Modeling Multivariate Spatial-Temporal Data with Latent Low-Dimensional Dynamics

Elynn Y. Chen [*1], Xin Yun [† 2], Rong Chen [‡3], and Qiwei Yao [§4]

[1]University of California, Berkeley

[2]School of Data Science, Fudan University

[3]Rutgers University

[4]London School of Economics and Political Science

January 31, 2020

Abstract

High-dimensional multivariate spatial-temporal data arise frequently in a wide range of applications; however, there are relatively few statistical methods that can simultaneously deal with spatial, temporal and variable-wise dependencies in large data sets. In this paper, we propose a new approach to utilize the correlations in variable, space and time to achieve dimension reduction and to facilitate spatial/temporal predictions in the high-dimensional settings. The multivariate spatial-temporal process is represented as a linear transformation of a lower-dimensional latent factor process. The spatial dependence structure of the factor process is further represented non-parametrically in terms of latent empirical orthogonal functions. The

[*]Supported in part by NSF Grants DMS-1803241.

[†]Equal contribution.

[‡]Supported in part by NSF Grants DMS-1503409, DMS-1737857, DMS-1803241 and IIS-1741390.

[§]Corresponding author. Email: Q.Yao@lse.ac.uk

low-dimensional structure is completely unknown in our setting and is learned entirely from data collected irregularly over space but regularly over time. We propose innovative estimation and prediction methods based on the latent low-rank structures. Asymptotic properties of the estimators and predictors are established. Extensive experiments on synthetic and real data sets show that, while the dimensions are reduced significantly, the spatial, temporal and variable-wise covariance structures are largely preserved. The efficacy of our method is further confirmed by the prediction performances on both synthetic and real data sets.

Keywords: High-dimensional data; Multivariate spatial temporal process; Factor analysis; Latent empirical orthogonal function.

1 Introduction

The increasing availability of multivariate data collected over geographic regions and time in various applications has created unique opportunities and challenges for practitioners seeking to capitalize on its full utility. For example, United States Environmental Protection Agency publishes daily, from more than 20,000 monitoring stations, a collection of environmental and meteorological measurements such as temperature, pressure, wind speed/direction and levels of various pollutants. Such data naturally constitute a tensor (multi-dimensional array) with three modes (dimensions) representing n spacial locations, T time points and p variables, respectively. Since physical processes rarely occur in isolation but rather influence and interact with one another, simultaneously modeling the dependencies among different variables, regions, and time points is of great potential to reduce dimensions, produce more accurate estimation/prediction and further provide a deeper understanding of real world phenomena. At the same time, methodological issues arise because these data exhibit complex multivariate spatial-temporal covariances that may involve potential dependencies between spatial locations, time points

and different processes.

Traditionally, researchers mainly restrict their analysis to only two dimensions while fixing the third: time series analysis applied to a slice of such data at one location focuses on temporal modeling and prediction (Tsay and Chen, 2018; Box et al., 2015; Tsay, 2014; Brockwell and Davis, 2013; Fan and Yao, 2005); spatial statistical models for a slice of such data at one time point address spatial dependence and prediction over unobserved locations (Cressie, 2015); and uni-variate spatial-temporal statistics concentrate on only one variable observed over space and time (Huang and Cressie, 1996; Cressie and Wikle, 2015; Lopes et al., 2008; Cressie and Johannesson, 2008).

In this paper, we propose a new class of multivariate spatial-temporal models that characterize spatial, temporal and variable dependence simultaneously. This is made possible by an innovative combination of multivariate factor models (Fan et al., 2018; Chang et al., 2015; Lam and Yao, 2012; Lam et al., 2011; Bai, 2003; Bai and Ng, 2002) and the method of latent empirical orthogonal functions (Monahan et al., 2009; Hannachi et al., 2007; Von Storch and Zwiers, 2001; Wilks, 1995). Specifically, the p-dimensional spatial-temporal process is represented as a linear combination of a r-dimensional latent common factor process ($r \ll p$), which captures the correlations among p variables. The factor spatial-temporal processes are further represented in terms of latent empirical orthogonal functions (EOFs), which captures the spatial dependencies. As we shall see later, the EOFs in our setting have a close relationship with the loading matrix in factor analysis. We refer to the EOFs in our setting as the spatial loading functions. The coefficients of spatial loading functions are time-varying random variables and thus capture the temporal dependence. We provide a detailed analysis of the covariance structure of the proposed model across variables, space and time in Section 2.1. It shows that the proposed model is a generalization of several low-rank models in the literature (Higdon, 2002; Wikle and Cressie, 1999; Kammann and Wand, 2003; Cressie et al., 2010; Banerjee et al., 2008; Finley et al., 2009; Tzeng and Huang, 2018). In addition, the low-dimensional

structure and the the spatial loading functions are completely unknown in our setting and is learned entirely from data collected irregularly over space but regularly over time.

The estimation builds upon the idea in Wang et al. (2019) and further incorporates non-parametric estimation for the spatial loading functions. Particularly, we assembled the observations from n discrete spatial locations as a time series of $n \times p$ matrices whose rows and columns correspond to n sampling sites and p variables, respectively. As a result, the model on the discrete sampling locations can be reformulated in a similar form as the matrix factor model and it is estimated with a variant procedure based on the whiteness of spacial nugget effects. We also proposed prediction method for new locations and time points. Thanks to the innovative combination of reduced-rank models of two aspects, our method is able to efficiently handle multivariate spatial-temporal data sets with large n (space), p (variable) and T (time points).

1.1 Related works

To overcome the computational burden with large spatial or spatial-temporal data sets, researchers have developed reduced-rank approximations for univariate processes. Higdon (2002) uses kernel convolution, Wikle and Cressie (1999); Kammann and Wand (2003); Cressie and Johannesson (2008) successfully reduces the computational cost of kriging by using a flexible family of non-stationary covariance functions constructed from low rank basis functions. Banerjee et al. (2008) and Finley et al. (2009) uses predictive process, and Tzeng and Huang (2018) uses thin-plate splines. See also reviews of low-rank representations for spatial processes in Wikle (2010); Cressie (2015); Cressie and Wikle (2015). Our method applies to multivariate processes and incorporates two aspects of dimension reductions. The first aspect is the variable-wise dimension reduction where the observed p-dimensional process is represented as a linear combination of r-dimensional latent factor process. Further, the latent factor process assumes a reduced-rank representation whose formulation is similar to the aforementioned reduced rank

approximation methods. However, the spatial loading functions is completely unknown. Moreover, we don't impose any distributional assumptions on the underlying process, nor any parametric forms on its covariance function.

For multivariate spatial data, Cook et al. (1994) introduced the concept of a spatially shifted factor and a single-factor shifted-lag model and Majure and Cressie (1997) discussed graphical methods for identifying shifts. Following the ideas of multiple-lag dynamic factor models that generalize static factor models in the time series setting, Christensen and Amemiya (2001, 2002, 2003) extended the shifted-lag model to a generalized shifted-factor model by adding multiple shifted-lags and developed a systematic statistical estimation, inference, and prediction procedure. However, they do not include the time dimension and their method is an analogy of the multiple-lag dynamic factor models applied in the spatial setting. Thus, their definition of factors is very different from ours. Moreover, the assumption that spatial processes are second-order stationary is required for the moment-based estimation procedure and the theoretical development.

Various multivariate spatial-temporal conditional auto-regressive models have also been proposed by Carlin et al. (2003); Congdon (2004); Pettitt et al. (2002); Zhu et al. (2005); Daniels et al. (2006); Tzala and Best (2008), among others. Most of these papers, however, focus on empirical applications and do not offer any theoretical guarantees. Also, their estimation methods necessitate assumptions on the distribution of the observations. Bradley et al. (2015) introduced a multivariate spatial-temporal mixed effects model to analyze high-dimensional multivariate data sets that vary over different geographic regions and time points. They adopt a reduced rank spatial structure (Wikle, 2010) and model temporal behavior via vector auto-regressive components. Their method only applies to low-dimensional multivariate observations because they model each variable separately. The cross-dependence structures of multiple processes are modeled jointly by Genton and Kleiber (2015); Bourotte et al. (2016). These approaches impose separability and various independence assumptions, which are not appropriate for

many settings, as these models fails to capture important interactions and dependencies between different variables, regions, and times (Stein, 2005). In addition, they assume the random effect term is common across all processes which is unrealistic especially in the case with a large number of variables. Our method can effectively deal with data sets with large n, p, and T by simultaneously modeling the variable-wise and spatial low-rankness. Besides, our modeling of the spatial dependence though latent factor processes is different from the aforementioned methods in that we impose no assumptions about the stationarity over space, nor the distribution of data, nor any restrictive form of spatial covariance functions.

1.2 Contribution

We propose a new class of models for large-scale multivariate spatial-temporal processes. The model characterizes spatial, temporal and variable-wise dependencies simultaneously. The spatial dimension n, the variable dimension p and the time dimension T can be very large at the same time. To our best knowledge, our model is the first to deal with spatial, temporal and variable-wise covariance simultaneously, while allowing large n, p and T. It provides a flexible and rich cross-covariance structure for these dimensions simultaneously.

We develop efficient estimation and prediction procedures and establish theoretical properties of the estimators and predictors. The estimation procedure is based on a novel reformulation of the discrete observations of the p-dimensional spatial-temporal process. We believe this formulation is quite general and flexible to be extended to enable more sophisticated analysis along space, time or variable dimensions.

1.3 Notation and Organization

When $\boldsymbol{A}$ is a square matrix, we denote by $tr(\boldsymbol{A})$, $\lambda_{max}(\boldsymbol{A})$ and $\lambda_{max}(\boldsymbol{A})$ the trace, maximum and minimum eigenvalues of the matrix $\boldsymbol{A}$, respectively. We use $\|\boldsymbol{A}\|_2 \triangleq \sqrt{\lambda_{max}(\boldsymbol{A}'\boldsymbol{A})}$

and $\|\boldsymbol{A}\|_F \triangleq \sqrt{tr(\boldsymbol{A}'\boldsymbol{A})}$ to denote the spectral and Frobenius norms of the matrix $\boldsymbol{A}$, respectively. $\|\boldsymbol{A}\|_{min}$ denotes the positive square root of the minimal eigenvalue of $\boldsymbol{A}'\boldsymbol{A}$ or $\boldsymbol{A}\boldsymbol{A}'$, whichever is a smaller matrix. For two sequences a_N and b_N, we write $a_N \asymp b_N$ if $a_N = O(b_N)$ and $b_N = O(a_N)$.

The remainder of the article is outlined as follows. Section 2 introduces the model settings. Section 3 discusses estimation procedures for loading matrix and loading functions. Section 4 discuss the procedures for kriging and forecasting over space and time, respectively. Section 5 presents the asymptotic properties of the estimators. Section 6 illustrates the proposed model and estimation scheme on a synthetic data set; and finally Section 7 applies the proposed method to a real data set. Technique proofs are relegated to the Appendix.

2 Model

Consider a multivariate spatial-temporal process $\widetilde{\boldsymbol{y}}_t(\boldsymbol{s}) \in \mathbb{R}^p$:

$$\widetilde{\boldsymbol{y}}_t(\boldsymbol{s}) = \boldsymbol{C}^\top(\boldsymbol{s})\boldsymbol{z}_t(\boldsymbol{s}) + \boldsymbol{\xi}_t(\boldsymbol{s}) + \boldsymbol{\varepsilon}_t(\boldsymbol{s}), \quad t = 0, \pm 1, \pm 2, \cdots, \ \boldsymbol{s} \in \mathcal{S} \subset \mathcal{R}^2. \tag{1}$$

The first mean process term with observable covariates $\boldsymbol{z}_t(\boldsymbol{s}) \in \mathbb{R}^m$ and unknown coefficient matrix $\boldsymbol{C}(\boldsymbol{s}) \in \mathbb{R}^{m\times p}$ captures the large-scale correlations. The second term $\boldsymbol{\xi}_t(\boldsymbol{s}) \in \mathbb{R}^p$ is the zero-mean latent spatial-temporal vector process that captures the medium or small-scale correlation structure. It satisfies the conditions

$$\mathbb{E}[\boldsymbol{\xi}_t(\boldsymbol{s})] = \boldsymbol{0}, \quad \mathrm{Cov}\Big[\boldsymbol{\xi}_{t_1}(\boldsymbol{u}), \boldsymbol{\xi}_{t_2}(\boldsymbol{v})\Big] = \Sigma_{\xi,|t_1-t_2|}(\boldsymbol{u},\boldsymbol{v}). \tag{2}$$

The additive error vector $\boldsymbol{\varepsilon}_t(\boldsymbol{s})$ is the unknown spatial nugget effects which are spatially uncorrelated but are allowed to be temporally correlated. It is also uncorrelated with the signal process. That is,

$$\mathbb{E}[\boldsymbol{\varepsilon}_t(\boldsymbol{s})] = \boldsymbol{0}, \quad \mathrm{Var}[\boldsymbol{\varepsilon}_t(\boldsymbol{s})] = \Sigma_\epsilon(\boldsymbol{s}), \quad \mathrm{Cov}\Big[\boldsymbol{\varepsilon}_{t_1}(\boldsymbol{u}), \boldsymbol{\varepsilon}_{t_2}(\boldsymbol{v})\Big] = \boldsymbol{0} \ \forall \ t_1, t_2, \ \boldsymbol{u} \neq \boldsymbol{v}, \tag{3}$$

$$\mathrm{Cov}\Big[\boldsymbol{\xi}_{t_1}(\boldsymbol{u}), \boldsymbol{\varepsilon}_{t_2}(\boldsymbol{v})\Big] = \boldsymbol{0} \ \forall \ t_1, t_2, \boldsymbol{u}, \boldsymbol{v}. \tag{4}$$

Given the observable covariates $\boldsymbol{z}_t(\boldsymbol{s}) \in \mathbb{R}^m$, the coefficients $\boldsymbol{C}(\boldsymbol{s})$ can be calculated by least square regression. To make the main idea clear, we focus on the zero-mean process $\boldsymbol{y}_t(\boldsymbol{s}) = \widetilde{\boldsymbol{y}}_t(\boldsymbol{s}) - \boldsymbol{C}^\top(\boldsymbol{s})\boldsymbol{z}_t(\boldsymbol{s})$ with out loss of generality. That is,

$$\boldsymbol{y}_t(\boldsymbol{s}) = \boldsymbol{\xi}_t(\boldsymbol{s}) + \boldsymbol{\varepsilon}_t(\boldsymbol{s}), \quad t = 0, \pm 1, \pm 2, \cdots, \ \boldsymbol{s} \in \mathcal{S} \subset \mathcal{R}^2. \tag{5}$$

Under the condition (2) and (3), $\boldsymbol{y}_t(\boldsymbol{s})$ is second-order stationary in time t. We have $\mathbb{E}\left[\boldsymbol{y}_t(\boldsymbol{s})\right] = \boldsymbol{0}$ and

$$\mathrm{Cov}\left[\boldsymbol{y}_{t_1}(\boldsymbol{u}), \boldsymbol{y}_{t_2}(\boldsymbol{v})\right] = \Sigma_{\xi,|t_1-t_2|}(\boldsymbol{u}, \boldsymbol{v}) + \Sigma_{\epsilon,|t_1-t_2|}(\boldsymbol{u}) \cdot \mathbb{1}(\boldsymbol{u} = \boldsymbol{v}),$$

where the covariance $\Sigma_{\xi,t}(\boldsymbol{u}, \boldsymbol{v})$ is assumed to be continuous in $\boldsymbol{u}$ and $\boldsymbol{v}$.

