_t(\mathbf{s}_j) - \bar{y}(\mathbf{s}_j)] - \mathrm{Cov}[y_t(\mathbf{s}_i), y_t(\mathbf{s}_j)]\Big\}^2 \\
&= O_p(p^2/n).
\end{aligned}$$

Thus, by (A.2), we have

$$\sup_j||\widehat{\mathbf{\Sigma}}^{(j)}(\widehat{\mathbf{\Sigma}}^{(j)})' - \mathbf{\Sigma}^{(j)}(\mathbf{\Sigma}^{(j)})'|| = O_p(p^{2-\delta}n^{-1/2}). \tag{A.3}$$

By (A.3), Lemma 3 can be shown similarly to Theorem 1 of Lam, Yao and Bathia (2011). We omit the details here. ■

Proof of Theorem 1(ii). Note that

$$\begin{aligned}
&\mathrm{E}\Big[\frac{1}{np}\sum_{t=1}^n\sum_{i=1}^p\{\widehat{\xi}_t(\mathbf{s}_i) - \xi_t(\mathbf{s}_i)\}^2\Big|\{\xi_t(\mathbf{s}_i),\, y_t(\mathbf{s}_i)\}\Big] \\
= &\frac{1}{npp_0}\sum_{t=1}^n\sum_{j=1}^{p_0}[\widehat{\mathbf{A}}_1^{(j)}\widehat{\mathbf{x}}_t^{(j)} - \mathbf{A}_1^{(j)}\mathbf{x}_t^{(j)}]'[\widehat{\mathbf{A}}_1^{(j)}\widehat{\mathbf{x}}_t^{(j)} - \mathbf{A}_1^{(j)}\mathbf{x}_t^{(j)}] \\
&+\frac{1}{npp_0}\sum_{t=1}^n\sum_{j=1}^{p_0}[\widehat{\mathbf{A}}_2^{(j)}\widehat{\mathbf{x}}_t^{*(j)} - \mathbf{A}_2^{(j)}\mathbf{x}_t^{*(j)}]'[\widehat{\mathbf{A}}_2^{(l)}\widehat{\mathbf{x}}_t^{*(j)} - \mathbf{A}_2^{(j)}\mathbf{x}_t^{*(j)}] \\
\equiv &\Sigma_1 + \Sigma_2.
\end{aligned}$$

By Lemma 3, we have

$$\begin{aligned}
\Sigma_1 &= \frac{1}{npp_0}\sum_{t=1}^{n}\sum_{j=1}^{p_0}\Big\{[\widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})' - \mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})']\mathbf{A}_1^{(j)}\mathbf{x}_t^{(j)} + \widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})'\varepsilon_{t,1}^{(j)}\Big\}' \\
&\quad \Big\{[\widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})' - \mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})']\mathbf{A}_1^{(j)}\mathbf{x}_t^{(j)} + \widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})'\varepsilon_{t,1}^{(j)}\Big\}' \\
&= \frac{1}{npp_0}\sum_{t=1}^{n}\sum_{j=1}^{p_0}(\mathbf{x}_t^{(j)})'(\widehat{\mathbf{A}}_1^{(j)})'[\widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})' - \mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})']'[\widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})' - \mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})']\mathbf{A}_1^{(j)}\mathbf{x}_t^{(j)} \\
&\quad + \frac{1}{npp_0}\sum_{t=1}^{n}\sum_{j=1}^{p_0}(\mathbf{x}_t^{(j)})'(\mathbf{A}_1^{(j)})'[\widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})' - \mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})']'\widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})'\varepsilon_{t,1}^{(j)} \\
&\quad + \frac{1}{npp_0}\sum_{t=1}^{n}\sum_{j=1}^{p_0}(\varepsilon_{t,1}^{(j)})'\widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})'[\widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})' - \mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})']\mathbf{A}_1^{(j)}\mathbf{x}_t^{(j)} \\
&\quad + \frac{1}{npp_0}\sum_{t=1}^{n}\sum_{j=1}^{p_0}(\varepsilon_{t,1}^{(j)})'\widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})'\widehat{\mathbf{A}}_1^{(j)}(\widehat{\mathbf{A}}_1^{(j)})'\varepsilon_{t,1}^{(j)} \\
&= O_p(p^{\delta}/n + p^{(\delta-1)/2}n^{-/2}) + \frac{1}{npp_0}\sum_{t=1}^{n}\sum_{j=1}^{p_0}(\varepsilon_{t,1}^{(j)})'\mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})'\varepsilon_{t,1}^{(j)}. \qquad \text{(A.4)}
\end{aligned}$$

