Factor Modelling for High-Dimensional Functional Time Series

Shaojun Guo[1], Xinghao Qiao[2], and Qingsong Wang[1]

[1]*Institute of Statistics and Big Data, Renmin University of China, P.R. China*

[2]*Department of Statistics, London School of Economics, U.K.*

Abstract

Many economic and scientific problems involve the analysis of high-dimensional functional time series, where the number of functional variables (p) diverges as the number of serially dependent observations (n) increases. In this paper, we present a novel functional factor model for high-dimensional functional time series that maintains and makes use of the functional and dynamic structure to achieve great dimension reduction and find the latent factor structure. To estimate the number of functional factors and the factor loadings, we propose a fully functional estimation procedure based on an eigenanalysis for a nonnegative definite matrix. Our proposal involves a weight matrix to improve the estimation efficiency and tackle the issue of heterogeneity, the rationality of which is illustrated by formulating the estimation from a novel regression perspective. Asymptotic properties of the proposed method are studied when p diverges at some polynomial rate as n increases. To provide a parsimonious model and enhance interpretability for near-zero factor loadings, we impose sparsity assumptions on the factor loading space and then develop a regularized estimation procedure with theoretical guarantees when p grows exponentially fast relative to n. Finally, we demonstrate that our proposed estimators significantly outperform the competing methods through both simulations and applications to a U.K. temperature dataset and a Japanese mortality dataset.

1 Introduction

Functional time series, which refers to a sequential collection of curves observed over time exhibiting serial dependence, has recently received a great deal of attention. Despite progress being made in this field, existing literature has focused on the estimation based on a single or fixed number of functional time series, see, e.g., Bosq (2000); Hörmann and Kokoszka (2010); Bathia et al. (2010); Panaretos and Tavakoli (2013); Hörmann et al. (2015); Aue et al. (2015); Li et al. (2020); Chen, Guo and Qiao (2020) and among many others.

With the rapid development in technology, datasets containing a large collection of functional time series are becoming increasingly available in various applications. Examples include intraday return density trajectories (Bathia et al.; 2010) for hundreds of stocks, annual temperature curves collected at a number of stations, daily energy consumption curves for thousands of households, age-specific mortality rates in different prefectures (Gao et al.; 2019), to list a few. Such data, which are referred to as high-dimensional functional time series, take the form of $\mathbf{Y}_t(\cdot) = \{Y_{t1}(\cdot), \ldots, Y_{tp}(\cdot)\}^{\mathrm{T}}$ $(t = 1, \ldots, n)$ defined on a compact interval $\mathcal{U}$, where the number of random functions (p) is comparable to, or even larger than the number of serially dependent observations (n). Under such scenario, not only p is large, but each $Y_{tj}(\cdot)$ is a functional object with serial dependence across observations, posing extra challenges. When jointly modelling the entire curve dynamics, it is of great interest and importance to explore the latent dynamic structure from a dimension-reduction viewpoint, while preserving the functional and time series structure. This motivates us to develop a functional factor model based on $\mathbf{Y}_1(\cdot), \ldots, \mathbf{Y}_n(\cdot)$.

In this paper, we deal with the factor modelling for high-dimensional functional time series, in which each functional datum $\mathbf{Y}_t(\cdot)$ arises as the sum of two unobservable components, one common $\mathbf{Z}_t(\cdot)$ and one idiosyncratic $\boldsymbol{\varepsilon}_t(\cdot)$. The co-movement of p functional time series

is assumed to be driven by a small number of functional factors and their inherited dynamic structures in $\mathbf{Z}_t(\cdot)$. Specifically, our proposed factor model admits the representation:

$$\mathbf{Y}_t(\cdot) = \mathbf{Z}_t(\cdot) + \boldsymbol{\varepsilon}_t(\cdot) = \mathbf{A}\mathbf{X}_t(\cdot) + \boldsymbol{\varepsilon}_t(\cdot), \quad t = 1, \ldots, n, \tag{1}$$

where $\mathbf{X}_t(\cdot) = \{X_{t1}(\cdot), \ldots, X_{tr}(\cdot)\}^{\mathrm{T}}$ is a set of latent functional factor time series with (unknown) number of factors $r < p$, $\mathbf{A} = (A_{ij})_{p \times r}$ is the factor loading matrix and the idiosyncratic component $\boldsymbol{\varepsilon}_t(\cdot)$ is a sequence of white noise process.

For multivariate scalar time series $\{\mathbf{Y}_t\}$, the factor model (1) degenerates to the classical dynamic factor model $\mathbf{Y}_t = \mathbf{A}\mathbf{X}_t + \boldsymbol{\varepsilon}_t$, where two different types of model assumptions are considered in econometrics and statistics literature. One type of models assumes that each common factor has non-trivial contribution to $\{\mathbf{Y}_t\}$ while the idiosyncratic noise $\{\boldsymbol{\varepsilon}_t\}$ is allowed to have weak cross-correlations and serial correlations (Chamberlain and Rothschild; 1983; Forni et al.; 2000; Bai and Ng; 2002; Stock and Watson; 2012; Fan et al.; 2013, 2016). However, the rigorous definition of the common and idiosyncratic components in those factor models can only be established asymptotically when the dimension of $\mathbf{Y}_t$ tends to infinity. Another type of models assumes that the common factors accommodate all dynamics of $\{\mathbf{Y}_t\}$, thus making the idiosyncratic component white noise with no serial correlations but allowing substantial contemporary cross-correlation among it; see Pena and Box (1987); Pan and Yao (2008); Lam et al. (2011); Lam and Yao (2012) and extensions to matrix-valued time series (Wang et al.; 2019; Chen, Tsay and Chen; 2020) and tensor-valued time series (Chen et al.; 2021). For functional time series, the factor modelling remains less touched in the literature. Hays et al. (2012) and Kokoszka et al. (2015) considered functional dynamic factor models for univariate functional time series. Gao et al. (2019) adopted factor modelling techniques to predict high-dimensional functional time series by fitting a factor model to estimated scores obtained via dynamic functional principal component analysis (Hörmann et al.; 2015). Following the framework in Bai and Ng (2002), a recent manuscript of Tavakoli et al. (2021) develops a high-dimensional functional factor model with scalar factors and functional factor loadings, which substantially differs from our model setup in (1).

Our paper is along the line of Lam and Yao (2012) in the model development. The

adopted white noise assumption is not the most ideal for the purpose to extract those asymptotically identifiable factors, however, since the white noise exhibits no serial correlations, the decomposition in (1) is unique, thus largely simplifying the tasks of model identification and inference. Moreover, unlike Bai and Ng (2002), we do not place any assumption on the covariance structure of the idiosyncratic noise. We propose to estimate both the number of factors and the factor loadings in terms of an eigenanalysis for a nonnegative definite matrix, which is the double integral and sum of weighted quadratic forms in autocovariance functions of observed curves from different time lags. In practical applications, it is common that many estimated factor loadings are close to zero especially when p is large in relation to n, This inspires us to add certain sparsity constraints on the factor loading space defined in terms of the factor loading matrix $\mathbf{A}$. Since functional sparsity patterns in autocovariance functions of $\mathbf{Y}_t(\cdot)$ inherits from the imposed sparsity structures in $\mathbf{A}$, we first propose Hilbert–Schmidt-norm based functional thresholding of entries of the sample auotocovariance functions, the consistencies of which are guaranteed under $\log p/n \to 0$. Furthermore, to recover sparse factor loading space and produce a parsimonious model with interpretable factors, we perform sparse principal component analysis (Vu and Lei; 2013) rather than standard eigenanalysis for our formed nonnegative definite matrix.

This paper makes useful contributions at multiple fronts. On the method side, our proposal has four main advantages. First, it involves a weight matrix to improve the estimation efficiency and tackle the issue of heterogeneity. We illustrate the rationality of its presence by formulating the estimation procedure based on a novel regression viewpoint. Our empirical results also demonstrate the uniform superiority of our method compared to the unweighted competitors. Furthermore, our strategy of incorporating the weight matrix is general and can be applied broadly to a range of dimension reduction problems, e.g., Bathia et al. (2010), Lam and Yao (2012) and Chang et al. (2018). Second, our estimation relies on the sample autocovariance functions of $\mathbf{Y}_t(\cdot)$. This not only gets rid of the impact from the noise $\boldsymbol{\varepsilon}_t(\cdot)$, but also makes the good use of the serial correlation information, which is most relevant in the time series modelling. Third, the double integral takes advantage of the functional

nature of the data by gathering the autocovariance information at each $(u, v) \in \mathcal{U}^2$ and then integrating over $\mathcal{U}^2$. By contrast, the estimation in Tavakoli et al. (2021) is based on the inner product between pairs of observed curves and hence ignores the covariance information. Moreover, our fully functional procedure avoids the information loss incurred by the dimension reduction step in Gao et al. (2019). Fourth, aided by the enforced sparsity, we enhance the model interpretability and enlarge the dimension of the feature space that our model can handle. Our regularized estimation procedure adopts techniques of functional thresholding and sparse principal component analysis, each of which leads to the estimation consistency and sparsity recovery at its own step.

On the theory side, we investigate convergence properties of relevant estimated terms under high-dimensional scaling, where p grows polynomially fast relative to n. With additional sparsity constraints and tail assumptions, we further provide convergence analysis of the regularized estimation in an ultra-high-dimensional regime, where p diverges at an exponential rate as n increases. Empirically, we demonstrate through an extensive set of simulations that our estimators substantially outperform the competitors. Moreover, we integrate our factor modelling framework into the challenging task of predicting high-dimensional functional time series and use a real data example to illustrate the superiority of our strategy over existing prediction methods.