Model (5) does not impose any stationary conditions over space. However, it requires that $\boldsymbol{y}_t(\boldsymbol{s})$ is second order stationary in time t to enable the learning of the dependence across different locations and times. In practice the data often show some trends and seasonal patterns in time. The existing de-trend and de-seasonality methods in time series analysis (Tsay and Chen, 2018; Tsay, 2014; Fan and Yao, 2005) can be applied to make each time series temporally stationary, including the inclusion of time trends in the mean term $\boldsymbol{C}'(\boldsymbol{s})\boldsymbol{z}_t(\boldsymbol{s})$.

2.1 The covariance structures across variables, space and time

To capture the correlation between the multiple processes, we assume that the latent spatial-temporal vector process are driven by a lower-dimension latent spatial-temporal factor process linearly in the form:

$$\boldsymbol{\xi}_t(\boldsymbol{s}) = \boldsymbol{B}\boldsymbol{f}_t(\boldsymbol{s}), \tag{6}$$

where $\boldsymbol{f}_t(\boldsymbol{s}) \in \mathbb{R}^r$ is the latent factor process ($r \ll p$) and $\boldsymbol{B}$ is the $p \times r$ loading matrix that characterized the correlation between multiple processes. Equation (6) is a generalization of the widely-used statistical factor models for high-dimensional data sets (Fan et al., 2018; Chang et al., 2015; Lam and Yao, 2012; Lam et al., 2011; Bai, 2003; Bai and Ng, 2002) to the spatial-temporal process.

To capture the spatial temporal correlations, we further assume a finite dimensional representation for $\boldsymbol{f}_t(\boldsymbol{s})$, that is, the latent $r \times 1$ factor process $\boldsymbol{f}_t(\boldsymbol{s})$ admits a finite functional structure,

$$\boldsymbol{f}_t(\boldsymbol{s}) = \sum_{j=1}^{d} a_j(\boldsymbol{s}) \boldsymbol{x}_{tj}, \tag{7}$$

where $a_j(\boldsymbol{s})$, $j \in [d]$ are deterministic and linearly 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 random vector $\boldsymbol{x}_{tj} \in \mathbb{R}^r$. Equation (7) models the latent factor process as the linear combination of random vectors with weight $a_j(\boldsymbol{s})$.

Functions $a_1(\cdot), \cdots, a_d(\cdot)$ are not uniquely defined by (7) even with known $\boldsymbol{f}_t$. Particularly, we can rewrite $\boldsymbol{f}_t(\boldsymbol{s}) = \sum_{j=1}^{d} a_j^*(\boldsymbol{s}) \boldsymbol{x}_{tj}^*$ where $a_j^*(\boldsymbol{s}) = c a_j(\boldsymbol{s})$ and $\boldsymbol{x}_{tj}^* = c^{-1} \boldsymbol{x}_{tj}$ for any scalar $c \neq 0$. There is no loss of generality in assuming that $a_1(\cdot), \cdots, a_d(\cdot)$ are orthonormal in the sense that

$$\left\langle a_i, a_j \right\rangle = \mathbb{1}(i = j),$$

as any set of linear independent functions in a Hilbert space can be standardized to this effect. The above identification condition is defined on the whole space. We will elaborate more on the model identification in the next section. Combining (6) and (7), we have

$$\boldsymbol{\xi}_t(\boldsymbol{s}) = \boldsymbol{B} \sum_{j=1}^{d} a_j(\boldsymbol{s}) \boldsymbol{x}_{tj} = \boldsymbol{B} \boldsymbol{X}_t' \boldsymbol{a}(\boldsymbol{s}), \tag{8}$$

where $\boldsymbol{X}_t = (\boldsymbol{x}_{t1}, \cdots, \boldsymbol{x}_{td})'$ and $\boldsymbol{a}(\boldsymbol{s}) = (a_1(\boldsymbol{s}), \cdots, a_d(\boldsymbol{s}))'$. Therefore, the latent spatial-temporal covariance of vector process $\boldsymbol{\xi}_{t_1}(\boldsymbol{u})$ and $\boldsymbol{\xi}_{t_2}(\boldsymbol{v})$ can be written as

$$\Sigma_{\xi,|t_1-t_2|}(\boldsymbol{u}, \boldsymbol{v}) = \mathrm{Cov}\left[\boldsymbol{B} \boldsymbol{X}_{t_1}' \boldsymbol{a}(\boldsymbol{u}), \boldsymbol{B} \boldsymbol{X}_{t_2}' \boldsymbol{a}(\boldsymbol{v})\right] = \boldsymbol{B} \Sigma_{f,|t_1-t_2|}(\boldsymbol{u}, \boldsymbol{v}) \boldsymbol{B}', \tag{9}$$

where

$$\Sigma_{f,|t_1-t_2|}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{d} \sum_{j=1}^{d} a_i(\boldsymbol{u}) a_j(\boldsymbol{v}) \Sigma_{x,ij,|t_1-t_2|}, \tag{10}$$

and $\Sigma_{x,ij,|t_1-t_2|} = \mathrm{Cov}\left[\boldsymbol{x}_{t_1 i}, \boldsymbol{x}_{t_2 j}\right] \in \mathbb{R}^{r \times r}$. Equation (9) captures the spatial-temporal dependence structure via the finite dimensional representation of latent factors in (7). Specifi-

cally, the covariance of factor $\Sigma_{f,|t_1-t_2|}$ is the linear combination of $\Sigma_{x,ij,|t_1-t_2|}$, which captures the time-dependence structure between t_1 and t_2. The weight $a_i(\boldsymbol{u})a_j(\boldsymbol{v})$ captures the spatial dependence between location $\boldsymbol{u}$ and $\boldsymbol{v}$.

Relation to the univariate reduced-rank models. In the special case where $\boldsymbol{f}_t(\boldsymbol{s})$ is a scalar, i.e. $r = 1$, the covariance of latent factor assumes the following structure

$$\sigma_{f,|t_1-t_2|}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{a}(\boldsymbol{u})^\top \Sigma_{x,|t_1-t_2|}\boldsymbol{a}(\boldsymbol{v}), \tag{11}$$

where $\Sigma_{x,|t_1-t_2|}$ is a $d \times d$ matrix consisting of $\Sigma_{x,ij,|t_1-t_2|}$ (which is a scalar when $r = 1$) for all $i, j \in [d]$. Spatial-temporal structure (11) corresponds to the low-rank empirical orthogonal function method in the literature of univariate geostatistics (Wikle and Cressie, 1999; Kammann and Wand, 2003; Cressie et al., 2010; Banerjee et al., 2008; Finley et al., 2009; Tzeng and Huang, 2018).

Relation to the multivariate reduced-rank models. In the case of known low-dimensional factor process $\boldsymbol{f}_t(\boldsymbol{s})$, the covariance of any pair of variables in $\boldsymbol{f}$ assumes the structure in (11). This corresponds to the low-rank approximation in the literature of multivariate geostatistics. In our setting, the latent factor process $\boldsymbol{f}_t(\boldsymbol{s})$ is unknown and needs to be estimated from an observed high-dimensional process $\boldsymbol{y}_t(\boldsymbol{s})$.

2.2 Discrete sample observations

Since we only observe discrete observations, we assume that we have a $n \times p$ matrix $\boldsymbol{\Xi}_t \triangleq [\xi_t(\boldsymbol{s}_1), \cdots, \xi_t(\boldsymbol{s}_n)]^\top$ where $\xi_t(\boldsymbol{s}_i) \in \mathbb{R}^p$ consists of values of $\xi_t(\boldsymbol{s})$ from the i-th sampling location. It follows from (8) that

$$\boldsymbol{\Xi}_t = \boldsymbol{A}\boldsymbol{X}_t\boldsymbol{B}', \tag{12}$$

where $\boldsymbol{A} = [a_j(\boldsymbol{s}_i)]_{ij}$, $i \in [n]$ and $j \in [d]$. We are interested in estimating the loading matrix $\boldsymbol{B}$, random matrix $\boldsymbol{X}_t$, the spatial loading function matrix $\boldsymbol{A}$, and the spatial loading functions $a_j(\boldsymbol{s})$ for $j \in [d]$.

Matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are not uniquely defined by (6). Specifically, we can rewrite $\boldsymbol{\Xi}_t = \boldsymbol{A}^*\boldsymbol{X}_t^*\boldsymbol{B}^{*'}$ where $\boldsymbol{A}^* = \boldsymbol{A}\boldsymbol{O}_1$, $\boldsymbol{B}^* = \boldsymbol{B}\boldsymbol{O}_2$, and $\boldsymbol{X}_t^* = \boldsymbol{O}_1^{-1}\boldsymbol{X}_t\boldsymbol{O}_2^{-1}$ for any invertible matrices $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$. To address this identification problem, we assume that columns of $\boldsymbol{A}$ ($\boldsymbol{B}$) are orthogonal.

Under the orthogonal assumption, the vector space spanned by the columns of $\boldsymbol{A}(\boldsymbol{s})$ and $\boldsymbol{B}$, denoted as $\mathcal{M}(\boldsymbol{A}(\boldsymbol{s}))$ and $\mathcal{M}(\boldsymbol{B})$, are uniquely defined. In this article, we estimate matrix representations $\boldsymbol{Q}_A$ and $\boldsymbol{Q}_B$ of $\mathcal{M}(\boldsymbol{A}(\boldsymbol{s}))$, $\mathcal{M}(\boldsymbol{B})$ instead of $\boldsymbol{A}(\boldsymbol{s})$ and $\boldsymbol{B}$ under the assumption that

$$\boldsymbol{Q}_A'\boldsymbol{Q}_A = \boldsymbol{I}_d, \quad \text{and} \quad \boldsymbol{Q}_B'\boldsymbol{Q}_B = \boldsymbol{I}_r, \tag{13}$$

and the corresponding $\boldsymbol{Z}_t$ such that (12) can be rewritten as

$$\boldsymbol{\Xi}_t = \boldsymbol{A}\boldsymbol{X}_t\boldsymbol{B}' = \boldsymbol{Q}_A\boldsymbol{Z}_t\boldsymbol{Q}_B'. \tag{14}$$

Given $\boldsymbol{Q}_A$, the kernel reproducing Hilbert space (KRHS) spanned by $a_1(\cdot), \cdots, a_d(\cdot)$ is also uniquely defined and we estimate a set of representative functions $q_{a,1}(\cdot), \cdots, a_{q,d}(\cdot)$. Therefore, the estimation of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{X}_t$ in the multivariate spatial-temporal model can be converted to the estimation of $\boldsymbol{Q}_A$, $\boldsymbol{Q}_B$, and $\boldsymbol{Z}_t$. Further we use the estimators to estimate the latent spatial-temporal covariance and make spatial-temporal predictions for large scale multi-variate spatial temporal data set. More details are discussed in the sequel.

3 Estimation

Let $\left\{\widetilde{\boldsymbol{y}}_t(\boldsymbol{s}_i), \boldsymbol{z}_t(\boldsymbol{s}_i)\right\}$, $i \in [n]$, $t \in [T]$ be the available observations over space and time, where $\widetilde{\boldsymbol{y}}_t(\boldsymbol{s}_i) \in \mathbb{R}^p$ and $\boldsymbol{z}_t(\boldsymbol{s}_i) \in \mathbb{R}^m$ is a vector of covariates observed at location $\boldsymbol{s}_i$ at time t. In this article, we restrict attention to the case where all variables have been measured at the same sample locations $\boldsymbol{s}_i$, $i \in [n]$.

In general cases where $\boldsymbol{C}(\boldsymbol{s}) \neq \boldsymbol{0}$, we can estimate $\widehat{\boldsymbol{C}}(\boldsymbol{s})$ by least square regression from the observations $\left\{\widetilde{\boldsymbol{y}}_t(\boldsymbol{s}_i), \boldsymbol{z}_t(\boldsymbol{s}_i)\right\}$. The following procedure can be applied to the residuals

$\widehat{\boldsymbol{y}}_t(\boldsymbol{s}_i) \triangleq \widetilde{\boldsymbol{y}}_t(\boldsymbol{s}_i) - \widehat{\boldsymbol{C}}^\top(\boldsymbol{s}_i)\boldsymbol{z}_t(\boldsymbol{s}_i)$. With out loss of generality, we consider a special case where $\boldsymbol{C}(\boldsymbol{s}) \equiv \boldsymbol{0}$ in (5). Now the observations are generated from the process

$$\boldsymbol{y}_t(\boldsymbol{s}) = \boldsymbol{\xi}_t(\boldsymbol{s}) + \boldsymbol{\varepsilon}_t(\boldsymbol{s}) = \boldsymbol{B}\boldsymbol{X}_t^\top \boldsymbol{a}(\boldsymbol{s}) + \boldsymbol{\varepsilon}_t(\boldsymbol{s}). \tag{15}$$

From (6), (7), and (12), we stack $\boldsymbol{y}_t(\boldsymbol{s}_i)$, $i \in [n]$ together as rows and get

$$\boldsymbol{Y}_t = \boldsymbol{\Xi}_t + \boldsymbol{E}_t = \boldsymbol{A}\boldsymbol{X}_t\boldsymbol{B}^\top + \boldsymbol{E}_t = \boldsymbol{Q}_A\boldsymbol{Z}_t\boldsymbol{Q}_B^\top + \boldsymbol{E}_t, \tag{16}$$

where $\boldsymbol{Y}_t = \big(\boldsymbol{y}_t(\boldsymbol{s}_1), \cdots, \boldsymbol{y}_t(\boldsymbol{s}_n)\big)$ and $\boldsymbol{E}_t = (\boldsymbol{\varepsilon}_t(\boldsymbol{s}_1), \cdots, \boldsymbol{\varepsilon}_t(\boldsymbol{s}_n))^\top$.

Note that $\boldsymbol{A}$ (or $\boldsymbol{B}$) has the same column space as $\boldsymbol{Q}_A$ (or $\boldsymbol{Q}_B$). They are different only up to a scalar factor or a rotation such that $\boldsymbol{A}$ satisfies Condition 5.5 in Section 5 while $\boldsymbol{Q}_A$ satisfies $\boldsymbol{Q}_A^\top\boldsymbol{Q}_A = \boldsymbol{I}_d$, and $\boldsymbol{B}$ satisfies Condition 5.4 while $\boldsymbol{Q}_B$ satisfies $\boldsymbol{Q}_B^\top\boldsymbol{Q}_B = \boldsymbol{I}_r$. In the following, we use the triplets $(\boldsymbol{Q}_A, \boldsymbol{Z}_t, \boldsymbol{Q}_B)$ and $(\boldsymbol{A}, \boldsymbol{X}_t, \boldsymbol{B})$ interchangeably.