Since $\mathrm{E}\Big(\frac{1}{p_0}\sum_{j=1}^{p_0}(\varepsilon_{t,1}^{(j)})'\mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})'\varepsilon_{t,1}^{(j)}\Big)^2 \le \frac{1}{p_0}\sum_{j=1}^{p_0}\mathrm{E}\Big[(\varepsilon_{t,1}^{(j)})'\mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})'\varepsilon_{t,1}^{(j)}\Big]^2 < \infty$, by Markov's inequality, it follows that

$$\frac{1}{np_0}\sum_{t=1}^{n}\sum_{j=1}^{p_0}(\varepsilon_{t,1}^{(j)})'\mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})'\varepsilon_{t,1}^{(j)} \xrightarrow{p} \mathrm{E}\Big[(\varepsilon_{t,1}^{(j)})'\mathbf{A}_1^{(j)}(\mathbf{A}_1^{(j)})'\varepsilon_{t,1}^{(j)}\Big].$$

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

(i) When $n \to \infty$, $\Sigma_1 = O_p(p^{\delta}/n + p^{(\delta-1)/2}n^{-/2} + p^{-1})$.

(ii) When $p^{1+\delta}/n \to 0$, $p\Sigma_1 \xrightarrow{p} \mathrm{E}\Big[(\varepsilon_{t,1}^{(1)})'\mathbf{A}_1^{(1)}(\mathbf{A}_1^{(1)})'\varepsilon_{t,1}^{(1)}\Big]$.

Similarly, the above properties hold also for Σ_2. Thus,

$$\mathrm{E}\Big[\frac{1}{np}\sum_{t=1}^{n}\sum_{i=1}^{p}\big\{\widehat{\xi}_t(\mathbf{s}_i) - \xi_t(\mathbf{s}_i)\big\}^2\Big|\{\xi_t(\mathbf{s}_i),\, y_t(\mathbf{s}_i)\}\Big] = O_p(p^{\delta}/n + p^{(\delta-1)/2}n^{-/2} + p^{-1}). \qquad \text{(A.5)}$$

Further, when $p^{1+\delta}/n \to 0$,

$$\mathrm{E}\Big[\frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{p}\big\{\widehat{\xi}_t(\mathbf{s}_i) - \xi_t(\mathbf{s}_i)\big\}^2\Big|\{\xi_t(\mathbf{s}_i),\, y_t(\mathbf{s}_i)\}\Big] = \mathrm{E}\Big[(\varepsilon_{t,1}^{(1)})'\mathbf{A}_1^{(1)}(\mathbf{A}_1^{(1)})'\varepsilon_{t,1}^{(1)} + (\varepsilon_{t,2}^{(1)})'\mathbf{A}_2^{(1)}(\mathbf{A}_2^{(1)})'\varepsilon_{t,2}^{(1)}\Big]$$

in probability. This completes the proof of Theorem 1(ii). ■

Lemma 4 *Let Condition 1 hold and* $pn^{-\beta/2} \to 0$. *Then*

$$\lim_{n\to\infty} P\left\{\min_{1\le i\le p} \lambda_{\min}[n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i)] \ge c_0/2\right\} = 1.$$

Proof. Let $z_t^j(\mathbf{s}_i)$, $j = 1, \cdots, m$ be the components of $\mathbf{z}_t(\mathbf{s}_i)$. Since $\{\mathbf{z}_t(\mathbf{s}_i)\}$ is a stationary α-mixing process satisfying Condition 1, it follows from Lemma 12.2.2 of Lin and Lu (1996) that for any $1 \le j, k \le m$,

$$\mathrm{E}\left|\frac{1}{n}\sum_{t=1}^{n}\{z_t^j(\mathbf{s}_i)z_t^k(\mathbf{s}_i) - \mathrm{E}[z_t^j(\mathbf{s}_i)z_t^k(\mathbf{s}_i)]\}\right|^{\beta} = O(n^{-\beta/2}). \tag{A.6}$$