The paper is set out as follows. In Section 2, we develop a fully functional procedure to estimate the proposed factor model and explain it from a regression perspective. The theoretical properties of the proposed estimators are investigated in Section 3. Under further sparsity assumptions, we develop the regularized estimation procedure and present the associated convergence analysis in Section 4. The finite-sample performance of our methods is illustrated through extensive simulation studies and two real data examples in Sections 5 and 6, respectively. Section 7 concludes the paper by discussing two extensions. We relegate all technical proofs to the Supplementary Material.

Notation. For any matrix $\mathbf{B}$, we let $\|\mathbf{B}\|_{\min} = \lambda_{\min}^{1/2}(\mathbf{B}^{\mathrm{T}}\mathbf{B})$ and $\|\mathbf{B}\| = \lambda_{\max}^{1/2}(\mathbf{B}^{\mathrm{T}}\mathbf{B})$. Let $L_2(\mathcal{U})$ denote the Hilbert space of squared integrable functions defined on $\mathcal{U}$ and $\otimes$ denotes

the Kronecker product. For $f \in L_2(\mathcal{U})$, we define its ℓ_2-norm by $\|f\| = \{\int f(u)^2 \mathrm{d}u\}^{1/2}$. For a Hilbert space $\mathbb{H} \subset L_2(\mathcal{U})$ and a matrix of bivariate functions $\mathbf{K} = \big(K_{ij}(\cdot, \cdot)\big)_{p \times q}$, we define the functional version of matrix ℓ_∞ norm by $\|\mathbf{K}\|_{\mathcal{S},\infty} = \max_{1 \leqslant i \leqslant p} \sum_{j=1}^{q} \|K_{ij}\|_{\mathcal{S}}$, where, for each $K_{ij} \in \mathbb{H} \otimes \mathbb{H}$, we denote its Hilbert–Schmidt norm by $\|K_{ij}\|_{\mathcal{S}} = \{\int \int K_{ij}(u, v)^2 \, \mathrm{d}u\mathrm{d}v\}^{1/2}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ or $b_n \gtrsim a_n$ if there exist a positive constant c such that $a_n/b_n \leqslant c$. We write $a_n \asymp b_n$ if and only if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously.

2 Methodology

2.1 Model setup and estimation procedure

We assume that the p-vector of functional time series $\mathbf{Y}_1(\cdot), \ldots, \mathbf{Y}_n(\cdot)$ satisfies the functional factor model in (1), where each $\mathbf{Y}_t(\cdot)$ is decomposed as the sum of two parts: a dynamic part driven by a r-vector of latent factor process $\mathbf{X}_t(\cdot)$ and a static component of white noise process $\boldsymbol{\varepsilon}_t(\cdot) = \{\varepsilon_{t1}(\cdot), \ldots, \varepsilon_{tp}(\cdot)\}^{\mathrm{T}}$ satisfying $E\{\boldsymbol{\varepsilon}_t(u)\} = \mathbf{0}$ and $\mathrm{Cov}\{\boldsymbol{\varepsilon}_{t+k}(u), \boldsymbol{\varepsilon}_t(v)\} = \mathbf{0}$ for any $u, v \in \mathcal{U}$ and $k \neq 0$. This formulation ensures that all linear dynamic structure of $\mathbf{Y}_t(\cdot)$ are determined by $\mathbf{X}_t(\cdot)$, while all white noise elements are absorbed into $\boldsymbol{\varepsilon}_t(\cdot)$. Note that $\{\mathbf{X}_t(\cdot)\}_{t=1}^n$ and $\{\boldsymbol{\varepsilon}_t(\cdot)\}_{t=1}^n$ are unobservable. We further assume that the rank of the factor loading matrix $\mathbf{A}$ is r. If this full rank condition is violated, model (1) can be expressed in terms of a lower-dimensional functional factor. When r is much smaller than p, we achieve an effective dimension reduction.

Remark 1. *In functional time series modelling, one crucial question is how to characterize the functional and dynamical structure. Our factor model* (1) *assumes that both are inherited from common factors* $\mathbf{X}_t(\cdot)$. *The recent manuscript of Tavakoli et al. (2021) provides another formulation of factor modeling for multivariate functional time series, treating the factor loading matrix as a functional object and the common factors as lower-dimensional scalar time series. Compared with* (1), *their model leads to excessive dimension reduction (from*

$p \times \infty$ to r) under the high-dimensional setting. We claim that our estimation framework can be applied to fit their model. See the detailed discussion in Section 2.3.

Similar to Lam and Yao (2012), there is an identifiable issue among $\mathbf{X}_t(\cdot)$ and $\mathbf{A}$, since model (1) remains unchanged if $\{\mathbf{A}, \mathbf{X}_t(\cdot)\}$ is replaced by $\{\mathbf{A}\boldsymbol{\Gamma}, \boldsymbol{\Gamma}^{-1}\mathbf{X}_t(\cdot)\}$ for any invertible matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times r}$. We may then assume that the columns of $\mathbf{A}$ are orthonormal, that is $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_r$ (an $r \times r$ identity matrix). With this constraint, $\mathbf{X}_t(\cdot)$ and $\mathbf{A}$ still can not be determined uniquely, but the linear space spanned by the columns of $\mathbf{A}$ (denoted by $\mathcal{C}(\mathbf{A})$) can. Hence we will focus on the estimation of the factor loading space $\mathcal{C}(\mathbf{A})$.

We next develop a fully functional procedure to estimate $\mathcal{C}(\mathbf{A})$ and the number of factors r. Under the assumption that $\{\mathbf{X}_t(\cdot)\}$ is weakly stationary, we define the following lag-k autocovariance functions:

$$
\begin{aligned}
\boldsymbol{\Sigma}_{yy}^{(k)}(u,v) &= \mathrm{Cov}\{\mathbf{Y}_{t+k}(u), \mathbf{Y}_t(v)\}, \quad \boldsymbol{\Sigma}_{xx}^{(k)}(u,v) = \mathrm{Cov}\{\mathbf{X}_{t+k}(u), \mathbf{X}_t(v)\}, \\
\boldsymbol{\Sigma}_{x\varepsilon}^{(k)}(u,v) &= \mathrm{Cov}\{\mathbf{X}_{t+k}(u), \boldsymbol{\varepsilon}_t(v)\}.
\end{aligned}
\tag{2}
$$

We further assume that the future white noise components are uncorrelated with the common factors up to the present. It then follows from (1) and (2) that

$$
\boldsymbol{\Sigma}_{yy}^{(k)}(u,v) = \mathbf{A}\boldsymbol{\Sigma}_{xx}^{(k)}(u,v)\mathbf{A}^{\mathrm{T}} + \mathbf{A}\boldsymbol{\Sigma}_{x\varepsilon}^{(k)}(u,v), \quad u,v \in \mathcal{U}, \ k \geqslant 1.
\tag{3}
$$

Given a prescribed and fixed integer $k_0 \geqslant 1$, we define a nonnegative definite matrix

$$
\mathbf{M} = \sum_{k=1}^{k_0} \int_{\mathcal{U}} \int_{\mathcal{U}} \boldsymbol{\Sigma}_{yy}^{(k)}(u,v)\mathbf{W}(u,v)\boldsymbol{\Sigma}_{yy}^{(k)}(u,v)^{\mathrm{T}} \, \mathrm{d}u\mathrm{d}v,
\tag{4}
$$

where the $p \times p$ weight matrix $\mathbf{W}(u,v)$ is nonnegative definite. In Section 2.2, we will suggest specific forms of $\mathbf{W}(u,v)$ and illustrate the rationality from a regression perspective. Let $\widetilde{\boldsymbol{\gamma}}_j$ be a unit eigenvector of $\mathbf{M}$ corresponding to the j-th largest eigenvalue of $\mathbf{M}$. We claim that the factor loading space $\mathcal{C}(\mathbf{A})$ can be recovered by $\mathrm{span}\{\widetilde{\boldsymbol{\gamma}}_1, \ldots, \widetilde{\boldsymbol{\gamma}}_r\}$ and $\mathrm{rank}(\mathbf{M}) = r$. To see this, let $\widetilde{\mathbf{A}}$ be a $p \times (p-r)$ orthogonal complement of matrix $\mathbf{A}$ such that $\widetilde{\mathbf{A}}^{\mathrm{T}}\mathbf{A} = \mathbf{0}$ and $\widetilde{\mathbf{A}}^{\mathrm{T}}\widetilde{\mathbf{A}} = \mathbf{I}_{p-r}$. It then follows from (3) and (4) that $\mathbf{M}\widetilde{\mathbf{A}} = \mathbf{0}$, which implies that the columns of $\widetilde{\mathbf{A}}$ are eigenvectors of $\mathbf{M}$ corresponding to $(p-r)$ zero eigenvalues and hence $\mathcal{C}(\mathbf{A})$ is spanned by the eigenvectors of $\mathbf{M}$ corresponding to r nonzero eigenvalues.