3.1 Partitioned spatial loading spaces $\mathcal{M}(\boldsymbol{A}_1)$ and $\mathcal{M}(\boldsymbol{A}_2)$

Note that the nugget effect $\boldsymbol{\varepsilon}_t(\boldsymbol{s})$ are uncorrelated over space. We exploit this fact to exclude the covariance term incurred by the nugget effect. Particularly, we divide n locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ into two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ with n_1 and n_2 elements respectively. Preferably, we set $n_1 \asymp n_2 \asymp n/2$ according to Theorem 5.7. Let $\boldsymbol{Y}_{lt}$ be a matrix consisting of $\boldsymbol{y}_t(\boldsymbol{s})$, $\boldsymbol{s} \in \mathcal{S}_l$, $l = 1, 2$ as rows. Then $\boldsymbol{Y}_{1t}$ and $\boldsymbol{Y}_{2t}$ are two matrices of dimension $n_1 \times p$ and $n_2 \times p$ respectively. It follows from (15) that

$$\boldsymbol{Y}_{1t} = \boldsymbol{\Xi}_{1t} + \boldsymbol{E}_{1t} = \boldsymbol{A}_1\boldsymbol{X}_t\boldsymbol{B}^\top + \boldsymbol{E}_{1t}, \qquad \boldsymbol{Y}_{2t} = \boldsymbol{\Xi}_{2t} + \boldsymbol{E}_{2t} = \boldsymbol{A}_2\boldsymbol{X}_t\boldsymbol{B}^\top + \boldsymbol{E}_{2t}, \tag{17}$$

where $\boldsymbol{A}_l$ is a $n_l \times d$ matrix, its rows are $(a_1(\boldsymbol{s}), \ldots, a_d(\boldsymbol{s}))$ at different locations $\boldsymbol{s} \in \mathcal{S}_l$ and $\boldsymbol{E}_{t,l}$ consists of $\boldsymbol{\varepsilon}_t(\boldsymbol{s})$ as rows with $\boldsymbol{s} \in \mathcal{S}_l$, $l = 1, 2$.

For model identification, we assume that the columns of $\boldsymbol{A}_l$, $l = 1, 2$ are orthogonal. Under this assumption, $\mathcal{M}(\boldsymbol{A}_1)$ and $\mathcal{M}(\boldsymbol{A}_2)$, which are the column spaces of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, are uniquely defined. This however implies that $\boldsymbol{X}_t$ in the second equation in (17) will be different from that in the first equation. Thus, we may rewrite (17) as

$$\boldsymbol{Y}_{1t} = \boldsymbol{\Xi}_{1t} + \boldsymbol{E}_{1t} = \boldsymbol{A}_1\boldsymbol{X}_t\boldsymbol{B}^\top + \boldsymbol{E}_{1t}, \qquad \boldsymbol{Y}_{2t} = \boldsymbol{\Xi}_{2t} + \boldsymbol{E}_{2t} = \boldsymbol{A}_2\boldsymbol{X}_t^*\boldsymbol{B}^\top + \boldsymbol{E}_{2t}, \tag{18}$$

where $\boldsymbol{X}_t^* = \boldsymbol{O}\boldsymbol{X}_t$ and $\boldsymbol{O}$ is an invertible $d \times d$ matrix.

Let $\boldsymbol{y}_{lt,\cdot j}$, $\boldsymbol{e}_{lt,\cdot j}$, and $\boldsymbol{b}_{j\cdot}$ be the j-th column of $\boldsymbol{Y}_{lt}$, $\boldsymbol{E}_{lt}$, and $\boldsymbol{B}$, $l = 1, 2$, $j \in [p]$, respectively. Define spatial-cross-covariance matrix between the i-th and j-th variables as

$$\boldsymbol{\Omega}_{A,ij} = \mathrm{Cov}\left[\boldsymbol{y}_{1t,\cdot i}, \boldsymbol{y}_{2t,\cdot j}\right] = \boldsymbol{A}_1 \mathrm{Cov}\left[\boldsymbol{X}_t \boldsymbol{b}_{i\cdot}', \boldsymbol{X}_t^* \boldsymbol{b}_{j\cdot}'\right] \boldsymbol{A}_2. \tag{19}$$

The covariance related to $\boldsymbol{e}_{1t,\cdot i}$ and $\boldsymbol{e}_{2t,\cdot j}$ are all zeros because they are spatial white noises and also uncorrelated with the signals. When $d \ll n$, it is reasonable to assume that $\mathrm{rank}\left[\boldsymbol{\Omega}_{A,ij}\right] = d$. Define

$$\boldsymbol{M}_{A_1} = \sum_{i=1}^{p}\sum_{j=1}^{p} \boldsymbol{\Omega}_{A,ij}\boldsymbol{\Omega}_{A,ij}^\top, \quad \text{and} \quad \boldsymbol{M}_{A_2} = \sum_{i=1}^{p}\sum_{j=1}^{p} \boldsymbol{\Omega}_{A,ij}^\top\boldsymbol{\Omega}_{A,ij}$$

$\boldsymbol{M}_{A_1}$ and $\boldsymbol{M}_{A_2}$ share the same d positive eigenvalues and $\boldsymbol{M}_{A_l}\boldsymbol{q} = \boldsymbol{0}$ for any vector $\boldsymbol{q}$ perpendicular to $\mathcal{M}(\boldsymbol{A}_l)$, $l = 1, 2$. Therefore, the columns of a matrix representation of $\mathcal{M}(\boldsymbol{A}_l)$, $l = 1, 2$, can be estimated as the d orthonormal eigenvectors of matrix $\boldsymbol{M}_{A_l}$ corresponding to largest d positive eigenvalues in the descending order.

Now we define the sample version of these quantities and introduce the estimation procedure. Suppose we have centered our observations $\boldsymbol{Y}_{1t}$ and $\boldsymbol{Y}_{2t}$, let $\widehat{\boldsymbol{\Omega}}_{A,ij}$ be the sample cross-space covariance of i-th and j-th variables and $\widehat{\boldsymbol{M}}_{A_l}$ be the sample version of $\boldsymbol{M}_{A_l}$, $l = 1, 2$, that is

$$\widehat{\boldsymbol{\Omega}}_{A,ij} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{Y}_{1t,\cdot i}\boldsymbol{Y}_{2t,\cdot j}^\top, \quad \widehat{\boldsymbol{M}}_{A_1} = \sum_{i=1}^{p}\sum_{j=1}^{p} \widehat{\boldsymbol{\Omega}}_{A,ij}\widehat{\boldsymbol{\Omega}}_{A,ij}^\top, \quad \widehat{\boldsymbol{M}}_{A_2} = \sum_{i=1}^{p}\sum_{j=1}^{p} \widehat{\boldsymbol{\Omega}}_{A,ij}^\top\widehat{\boldsymbol{\Omega}}_{A,ij}. \tag{20}$$

A natural estimator for a matrix representation of $\mathcal{M}(\boldsymbol{A}_l)$, under the constraint that $\boldsymbol{Q}_{A,l}^\top\boldsymbol{Q}_{A,l} = \boldsymbol{I}_d$, is defined as

$$\widehat{\boldsymbol{Q}}_{A,l} = \{\widehat{\boldsymbol{q}}_{A,l1}, \cdots, \widehat{\boldsymbol{q}}_{A,ld}\}, \quad l = 1, 2, \tag{21}$$

where $\widehat{\boldsymbol{q}}_{A,lj}$ is the eigenvector of $\widehat{\boldsymbol{M}}_{A_l}$ corresponding to its j-th largest eigenvalue. Matrix $\widehat{\boldsymbol{Q}}_{A,l}$ estimates $\boldsymbol{A}_l(\boldsymbol{s})$ up to a scalar factor while sharing the same column space. However such an estimator ignores the fact that $\boldsymbol{\xi}_t(\boldsymbol{s})$ is continuous over the set $\mathcal{S}$. Section 3.4 estimates a refined spatial loading matrix $\widehat{\boldsymbol{Q}}_A$ and further estimates the loading function

$\widehat{\boldsymbol{Q}}_A(\boldsymbol{s})$, which estimates $A(\boldsymbol{s})$ up to a scalar factor.

3.2 Variable loading space $\mathcal{M}(\boldsymbol{B})$

To estimate the $p \times r$ variable loading matrix $\boldsymbol{B}$, we again utilize the spatial whiteness properties of the nugget effect. Recall that in Section 3.1, the entire set of n sampled locations are divided into two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ of size n_1 and n_2, where $n_1 \asymp n_2 \asymp \frac{n}{2}$. We keep only $m = \lfloor \frac{n}{2} \rfloor$ in each of $\mathcal{S}_1$ and $\mathcal{S}_2$ to calculate $\boldsymbol{B}$. When n is even, we make use of all sampled locations, while when n is odd, one of the sampled locations is dropped randomly.

We reuse the notation in equation (17) for the observations in $\mathcal{S}_1$ and $\mathcal{S}_2$ and rewrite it as (18) for model identification, except for now $\boldsymbol{Y}_{1t}$ and $\boldsymbol{Y}_{2t}$ are two matrices of same dimension $m \times p$. Let $\boldsymbol{y}_{lt,i\cdot}$, $\boldsymbol{e}_{lt,i\cdot}$, and $\boldsymbol{a}_{l,i\cdot}$ be the i-th row of $\boldsymbol{Y}_{lt}$, $\boldsymbol{E}_{lt}$, and $\boldsymbol{A}_l$, $l = 1, 2$, respectively. Define the covariance matrix of p variables sampled at the i-th location in $\mathcal{S}_1$ and j-th location in $\mathcal{S}_2$ as

$$\boldsymbol{\Omega}_{B,ij} = \mathrm{Cov}\left[\boldsymbol{y}_{1t,i\cdot}, \boldsymbol{y}_{2t,j\cdot}\right] = \boldsymbol{B}\,\mathrm{Cov}\left[\boldsymbol{X}_t^\top \boldsymbol{a}_{1,i\cdot}, \boldsymbol{X}_t^{*\top} \boldsymbol{a}_{2,j\cdot}\right]\boldsymbol{B}^\top.$$

When $r \ll p$, it is reasonable to assume that $\mathrm{rank}\left[\boldsymbol{\Omega}_{B,ij}\right] = r$. Let

$$\boldsymbol{M}_B = \sum_{i=1}^{m}\sum_{j=1}^{m} \boldsymbol{\Omega}_{B,ij}\boldsymbol{\Omega}_{B,ij}^\top. \tag{22}$$

Then, $\boldsymbol{M}_B$ has r positive eigenvalues and $\boldsymbol{M}_B\boldsymbol{q} = \boldsymbol{0}$ for any vector $\boldsymbol{q}$ perpendicular to $\mathcal{M}(\boldsymbol{B})$. Therefore, the columns of a matrix representation of $\mathcal{M}(\boldsymbol{B})$ can be estimated as the r orthonormal eigenvectors of matrix $\boldsymbol{M}_B$ corresponding to the largest r positive eigenvalues in the descending order.

Define the sample version of $\boldsymbol{\Omega}_{B,ij}$ and $\boldsymbol{M}_B$ for centered observation $\boldsymbol{Y}_t$ as

$$\widehat{\boldsymbol{\Omega}}_{B,ij} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}_{1t,i\cdot}\boldsymbol{y}_{2t,j\cdot}^\top, \quad \widehat{\boldsymbol{M}}_B = \sum_{i=1}^{m}\sum_{j=1}^{m} \widehat{\boldsymbol{\Omega}}_{B,ij}\widehat{\boldsymbol{\Omega}}_{B,ij}^\top. \tag{23}$$

A natural estimator for a matrix representation of $\mathcal{M}(\boldsymbol{B})$ under constraint (13) is defined as

$$\widehat{\boldsymbol{Q}}_B = \{\widehat{\boldsymbol{q}}_{B,1}, \cdots, \widehat{\boldsymbol{q}}_{B,r}\},$$

where $\widehat{\boldsymbol{q}}_{B,i}$ is the eigenvector of $\widehat{\boldsymbol{M}}_B$ corresponding to its i-th largest eigenvalue. Matrix $\widehat{\boldsymbol{Q}}_B$ estimates $\boldsymbol{B}$ up to a scalar factor while sharing the same column space.

The above estimation procedure assumes that the latent dimensions $d \times r$ are known. However, in practice we need to estimate d and r as well. Two methods of estimating the latent dimension are (a) the eigenvalue ratio-based estimator, similar to those defined in Lam and Yao (2012); Wang et al. (2019); (b) the Scree plot which is standard in principal component analysis. Let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_r \geq 0$ be the ordered eigenvalues of $\widehat{\boldsymbol{M}}_B$. The ratio-based estimator for r is defined as

$$\widehat{r} = \underset{1 \leq j \leq r_{\max}}{\arg\max} \frac{\widehat{\lambda}_j}{\widehat{\lambda}_{j+1}}, \tag{24}$$

where $r \leq r_{\max} \leq p$ is an integer. In practice we may take $r_{\max} = \lceil p/2 \rceil$ or $r_{\max} = \lceil p/3 \rceil$. Ratio estimators $\widehat{d}_1$ and $\widehat{d}_2$ is defined similarly with respect to $\widehat{\boldsymbol{M}}_{A_1}$ and $\widehat{\boldsymbol{M}}_{A_2}$, respectively. We set $\widehat{d} = \max\{\widehat{d}_1, \widehat{d}_2\}$. Chen et al. (2019) shows that eigen-ratio estimators $\widehat{d}$ and $\widehat{r}$ are consistent under a similar setting.

3.3 Signal matrix $\boldsymbol{\Xi}_t$

By (17), the estimators of two representations of the rotated latent matrix factor $\boldsymbol{Z}_t$, $t \in [T]$, are defined as

$$\widehat{\boldsymbol{Z}}_{1t} = \widehat{\boldsymbol{Q}}_{A,1}^\top \boldsymbol{Y}_{1t} \widehat{\boldsymbol{Q}}_B, \qquad \widehat{\boldsymbol{Z}}_{2t} = \widehat{\boldsymbol{Q}}_{A,2}^\top \boldsymbol{Y}_{2t} \widehat{\boldsymbol{Q}}_B. \tag{25}$$

The latent signal process are estimated by

$$\widehat{\boldsymbol{\Xi}}_t = \begin{bmatrix} \widehat{\boldsymbol{\Xi}}_{1t} \\ \widehat{\boldsymbol{\Xi}}_{2t} \end{bmatrix}, \tag{26}$$

where

$$\widehat{\boldsymbol{\Xi}}_{1t} = \widehat{\boldsymbol{Q}}_{A,1} \widehat{\boldsymbol{Z}}_{1t} \widehat{\boldsymbol{Q}}_B^\top = \widehat{\boldsymbol{Q}}_{A,1} \widehat{\boldsymbol{Q}}_{A,1}^\top \boldsymbol{Y}_{1t} \widehat{\boldsymbol{Q}}_B \widehat{\boldsymbol{Q}}_B^\top, \qquad \widehat{\boldsymbol{\Xi}}_{2t} = \widehat{\boldsymbol{Q}}_{A,2} \widehat{\boldsymbol{Z}}_{2t} \widehat{\boldsymbol{Q}}_B^\top = \widehat{\boldsymbol{Q}}_{A,2} \widehat{\boldsymbol{Q}}_{A,2}^\top \boldsymbol{Y}_{2t} \widehat{\boldsymbol{Q}}_B \widehat{\boldsymbol{Q}}_B^\top.$$

Equation (25) provides two estimates of $\boldsymbol{Z}_t$ based on two partitioned sets of locations.

Section 3.4 will re-estimate a unified version of latent factor matrix $\boldsymbol{Z}_t$ from all sampling locations. Estimator of the latent signal process will also be re-estimated from all sampling locations.