Since m is finite, it follows that

$$\mathrm{E}||n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i) - \mathrm{Var}(\mathbf{z}_t(\mathbf{s}_i))||_F^{\beta} = O(n^{-\beta/2}), \tag{A.7}$$

where $||\cdot||_F$ denotes the Frobenius norm. Now suppose that $\min_{1\le i\le p} \lambda_{\min}[n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i)] < c_0/2$. Since $\min_{1\le i\le p} \lambda_{\min}[\mathrm{Var}(\mathbf{z}_t(\mathbf{s}_i))] > c_0$, and

$$\mathrm{Var}(\mathbf{z}_t(\mathbf{s}_i)) = [\mathrm{Var}(\mathbf{z}_t(\mathbf{s}_i)) - n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i)] + n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i),$$

it must hold that

$$\max_{1\le i\le p} ||n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i) - \mathrm{Var}(\mathbf{z}_t(\mathbf{s}_i))||_F \ge c_0/2. \tag{A.8}$$

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

$$\begin{aligned} P\{\max_{1\le i\le p} ||n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i) - \mathrm{Var}(\mathbf{z}_t(\mathbf{s}_i))||_F \ge c_0/2\} &\le \sum_{i=1}^{p}(c_0/2)^{-\beta}\mathrm{E}||n^{-1}\mathbf{z}(\mathbf{s}_i)'\mathbf{z}(\mathbf{s}_i) - \mathrm{Var}(\mathbf{z}(\mathbf{s}_i))||_F^{\beta} \\ &= O_p(pn^{-\beta/2}) = o(1). \end{aligned} \tag{A.9}$$

This implies that $P\{\min_{1\le i\le p} \lambda_{\min}[n^{-1}\mathbf{z}(\mathbf{s}_i)'\mathbf{z}(\mathbf{s}_i)] < c_0/2\} = o(1)$. Thus, we have proved Lemma 4. ■

Proof for the convergence rate of $\widehat{\boldsymbol{\beta}}(\mathbf{s}_0)$. Let $e_t(\mathbf{s}) = y_t(\mathbf{s}) - \mathbf{z}_t(\mathbf{s})'\boldsymbol{\beta}(\mathbf{s})$ and $w_i = K_h(\mathbf{s}_i - \mathbf{s}_0)/\sum_{i=1}^{p} K_h(\mathbf{s}_i - \mathbf{s}_0)$. Then $\mathbf{e}(\mathbf{s}) = (e_1(\mathbf{s}), \cdots, e_n(\mathbf{s}))'$ and

$$\widehat{\boldsymbol{\beta}}(\mathbf{s}_0) = \sum_{i=1}^{p}[\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i)]^{-1}[\mathbf{Z}(\mathbf{s}_i)'\mathbf{e}(\mathbf{s}_i)]w_i + \sum_{i=1}^{p}\boldsymbol{\beta}(\mathbf{s}_i)w_i \equiv I_1 + I_2.$$

For any twice differentiable function $g(\mathbf{s}) = g(s_1, s_2)$, $\mathbf{s} = (s_1, s_2) \in \mathcal{R}^2$, define $g_{1\cdot}(\mathbf{s}) = \partial g(\mathbf{s})/\partial s_1$, $g_{\cdot 2}(\mathbf{s}) = \partial g(\mathbf{s})/\partial s_2$, $g_{11}(\mathbf{s}) = \partial^2 g(\mathbf{s})/(\partial s_1)^2$ and $g_{22}(\mathbf{s}) = \partial^2 g(\mathbf{s})/(\partial s_2)^2$. Under Conditions 3, 4 and Taylor's expansion, it can be shown that as $p \to \infty$,

$$\begin{aligned} I_2 - \boldsymbol{\beta}(\mathbf{s}_0) &= \sum_{i=1}^{p} (\boldsymbol{\beta}(\mathbf{s}_i) - \boldsymbol{\beta}(\mathbf{s}_0)) w_i \\ &= \frac{h^2}{f(\mathbf{s}_0)} [\boldsymbol{\beta}_{1\cdot}(\mathbf{s}_0) f_{1\cdot}(\mathbf{s}_0) + \frac{1}{2} f(\mathbf{s}_0) \boldsymbol{\beta}_{11}(\mathbf{s}_0)] \int_R \int_R x^2 K(x,y)\, dxdy \\ &\quad + \frac{h^2}{f(\mathbf{s}_0)} [\boldsymbol{\beta}_{\cdot 2}(\mathbf{s}_0) f_{\cdot 2}(\mathbf{s}_0) + \frac{1}{2} f(\mathbf{s}_0) \boldsymbol{\beta}_{22}(\mathbf{s}_0)] \int_R \int_R y^2 K(x,y)\, dxdy + o(h^2). \quad \text{(A.10)} \end{aligned}$$