Remark 2. *(i) We take the double integral and the sum in the definition of* **M** *in (4) to accumulate the information of autocovariance functions as much as possible from each* $(u, v) \in \mathcal{U}^2$ *and from different time lags, whereas fixing at certain* (u, v) *or time lag may lead to spurious estimation results.*

(ii) Replacing $\mathbf{W}(u, v)$ *in (4) by its sample version* $\widehat{\mathbf{W}}(u, v)$, *we define*

$$\widetilde{\mathbf{M}} = \sum_{k=1}^{k_0} \int_{\mathcal{U}} \int_{\mathcal{U}} \mathbf{\Sigma}_{yy}^{(k)}(u, v) \widehat{\mathbf{W}}(u, v) \mathbf{\Sigma}_{yy}^{(k)}(u, v)^{\mathrm{T}} \, \mathrm{d}u \mathrm{d}v. \tag{5}$$

Let $\nu_1 \geqslant \cdots \geqslant \nu_p \geqslant 0$ *be eigenvalues of* $\widetilde{\mathbf{M}}$ *with the corresponding unit eigenvectors* $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$, *which differ from the eigenpairs of* **M**. *By the same arguments below (4), we can easily show that* $\mathcal{C}(\mathbf{A}) = \mathcal{C}(\mathbf{K}) = span\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r\}$ *with* $\mathbf{K} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r) \in \mathbb{R}^{p \times r}$ *and* $\nu_1 \geqslant \cdots \geqslant \nu_r > 0 = \nu_{r+1} = \cdots = \nu_p$.

(iii) For the choice of the maximum lag k_0, *we tend to select a small value in practice as the strongest autocorrelations usually appear at small lags while larger* k_0 *will add more noise to estimate* **M**.

To estimate $\mathcal{C}(\mathbf{A})$, we need to carry out an eigenanalysis of the natural estimator for **M**,

$$\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \int_{\mathcal{U}} \int_{\mathcal{U}} \widehat{\mathbf{\Sigma}}_{yy}^{(k)}(u, v) \widehat{\mathbf{W}}(u, v) \widehat{\mathbf{\Sigma}}_{yy}^{(k)}(u, v)^{\mathrm{T}} \, \mathrm{d}u \mathrm{d}v, \tag{6}$$

where

$$\widehat{\mathbf{\Sigma}}_{yy}^{(k)}(u, v) = \frac{1}{n-k} \sum_{t=1}^{n-k} \{\mathbf{Y}_{t+k}(u) - \bar{\mathbf{Y}}(u)\} \{\mathbf{Y}_t(v) - \bar{\mathbf{Y}}(v)\}^{\mathrm{T}}, \quad u, v \in \mathcal{U}, \tag{7}$$

is the estimator for $\mathbf{\Sigma}_{yy}^{(k)}(u, v)$ and $\bar{\mathbf{Y}}(\cdot) = n^{-1} \sum_{t=1}^n \mathbf{Y}_t(\cdot)$. Performing eigen-decomposition of $\widehat{\mathbf{M}}$ leads to estimated eigenvalues $\hat{\nu}_1 \geqslant \cdots \geqslant \hat{\nu}_p$ and the corresponding estimated eigenvectors $\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_p$. Let $\widehat{\mathbf{K}} = (\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_r)$. Then $\mathcal{C}(\widehat{\mathbf{K}}) = \mathrm{span}\{\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_r\}$ forms the estimate of $\mathcal{C}(\mathbf{A})$.

We have developed the estimation of model (1) assuming that the number of factors r is known or can be identified correctly. In practice, r is unknown and there is vast literature on the topic of determining it, see, e.g., Bai and Ng (2002); Onatski (2010); Lam and Yao (2012); Ahn and Horenstein (2013) and Fan et al. (2020), to quote a few. Here we take the

commonly adopted ratio-based estimator (Lam and Yao; 2012) for r as:

$$\hat{r} = \operatorname*{argmin}_{1 \leqslant j \leqslant c_r p} \hat{\nu}_{j+1}/\hat{\nu}_j, \tag{8}$$

where $c_r \in (0, 1)$ is a prespecified constant. In empirical studies, we take $c_r = 0.75$ to avoid the fluctuations due to the ratios of extreme small values.

2.2   Weight matrix

Under the nonnegative-definiteness constraint, we suggest the following weight matrix:

$$\mathbf{W}(u, v) = \mathbf{Q}\{\mathbf{Q}^{\mathrm{T}}\mathbf{\Sigma}_{yy}^{(0)}(v, v)\mathbf{Q}\}^{-1}\mathbf{Q}^{\mathrm{T}}, \quad u \in \mathcal{U}, \tag{9}$$

where $\mathbf{Q} \in \mathbb{R}^{p \times q}$ is a full-rank matrix and q is larger than r but much smaller than p. The entries of $\mathbf{Q}$ are independently generated from $N(0, 1)$. Then we can obtain the sample version of $\mathbf{W}(u, v)$ by

$$\widehat{\mathbf{W}}(u, v) = \mathbf{Q}\{\mathbf{Q}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{yy}^{(0)}(v, v)\mathbf{Q}\}^{-1}\mathbf{Q}^{\mathrm{T}}. \tag{10}$$

From a regression perspective, we next provide an intuitive explanation why the weight matrix takes the suggested form in (9). Despite the unobservable common factors $\mathbf{X}_t(\cdot)$, we can simply view (1) fixing at $u \in \mathcal{U}$ as a multiple linear regression with multivariate responses $\mathbf{Y}_t(u)$ and the coefficient matrix $\mathbf{A}$ to be estimated. To construct the covariate vector from observed data in a regression setup, we assume that there exists some orthogonal matrix $\mathbf{\Phi} \in \mathbb{R}^{q \times r}$ with $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} = \mathbf{I}_r$ such that $\widetilde{\mathbf{X}}_t(\cdot) = \mathbf{\Phi}\mathbf{X}_t(\cdot)$ can be represented by the linear combination of past observed curves with the addition of some random noise:

$$\widetilde{\mathbf{X}}_t(\cdot) = \mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(\cdot) + \mathbf{e}_{tk}(\cdot), \quad k = 1, \ldots, k_0,$$

where $\mathbf{e}_{tk}(\cdot)$ is the white noise process, independent of $\mathbf{Y}_{t-1}(\cdot), \ldots, \mathbf{Y}_{t-k_0}(\cdot)$. Then model (1) can be rewritten as

$$\mathbf{Y}_t(\cdot) = \mathbf{B}\widetilde{\mathbf{X}}_t(\cdot) + \boldsymbol{\varepsilon}_t(\cdot) = \mathbf{B}\mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(\cdot) + \widetilde{\boldsymbol{\varepsilon}}_{tk}(\cdot), \quad t = k + 1, \ldots, n, \tag{11}$$

where $\widetilde{\boldsymbol{\varepsilon}}_{tk}(\cdot) = \boldsymbol{\varepsilon}_t(\cdot) + \mathbf{B}\mathbf{e}_{tk}(\cdot)$ and $\mathbf{B} = \mathbf{A}\mathbf{\Phi}^{\mathrm{T}}$ is a $p \times q$ matrix satisfying $\operatorname{rank}(\mathbf{B}) = r$ and $\mathcal{C}(\mathbf{B}) = \mathcal{C}(\mathbf{A})$. Now we can treat (11) fixing at $u \in \mathcal{U}$ as a linear regression model with

observed covariate vectors $\{\mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(u)\}_{t=k+1}^{n}$ and unknown low-rank coefficient matrix $\mathbf{B}$. Based on (11) and the population moment equation $\mathrm{Cov}\{\tilde{\varepsilon}_{tk}(u), \mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(v)\} = 0$ for $u, v \in \mathcal{U}$, we can solve $\mathbf{B}$ by

$$\mathbf{B} = \mathrm{Cov}\Big\{\mathbf{Y}_t(u), \mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(v)\Big\}\mathrm{Cov}\Big\{\mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(u), \mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(v)\Big\}^{-1} \tag{12}$$

To simplify our subsequent derivation, we assume that $\mathbf{X}_t(\cdot)$ have mean zero and $\mathrm{Cov}\{\tilde{\varepsilon}_{tk}(u)\} = \mathbf{I}_p$ for all t, k and $u \in \mathcal{U}$. Replacing the covariance terms in (12) by their sample versions, we obtain the estimator for $\mathbf{B}$ by

$$\widehat{\mathbf{B}} = \Big\{ \sum_{t=k+1}^{n} \mathbf{Y}_t(u)\mathbf{Y}_{t-k}^{\mathrm{T}}(v)\mathbf{Q}\Big\}\Big\{ \sum_{t=k+1}^{n} \mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(u)\mathbf{Y}_{t-k}^{\mathrm{T}}(v)\mathbf{Q}\Big\}^{-1}. \tag{13}$$