To mitigate the estimation error associated with the random partition of the location set, one could again carry out the estimation procedure with multiple random partitions and return the average estimates, similar to those done in Huang et al. (2016). To keep the core idea clear, we do not consider random partitions in this paper. The results for the average estimates from random partitions can be derived similarly to Huang et al. (2016) based on the results of the present paper.

3.4 Spatial loading space $\mathcal{M}(\boldsymbol{A})$ and loading function $\boldsymbol{Q}_A(\boldsymbol{s})$

The procedure in Section 3.1 only estimates the spatial loading matrices $\widehat{\boldsymbol{Q}}_{A,1}$ and $\widehat{\boldsymbol{Q}}_{A,2}$ on two partitioned set of sampling locations. Estimate loading functions from $\widehat{\boldsymbol{Q}}_{A,1}$ and $\widehat{\boldsymbol{Q}}_{A,2}$ separately will result in inefficient use of sampling locations. In addition, equation (25) gives estimators for two different representations of the latent matrix factor $\boldsymbol{Z}_t$. To get estimators of the $n\times d$ spatial loading matrix $\boldsymbol{Q}_A$ for all sampling locations and $\boldsymbol{Z}_t$, we use the estimated $\widehat{\boldsymbol{\Xi}}_t$ to re-estimate $\widehat{\boldsymbol{Q}}_A$ and $\widehat{\boldsymbol{Z}}_t$.

Recall that the population signals process is $\boldsymbol{\xi}_t(\boldsymbol{s}) = \boldsymbol{B}\boldsymbol{X}_t^\top \boldsymbol{Q}_A(\boldsymbol{s}) = \boldsymbol{Q}_B \boldsymbol{Z}_t^\top \boldsymbol{q}_a(\boldsymbol{s})$ and the $n\times p$ matrix $\boldsymbol{\Xi}_t = \boldsymbol{A}\boldsymbol{X}_t\boldsymbol{B}^\top = \boldsymbol{Q}_A\boldsymbol{Z}_t\boldsymbol{Q}_B^\top$ is the signal matrix at discretized sampling locations at each time t. To reduce dimension, we use $\boldsymbol{\Psi}_t = \boldsymbol{Q}_A\boldsymbol{Z}_t \in \mathbb{R}^{n\times r}$, rather than $\boldsymbol{\Xi}_t \in \mathbb{R}^{n\times p}$. Define

$$\boldsymbol{M}_A = \sum_{j=1}^{r} \mathrm{Cov}\left[\boldsymbol{\Psi}_{t,\cdot j}, \boldsymbol{\Psi}_{t,\cdot j}\right] = \boldsymbol{Q}_A \sum_{j=1}^{r} \mathrm{Cov}\left[\boldsymbol{Z}_{t,\cdot j}, \boldsymbol{Z}_{t,\cdot j}\right] \boldsymbol{Q}_A^\top.$$

However, true $\boldsymbol{\Xi}_t$ or $\boldsymbol{\Psi}_t$ are not observable. We estimate $\widehat{\boldsymbol{\Xi}}_t$ from (26) and obtain

$$\widehat{\boldsymbol{\Psi}}_t = \widehat{\boldsymbol{\Xi}}_t \widehat{\boldsymbol{Q}}_B.$$

From estimated values, we defined the estimated version of $\boldsymbol{M}_A$ as

$$\widehat{\boldsymbol{M}}_A = \frac{1}{T}\sum_{t=1}^{T} \widehat{\boldsymbol{\Psi}}_t \widehat{\boldsymbol{\Psi}}_t^\top,$$

where $\widehat{\boldsymbol{\Psi}}$ is chosen over $\widehat{\boldsymbol{\Xi}}$ because $\widehat{\boldsymbol{\Psi}}$ has the same estimation error bound but is of lower dimension.

A natural estimator of a matrix representation of $\mathcal{M}(\boldsymbol{A})$ under constraint (13) is defined as

$$\widehat{\boldsymbol{Q}}_A = \{\widehat{\boldsymbol{q}}_{A,1}, \cdots, \widehat{\boldsymbol{q}}_{A,n}\},$$

where $\widehat{\boldsymbol{q}}_{A,i}$ is the eigenvector of $\widehat{\boldsymbol{M}}_A$ corresponding to its i-th largest eigenvalue. Matrix $\widehat{\boldsymbol{Q}}_A$ estimates $\boldsymbol{A}$ up to a scalar factor while sharing the same column space.

The estimator of the rotated latent factor matrix $\boldsymbol{Z}_t$ is obtained as

$$\widehat{\boldsymbol{Z}}_t = \widehat{\boldsymbol{Q}}_A^\top \widehat{\boldsymbol{\Psi}}_t. \tag{27}$$

Once $\widehat{\boldsymbol{Q}}_A$ is estimated, we estimate loading functions $q_{a,j}(\boldsymbol{s})$ from the estimated n observations in column $\widehat{\boldsymbol{Q}}_{A,\cdot j}$ by the sieve approximation. Any set of bivariate basis functions can be chosen. In our procedure, we consider the tensor product linear sieve space Θ_n, which is constructed as a tensor product space of some commonly used univariate linear approximating spaces, such as B-spline, orthogonal wavelets and polynomial series. Then for each $j \le d$,

$$q_{a,j}(\boldsymbol{s}) = \sum_{i=1}^{J_n} \beta_{i,j} u_i(\boldsymbol{s}) + r_j(\boldsymbol{s}).$$

Here $\beta_{i,j}$ are the sieve coefficients of i basis function $u_i(\boldsymbol{s})$ corresponding to the j-th factor loading function; $r_j(\boldsymbol{s})$ is the sieve approximation error; J_n represents the number of sieve terms which grows slowly as n goes to infinity. We estimate $\widehat{\beta}_{i,j}$ and the loading functions are approximated by $\widehat{q}_{a,j}(\boldsymbol{s}) = \sum_{i=1}^{J_n} \widehat{\beta}_{i,j} u_i(\boldsymbol{s})$.

4 Prediction

4.1 Spatial Prediction

A major focus of spatial-temporal data analysis is the prediction of variable of interest over new locations. For some new location $\boldsymbol{s}_0 \in \mathcal{S}$ and $\boldsymbol{s}_0 \neq \boldsymbol{s}_i$, $i \in [n]$, we aim to predict the unobserved value $\boldsymbol{y}_t(\boldsymbol{s}_0)$ observations $\boldsymbol{Y}_t$, $t = [T]$. By (15), we have $\boldsymbol{y}_t(\boldsymbol{s}_0) = \boldsymbol{\xi}_t(\boldsymbol{s}_0) + \boldsymbol{\varepsilon}_t(\boldsymbol{s}_0) = \boldsymbol{Q}_B \boldsymbol{Z}_t' \boldsymbol{q}_a(\boldsymbol{s}_0) + \boldsymbol{\varepsilon}_t(\boldsymbol{s}_0)$. As recommended by Cressie and Wikle (2015), we predict $\boldsymbol{\xi}_t(\boldsymbol{s}_0) = \boldsymbol{Q}_B \boldsymbol{Z}_t' \boldsymbol{q}_a(\boldsymbol{s}_0)$ instead of $\boldsymbol{y}_t(\boldsymbol{s}_0)$ directly. Thus, a natural estimator is

$$\widehat{\boldsymbol{\xi}}_t(\boldsymbol{s}_0) = \widehat{\boldsymbol{Q}}_B \widehat{\boldsymbol{Z}}_t' \widehat{\boldsymbol{q}}_a(\boldsymbol{s}_0), \tag{28}$$

where $\widehat{\boldsymbol{Q}}_B$, $\widehat{\boldsymbol{Z}}_t$ and $\widehat{\boldsymbol{q}}_a(\boldsymbol{s})$ are estimated following procedures in Section 3.

For univariate spatial temporal process, Huang et al. (2016) propose the kriging with kernel smoothing for spatial prediction. This method can be extended to our case by applying kriging with kernel smoothing for each one of the multivariate spatial temporal process. We implement both our spatial prediction based on (28) and kriging with kernel smoothing for each one of the multivariate spatial temporal process. Empirical results on synthetic as well as real data show that our method performance better than the kriging with kernel smoothing method.

4.2 Temporal Prediction

Temporal prediction focuses on predict the future values $\boldsymbol{y}_{t+h}(\boldsymbol{s}_1), \ldots, \boldsymbol{y}_{t+h}(\boldsymbol{s}_n)$ for some $h \geq 1$. By (15), we have $\boldsymbol{y}_{t+h}(\boldsymbol{s}) = \boldsymbol{\xi}_{t+h}(\boldsymbol{s}) + \boldsymbol{\varepsilon}_{t+h}(\boldsymbol{s}) = \boldsymbol{Q}_B \boldsymbol{Z}_{t+h}' \boldsymbol{q}_a(\boldsymbol{s}) + \boldsymbol{\varepsilon}_{t+h}(\boldsymbol{s})$. Since $\boldsymbol{\varepsilon}_{t+h}(\boldsymbol{s})$ is unpredictable white noise, the ideal predictor for $\boldsymbol{y}_{t+h}(\boldsymbol{s})$ is that for $\boldsymbol{\xi}_{t+h}(\boldsymbol{s})$. Thus, we focus on predict $\boldsymbol{\xi}_{t+h}(\boldsymbol{s}) = \boldsymbol{Q}_B \boldsymbol{Z}_{t+h}' \boldsymbol{q}_a(\boldsymbol{s})$. The temporal dynamics of the $\boldsymbol{\xi}_{t+h}(\boldsymbol{s})$ present in a lower dimensional matrix factor $\boldsymbol{Z}_{t+h}$, thus a more effective approach is to predict $\boldsymbol{Z}_{t+h}$ based on $\boldsymbol{Z}_{t-l}, \ldots, \boldsymbol{Z}_t$ where l is a prescribed integer. Time series analysis (Tsay, 2014; Tsay and Chen, 2018) can be applied to $\boldsymbol{Z}_t$ under general settings. We use the auto-regression

of order one (AR(1)) and take $l = 1$ to illustrate the idea.

Since the latent factor matrix time series $\boldsymbol{Z}_t \in \mathbb{R}^{d\times r}$ is of low-dimension, a straight forward method for predicting $\boldsymbol{Z}_{t+h}$ is applying the multivariate time series analysis techniques to $\textsc{Vec}(\boldsymbol{Z}_t)$. Under vector auto-regressive model of order 1 – VAR(1), we have

$$\textsc{Vec}(\boldsymbol{Z}_t) = \boldsymbol{\Phi}\,\textsc{Vec}(\boldsymbol{Z}_{t-1}) + \boldsymbol{u}_t,$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{dr\times dr}$ is the coefficient matrix of the VAR(1). Following the vector time series analysis (Tsay, 2014; Tsay and Chen, 2018), we obtain estimators $\widehat{\boldsymbol{\Phi}}$. A h-step forward prediction is given by

$$\widehat{\boldsymbol{Z}}_{t+h}^{VAR} = \textsc{Mat}\left(\widehat{\boldsymbol{\Phi}}^h \textsc{Vec}\left(\widehat{\boldsymbol{Z}}_t\right)\right). \tag{29}$$

To preserve the matrix structure intrinsic to $\boldsymbol{Z}_t$, we model $\{\boldsymbol{Z}_t\}_{1:T}$ as the matrix auto-regressive model of order 1 – MAR(1) (Yang et al., 2017). Mathematically,

$$\boldsymbol{Z}_t = \boldsymbol{\Phi}_R\,\boldsymbol{Z}_{t-1}\,\boldsymbol{\Phi}_C + \boldsymbol{U}_t,$$

where $\boldsymbol{\Phi}_R \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\Phi}_C \in \mathbb{R}^{r\times r}$ are row and column coefficient matrices, respectively. The covariance structure of the matrix white noise $\boldsymbol{U}_t$ is not restricted. Thus, $\mathbf{vec}(\boldsymbol{U}_t) \sim \mathcal{N}(\mathbf{0}, \Sigma_U)$ where Σ_U is an arbitrary covariance matrix. Matrix $\boldsymbol{\Phi}_R$ captures the auto-correlations between the spatial latent factors and $\boldsymbol{\Phi}_C$ captures the auto-correlations between the variable latent factors. Following the generalized iterative method proposed in Yang et al. (2017), we obtain estimators $\widehat{\boldsymbol{\Phi}}_R$ and $\widehat{\boldsymbol{\Phi}}_C$. A h-step forward prediction is given by

$$\widehat{\boldsymbol{Z}}_{t+h}^{MAR} = \widehat{\boldsymbol{\Phi}}_R^h\,\widehat{\boldsymbol{Z}}_t\,\widehat{\boldsymbol{\Phi}}_C^h. \tag{30}$$

Having an estimator $\widehat{\boldsymbol{Z}}_{t+h}$ from either vector AR(1) (29) or matrix AR(1) (30), we obtain the prediction for $\boldsymbol{y}_{t+h}(\boldsymbol{s})$ by

$$\widehat{\xi}_{t+h}(\boldsymbol{s}) = \widehat{Q}_B\,\widehat{\boldsymbol{Z}}'_{t+h}\,\widehat{\boldsymbol{q}}_a(\boldsymbol{s}), \tag{31}$$

where $\widehat{Q}_B$, $\widehat{\boldsymbol{Z}}_t$ and $\widehat{\boldsymbol{q}}_a(\boldsymbol{s})$ are estimated following procedures in Section 3.

The advantage of MAR(1) over VAR(1) is that the number of unknowns in $\boldsymbol{\Phi}_R \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\Phi}_C \in \mathbb{R}^{r\times r}$ is smaller than that in $\boldsymbol{\Phi} \in \mathbb{R}^{dr\times dr}$. This is especially important in high-

dimensional setting. Since the latent matrix factor $\boldsymbol{Z}_t$ is of low-dimension in our case, they have similar performance as shown in the simulation.

5 Asymptotic properties

In this section, we investigate the rates of convergence for the estimators under the setting that n, p and T all go to infinity while d and r are fixed and the factor structure does not change over time.

Assumption 5.1. ***Alpha-mixing.*** $\{\text{vec}[\boldsymbol{X}_t], t = 0, \pm 1, \pm 2, \cdots\}$ *is* α*-mixing. Specifically, for some* $\gamma > 2$*, the mixing coefficients satisfy the condition that*

$$\sum_{h=1}^{\infty} \alpha(h)^{1-2/\gamma} < \infty,$$

where $\alpha(h) = \sup_{\tau} \sup_{A \in \mathcal{F}_{-\infty}^{\tau}, B \in \mathcal{F}_{\tau+h}^{\infty}} |P(A \cap B) - P(A)P(B)|$ *and* $\mathcal{F}_{\tau}^{s}$ *is the* σ*-field generated by* $\{vec(\boldsymbol{X}_t) : \tau \leq t \leq s\}$.