As for I_1, by Hölder's inequality, it follows that

$$||I_1|| \leq \Big(\max_{1 \leq i \leq p} \|[\mathbf{Z}(\mathbf{s}_i)' \mathbf{Z}(\mathbf{s}_i)/n]^{-1}\|_F \Big) \sum_{i=1}^{p} \|\mathbf{Z}(\mathbf{s}_i)' \mathbf{e}(\mathbf{s}_i)/n\| w_i. \quad \text{(A.11)}$$

By Lemma 4, we have $\max_{1 \leq i \leq p} \lambda_{max}\{(n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i))^{-1}\} \leq 2/c_0$ holds in probability. Since the dimension of $\mathbf{z}_t(\mathbf{s})$ is fixed, it follows that

$$\max_{1 \leq i \leq p} ||(n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{Z}(\mathbf{s}_i))^{-1}||_F \leq c_1 \quad \text{(A.12)}$$

holds in probability for some positive constant c_1. On the other hand, it is easy to get that

$$\max_{1 \leq i \leq p} \mathrm{E}||n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{e}(\mathbf{s}_i)|| = O(n^{-1/2}),$$

hence,

$$\mathrm{E}\Big[\sum_{i=1}^{p} ||n^{-1}\mathbf{Z}(\mathbf{s}_i)'\mathbf{e}(\mathbf{s}_i)|| w_i \Big] = O(n^{-1/2}). \quad \text{(A.13)}$$

It follows from (A.11), (A.12) and (A.13) that

$$||I_1|| = O_p(n^{-1/2}). \quad \text{(A.14)}$$

Thus, by (A.10) and (A.13), we have $|\widehat{\boldsymbol{\beta}}(\mathbf{s}_0) - \boldsymbol{\beta}(\mathbf{s}_0)| = O_p(h^2 + n^{-1/2})$. ■

Proof of Theorem 2. Let $\mathbf{x}_t^o = \mathbf{x}_t I(\mathbf{s}_i \in \mathcal{S}_1) + \mathbf{x}_t^* I(\mathbf{s}_i \in \mathcal{S}_2)$. Then

$$\begin{aligned} \widehat{\xi}_t(\mathbf{s}_0) - \xi_t(\mathbf{s}_0) &= \sum_{i=1}^{p} (\widehat{\mathbf{a}}'(\mathbf{s}_i)\widehat{\mathbf{x}}_t^o - \mathbf{a}'(\mathbf{s}_i)\mathbf{x}_t^o) w_i + \sum_{j=1}^{d} \sum_{i=1}^{p} (a_j(\mathbf{s}_i) - a_j(\mathbf{s}_0)) x_{tj}^o w_i \\ &\equiv J_1 + J_2. \quad \text{(A.15)} \end{aligned}$$

Similar to (A.10), we have

$$\left|\sum_{i=1}^{p}[(a_j(\mathbf{s}_i)-a_j(\mathbf{s}_0))/||a(\mathbf{s}_0)||]w_i\right| = O(h^2),$$

which implies that

$$|J_2| = O(h^2)(||\mathbf{a}(\mathbf{s}_0)||)\sum_{j=1}^{d}|x_{tj}^o| = O(dh^2\cdot||\mathbf{a}(\mathbf{s}_0)||\cdot||\mathbf{x}_t^o||) = O_p(h^2), \tag{A.16}$$

where we use the fact that $||\mathbf{x}_t^o|| = O_p(p^{(1-\delta)/2})$ and $||\mathbf{a}(\mathbf{s}_0)|| = O(p^{(\delta-1)/2})$, which is followed by $\lambda_{\min}\{\mathrm{E}(\mathbf{x}_t^o(\mathbf{x}_t^o)')\} \asymp p^{1-\delta}$ and

$$(||\mathbf{a}(\mathbf{s}_0)||^2)\lambda_{\min}\{\mathrm{E}(\mathbf{x}_t^o(\mathbf{x}_t^o)')\} \le (||\mathbf{a}(\mathbf{s}_0)||^2)\mathrm{E}[(\mathbf{a}'(\mathbf{s}_0)/||\mathbf{a}(\mathbf{s}_0)||)\mathbf{x}_t^o]^2 = \mathrm{E}[\mathbf{a}'(\mathbf{s}_0)\mathbf{x}_t^o]^2 = \mathrm{E}y_t^2(\mathbf{s}_0) < \infty.$$