Denote the j-th row vector of $\widehat{\mathbf{B}}$ by $\widehat{\mathbf{b}}_j \in \mathbb{R}^q$ with invertible covariance matrix $\mathrm{Cov}(\widehat{\mathbf{b}}_j) \in \mathbb{R}^{q\times q}$. Our target is to identify the rank of $\widehat{\mathbf{B}}$ and to recover the space spanned by columns of $\widehat{\mathbf{B}}$, which can be carried out in terms of an eigenanalysis for nonnegative-definite matrix $\widehat{\mathbf{B}}\{\mathrm{Cov}(\widehat{\mathbf{b}}_j)\}^{-1}\widehat{\mathbf{B}}^{\mathrm{T}} = \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^{\mathrm{T}}$ with $\widetilde{\mathbf{B}} = \widehat{\mathbf{B}}\{\mathrm{Cov}(\widehat{\mathbf{b}}_j)\}^{-1/2}$. The presence of $\mathrm{Cov}(\widehat{\mathbf{b}}_j)$ accounts for the heterogeneous variance-covariance effect and leads to the scaled low-rank coefficient matrix $\widetilde{\mathbf{B}}$ with its j-th row vector $\widetilde{\mathbf{b}}_j$ satisfying $\mathrm{Cov}(\widetilde{\mathbf{b}}_j) = \mathbf{I}_q$, $\mathrm{rank}(\widetilde{\mathbf{B}}) = \mathrm{rank}(\widehat{\mathbf{B}})$ and $\mathcal{C}(\widetilde{\mathbf{B}}) = \mathcal{C}(\widehat{\mathbf{B}})$. It follows from (13) that

$$\begin{aligned}
\mathrm{Cov}(\widehat{\mathbf{b}}_j) = &\Big\{ \sum_{t=k+1}^{n} \mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(v)\mathbf{Y}_{t-k}^{\mathrm{T}}(u)\mathbf{Q}\Big\}^{-1}\Big\{ \sum_{t=k+1}^{n} \mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(v)\mathbf{Y}_{t-k}^{\mathrm{T}}(v)\mathbf{Q}\Big\} \\
&\cdot \Big\{ \sum_{t=k+1}^{n} \mathbf{Q}^{\mathrm{T}}\mathbf{Y}_{t-k}(u)\mathbf{Y}_{t-k}^{\mathrm{T}}(v)\mathbf{Q}\Big\}^{-1}.
\end{aligned} \tag{14}$$

Combining (13) and (14) yields that

$$\widehat{\mathbf{B}}\big\{\mathrm{Cov}(\widehat{\mathbf{b}}_j)\big\}^{-1}\widehat{\mathbf{B}}^{\mathrm{T}} = \widehat{\mathbf{\Sigma}}_{yy}^{(k)}(u,v)\mathbf{Q}\{\mathbf{Q}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{yy}^{(0)}(v,v)\mathbf{Q}\}^{-1}\mathbf{Q}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{yy}^{(k)}(u,v)^{\mathrm{T}}. \tag{15}$$

We integrate the right side of (15) over $(u, v) \in \mathcal{U}^2$ and sum it over time lags $k = 1, \ldots, k_0$, thus obtaining (6) with the choice of $\widehat{\mathbf{W}}(u, v)$ in (10).

An alternative choice of the weight matrix is $\mathbf{W}_1(u, v) = \mathbf{I}_p$, where the homogeneous weights are assigned. As a consequence, we need to perform an eigenanalysis on the estimated

nonnegative-definite matrix:

$$\widehat{\mathbf{M}}_1 = \sum_{k=1}^{k_0} \int_{\mathcal{U}} \int_{\mathcal{U}} \widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}(u,v) \widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}(u,v)^{\mathrm{T}} \, \mathrm{d}u\mathrm{d}v, \tag{16}$$

which can be further simplified by integrating along the diagonal path $u = v \in \mathcal{U}$:

$$\widehat{\mathbf{M}}_2 = \sum_{k=1}^{k_0} \int_{\mathcal{U}} \widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}(u,u) \widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}(u,u)^{\mathrm{T}} \, \mathrm{d}u. \tag{17}$$

Remark 3. *It is worth noting that, without the double or single integral, the estimators in (16) and (17) coincide with the proposed method for multivariate scalar time series in Lam et al. (2011) and Lam and Yao (2012). Compared with the weighted estimator $\widehat{\mathbf{M}}$, the performance of the unweighted estimator $\widehat{\mathbf{M}}_1$ is expected to deteriorate especially for the heterogeneous case. Moreover, due to the loss of autocovariance information for $u \neq v$ incurred by the single integral, we expect that $\widehat{\mathbf{M}}_1$ outperforms $\widehat{\mathbf{M}}_2$. We will compare the sample performance of $\widehat{\mathbf{M}}$, $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$ in our simulations.*

2.3 Relationship to relevant work

Our estimation procedure can be extended easily to fit the factor model for multivariate scalar time series (Lam and Yao; 2012) and, with more efforts, to deal with the factor modelling for multivariate matrix-valued time series (Wang et al.; 2019; Chen et al.; 2021). Despite the integration step being no longer needed, we can still incorporate the suggested weight matrix to account for the heterogeneous effect.

It is also possible to apply our method to estimate the functional factor model (Tavakoli et al.; 2021) in the form of

$$\mathbf{Y}_t(u) = \mathbf{A}(u)\mathbf{X}_t + \boldsymbol{\varepsilon}_t(u), \quad u \in \mathcal{U}, \ t = 1, \ldots, n, \tag{18}$$

where $\mathbf{X}_t = (X_{t1}, \ldots, X_{tr})^{\mathrm{T}}$ is a set of latent factor time series and $\mathbf{A}(\cdot) = \{\mathbf{A}_{jl}(\cdot)\}_{p \times r}$ is the unknown functional factor loading matrix. To comply with our estimation framework, we assume that the idiosyncratic component $\boldsymbol{\varepsilon}_t(\cdot)$ is white noise. Given an orthonormal basis $\{\phi_i(\cdot)\}_{i=1}^{\infty}$, we expand the functional objects in (18) by $Y_{tj}(u) = \sum_{i=1}^{\infty} \xi_{tji} \phi_i(u)$, $A_{jl}(u) =$

$\sum_{i=1}^{\infty} a_{jli}\phi_i(u)$, $\varepsilon_{tj}(u) = \sum_{i=1}^{\infty} \epsilon_{tji}\phi_i(u)$ and hence (18) can be rewritten as

$$\xi_{tji} = \sum_{l=1}^{r} a_{jli}X_{tl} + \epsilon_{tji}, \quad i = 1, \ldots, \infty. \tag{19}$$

To simplify notation, we assume the same truncated dimension M across j and stack the basis coefficients ξ_{tji}, a_{jli} and ϵ_{tji} for $j = 1, \ldots, p$ and $i = 1, \ldots, M$ to pM-dimensional vectors $\boldsymbol{\xi}_t, \mathbf{a}_l$ and $\boldsymbol{\epsilon}_t$, respectively. As a result, (19) can be equivalently represented as the following factor model for pM-dimensional vector time series $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$:

$$\boldsymbol{\xi}_t = \boldsymbol{\Lambda}\mathbf{X}_t + \tilde{\boldsymbol{\epsilon}}_t, \tag{20}$$

where the factor loading matrix $\boldsymbol{\Lambda} = (\mathbf{a}_1, \ldots, \mathbf{a}_r) \in \mathbb{R}^{pM \times r}$ and the noise vector $\tilde{\boldsymbol{\epsilon}}_t \in \mathbb{R}^{pM}$ is decomposed as the sum of random and truncation errors. Then we can integrate our suggested weight matrix into the estimation of r and $\mathcal{C}(\boldsymbol{\Lambda})$ for the factor model (20). Hence the space spanned by columns of $\mathbf{A}(\cdot)$ can be recovered accordingly.

3 Theoretical properties

In this section, we study asymptotic properties of the proposed method under a high-dimensional regime, where p and n tend to infinity together and r is fixed. Before presenting the theoretical results, we impose some regularity conditions.

Condition 1. *(i) The latent functional factor process $\{\mathbf{X}_t(\cdot)\}$ is weakly stationary with $\mathbb{E}(\|X_{tj}\|^4) = O(1)$ for $j = 1, \ldots, r$; (ii) There exists at least one $k \in \{1, \ldots, k_0\}$ such that the rank of $\iint \boldsymbol{\Sigma}_{xx}^{(k)}(u,v)\boldsymbol{\Sigma}_{xx}^{(k)}(u,v)^{\mathrm{T}} \,\mathrm{d}u\mathrm{d}v$ is r.*

Condition 2. *(i) The idiosyncratic component $\{\boldsymbol{\varepsilon}_t(\cdot)\}$ is white noise with $\max_j \mathbb{E}(\|\varepsilon_{tj}\|^2) = O(1)$; (ii) $\inf_{v \in \mathcal{U}} \lambda_{\min}\{\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{(0)}(v,v)\}$ is bounded away from zero.*

Condition 3. *There exists some constant $\delta \in [0,1]$ such that $\|\mathbf{A}\|^2 \asymp p^{1-\delta} \asymp \|\mathbf{A}\|_{\min}^2$.*

Condition 4. *(i) For $k = 1, \ldots, k_0$, $\|\boldsymbol{\Sigma}_{x\varepsilon}^{(k)}\|_{\mathcal{S},\infty} = o(p^{(1-\delta)/2})$; (ii) $Cov\{\mathbf{X}_t(u), \boldsymbol{\varepsilon}_{t+k}(v)\} = \mathbf{0}$ for any $k \geq 0$ and $(u,v) \in \mathcal{U}^2$.*

Conditions 1 and 2 contain some standard finite moment assumptions in functional data analysis literature. The parameter δ in Condition 3 can be viewed as the strength of factors with smaller values yielding stronger factors. It measures the relative growth rate of the amount of information contained in $\mathbf{Z}_t(\cdot)$ as the dimension p increases, compared to that in $\boldsymbol{\varepsilon}_t(\cdot)$. When $\delta = 0$, Condition 3 corresponds to the pervasiveness assumption in Fan et al. (2013), which means that all factors are strong. When $\delta > 0$, the factors are termed as weak factors. We refer to Lam and Yao (2012) for more detailed discussion on the factor strength. Condition 4(i) requires that the correlation between $\mathbf{X}_{t+k}(\cdot)$ and $\boldsymbol{\varepsilon}_t(\cdot)$ is not too strong, while Condition 4(ii) assumes that the future idiosyncratic components are uncorrelated with the common factors up to the present.