Assumption 5.2. *Let* $X_{t,ij}$ *be the* ij*-th entry of* $\boldsymbol{X}_t$*. Then,* $E(\left|X_{t,ij}\right|^2) \leq C$ *for any* $i = 1, \ldots, d$*,* $j = 1, \ldots, r$ *and* $t = 1, \ldots, T$*, where* C *is a positive constant and* γ *is given in Condition 5.1.*

Assumption 5.1 requires the random vector $\text{vec}[\boldsymbol{X}_t]$ be α-mixing – weaker than stationarity. Each entry of covariance matrix $\text{Var}[\text{vec}[\boldsymbol{X}_t]]$ is bounded according to Assumption 5.2. There is no further requirement on the temporal dependence structure on $\boldsymbol{X}_t$, i.e., $\text{Cov}[\text{vec}(\boldsymbol{X}_{t_1}), \text{vec}(\boldsymbol{X}_{t_2})']$, $t_1 \neq t_2$. This is weaker than that required in Wang et al. (2019). The following three assumptions control the signal noise ratio. Matrix $\boldsymbol{\Xi}_t$ can be seen as the signal of the observation $\boldsymbol{Y}_t$ and $\boldsymbol{E}_t$ as the noise. Assumption 5.3 control the noise strength by bounding each entry of spatial covariance matrix of noise $\boldsymbol{E}_t$. The signal strength is measured jointly by the L_2-norm $\|\boldsymbol{A}\|_2^2$ and $\|\boldsymbol{B}\|_2^2$, which correspond to the spatial and variable strengthes, respectively.

Assumption 5.3. ***Noise strength.*** *Each entry of* $\text{Var}[\text{vec}[\boldsymbol{E}_t]]$ *remains bounded as* n *and* p *increase to infinity.*

Assumption 5.4. ***Variable factor strength.*** *There exists a constant* $\gamma \in [0,1]$ *such that* $\|\boldsymbol{B}\|_{min}^2 \asymp p^{1-\gamma} \asymp \|\boldsymbol{B}\|_2^2$ *as* p *goes to infinity and* r *is fixed.*

Assumption 5.5. ***Spatial factor strength.*** *For any partition* $\{\mathcal{S}_1, \mathcal{S}_2\}$ *of locations* $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, *we have* $\|\boldsymbol{A}_1\|_{min}^2 \asymp n_1 \asymp \|\boldsymbol{A}_1\|_2^2$ *and* $\|\boldsymbol{A}_2\|_{min}^2 \asymp n_2 \asymp \|\boldsymbol{A}_2\|_2^2$ *for any* $\boldsymbol{s} \in \mathcal{S}$, *where* n_1 *and* n_2 *are the number of locations in sets* $\mathcal{S}_1$ *and* $\mathcal{S}_2$ *respectively.*

This assumption is satisfied automatically under Assumption 5.6 with randomly sampled $\mathcal{S}_1$ and $\mathcal{S}_2$. Assumption 5.6 further guarantee the accuracy of sieve approximation of loading function $\boldsymbol{a}_j(\boldsymbol{s})$, $j = 1, 2, \ldots, d$.

Assumption 5.6. ***Loading functions belongs to Hölder class.*** *For* $j = 1, \ldots, d$, *the loading functions* $\boldsymbol{a}_j(\boldsymbol{s})$, $\boldsymbol{s} \in \mathcal{S} \in \mathbb{R}^2$ *belongs to a Hölder class* $\mathcal{A}_c^\kappa(\mathcal{S})$ *(κ-smooth) defined by*

$$\mathcal{A}_c^\kappa(\mathcal{S}) = \left\{ a \in \mathcal{C}^m(\mathcal{S}) : \sup_{[\eta] \le m} \sup_{\boldsymbol{s} \in \mathcal{S}} |D^\eta a(\boldsymbol{s})| \le c, \text{ and } \sup_{[\eta]=m} \sup_{\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}} \frac{|D^\eta a(\boldsymbol{u}) - D^\eta a(\boldsymbol{v})|}{\|\boldsymbol{u} - \boldsymbol{v}\|_2^\alpha} \le c \right\},$$

for some positive number c. *Here,* $\mathcal{C}^m(\mathcal{S})$ *is the space of all* m*-times continuously differentiable real-value functions on* $\mathcal{S}$. *The differential operator* D^η *is defined as* $D^\eta = \frac{\partial^{[\eta]}}{\partial s_1^{\eta_1} \partial s_2^{\eta_2}}$ *and* $[\eta] = \eta_1 + \eta_2$ *for non-negative integers* η_1 *and* η_2.

Theorem 5.7 and 5.8 present the error bounds for the estimated spatial loading spaces $\mathcal{M}(\boldsymbol{A}_l)$, $l = 1, 2$, on partitioned sampling locations and for estimated variable loading space $\mathcal{M}\left(\widehat{\boldsymbol{B}}\right)$, respectively. Asymptotically, the bounds are the similar to those derived under the time series settings in Wang et al. (2019) and Chen et al. (2019). Indeed, when we only consider the samples from discrete locations with spatial white noises, the estimation of model (16) is similar to that of the matrix-variate time series with temporal white noise.

Theorem 5.7. *Under Assumption 5.1-5.6 and* $p^\gamma T^{-1/2} = o(1)$, *we have*

$$\mathcal{D}\left(\mathcal{M}(\widehat{\boldsymbol{A}}_i), \mathcal{M}(\boldsymbol{A}_i)\right) = \left\|\widehat{\boldsymbol{Q}}_{A,l} - \boldsymbol{Q}_{A,l}\right\| = \mathcal{O}_p\left(\sqrt{n_1 n_2^{-1} p^\gamma + n_1^{-1} n_2 p^\gamma + p^{2\gamma}}\, T^{-1/2}\right).$$

If $n_1 \asymp n_2 \asymp n$, *we have*

$$\mathcal{D}\left(\mathcal{M}(\widehat{\boldsymbol{A}}_i), \mathcal{M}(\boldsymbol{A}_i)\right) = \left\|\widehat{\boldsymbol{Q}}_{A,l} - \boldsymbol{Q}_{A,l}\right\| = \mathcal{O}_p\left(p^{\gamma} T^{-1/2}\right).$$

Theorem 5.8. *Under Assumption 5.1-5.6 and* $p^{\gamma} T^{-1/2} = o(1)$*, we have*

$$\mathcal{D}\left(\mathcal{M}(\widehat{\boldsymbol{B}}), \mathcal{M}(\boldsymbol{B})\right) = \left\|\widehat{\boldsymbol{Q}}_{B} - \boldsymbol{Q}_{B}\right\| = \mathcal{O}_p\left(p^{\gamma} T^{-1/2}\right).$$

When p is fixed, the convergence rate of $\boldsymbol{A}_i$ and $\boldsymbol{B}$ are $\sqrt{T}$, $i = 1, 2$. If dimension p increases, the estimations of $\boldsymbol{A}_i$ and $\boldsymbol{B}$ become more difficult. The noise term is of order np. The signal contribute the accuracy of $\widehat{\boldsymbol{A}}_i$ and $\boldsymbol{B}$ with $np^{1-\gamma}$, which is affected by the variable strength γ. If γ is small (strong variable factor strength), the convergence speed of $\boldsymbol{A}_i$ and $\boldsymbol{B}$ is faster. Specifically, the convergence rate of $\boldsymbol{A}_i$ and $\boldsymbol{B}$ are not affected by n. The noise term and the signal contribution both have order n.

Theorem 5.9. *Under Assumption 5.1-5.6 and* $p^{\gamma} T^{-1/2} = \mathrm{o}(1)$*, if* $n_1 \asymp n2 \asymp n$*, then*

$$\frac{1}{\sqrt{np}} \left\|\widehat{\boldsymbol{\Xi}}_{it} - \boldsymbol{\Xi}_{it}\right\|_2 = \mathcal{O}_p\left(p^{\gamma/2} T^{-1/2} + n^{-1/2} p^{-1/2}\right),$$

for $i = 1, 2$*, and*

$$\frac{1}{\sqrt{np}} \left\|\widehat{\boldsymbol{\Xi}}_{t} - \boldsymbol{\Xi}_{t}\right\|_2 = \mathcal{O}_p\left(p^{\gamma/2} T^{-1/2} + n^{-1/2} p^{-1/2}\right)$$

Theorem 5.9 presents the error bound for estimated signal $\widehat{\boldsymbol{\Xi}}_{it}$ as in (26) for each partition and $\widehat{\boldsymbol{\Xi}}_t$ for all sampling locations. The error of estimated signal $\widehat{\boldsymbol{\Xi}}_{it}$ is contributed by the noise $\boldsymbol{E}_t$, and the estimation error for $\boldsymbol{Q}_A$ and $\boldsymbol{Q}_B$. In the proof of this theorem, we show that $p^{\gamma/2} T^{-1/2}$ comes from the estimation error for $\boldsymbol{Q}_A$ and $\boldsymbol{Q}_B$ in Theorem 5.7 and 5.8. Since we use the sample $\boldsymbol{Y}_t$ instead of $\boldsymbol{\Xi}_t$, $n^{-1/2} p^{-1/2}$ comes from the noise $\boldsymbol{E}_t$, which is a $p \times n$ matrix. Theorem 5.10 presents the error bond for re-estimated spatial loading space $\mathcal{M}\left(\widehat{\boldsymbol{A}}\right)$ from estimated $\widehat{\boldsymbol{\Xi}}_t$ and $\widehat{\boldsymbol{Q}}_B$ of the first step.

Theorem 5.10. *Under Assumption 5.1-5.6 and* $p^{\gamma} T^{-1/2} = \mathrm{o}(1)$*, if* $n_1 \asymp n2 \asymp n$*, then*

$$\left\|\widehat{\boldsymbol{Q}}_{A} - \boldsymbol{Q}_{A}\right\|_2 = \mathcal{O}_p\left(p^{\gamma} T^{-1/2} + n^{-1/2} p^{\gamma/2 - 1/2}\right).$$

Re-estimation introduces the noise error from $\boldsymbol{E}_t$. Comparing to the result in Theorem 5.7, the re-estimated loading matrix $\widehat{\boldsymbol{Q}}_A$ has an extra error term $n^{-1/2}p^{\gamma/2-1/2}$, which results from the noise error $n^{-1}p^{-1}$ of $\widehat{\boldsymbol{\Xi}}_t$ that appears in Theorem 5.9. Simulations in Section 6 show that the differences between the re-estimator and first estimator of $\boldsymbol{A}$ are also negligible with finite n, p, T.

Let $\Delta_{npT} = p^{\gamma}T^{-1}+n^{-1}p^{-1}$ represent the estimation error from the first-step estimation. Note that under the identification constraint that $\boldsymbol{Q}_A$ and $\boldsymbol{Q}_B$ are orthonormal matrices, $\|\boldsymbol{Z}_t\|$ is of order $np^{1-\gamma}$. Theorem 5.11 shows the normalized error bound of $\boldsymbol{Z}_t$.

Theorem 5.11. *Under Assumption 5.1-5.6, the estimator of rotated latent factor matrix $\boldsymbol{Z}_t$ satisfies*

$$\frac{1}{np}\left\|\widehat{\boldsymbol{Z}}_t - \boldsymbol{Z}_t\right\|_2^2 = \mathcal{O}_p\left(\Delta_{npT} + p^{\gamma}\Delta_{npT}^2\right).$$

Theorem 5.12 presents the space kriging error bound based on sieve approximated function $\widehat{\boldsymbol{A}}(\boldsymbol{s})$. In the proof of Theorem 5.12, we decompose the error of $\widehat{\xi}_t(\boldsymbol{s}_0)$ and show that it is dominated by three parts. $J_n^{-2\kappa}p^{-\gamma}$ is roughly the error of $\boldsymbol{q}_a(\boldsymbol{s}_0)$, which includes the sieve approximation error and estimation error. $\Delta_{npT}p^{\gamma} + \Delta_{npT}^2$ comes from the error of $\widehat{\boldsymbol{Z}_t}$, and $p^{\gamma}T^{-1}$ is the error of $\widehat{\boldsymbol{Q}}_B$.

Theorem 5.12. *Under Assumption 5.1-5.6, for a new site $\boldsymbol{s}_0 \in \mathcal{S}$*

$$\frac{1}{p}\left\|\widehat{\xi}_t(\boldsymbol{s}_0) - \xi_t(\boldsymbol{s}_0)\right\|_2^2 \;=\; \mathcal{O}_p\left(J_n^{-2\kappa}p^{-\gamma} + \Delta_{npT} + p^{\gamma}\Delta_{npT}^2 + p^{\gamma}T^{-1}\right).$$

6 Simulation

In this section we study the numerical performance of the proposed method on synthetic data sets. We let $\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n$ be drawn randomly from the uniform distribution on $[-1,1]^2$ and the observed data $\boldsymbol{y}_t(\boldsymbol{s})$ be generated according to Model (15):

$$\boldsymbol{y}_t(\boldsymbol{s}) = \xi_t(\boldsymbol{s}) + \varepsilon_t(\boldsymbol{s}) = \boldsymbol{B}\boldsymbol{X}_t'\boldsymbol{a}(\boldsymbol{s}) + \varepsilon_t(\boldsymbol{s}).$$

The dimensions of $\boldsymbol{X}_t$ are chosen to be $d = 3$, $r = 2$, and are fixed in all simulations. The latent factor $\boldsymbol{X}_t$ is generated from the Gaussian matrix time series (30):

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_R \boldsymbol{X}_{t-1} \boldsymbol{\Phi}_C + \boldsymbol{U}_t,$$

where $\boldsymbol{\Phi}_R = diag(0.7, 0.8, 0.9)$, $\boldsymbol{\Phi}_C = diag(0.8, 0.6)$ and the entries of $\boldsymbol{U}_t$ are white noise Gaussian process with mean $\mathbf{0}$ and covariance structure such that $\Sigma_U = \mathrm{Cov}\, vec(\boldsymbol{U}_t)$. Here we use $\Sigma_U = \boldsymbol{I}_{dr}$. Alternatively, we could use Kronecker product covariance structure $\Sigma_U = \Sigma_C \otimes \Sigma_R$ or arbitrary covariance matrix Σ_U. As shown in Yang et al. (2017) and from our own experiments, this setting does not affect much on the results.

The entries of $\boldsymbol{B}$ is independently sampled from the uniform distribution $\mathcal{U}(-1,1) \cdot p^{\gamma/2}$. The nugget process $\boldsymbol{\varepsilon}_t(\boldsymbol{s})$ are independent and normal with mean $\mathbf{0}$ and the covariance $(1 + s_1^2 + s_2^2)/2\sqrt{3} \cdot \boldsymbol{I}_p$. The basis functions $a_j(\boldsymbol{s})$'s are designed to be

$$a_1(\boldsymbol{s}) = (s_1 - s_2)/2, \quad a_2(\boldsymbol{s}) = \cos\left(\pi\sqrt{2(s_1^2 + s_2^2)}\right), \quad a_3(\boldsymbol{s}) = 1.5 s_1 s_2. \tag{32}$$

With the above generating model setting, the signal-noise-ratio of p-dimensional variable, which is defined as

$$SNR \equiv \frac{\int_{\boldsymbol{s}\in[-1,1]^2} Trace[Cov(\boldsymbol{\xi}_t(\boldsymbol{s}))]\, d\boldsymbol{s}}{\int_{\boldsymbol{s}\in[-1,1]^2} Trace[Cov(\boldsymbol{\varepsilon}_t(\boldsymbol{s}))]\, d\boldsymbol{s}} \approx 2.58.$$

We run 200 simulations for each combination of $n = 50, 100, 200, 400$, $p = 10, 20, 40$, and $T = 60, 120, 240$. With each simulation, we calculate $\widehat{d}$, $\widehat{r}$, $\widehat{\boldsymbol{A}}_1$, $\widehat{\boldsymbol{A}_2}$, $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Xi}}_t$, re-estimate $\widehat{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{\Xi}}_t$, then use $\widehat{\boldsymbol{A}}$ to get approximated $\widehat{a_j}(\boldsymbol{s})$ following the estimation procedure described in Section 3.