By (iii) of Proposition 3 and the same arguments as in Theorem 2.2 of Chang, Guo and Yao (2015), we have for $i = 1, 2$,

$$p^{-1/2}||\widehat{\mathbf{A}}_i\widehat{\mathbf{x}}_t^o - \mathbf{A}_i\mathbf{x}_t^o|| = O_p(||\widehat{\mathbf{A}}_i - \mathbf{A}_i|| + n^{-1/2} + p^{-1/2}) = O_p(n^{-1/2}p^{\delta} + p^{-1/2}). \tag{A.17}$$

It follows from Hölder inequality and (A.17) that

$$\begin{aligned} J_1 &\le \Big\{\sum_{i=1}^{p}[(\widehat{\mathbf{a}}'(\mathbf{s}_i)\widehat{\mathbf{x}}_t^o - \mathbf{a}'(\mathbf{s}_i)\mathbf{x}_t^o)]^2\Big\}^{1/2}\Big(\sum_{i=1}^{p}w_i^2\Big)^{1/2} \\ &\le (\sum_{i=1}^{2}||\widehat{\mathbf{A}}_i\widehat{\mathbf{x}}_t^o - \mathbf{A}_i\mathbf{x}_t^o||)\Big(\sum_{i=1}^{p}w_i^2\Big)^{1/2} \\ &= O_p\{p^{1/2}(n^{-1/2}p^{\delta}+p^{-1/2})\}O((ph)^{-1/2}) = O_p\{p^{\delta}(nh)^{-1/2} + (ph)^{-1/2}\}. \end{aligned} \tag{A.18}$$

Thus, (ii) follows from (A.16) and (A.18). Similarly, we can show that (A.18) holds also for $\widetilde{\boldsymbol{\xi}}_t(\mathbf{s}_0)$.

■

Proof of Theorem 3. For simplicity, we only show the case with spatial points over $\mathcal{S}_1$, i.e., $\mathbf{y}_{t1} = \mathbf{A}_1\mathbf{x}_t + \boldsymbol{\varepsilon}_{t,1}$. For points over $\mathcal{S}_2$ can be shown similarly. Let $\widehat{\boldsymbol{\Sigma}}_{\varepsilon}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(\boldsymbol{\varepsilon}_{t+k,1} - \bar{\boldsymbol{\varepsilon}}_1)(\boldsymbol{\varepsilon}_{t,1} - \bar{\boldsymbol{\varepsilon}}_1)'$, $\widehat{\boldsymbol{\Sigma}}_{x\varepsilon}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(\mathbf{x}_{t+k} - \bar{\mathbf{x}})(\boldsymbol{\varepsilon}_{t,1} - \bar{\boldsymbol{\varepsilon}}_1)'$, $\widehat{\boldsymbol{\Sigma}}_{\varepsilon x}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(\boldsymbol{\varepsilon}_{t+k,1} - \bar{\boldsymbol{\varepsilon}}_1)(\mathbf{x}_t - \bar{\mathbf{x}})'$ and $\widehat{\boldsymbol{\Sigma}}_{xx}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(\mathbf{x}_{t+k} - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})'$. It follows that for any k,