Condition 5. *Let $\widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}(u,v) = \big\{\widehat{\Sigma}_{yy,ij}^{(k)}(u,v)\big\}_{p\times p}$ and $\boldsymbol{\Sigma}_{yy}^{(k)}(u,v) = \big\{\Sigma_{yy,ij}^{(k)}(u,v)\big\}_{p\times p}$ for $u,v \in$ $\mathcal{U}$. Then (i) For $k = 1,\ldots,k_0$, $\max_{i,j}\mathbb{E}\big\{\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}\big\} = O(n^{-1/2})$; (ii) $\boldsymbol{\Sigma}_{yy}^{(0)}(u,v)$ is Lipschitz-continuous over $(u,v) \in \mathcal{U}^2$; (iii) $\max_{i,j}\mathbb{E}\big\{\sup_{(u,v\in\mathcal{U}^2)}\big|\widehat{\Sigma}_{yy,ij}^{(0)}(u,v) - \Sigma_{yy,ij}^{(0)}(u,v)\big|\big\} = O\big\{n^{-1/2}(\log n)^{1/2}\big\}$.*

There are several sufficient conditions that have been commonly imposed in functional time series literature when p is fixed and can lead to the result in Condition 5(i), which further implies the standard root-n rate for entrywise sample autocovariance functions, i.e., $\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}} = O_P(n^{-1/2})$. The key requirement to establish the consistency is to control the temporal dependence in $\{Y_{tj}(\cdot)\}$ for $j = 1,\ldots,p$. Examples include strong mixing conditions (Bosq; 2000; Chen, Guo and Qiao; 2020), cumulant mixing conditions (Panaretos and Tavakoli; 2013) and L^q-m-approximability (Hörmann and Kokoszka; 2010; Hörmann et al.; 2015). With the additional tail assumption and Lipschitz-continuity in Condition 5(ii), it is not difficult to apply the partition technique that reduces the problem from supremum over $\mathcal{U}^2$ to the maximum over a grid of pairs, and hence the uniform convergence rate of $n^{-1/2}(\log n)^{1/2}$ in Condition 5(iii) can be achieved. See the same uniform rate with detailed proof under an independent setting in Qiao et al. (2020).

Now we are ready to present theorems about the rates of convergence for estimators of the factor loading space $\mathcal{C}(\mathbf{K}) = \mathcal{C}(\mathbf{A})$ and the eigenvalues $\{\nu_j\}_{j=1}^p$. To measure the accuracy

in estimating $\mathcal{C}(\mathbf{K})$, we use the metric of the distance between $\mathcal{C}(\mathbf{K})$ and $\mathcal{C}(\widehat{\mathbf{K}})$. For two orthogonal matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ of dimensions $p \times r_1$ and $p \times r_2$, respectively, we define

$$\mathcal{D}\Big(\mathcal{C}(\mathbf{K}_1), \mathcal{C}(\mathbf{K}_2)\Big) = \Big\{1 - \frac{1}{\max(r_1, r_2)} \mathrm{tr}(\mathbf{K}_1\mathbf{K}_1^{\mathrm{T}}\mathbf{K}_2\mathbf{K}_2^{\mathrm{T}})\Big\}^{1/2}.$$

This distance ranges between 0 and 1. It equals 0 if and only if $\mathcal{C}(\mathbf{K}_1) = \mathcal{C}(\mathbf{K}_2)$, and 1 if and only if $\mathbf{K}_1$ and $\mathbf{K}_2$ are orthogonal. See also Pan and Yao (2008) and Chang et al. (2018).

Theorem 1. *Let Conditions 1–5 hold and $p^\delta n^{-1/2} = o(1)$. Then as $p, n \to \infty$, it holds that:*

$$\mathcal{D}\Big(\mathcal{C}(\mathbf{K}), \mathcal{C}(\widehat{\mathbf{K}})\Big) = O_P(p^\delta n^{-1/2}).$$

Remark 4. *Theorem 1 implies that, as the factors become stronger with smaller δ, we obtain more efficient estimator for the factor loading space. When the dimension p is fixed, the $\sqrt{n}$ rate is attained. When all factors are strong (i.e., $\delta = 0$), $\sqrt{n}$ rate still retains, since the signal is as strong as the noise and hence enhancing p will not affect the estimation efficiency, circumventing the phenomenon of "curse of dimensionality". For weak factors (i.e., $\delta > 0$), the noise increases faster than the signal, and the increase in p will result in a slower convergence rate.*

Theorem 2. *Let Conditions 1–5 hold and $p^\delta n^{-1/2} = o(1)$. Then as $p, n \to \infty$, the following assertions hold:*

(i) $|\hat{\nu}_j - \nu_j| = O_P(p^{2-\delta}n^{-1/2})$ for $j = 1, \ldots, r$, and

   $\hat{\nu}_j = O_P(p^2 n^{-1})$ for $j = r+1, \ldots, p$;

(ii) $\hat{\nu}_{j+1}/\hat{\nu}_j \asymp 1$ for $j = 1, \ldots, r-1$, and $\hat{\nu}_{r+1}/\hat{\nu}_r = O_P(p^{2\delta}n^{-1})$.

Remark 5. *We observe from part (i) of Theorem 2 that estimators for nonzero eigenvalues converge at a slower rate than those for zero eigenvalues. Part (ii) of Theorem 2 implies that the eigen-ratio $\hat{\nu}_{j+1}/\hat{\nu}_j$ will drop steeply at $j = r$, thus providing partial theoretical support for the proposed ratio-based estimator $\hat{r}$ in (8). When all factors are strong (i.e., $\delta = 0$), $\hat{\nu}_{r+1}/\hat{\nu}_r = O_P(n^{-1})$, suggesting that $\hat{r}$ may not suffer from the increase in p. In fact, these results are consistent to those established in Lam and Yao (2012). In a similar fashion, we*

can obtain improved rates as p increases under additional assumptions. Although we do not theoretically pursue such "blessing of dimensionality" property here, our simulation results lend empirical support for this phenomenon.

4 Sparse factor model

Despite the phenomenon of "curve of dimensionality" being avoided when all factors are strong, our method does not guarantee a parsimonious and interpretable model in the presence of weak factors. In real applications, it is quite common that many estimated factor loadings are close to zero especially when p is large, see the gene expression study in Carvalho et al. (2008) and the sea surface air pressure records example in Lam and Yao (2012). Such phenomenon motivates us to propose a sparse factor model by imposing sparsity assumptions on the factor loading space $\mathcal{C}(\mathbf{A})$ of (1). In this section, we target to develop the regularized estimation under sparsity constraints, which not only leads to the enhanced intepretability in practice, but also theoretically enlarges the dimension of the feature space that our proposed sparse factor model can handle compared with the nonsparse factor model with weak factors.

We consider two complementary notions of subspace sparsity defined in terms of the factor loading matrix $\mathbf{A}$: row sparsity and column sparsity, which are consistent with the definitions in Vu and Lei (2013). Intuitively, the row sparsity entails that only a small subset of components in $\mathbf{Y}_t(\cdot)$ are driven by the common factors $\mathbf{X}_t(\cdot)$, thus making the row sparse factor loading space generated by a small number of variables, independent of the choice of the basis. The column sparsity, on the other hand, corresponds to the case that each common factor $X_{tj}(\cdot)$ has impact on only a small fraction of components of $\mathbf{Y}_t(\cdot)$ and hence the column sparse factor loading space has orthonormal basis consisting of sparse vectors. We begin by introducing parameter spaces of "approximately column sparse" and "approximately row sparse" factor loading matrices respectively defined in Conditions 6 and 6′ below. Let $\mathbf{a}_i$ be the i-th row vector of $\mathbf{A} = (A_{ij})_{p \times r}$.

Condition 6 (Column sparsity). *The matrix* $\mathbf{A}$ *is from the class*

$$\mathcal{V}(\tau, c_1(p), L) = \left\{ \mathbf{A} : \max_{1 \leqslant j \leqslant r} \sum_{i=1}^{p} |A_{ij}|^\tau \leqslant c_1(p), \max_{1 \leqslant i \leqslant p, 1 \leqslant j \leqslant r} |A_{ij}| \leqslant L \right\}, \quad \tau \in [0, 1).$$

Condition 6' (Row sparsity). *The matrix* $\mathbf{A}$ *is from the class*

$$\mathcal{V}^*(\tau, c_1(p), L) = \left\{ \mathbf{A} : \sum_{i=1}^{p} \|\mathbf{a}_i\|^\tau \leqslant c_1(p), \max_{1 \leqslant i \leqslant p, 1 \leqslant j \leqslant r} |A_{ij}| \leqslant L \right\}, \quad \tau \in [0, 1).$$

The parameters $c_1(p)$ and τ together control the column or row sparsity in the factor loading matrix $\mathbf{A}$. In the special case of $\tau = 0$, $\mathcal{V}(0, c_1(p), L)$ and $\mathcal{V}^*(0, c_1(p), L)$ correspond to the truly sparse situations, in which $\mathbf{A}$ has at most $c_1(p)$ nonzero entries on each column under Condition 6 or $c_1(p)$ nonzero row vectors under Condition 6'. By Hölder's inequality, it is easy to check that $\mathcal{V}^*(\tau, c_1(p), L) \subseteq \mathcal{V}(\tau, c_{r,\tau} c_1(p), L)$, where $c_{r,\tau} = 1$ for $\tau = 0$ and $r^{1-\tau/2}$ for $\tau \in (0, 1)$. Provided that r is fixed, our subsequent analysis will focus on the class of column sparse factor loading matrices, $\mathcal{V}(\tau, c_1(p), L)$.