Table 1 presents the relative frequencies of estimated rank pairs over 200 simulations. The columns corresponding to the true rank pair $(3, 2)$ is highlighted.

Specifically, we show the estimated performance of spatial loading matrix $\boldsymbol{A}$, spatial-temporal covariance $\Sigma_{\xi,|t_1-t_2|}(\boldsymbol{u}, \boldsymbol{v})$ and latent factor $\boldsymbol{f}_t(\boldsymbol{s})$. Let $p = 40$, $n = 400$ and $T = 240$. Figure 1 presents the true surface of loading function $\boldsymbol{a}_1(\boldsymbol{s})$, $\boldsymbol{a}_2(\boldsymbol{s})$, $\boldsymbol{a}_3(\boldsymbol{s})$ in (32) on the top, and the fitted surface of $\widehat{\boldsymbol{a}}_1(\boldsymbol{s})$, $\widehat{\boldsymbol{a}}_2(\boldsymbol{s})$, $\widehat{\boldsymbol{a}}_3(\boldsymbol{s})$ on the bottom, which are all quite close with the true surface in shape. Figure 2 presents one example of the sample temporal covari-

ance $\widetilde{\Sigma}_{\xi,|t_1-t_2|}(\boldsymbol{s}_1,\boldsymbol{s}_2)$ (top three) and estimated temporal covariance $\widehat{\Sigma}_{\xi,|t_1-t_2|}(\boldsymbol{s}_1,\boldsymbol{s}_2)$ (bottom three) of $\Sigma_{\xi,|t_1-t_2|}(\boldsymbol{s}_1,\boldsymbol{s}_1)$ with time lag $t_1-t_2=0,1,2$ and $\boldsymbol{s}_1$ is randomly selected. Our proposed model and estimation method can duplicate the temporal dependence structure very well. Spatial covariance also shows the similar result. Figure 3 presents the true factor $\boldsymbol{f}_t(\boldsymbol{s})$ and the estimated factor $\widehat{\boldsymbol{f}}_t(\boldsymbol{s})$ by proposed method. We can see that they are very close.

The performance of correctly estimating the loading spaces are measured by the space distance between the estimated and true loading matrices $\widehat{\boldsymbol{A}}$ and $\boldsymbol{A}$, which is defined as

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

It can be shown that $\mathcal{D}(\mathcal{M}(\widehat{\boldsymbol{A}}),\mathcal{M}(\boldsymbol{A}))$ takes its value in $[0,1]$, it equals to 0 if and only if $\mathcal{M}(\widehat{\boldsymbol{A}})=\mathcal{M}(\boldsymbol{A})$, and equals to 1 if and only if $\mathcal{M}(\widehat{\boldsymbol{A}})\perp\mathcal{M}(\boldsymbol{A})$.

Figure 4 presents the box plot of the average space distance

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

and compare it with the box plot of space distance between re-estimated $\widehat{\boldsymbol{A}}$ and the truth $\boldsymbol{A}$. Figure 5 presents the box plot of the space distance between $\widehat{\boldsymbol{B}}$ and the truth $\boldsymbol{B}$.

Define the mean squared error of estimated signals $\widehat{\xi}$ as

$$MSE(\widehat{\xi})=\frac{1}{npT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\|\widehat{\xi}_t(\boldsymbol{s}_i)-\xi_t(\boldsymbol{s}_i)\right\|_2^2.$$

We compare the mean square error between first estimated $\widehat{\boldsymbol{\Xi}}_t$ defined in (26) and re-estimated $\widetilde{\boldsymbol{\Xi}}_t$ defined as

$$\widetilde{\boldsymbol{\Xi}}=\left[\widetilde{\boldsymbol{\Xi}}_1,\cdots,\widetilde{\boldsymbol{\Xi}}_T\right]=\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{B}}'.$$

The box plots of $MSE(\widehat{\xi})$ and $MSE(\widetilde{\xi})$ are in Figure 7. Re-estimated provides much more accurate estimate for $\xi_t(\boldsymbol{s}_j)$ than $\widetilde{\xi}_t(\boldsymbol{s}_j)$ does.

To demonstrate the performance of spatial prediction, we generate data at a set $\mathcal{S}_0$ of 50 new locations randomly sampled from $\mathcal{U}[-1,1]^2$. For each $t=1,\ldots,T$, we calculate the spatial prediction $\widehat{\boldsymbol{y}}_t(\cdot)=\widehat{\xi}_t(\cdot)$ defined in (28) for each location in $\mathcal{S}_0$. The mean squared

Table 1: Relative frequency of estimated rank pair $(\widehat{d},\widehat{r})$ over 200 simulations. The columns correspond to the true value pair $(3,2)$ are highlighted. Blank cell represents zero value.

$(\widehat{d},\widehat{r})$			$\gamma=0$				$\gamma=0.5$					
T	p	n	(3,2)	(3,1)	(2,2)	(1,2)	(3,2)	(3,1)	(2,2)	(2,1)	(1,2)	(1,1)
60	10	50	0.77	0.01	0.04	0.19	0.11	0.02	0.12	0.02	0.61	0.14
120	10	50	1.00			0.01	0.42		0.08		0.51	
240	10	50	1.00				0.91		0.01		0.09	
60	20	50	0.86		0.02	0.13	0.02		0.10		0.88	0.01
120	20	50	1.00				0.08		0.04		0.88	
240	20	50	1.00				0.49		0.01		0.50	
60	40	50	0.96		0.01	0.04	0.03		0.09		0.88	0.01
120	40	50	1.00				0.02		0.07		0.91	
240	40	50	1.00				0.32		0.01		0.68	
60	10	100	0.98		0.02		0.65	0.10	0.18	0.04	0.03	0.01
120	10	100	1.00				0.99	0.01	0.01			
240	10	100	1.00				1.00					
60	20	100	1.00				0.73		0.22		0.06	
120	20	100	1.00				0.97		0.04			
240	20	100	1.00				1.00					
60	40	100	1.00				0.72		0.24		0.05	
120	40	100	1.00				0.96		0.04			
240	40	100	1.00				1.00					
60	10	200	1.00				0.80	0.15	0.02	0.01	0.03	
120	10	200	1.00				1.00	0.01				
240	10	200	1.00				1.00					
60	20	200	1.00				0.94		0.02		0.04	
120	20	200	1.00				1.00					
240	20	200	1.00				1.00					
60	40	200	1.00				0.97		0.01		0.03	
120	40	200	1.00				1.00					
240	40	200	1.00				1.00					
60	10	400	1.00				0.89	0.10			0.02	
120	10	400	1.00				1.00	0.01				
240	10	400	1.00				1.00					
60	20	400	1.00				1.00				0.01	
120	20	400	1.00				1.00					
240	20	400	1.00				1.00					
60	40	400	1.00				1.00				0.01	
120	40	400	1.00				1.00					
240	40	400	1.00				1.00					

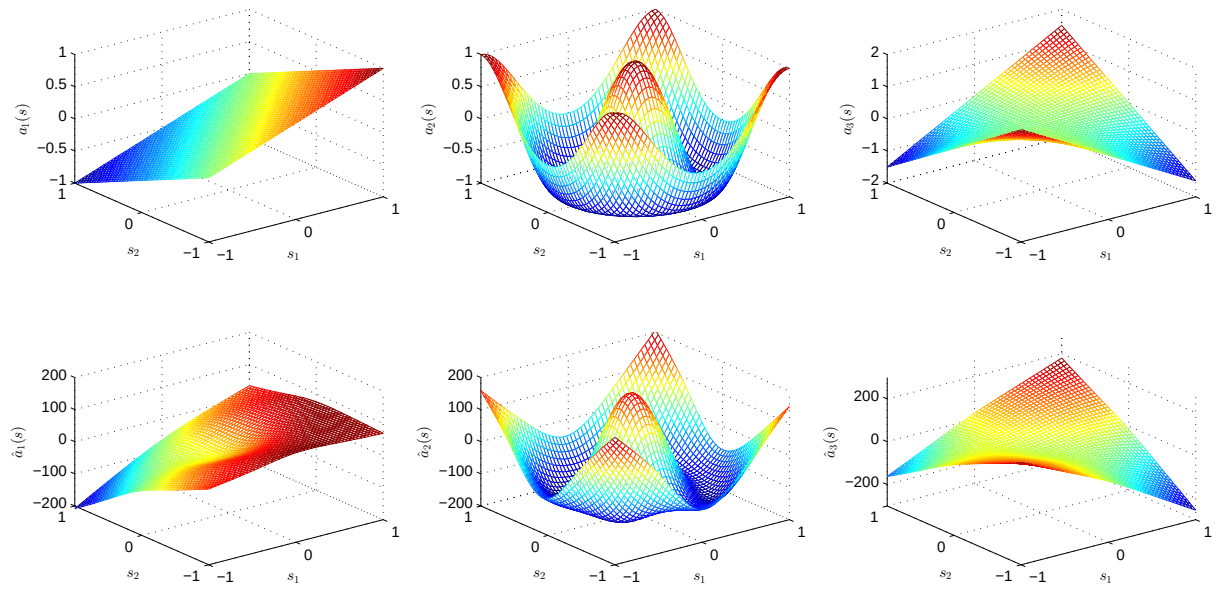


Figure 1: True surface of loading function $\boldsymbol{a}_j(\boldsymbol{s})$ (top three) and fitted surface of $\widehat{\boldsymbol{a}}_j(\boldsymbol{s})$ (bottom three), $j = 1, 2, 3$. Let $n = 400$, $p = 40$, and $T = 240$.

spatial prediction error is calculated as

$$MSPE(\widehat{\boldsymbol{y}}) = \frac{1}{50pT} \sum_{t=1}^{T} \sum_{s_0 \in \mathcal{S}_0} \left\| \widehat{\boldsymbol{y}}_t(\boldsymbol{s}_0) - \boldsymbol{\xi}_t(\boldsymbol{s}_0) \right\|_2^2.$$

To demonstrate the performance of temporal forecasting, we generate $\boldsymbol{X}_{T+h}$ according to the matrix time series (30) for $h = 1, 2$ and compute both the one-step-ahead and two-step-ahead predictions at time T. The mean square temporal prediction error is computed as

$$MSPE(\widehat{\boldsymbol{y}}_{T+h}) = \frac{1}{np} \sum_{j=1}^{n} \left\| \widehat{\boldsymbol{y}}_{T+h}(\boldsymbol{s}_j) - \boldsymbol{\xi}_{T+h}(\cdot) \right\|_2^2.$$

Figure 7 presents box-plots of the spatial prediction measured by average MSPE for 50 new locations. The results are based on 200 iterations. Figure 8 compares the MSPEs using matrix time series MAR(1) and vectorized time series VAR(1) estimates.

The means and standard errors of the MSPEs from 200 simulations for each model setting are reported in Table 6 in Appendix B. It also reports the means and standard errors of the MSPEs using matrix time series MAR(1) and vectorized time series VAR(1)

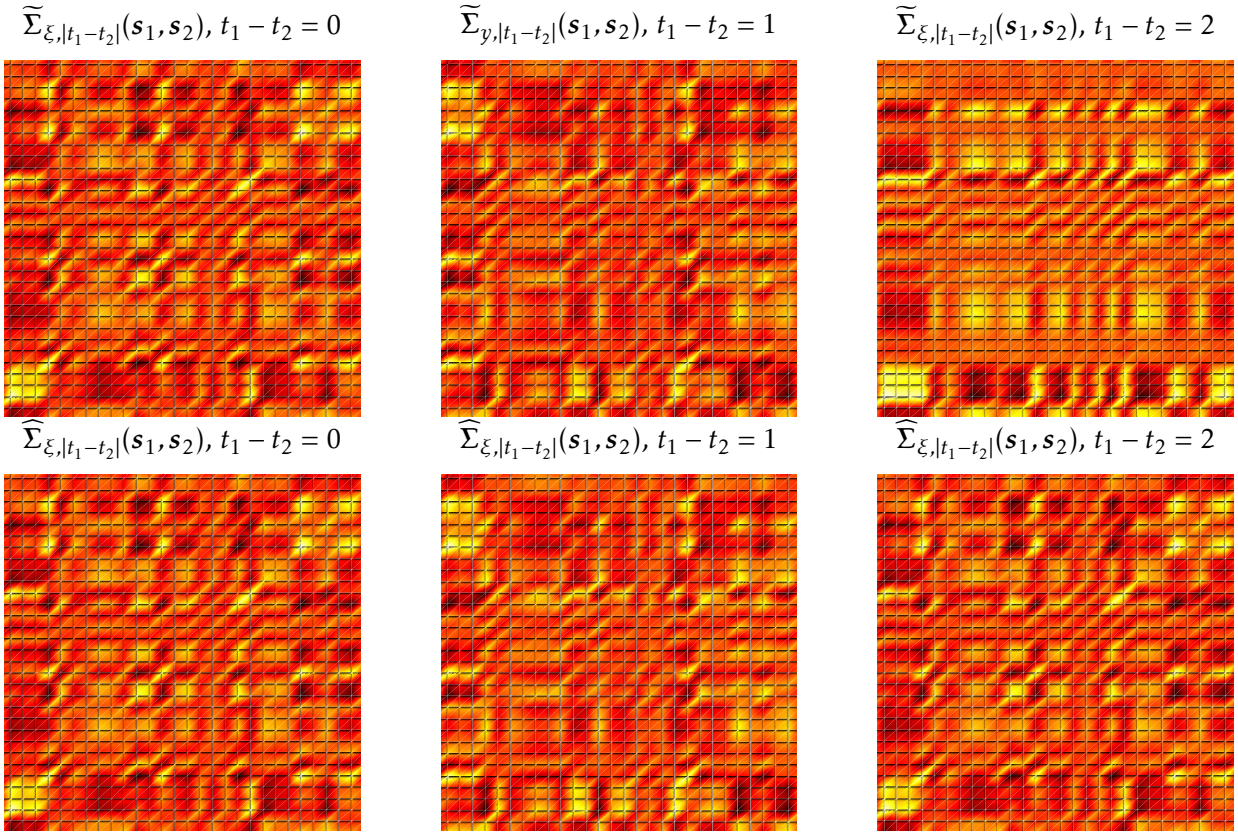


Figure 2: The sample temporal covariance $\widetilde{\Sigma}_{\xi,|t_1-t_2|}$ and estimated temporal covariance matrix $\widehat{\Sigma}_{\xi,|t_1-t_2|}$. Here, the time lag $t_1 - t_2 = 0, 1, 2$. $n = 400$, $p = 40$, and $T = 240$. s_1 is randomly generated from $[-1, 1]^2$.

estimates.