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_x(k) - \boldsymbol{\Sigma}_x(k) &= (\widehat{\mathbf{A}}_1' - \mathbf{A}_1')\mathbf{A}_1\widehat{\boldsymbol{\Sigma}}_{xx}(k)\mathbf{A}_1'\widehat{\mathbf{A}}_1 + \widehat{\boldsymbol{\Sigma}}_{xx}(k)\mathbf{A}_1'(\widehat{\mathbf{A}}_1 - \mathbf{A}_1) + (\widehat{\boldsymbol{\Sigma}}_{xx}(k) - \boldsymbol{\Sigma}_x(k)) \\ &\quad + \widehat{\mathbf{A}}_1'\widehat{\boldsymbol{\Sigma}}_{\varepsilon}(k)\widehat{\mathbf{A}}_1 + \widehat{\mathbf{A}}_1'\mathbf{A}_1\widehat{\boldsymbol{\Sigma}}_{x\varepsilon}(k)\widehat{\mathbf{A}}_1 + \widehat{\mathbf{A}}_1'\widehat{\boldsymbol{\Sigma}}_{\varepsilon x}(k)\mathbf{A}_1'\widehat{\mathbf{A}}_1 \\ &\equiv \sum_{j=1}^{6}L_j. \end{aligned}$$

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

$$||L_1|| + ||L_2|| = O(||\widehat{\mathbf{A}}_1 - \mathbf{A}_1|| \cdot ||\widehat{\boldsymbol{\Sigma}}_{xx}(k)||) = O_p(n^{-1/2}p^{\delta}p^{1-\delta}) = O_p(n^{-1/2}p). \tag{A.19}$$

By (A.1) of Lam and Yao (2012), we have

$$||L_3|| \leq ||\widehat{\boldsymbol{\Sigma}}_{xx}(k) - \boldsymbol{\Sigma}_x(k)||_F = O(d||\widehat{\boldsymbol{\Sigma}}_{xx}(k) - \boldsymbol{\Sigma}_x(k)||) = O(p^{1-\delta}n^{-1/2}). \tag{A.20}$$

It is easy to get that

$$||\widehat{\boldsymbol{\Sigma}}_{x\varepsilon}(k)|| = O_p(p^{1-\delta/2}n^{-1/2}) = ||\widehat{\boldsymbol{\Sigma}}_{\varepsilon x}(k)||, \text{ and } ||\widehat{\boldsymbol{\Sigma}}_{\varepsilon}(k)|| = O_p(pn^{-1/2}),$$

see for example Lemma 2 of Lam, Yao and Bathia (2011). Thus,

$$||L_4|| + ||L_5|| + ||L_6|| = O_p(pn^{-1/2}). \tag{A.21}$$

Combining (A.19), (A.20) and (A.21) yield that for any $0 \leq k \leq j_0$,

$$||\widehat{\boldsymbol{\Sigma}}_x(k) - \boldsymbol{\Sigma}_x(k)|| = O_p(pn^{-1/2}). \tag{A.22}$$

Thus, by $p^{\delta}n^{-1/2} = o(1)$, we get $pn^{-1/2} = o(p^{1-\delta})$ and in probability,

$$||\widehat{\boldsymbol{\Sigma}}_x(k)||_{\min} \asymp ||\boldsymbol{\Sigma}_x(k)||_{\min} \asymp p^{1-\delta} \asymp ||\boldsymbol{\Sigma}_x(k)|| \asymp ||\widehat{\boldsymbol{\Sigma}}_x(k)||. \tag{A.23}$$

Since j_0 is fixed, from (A.22) it follows that

$$||\widehat{\mathbf{R}}_{j_0} - \mathbf{R}_{j_0}|| \asymp ||\mathbf{W}_{j_0} - \widehat{\mathbf{W}}_{j_0}|| \asymp O_p(pn^{-1/2}) \tag{A.24}$$

and from (A.23) it follows that

$$\|\mathbf{R}_{j_0}\| = O(p^{1-\delta}) \text{ and } ||\widehat{\mathbf{W}}_{j_0}^{-1}|| \asymp ||\mathbf{W}_{j_0}^{-1}|| \asymp O_p(p^{\delta-1}). \tag{A.25}$$