Our estimation procedure in Section 2 is developed based on the estimation of $\Sigma_{yy}^{(k)}$'s. When p grows larger than $n^{1/2}$, it follows from Condition 5 that the sample autocovariance functions $\widehat{\Sigma}_{yy}^{(k)}$'s are no longer consistent. Nevertheless, under the sparse factor model setup, the decomposition of $\Sigma_{yy}^{(k)}$ in (3) suggests that our enforced column sparsity in $\mathbf{A}$ is inherited by the functional sparsity structure in $\Sigma_{yy}^{(k)}$, thus enabling us to possibly construct the threhsolding-based estimator for $\Sigma_{yy}^{(k)}$ to ensure the consistency. Before introducing the estimator, we slightly modify Condition 4 to accommodate the sparse setting and present a lemma that reveals the functional sparsity pattern in $\Sigma_{yy}^{(k)}$.

Condition 4'. *(i) For* $k = 1, \ldots, k_0$, *let* $\Sigma_{x\varepsilon}^{(k)}(u, v) = \{\Sigma_{x\varepsilon,ij}^{(k)}(u, v)\}_{p \times p}$ *for* $u, v \in \mathcal{U}$ *and* $\max_{1 \leqslant l \leqslant r} \sum_{j=1}^{p} \|\Sigma_{x\varepsilon,lj}^{(k)}\|_{\mathcal{S}}^\tau \leqslant c_2(p)$; *(ii)* $Cov\{\mathbf{X}_t(u), \varepsilon_{t+k}(v)\} = 0$ *for any integer* $k \geqslant 0$ *and* $u, v \in \mathcal{U}$.

Condition 4'(i) requires relatively weak correlations between $\mathbf{X}_{t+k}(\cdot)$ and $\varepsilon_t(\cdot)$. As long as $c_2(p) = o(p^{(1-\delta)/2})$, Condition 4(i) follows directly from Condition 4'(i). Moreover, Condition 4'(i) relies on the Hilbert–Schmidt norm to encourage the functional sparsity in $\Sigma_{x\varepsilon}^{(k)}$, i.e., each common factor $X_{(t+k)l}(\cdot)$ is only correlated with a few components of $\varepsilon_t(\cdot)$. With imposed sparsity constraints in Conditions 4' and 6, it can be inferred from (3) that $\Sigma_{yy}^{(k)}$ is functional sparse as justified in Lemma 1 below.

16

Lemma 1. *Under Conditions 1, 2, 4′ and 6, it holds that, for $k = 1, \ldots, k_0$,*

$$\max_{1 \leqslant j \leqslant p} \sum_{i=1}^{p} \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}^{\tau} \lesssim c_1(p) \quad and \quad \max_{1 \leqslant i \leqslant p} \sum_{j=1}^{p} \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}^{\tau} \lesssim c_1(p) + c_2(p).$$

Lemma 1 shows that the functional sparsity patterns in columns/rows of $\mathbf{\Sigma}_{yy}^{(k)}$ are determined by parameters $c_1(p)$ and $c_2(p)$ with smaller values yielding functional sparser $\mathbf{\Sigma}_{yy}^{(k)}$. To obtain a functional sparse estimator for $\mathbf{\Sigma}_{yy}^{(k)}$, we apply the hard functional thresholding rule, which combines functional versions of hard thresholding and shrinkage based on the Hilbert-Schmidt norm of functions, on entries of the sample autocovariance function $\widehat{\mathbf{\Sigma}}_{yy}^{(k)}$. Then the functional thresholding estimator is constructed as

$$\mathcal{T}_{\eta_k}(\widehat{\mathbf{\Sigma}}_{yy}^{(k)})(u,v) = \left[\widehat{\Sigma}_{yy,ij}^{(k)}(u,v) I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}} \geqslant \eta_k\}\right]_{p \times p}, \quad u,v \in \mathcal{U}, \tag{21}$$

where $I(\cdot)$ is the indicator function and $\eta_k \geqslant 0$ is the thresholding parameter. Under mild regularity conditions, it follows from Lemma B3 in the Supplementary Material that

$$\max_{1 \leqslant i,j \leqslant p} \|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}} = O_P\left(\mathcal{M}_y \sqrt{\frac{\log p}{n}}\right), \tag{22}$$

where $\mathcal{M}_y$ is the functional stability measure (Guo and Qiao; 2021; Fang et al.; 2021) defined in (A.1) in Appendix A. The rate under the functional version of ℓ_∞ norm in (22) plays a crucial role in our theoretical analysis under an ultra-high-dimensional regime and, in particular, suggests us to set the thresholding level as $\eta_k \asymp \mathcal{M}_y(n^{-1} \log p)^{1/2}$. Replacing $\widehat{\mathbf{\Sigma}}_{yy}^{(k)}$ in (6) with $\mathcal{T}_{\eta_k}(\widehat{\mathbf{\Sigma}}_{yy}^{(k)})$, we obtain the corresponding estimator for $\mathbf{M}$ by

$$\widetilde{\mathbf{M}} = \sum_{k=1}^{k_0} \int \int \mathcal{T}_{\eta_k}(\widehat{\mathbf{\Sigma}}_{yy}^{(k)})(u,v) \widehat{\mathbf{W}}(u,v) \mathcal{T}_{\eta_k}(\widehat{\mathbf{\Sigma}}_{yy}^{(k)})(u,v)^{\mathrm{T}} \, du dv. \tag{23}$$

To recover the column sparsity structure in $\mathcal{C}(\mathbf{A})$, we perform sparse *principal component analysis* (PCA) (Vu and Lei; 2013) on $\widetilde{\mathbf{M}}$ rather than the standard eigenanalysis for $\widehat{\mathbf{M}}$ in Section 2.1. For matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ with the same dimension, let $\langle \mathbf{A}_1, \mathbf{A}_2 \rangle := \text{trace}(\mathbf{A}_1^{\mathrm{T}} \mathbf{A}_2)$. We define $\widetilde{\mathbf{K}} \in \mathbb{R}^{p \times r}$ as a solution to the following constraint optimization problem:

$$\widetilde{\mathbf{K}} = \arg\max_{\mathbf{K}=(K_{jl})_{p \times r}} \langle \widetilde{\mathbf{M}}, \mathbf{K}\mathbf{K}^{\mathrm{T}} \rangle \text{ subject to } \mathbf{K}^{\mathrm{T}}\mathbf{K} = \mathbf{I}_r, \ \max_{1 \leqslant l \leqslant r} \sum_{j=1}^{p} |K_{jl}|^{\tau} \leqslant C_{\tau}, \tag{24}$$

17

where $C_\tau > 0$ is a regularization parameter. Alternatively, to estimate the row sparse factor loading space, we can substitute the second constraint in (24) by $\sum_{i=1}^p \|\mathbf{k}_i\|^\tau \leqslant \widetilde{C}_\tau$, where $\mathbf{k}_i$ denotes the i-th row vector of $\mathbf{K}$ and $\widetilde{C}_\tau > 0$ is a regularization parameter. It is worth noting that, without the sparsity constraint in (24), the optimization problem degenerates to the ordinary PCA. Despite being challenging to solve (24) due to the non-convex constraint, some efficient and computationally tractable algorithms have been developed, see, e.g., under the truly sparse case ($\tau = 0$), the combinatorial approaches (Moghaddam et al.; 2006; d'Aspremont et al.; 2008; Mackey; 2009), the semi-definite relaxation (d'Aspremont et al.; 2007) and its variants, and the random-projection-based method (Gataric et al.; 2020). We refer to Zou and Xue (2018) for an extensive review on recent developments for sparse PCA.

We now present the asymptotic analysis of $\mathcal{C}(\widetilde{\mathbf{K}})$ in the following theorem.