7 Real Data Applications

In this section, we apply the proposed method to the Comprehensive Climate Data Set (CCDS) – a collection of climate records of North America. The data set was compiled from five federal agencies sources by Lozano et al. (2009)[1]. It contains monthly observations of 17 climate variables spanning from 1990 to 2001 on a 2.5×2.5 degree grid for latitudes in $(30.475, 50.475)$, and longitudes in $(-119.75, -79.75)$. The total number of observation locations is 125 and the whole time series spans from January, 1991 to

[1] http://www-bcf.usc.edu/~liu32/data/NA-1990-2002-Monthly.csv

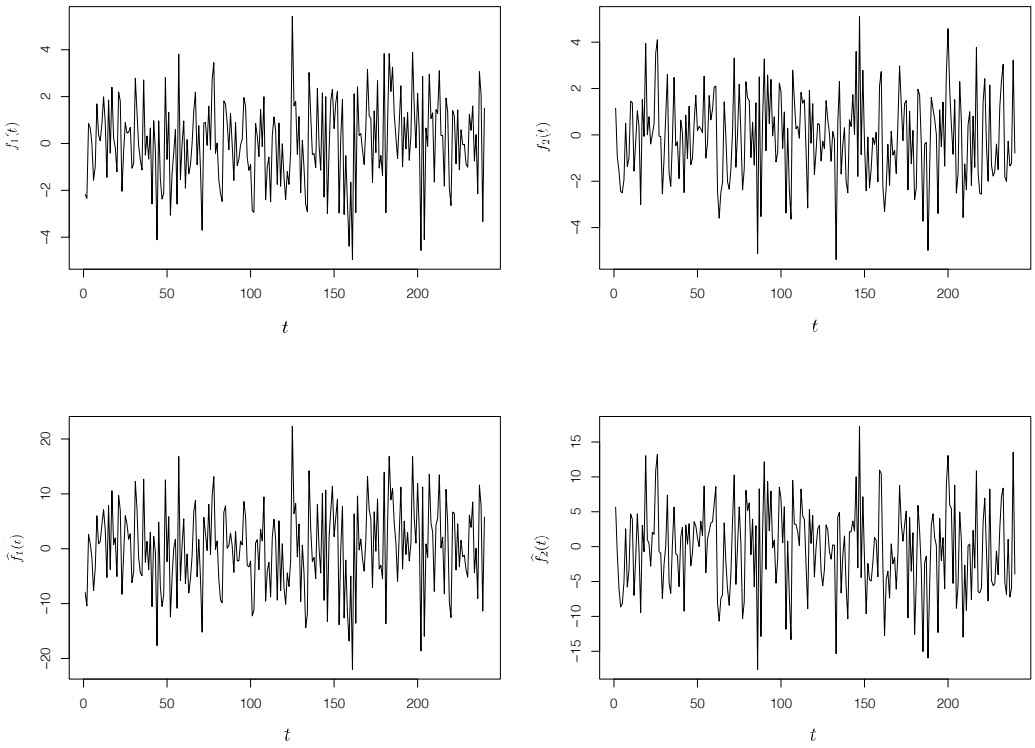


Figure 3: True factor $\boldsymbol{f}_t(\boldsymbol{s})$ (on the top) and the estimated factor $\widehat{\boldsymbol{f}}_t(\boldsymbol{s})$ (on the bottom) by proposed method. $n = 400$, $p = 40$, and $T = 240$. $\boldsymbol{s}$ is randomly generated from $[-1, 1]^2$

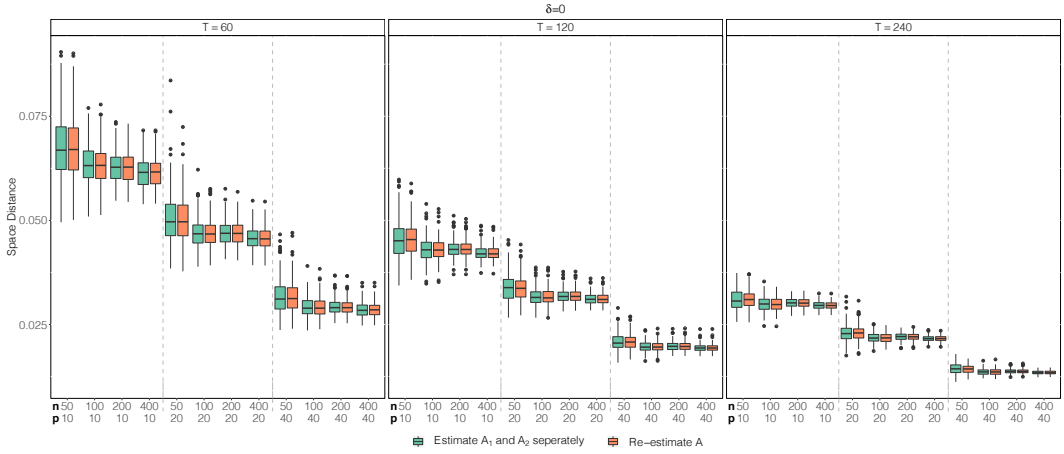


Figure 4: Box-plots of the estimation accuracy measured by $\mathcal{D}(\widehat{\boldsymbol{A}}(\boldsymbol{s}), \boldsymbol{A}(\boldsymbol{s}))$ for the case of orthogonal constraints. Gray boxes represent the average of $\mathcal{D}(\widehat{\boldsymbol{A}}_1(\boldsymbol{s}), \boldsymbol{A}_1(\boldsymbol{s}))$ and $\mathcal{D}(\widehat{\boldsymbol{A}}_2(\boldsymbol{s}), \boldsymbol{A}_2(\boldsymbol{s}))$. The results are based on 200 iterations. See Table 5 in Appendix B for mean and standard deviations of the spatial distance.

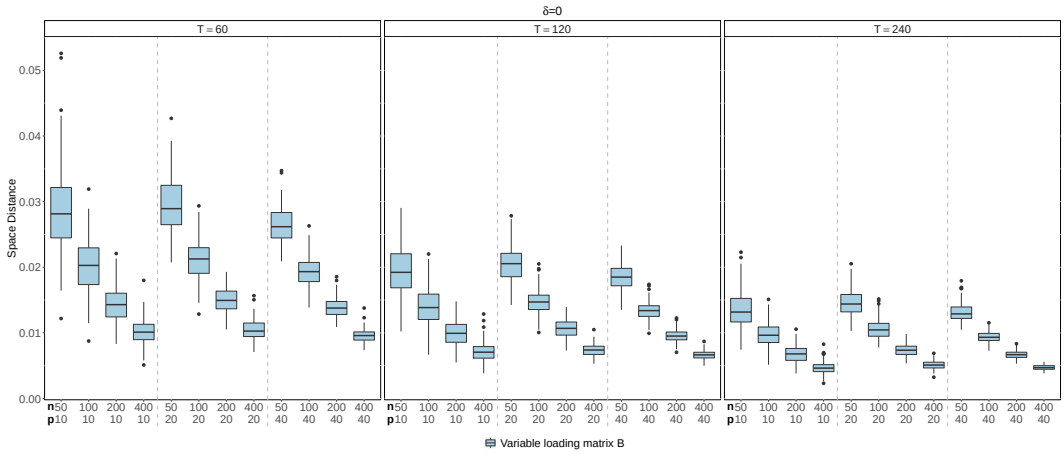


Figure 5: Box-plots of the estimation accuracy of variable loading matrix measured by $\mathcal{D}(\widehat{\boldsymbol{B}}, \boldsymbol{B})$. The results are based on 200 iterations. See Table 5 in Appendix B for mean and standard deviations of the spatial distance.

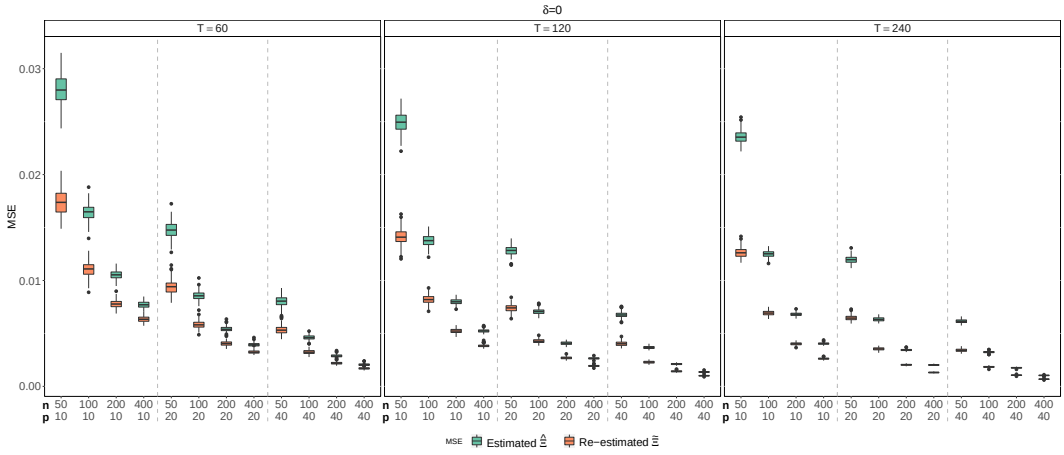


Figure 6: Box-plots of the estimation of signals MSE. Gray boxes represent the our procedure. The results are based on 200 iterations. See Table 5 in Appendix B for mean and standard deviations of the MSE.

December, 2002. We use a subset of the original data set because of the data quality. It contains measurements of 16 variables at all the locations range from January, 1992 to December, 2002. Thus, the dimensions our our data set are 125 (locations) × 16 (variables) × 132 (time points). Table 2 lists the variables used in our analysis. Detailed information about data is given in Lozano et al. (2009).

We first remove seasonal patterns in this data set by taking difference between the same month in consequent years. We then centralize and standardize each series to have

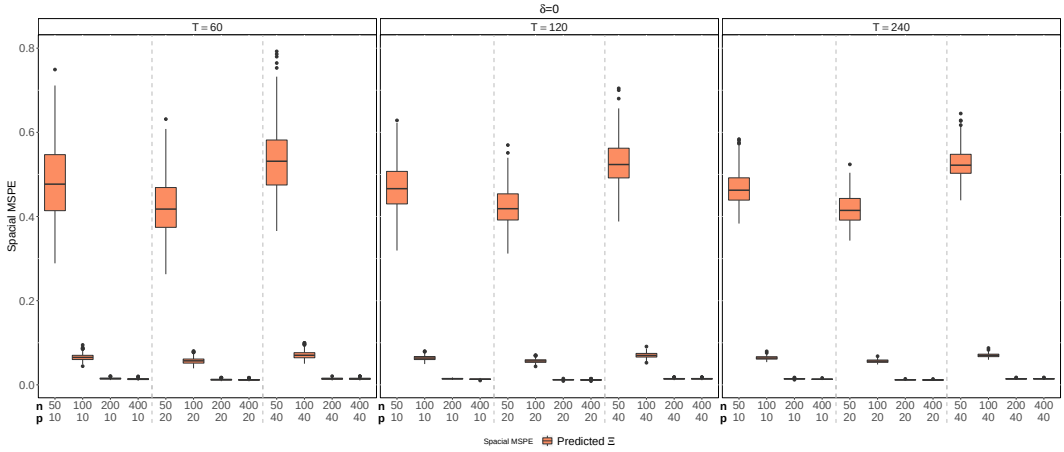


Figure 7: Box-plots of the spatial prediction measured by average MSPE for 50 new locations. Colored boxes represent the our model. The results are based on 200 iterations. See Table 6 in Appendix B for mean and standard deviations of the MSPE.

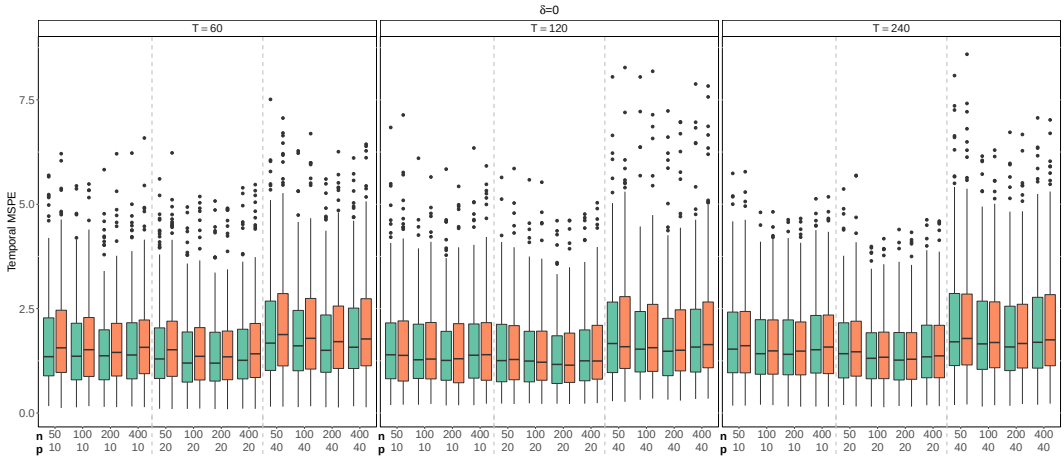


Figure 8: Box-plots of the one step ahead forecasting accuracy measured by MSPE. Gray boxes represent the MAR(1) model. The results are based on 200 iterations. See Table 6 in Appendix B for mean and standard deviations of the MSPE.

zero mean and unit variance before further investigation.

To estimate the latent dimensions, we combine the method of the scree plots and the eigen-ratio method. Figure 9 shows the scree plots and the eigen-ratio plots of the latent spatial and variable dimensions. Scree plots show that, in order to achieve 90% variance, we need to have latent spatial dimension $\widehat{d} = 6$ and latent variable dimension $\widehat{r} = 6$. Eigen-ratio (24) estimates latent spatial dimension $\widehat{d} = 12$ and latent variable dimension $\widehat{r} = 4$. Due to the dominance of the largest factors and weak signal in real data, the estimate

Table 2: Variables and data sources in the Comprehensive Climate Data Set (CCDS)

Variables (Short name)	Variable group	Type	Source
Methane (CH4)	CH_4	Greenhouse Gases	NOAA
Carbon-Dioxide (CO2)	CO_2		
Hydrogen (H2)	H_2		
Carbon-Monoxide (CO)	CO		
Temperature (TMP)	TMP	Climate	CRU
Temp Min (TMN)	TMP		
Temp Max (TMX)	TMP		
Precipitation (PRE)	PRE		
Vapor (VAP)	VAP		
Cloud Cover (CLD)	CLD		
Wet Days (WET)	WET		
Frost Days (FRS)	FRS		
Global Horizontal (GLO)	SOL	Solar Radiation	NCDC
Direct Normal (DIR)	SOL		
Global Extraterrestrial (ETR)	SOL		
Direct Extraterrestrial (ETRN)	SOL		

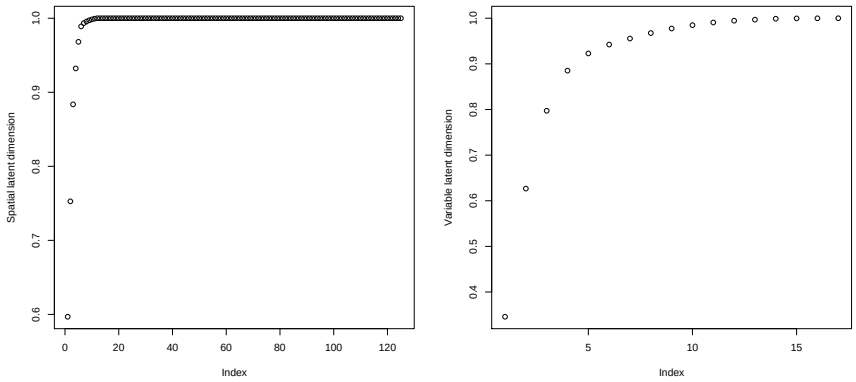
by (24) tends to be less useful than the one given by the scree plot. In the following, we choose $(\widehat{d},\widehat{r}) = (6,6)$ as the latent dimensions.