Since

$$||\widehat{\mathbf{x}}_t - \mathbf{x}_t|| = ||(\widehat{\mathbf{A}}_1 - \mathbf{A}_1)'\mathbf{A}\mathbf{x}_t + (\widehat{\mathbf{A}}_1 - \mathbf{A}_1)'\boldsymbol{\varepsilon}_{t,1} + \mathbf{A}_1'\boldsymbol{\varepsilon}_{t,1}|| = O_p(p^{(1+\delta)/2}n^{-1/2} + 1) = O_p(p^{(1-\delta)/2}),$$

it follows that $\|\widehat{\mathbf{X}}\widehat{\mathbf{X}}'\| = O_p(p^{1-\delta})$. Note that

$$\begin{aligned}
\widehat{\mathbf{x}}_{n+j}^r - \mathbf{x}_n(j) &= \widehat{\mathbf{R}}_{j_0}\widehat{\mathbf{W}}_{j_0}^{-1}\widehat{\mathbf{X}} - \mathbf{R}_{j_0}\mathbf{W}_{j_0}^{-1}\mathbf{X} \\
&= (\widehat{\mathbf{R}}_{j_0} - \mathbf{R}_{j_0})\widehat{\mathbf{W}}_{j_0}^{-1}\widehat{\mathbf{X}} + \mathbf{R}_{j_0}\widehat{\mathbf{W}}_{j_0}^{-1}(\mathbf{W}_{j_0} - \widehat{\mathbf{W}}_{j_0})\mathbf{W}_{j_0}^{-1}\widehat{\mathbf{X}} + \mathbf{R}_{j_0}\mathbf{W}_{j_0}^{-1}(\widehat{\mathbf{X}} - \mathbf{X}).
\end{aligned}$$

By (A.24) and (A.25), we have

$$||(\widehat{\mathbf{R}}_{j_0} - \mathbf{R}_{j_0})\widehat{\mathbf{W}}_{j_0}^{-1}\widehat{\mathbf{X}}||^2 \;=\; O(||\widehat{\mathbf{R}}_{j_0} - \mathbf{R}_{j_0}||^2 \cdot ||\widehat{\mathbf{W}}_{j_0}^{-1}||^2 \cdot ||\widehat{\mathbf{X}}\widehat{\mathbf{X}}'||) = O_p(p^{1+\delta}n^{-1}). \quad \text{(A.26)}$$

Similarly,

$$\begin{aligned} ||\mathbf{R}_{j_0}\widehat{\mathbf{W}}_{j_0}^{-1}(\mathbf{W}_{j_0} - \widehat{\mathbf{W}}_{j_0})\mathbf{W}_{j_0}^{-1}\widehat{\mathbf{X}}||^2 &= O(||\mathbf{R}_{j_0}||^2 \cdot ||\widehat{\mathbf{W}}_{j_0}^{-1}||^2 \cdot ||\mathbf{W}_{j_0} - \widehat{\mathbf{W}}_{j_0}||^2 \cdot ||\mathbf{W}_{j_0}^{-1}||^2 \cdot \|\widehat{\mathbf{X}}\widehat{\mathbf{X}}'\|) \\ &= O_p(p^{1+\delta}n^{-1}). \end{aligned} \quad \text{(A.27)}$$

On the other hand, by (A.25) and $||\widehat{\mathbf{x}}_t - \mathbf{x}_t|| = O_p(p^{(1+\delta)/2}n^{-1/2} + 1)$, we have

$$||\mathbf{R}_{j_0}\mathbf{W}_{j_0}^{-1}(\widehat{\mathbf{X}} - \mathbf{X})|| = O_p(p^{(1+\delta)/2}n^{-1/2} + 1). \quad \text{(A.28)}$$

Thus,

$$||\widehat{\mathbf{x}}_{n+j}^r - \mathbf{x}_n(j)|| = O_p(p^{(1+\delta)/2}n^{-1/2} + 1)$$

hold and (a) of Theorem 3 is proved.

As for Conclusion (b), by Conclusion (a) and (iii) of Proposition 3, we have

$$\begin{aligned} ||\widehat{\mathbf{y}}_{n+j}^r - \mathbf{y}_n(j)|| &= ||\widehat{\mathbf{A}}\widehat{\mathbf{x}}_{n+j}^r - \mathbf{A}\mathbf{x}_n(j)|| \\ &\le ||(\widehat{\mathbf{A}} - \mathbf{A})\mathbf{x}_n(j)|| + ||\widehat{\mathbf{A}}(\widehat{\mathbf{x}}_{n+j}^r - \mathbf{x}_n(j))|| \\ &= O_p\{p^{\delta}n^{-1/2}p^{1/2-\delta/2} + p^{(1+\delta)/2}n^{-1/2} + 1\} = O_p(p^{(1+\delta)/2}n^{-1/2} + 1). \end{aligned}$$

This gives (b) as desired and completes the proof of Theorem 3. ■