Theorem 3. *Let Conditions 1, 2, 4′, 6 and 7–8 in Appendix A hold and $\mathcal{M}_y^2 \log p n^{-1} = o(1)$. Then as $p, n \to \infty$, it holds that:*

$$\nu_r \mathcal{D}\Big(\mathcal{C}(\mathbf{K}), \mathcal{C}(\widetilde{\mathbf{K}})\Big) = O_P\left[c_1(p)\{c_1(p) + c_2(p)\}\mathcal{M}_y^{1-\tau}\Big(\frac{\log p}{n}\Big)^{\frac{1-\tau}{2}}\right].$$

Remark 6. *(i) The convergence rate of $\mathcal{D}\big(\mathcal{C}(\mathbf{K}), \mathcal{C}(\widetilde{\mathbf{K}})\big)$ is governed by both dimensionality parameters $\{n, p, c_1(p), c_2(p)\}$ and internal parameters $(\mathcal{M}_y, \nu_r, \tau)$. It is easy to see that the rate is better when ν_r is large and $\{c_1(p), c_2(p), \mathcal{M}_y, \tau\}$ are small.*

(ii) Under the truly sparse case ($\tau = 0$) with $c_1(p) \gtrsim c_2(p)$, and $|A_{ij}| \asymp \gamma$ for (i, j) such that $A_{ij} \neq 0$, it follows from the framework in Section 3 that $c_1(p)\gamma^2 \asymp p^{1-\delta}$ under Condition 3, which together with Theorem 3 imply that

$$\nu_r \mathcal{D}\big(\mathcal{C}(\mathbf{K}), \mathcal{C}(\widetilde{\mathbf{K}})\big) = O_P\{p^{2-2\delta}\gamma^{-4}\mathcal{M}_y(\log p)^{1/2}n^{-1/2}\}. \tag{25}$$

By comparison, Theorem 1 and (S.3) of the Supplementary Material lead to $\nu_r \mathcal{D}\big(\mathcal{C}(\mathbf{K}), \mathcal{C}(\widetilde{\mathbf{K}})\big) = O_P\big(p^{2-\delta}n^{-1/2}\big)$, which is slower than the rate in (25) for larger values of γ or δ (i.e., smaller values of $c_1(p)$ provided that $\gamma \asymp 1$). Hence, when the magnitudes of nonzero entries in $\mathbf{A}$ become larger or the factors are weaker in the sense of Condition 3 (i.e., $\mathbf{A}$ is sparser in the sense of Condition 6), our regularized estimation benefits more from the imposed sparsity and enjoys faster convergence rate than the ordinary method in Section 2.

5 Simulation studies

5.1 Setup

We illustrate the finite-sample performance of our proposed methods through extensive simulations. Sections 5.2 and 5.3 consider scenarios where the factor loading matrix $\mathbf{A}$ is ordinary and sparse, respectively.

In each simulated scenario, we generate p-vector of functional time series by

$$\mathbf{Y}_t(\cdot) = \kappa_0 \mathbf{A}\widetilde{\mathbf{X}}_t(\cdot) + \boldsymbol{\varepsilon}_t(\cdot), \quad t = 1, \ldots, n, \tag{26}$$

where the parameter $\kappa_0 > 0$ controls the strength of common factors $\mathbf{X}_t(\cdot) = \kappa_0 \widetilde{\mathbf{X}}_t(\cdot)$ and the entries of $\mathbf{A} \in \mathbb{R}^{p \times r}$ are sampled from $\text{Uniform}[-\sqrt{3}p^{-\delta/2}, \sqrt{3}p^{-\delta/2}]$ with $\delta \in [0, 1]$. Hence Condition 3 is satisfied, in which $\delta = 0$ (or $\delta > 0$) corresponds to the case of strong (or weak) factors. To mimic the infinite-dimensionality of functional data, we generate each scaled latent factor by $\widetilde{X}_{tl}(\cdot) = \sum_{i=1}^{50} \xi_{tli}\phi_i(\cdot)$ for $l = 1, \ldots, r$ over $\mathcal{U} = [0, 1]$, where $\{\phi_i(\cdot)\}_{i=1}^{50}$ is a 50-dimensional Fourier basis function and the basis coefficients $\boldsymbol{\xi}_{ti} = (\xi_{t1i}, \ldots, \xi_{tri})^{\mathrm{T}}$ are generated from a vector autoregressive model, $\boldsymbol{\xi}_{ti} = \mathbf{V}\boldsymbol{\xi}_{(t-1)i} + \boldsymbol{\epsilon}_i$ with $\mathbf{V} = (\rho^{|l-l'|+1})_{1 \leqslant l, l' \leqslant r}$ and the innovation $\boldsymbol{\epsilon}_{ti} = (\epsilon_{t1i}, \ldots, \epsilon_{tri})^{\mathrm{T}}$ consisting of independent $N(0, i^{-2\iota})$ components. We set $\rho = 0.45$ and $\iota = 0.75$. For the idiosyncratic component $\boldsymbol{\varepsilon}_t(\cdot)$, we consider the following three scenarios.

SCENARIO 1. For each $j = 1, \ldots, p$, we generate $\varepsilon_{tj}(\cdot) = \sum_{i=1}^{20} 2^{-(i+1)}Z_{tji}\phi_i(\cdot)$, where Z_{tji}'s are independent standard normal.

SCENARIO 2. We generate $\boldsymbol{\varepsilon}_t(\cdot) = \mathbf{H}\tilde{\boldsymbol{\varepsilon}}_t(\cdot)$, where $\mathbf{H} = 5^{-1}\text{diag}(h_1, \ldots, h_p)$, h_j's are sampled uniformly from $\{1, \ldots, 10\}$ and each $\tilde{\varepsilon}_{tj}(\cdot)$ is generated in the same way as $\varepsilon_{tj}(\cdot)$ in Scenario 1.

SCENARIO 3. We fix $\kappa_0 = 1$ and generate $\varepsilon_{tj}(\cdot) = \kappa_1\varepsilon_{t0}(\cdot) + \tilde{\varepsilon}_{tj}(\cdot)$, where $\kappa_1 > 0$, $\varepsilon_{t0}(\cdot) = \sum_{i=1}^{4} \widetilde{Z}_{ti}\phi_i(\cdot)$ and each $\widetilde{Z}_{ti}$ is sampled independently from $N(0, 1)$.

In Scenario 1, each $\varepsilon_{tj}(\cdot)$ is white noise with the identical covariance function across j, while Scenario 2 corresponds to the heterogeneous case with different covariance functions. Scenario 3 consists of $(r+1)$ actual common factors, among which the additional factor $\varepsilon_{t0}(\cdot)$ is independent of other factors in $\mathbf{X}_t(\cdot)$, and its signal strength is determined by the parameter κ_1. The underlying factor loading matrix is $(\mathbf{A}, \mathbf{1}_p)$, where $\mathbf{1}_p$ denotes the p-vector of ones. For simplicity, we still denote the factor loading matrix by $\mathbf{A}$.

5.2   Ordinary case

We compare our proposed method based on $\widehat{\mathbf{M}}$ with two competing methods based on $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$. Since our simulation results suggest that $\hat{r}$ and $\mathcal{C}(\widehat{\mathbf{K}})$ are insensitive to the choice of q and the maximum lag k_0, we set $q = 12$ and $k_0 = 4$ in our simulations. In each setting, we generate $n = 100$ serially dependent observations of $p = 50, 100, 200$ functional variables based on $r = 4$ functional factors. We ran each simulation 100 times. The sample performance of three approaches is examined in terms of their abilities of correctly identifying the number of factors and the estimation accuracy in recovering the factor loading space, respectively measured by the relative frequency estimate for $P(\hat{r} = r)$ and $\mathcal{D}\big(\mathcal{C}(\mathbf{K}), \mathcal{C}(\widehat{\mathbf{K}})\big)$ using the correct r. For each of three comparison methods, Figures 1 and 2 plot average relative frequencies $\hat{r} = r$ as the factor strength increases when the factors are strong (i.e., $\delta = 0$) and weak with $\delta = 0.5$, respectively. Figures 3 and 4 plot the corresponding average estimation errors for $\mathcal{C}(\mathbf{K})$.

Several conclusions can be drawn from Figures 1–4. First, our proposed method based on $\widehat{\mathbf{M}}$ provides highly significant improvements in accuracies for the identification of r and the recovery of $\mathcal{C}(\mathbf{K})$ over the competing methods in all scenarios we consider. The improvement of our method involving the weight matrix is larger for the heterogeneous case and even more substantial for the case when the factors are weak. Between two competitors, the method based on $\widehat{\mathbf{M}}_1$ outperforms that based on $\widehat{\mathbf{M}}_2$, providing empirical evidence for Remark 3. Second, the estimation for r and $\mathcal{C}(\mathbf{K})$ performs better as the strength of factors increases (i.e., κ_0 or κ_1 increases or δ decreases), which is in line with our theoretical results in Section 3.

Figure 1: *Scenario 1 (top row), Scenario 2 (middle row) and Scenario 3 (bottom row) for $p = 50$, 100 and 200 when the factors are strong: Plots of average relative frequency estimates for $P(\hat{r} = r)$ against the factor strength (κ_0 or κ_1) for three methods based on $\widehat{\mathbf{M}}$ (black solid), $\widehat{\mathbf{M}}_1$ (red dashed) and $\widehat{\mathbf{M}}_2$ (cyan dash dotted).*