For kriging in space, we compare the performance kriging with kernel smoothing and prediction with functional $\boldsymbol{A}(\boldsymbol{s})$. We randomly pick a portion of locations (10%, 25% and 33% of all locations) and eliminate the measurements of all variables over the whole time span. Then, we produce the estimates for all variables of each timestamp. We repeat the procedure for 100 times. Table 3 report the average prediction RMSEs for all timestamps and 10 random sets of missing locations. It shows that the prediction by the proposed prediction with functional estimation $\boldsymbol{A}(\boldsymbol{s})$ performs much better than kriging with kernel smoothing.

We also compare the sample spatial-temporal covariance of the real data $\boldsymbol{y}_t$, and estimated spatial-temporal covariance of $\widehat{\xi_t}$ with the reduced rank structure in the proposed model with time lag $t_1 - t_2 = 0, 1, 2$ and two randomly selected location $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ in Figure 10. In this data, we only observe the $\boldsymbol{y}_t$, which include the noise $\boldsymbol{\varepsilon}_t$. It shows that the co-variance structure of the real data is largely preserved with the reduced rank approximation even when the dimension reduction is significant.

For temporal forecasting, we are interested in forecasting values in year 2001 and

(a) Scree plots



(b) Eigen-ratio plots

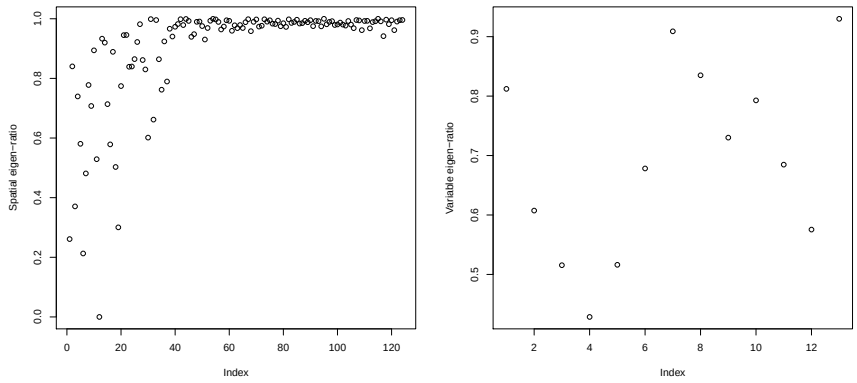


Figure 9: Latent dimensions

Table 3: Means of standard errors of MSPE by the proposed method for CCDS dataset. Results are based 100 simulations.

% Testing Sites	33%	25%	10%
# Training / Testing Sites	84 / 41	94 / 31	113 / 12
Kriging with kernel smoothing	0.580 (0.028)	0.578 (0.027)	0.572 (0.035)
Prediction with functional $\boldsymbol{A}(\boldsymbol{s})$	0.314 (0.011)	0.312 (0.009)	0.309 (0.013)

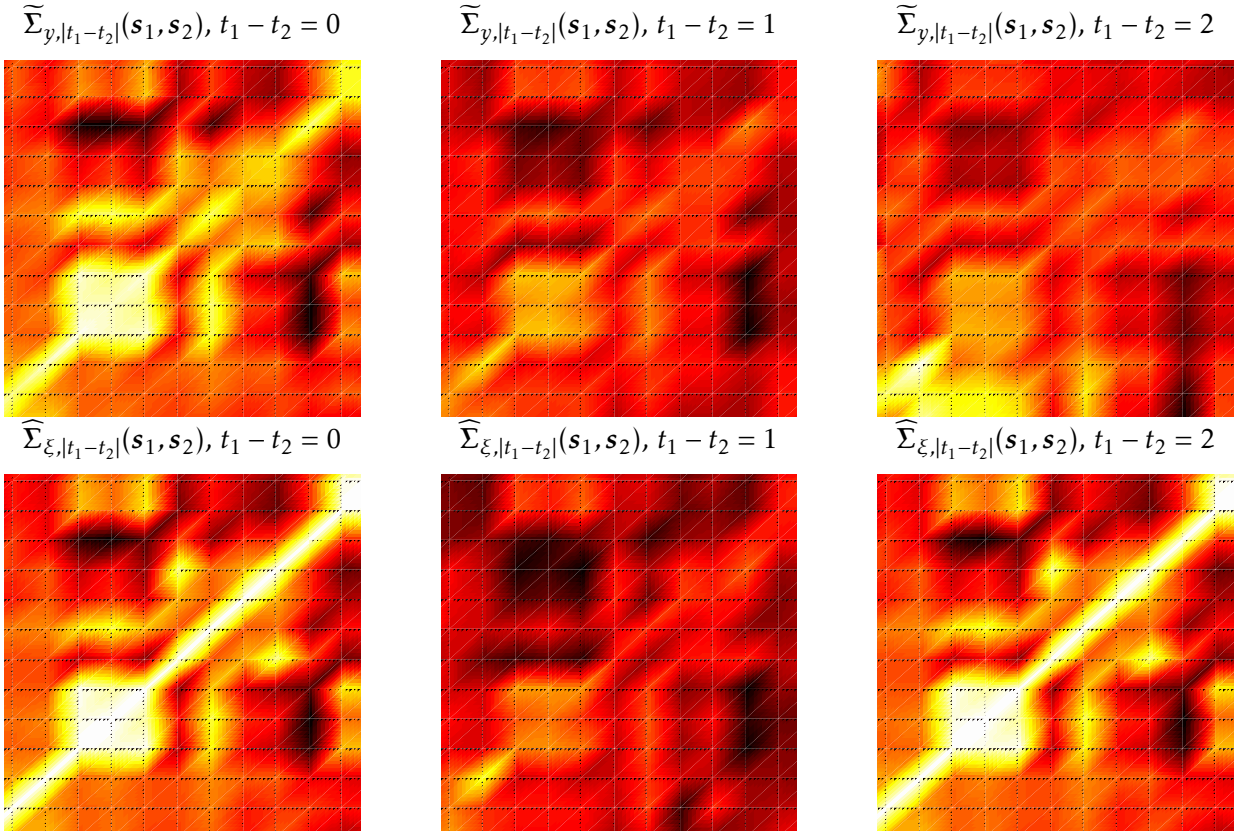


Figure 10: Heat map of the spacial-temporal covariances of $\boldsymbol{y}$ and $\widehat{\boldsymbol{\xi}}$ on the testing set. Top three heat maps corresponds to the sample spacial-temporal covariance of $\boldsymbol{y}_t(\boldsymbol{s})$. Bottom three heat maps corresponds to the estimated spacial-temporal covariance of $\widehat{\boldsymbol{\xi}}_t(\boldsymbol{s})$ with the reduced rank structure. The time lags are chosen $t_1 - t_2 = 0, 1, 2$. The testing sites is 10%. $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are randomly chosen in 12 test points. The covariances based on the low rank model are very close to the sample covariances of the full data.

2002. We experiment with two different length of training data – 5 and 9 years – respectively. For each setting, we estimate the loading matrices and factor matrices using the training data and make 1-step and 2-step prediction. We move forward with one month for both training and testing data and repeat the process until we reach 2002-12. For example with 5 training years, we start with estimation with 5 years training data from 1996-01 to 2000-12 and make 1-step prediction on 2001-01 and 2-step prediction on 2001-02. Then we move forward with one month – estimation with training data from 1996-02 to 2001-01 and prediction on the month 2001-02 and 2001-03. We repeat this process until the last estimation with 1998-11 to 2002-10 data and prediction on 2002-11

and 2002-12. So in total we have 23 predictions for 1-step and 2-step forecasting each for a given length of training set. With latent matrix time series, we predict each individual time series using *auto.arima* and *forecast* functions in the R *forecast* package. This is feasible because the latent factor matrix is low dimensional. With original matrix time series of 125×16 dimension, the computational cost is much higher. Table 4 reports the mean and standard deviation of the mean squared prediction errors. As shown by the results, temporal prediction is much harder than spatial prediction.

Table 4: Means (standard deviation) of MSPE by the proposed method for CCDS dataset.

Training Years	5	9
1-step MSPE	0.633 (0.181)	0.574 (0.141)
2-step MSPE	0.682 (0.225)	0.623 (0.190)
Time (min)	0.56 (0.04)	1.55 (0.20)

8 Summary

In this paper, we study the problem of large-scale multivariate spatial-temporal data analysis with a focus on dimension reduction and spatial/temporal forecasting. We propose a new class of multivariate spatial-temporal models that model spatial, temporal and multivariate dependencies simultaneously. This is made possible by an innovative combination of the multivariate factor analysis with the method of empirical orthogonal functions. For estimation, we assembled the observations from discrete spatial locations as a time series of matrices whose rows and columns correspond to sampling sites and variables, respectively. The matrix structure of observations is well preserved through the matrix factor model reformulation, while further incorporating the functional structure of the spatial process and dynamics of the latent matrix factor. We proposed methods of prediction over space and time based on the estimated latent structure. We established theoretical properties of the estimators and predictors. We validate the correctness and efficiency of our proposed method on both the synthetic and real application data sets.

For future work, we are interested in incorporating time-variant loading matrices to deal with possible structural changes. To improve the performance of spatial prediction, it is of great interest to investigate different ways to include spatial variograms. Since we use a two-step method to estimate the loading functions, possibly ways to estimate loading functions directly in one-step would also be an interesting direction for future research.

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

Bourotte, M., D. Allard, and E. Porcu (2016). A flexible class of non-separable cross-covariance functions for multivariate space–time data. *Spatial Statistics 18*, 125–146.

Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Bradley, J. R., S. H. Holan, C. K. Wikle, et al. (2015). Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *The Annals of Applied Statistics 9*(4), 1761–1791.

Brockwell, P. J. and R. A. Davis (2013). *Time series: theory and methods*. Springer Science & Business Media.

Carlin, B. P., S. Banerjee, et al. (2003). Hierarchical multivariate car models for spatio-temporally correlated survival data. *Bayesian Statistics 7*, 45–63.

Chang, J., B. Guo, and Q. Yao (2015). High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. *Journal of Econometrics 189*(2), 297–312.

Chen, E. Y., J. Fan, and E. Li (2019). Statistical inference for low rank matrix-variate data. *Working paper*.

Chen, E. Y., R. S. Tsay, and R. Chen (2019). Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association*.

Christensen, W. F. and Y. Amemiya (2001). Generalized shifted-factor analysis method for multivariate geo-referenced data. *Mathematical Geosciences 33*(7), 801.

Christensen, W. F. and Y. Amemiya (2002). Latent variable analysis of multivariate spatial data. *Journal of the American Statistical Association 97*(457), 302–317.

Christensen, W. F. and Y. Amemiya (2003). Modeling and prediction for multivariate spatial factor analysis. *Journal of Statistical Planning and Inference 115*(2), 543–564.

Congdon, P. (2004). A multivariate model for spatio-temporal health outcomes with an application to suicide mortality. *Geographical Analysis 36*(3), 234–258.

Cook, D., N. Cressie, J. Majure, and J. Symanzik (1994). Some dynamic graphics for spatial data (with multiple attributes) in a GIS. In *Compstat*, pp. 105–119. Springer.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Cressie, N. and G. Johannesson (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 209–226.

Cressie, N., T. Shi, and E. L. Kang (2010). Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics 19*(3), 724–745.

Cressie, N. and C. K. Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Daniels, M. J., Z. Zhou, and H. Zou (2006). Conditionally specified space-time models for multivariate processes. *Journal of Computational and Graphical Statistics 15*(1), 157–177.

Fan, J., Y. Liao, and W. Wang (2016). Projected principal component analysis in factor models. *The Annals of Statistics 44*(1), 219.

Fan, J., K. Wang, Y. Zhong, and Z. Zhu (2018). Robust high dimensional factor models with applications to statistical machine learning. *arXiv e-prints*, arXiv:1808.03889.

Fan, J. and Q. Yao (2005). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Series in Statistics. Springer.

Finley, A. O., H. Sang, S. Banerjee, and A. E. Gelfand (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis 53*(8), 2873–2884.

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

Hannachi, A., I. Jolliffe, and D. Stephenson (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society 27*(9), 1119–1152.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues*, pp. 37–56. Springer.

Huang, D., Q. Yao, and R. Zhang (2016). Krigings over space and time based on latent low-dimensional structures. *arXiv preprint arXiv:1609.06789*.

Huang, H.-C. and N. Cressie (1996). Spatio-temporal prediction of snow water equivalent using the kalman filter. *Computational Statistics & Data Analysis 22*(2), 159–175.

Kammann, E. and M. P. Wand (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 52*(1), 1–18.

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

Lam, C., Q. Yao, and N. Bathia (2011). Estimation of latent factors for high-dimensional time series. *Biometrika 98*(4), 901–18.

Lopes, H. F., E. Salazar, D. Gamerman, et al. (2008). Spatial dynamic factor analysis. *Bayesian Analysis 3*(4), 759–792.

Lozano, A. C., H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe (2009). Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 587–596. ACM.

Majure, J. J. and N. Cressie (1997). Dynamic graphics for exploring spatial dependence in multivariate spatial data. *Geographical Systems 4*(2), 131–158.

Merikoski, J. K. and R. Kumar (2004). Inequalities for spreads of matrix sums and products. *Applied Mathematics E-Notes 4*, 150–159.

Monahan, A. H., J. C. Fyfe, M. H. Ambaum, D. B. Stephenson, and G. R. North (2009). Empirical orthogonal functions: The medium is the message. *Journal of Climate 22*(24), 6501–6514.

Pettitt, A. N., I. S. Weir, and A. G. Hart (2002). A conditional autoregressive gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing 12*(4), 353–367.

Schumaker, L. (2007). *Spline functions: basic theory*. Cambridge University Press.

Stein, M. L. (2005). Space–time covariance functions. *Journal of the American Statistical Association 100*(469), 310–321.

Tsay, R. S. (2014). *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons.

Tsay, R. S. and R. Chen (2018). *Nonlinear time series analysis*, Volume 891. Wiley.

Tzala, E. and N. Best (2008). Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical Methods in Medical Research 17*(1), 97–118.

Tzeng, S. and H.-C. Huang (2018). Resolution adaptive fixed rank kriging. *Technometrics 60*(2), 198–208.

Von Storch, H. and F. W. Zwiers (2001). *Statistical analysis in climate research*. Cambridge university press.

Wang, D., X. Liu, and R. Chen (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics 208*(1), 231–248.

Wikle, C. K. (2010). *Low-rank representations for spatial processes*, pp. 107–118. CRC Press.

Wikle, C. K. and N. Cressie (1999). A dimension-reduced approach to space-time kalman filtering. *Biometrika 86*(4), 815–829.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences*. Academic Press.

Yang, D., X. Han, and R. Chen (2017). Autoregressive models for matrix-valued time series. *Working paper*.

Zhu, J., J. Eickhoff, and P. Yan (2005). Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. *Biometrics 61*(3), 674–683.