In Figure 1, our method makes the sharpest progress as the factors become stronger, while the two comparison methods require much higher strength of factors to compete. Third, when the factors are strong, we observe the phenomenon of "blessing of dimensionality" in Figures 1 and 2 in the sense that the estimation improves as p increases from 50 to 200. The improvement is due to the increase of the information from added components on the factors. Under Scenario 3 with an extra factor independent of the others, the information and noise on the factors increase simultaneously as p grows, and hence the estimation does
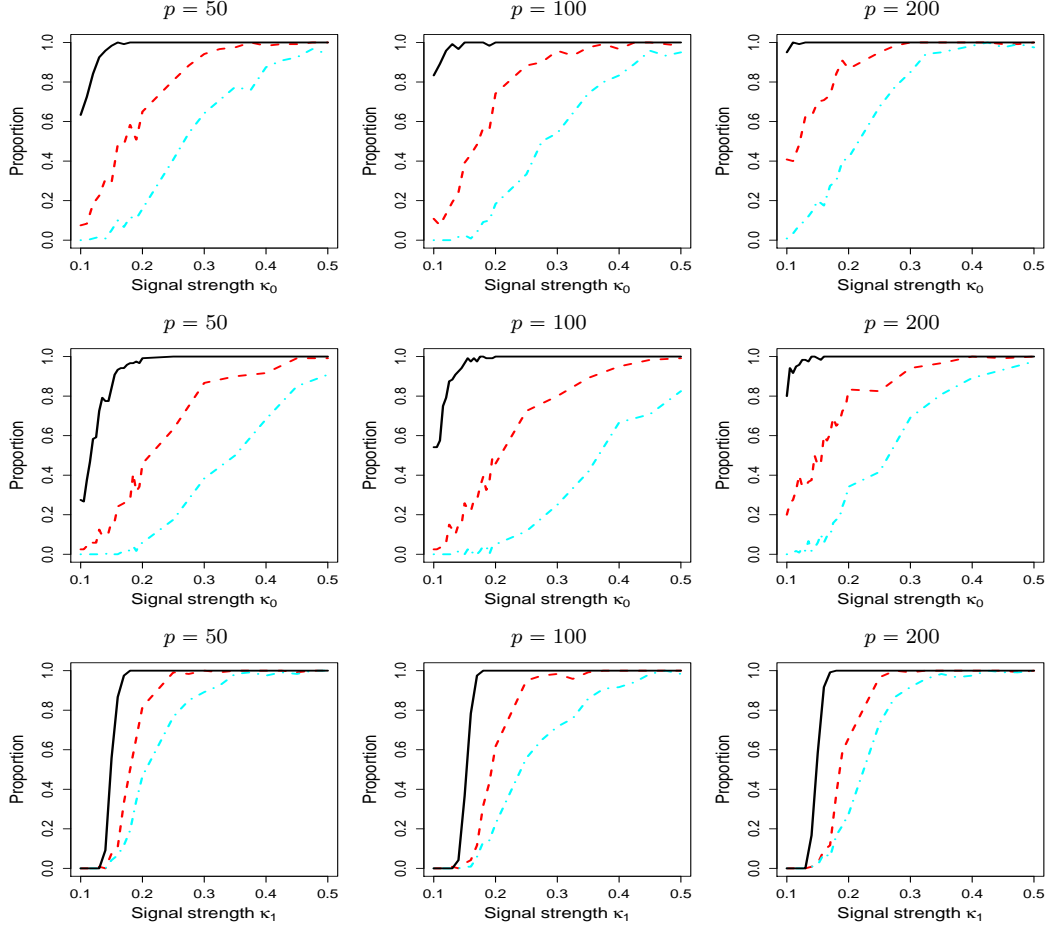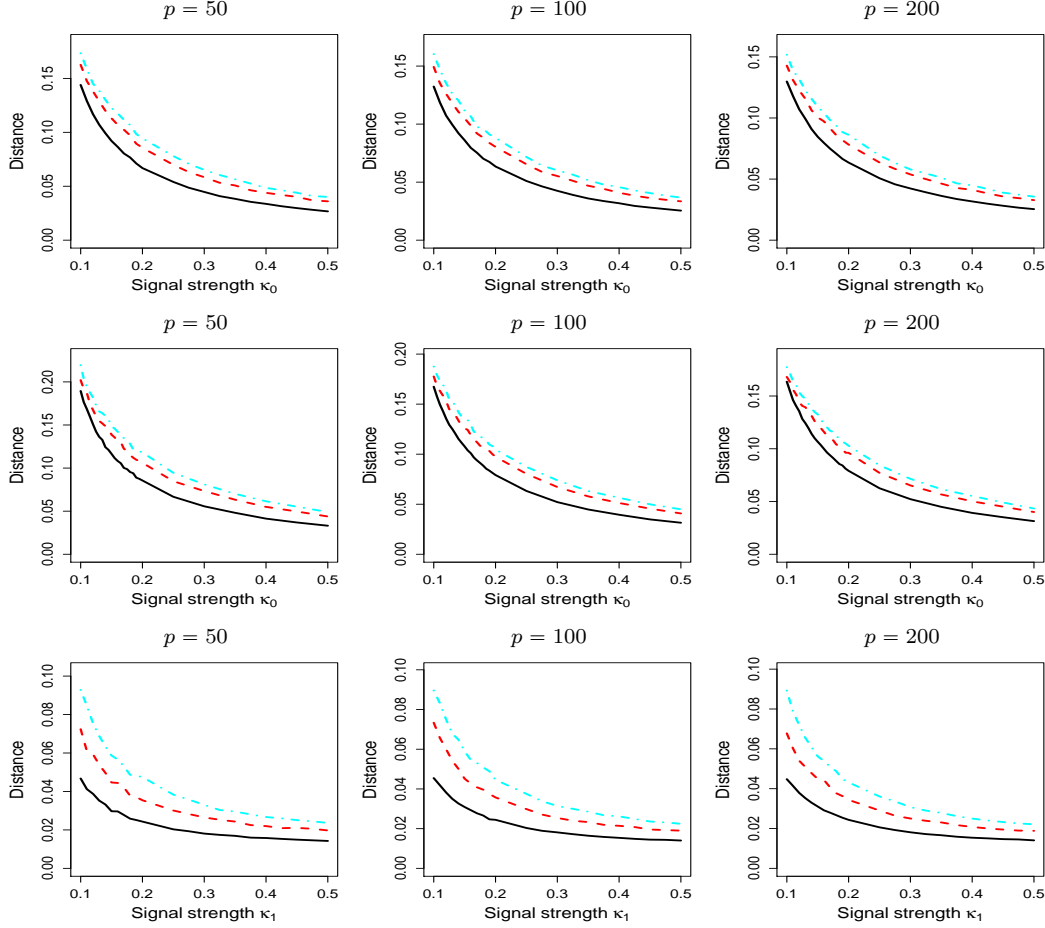
Figure 2: *Scenario 1 (top row), Scenario 2 (middle row) and Scenario 3 (bottom row) for* $p = 50, 100, 200$ *when the factors are strong: Plots of average* $\mathcal{D}\big(\mathcal{C}(\mathbf{K}), \mathcal{C}(\widehat{\mathbf{K}})\big)$ *against* κ_0 *or* κ_1 *for three methods based on* $\widehat{\mathbf{M}}$ *(black solid),* $\widehat{\mathbf{M}}_1$ *(red dashed) and* $\widehat{\mathbf{M}}_2$ *(cyan dash dotted).*

not necessarily improve. Fourth, under the weak factor setting, many entries of $\mathbf{A}$ are quite close to zero. As p enlarges, the enhanced information on the factors is accompanied with the increase of noise. E.g., in Figure 3, we observe that enlarging p does not necessarily lead to the increased estimation accuracy.

5.3   Sparse Case

In this section, we conduct simulations to evaluate the performance of the *functional-thresholding-and-sparse-PCA-based approach* (TSPCA) developed in Section 4 to estimate
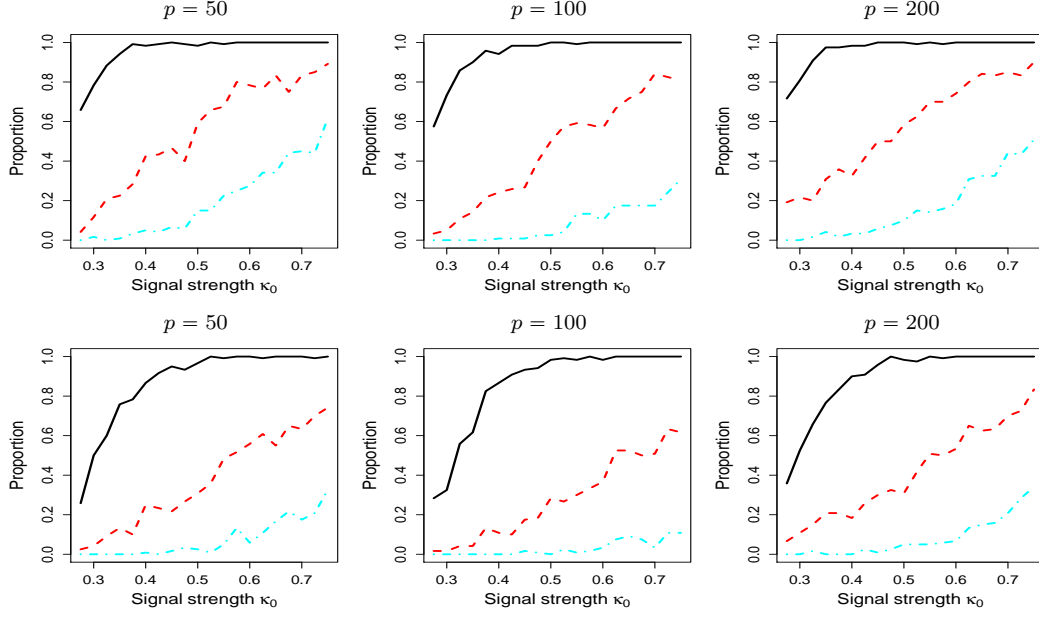
Figure 3: *Scenario 1 (top row) and Scenario 2 (bottom row) for $p = 50$, 100 and 200 with correct r when the factors are weak: Plots of relative frequency estimates for $P(\hat{r} = r)$ against κ_0 for three methods based on $\widehat{\mathbf{M}}$ (black solid), $\widehat{\mathbf{M}}_1$ (red dashed) and $\widehat{\mathbf{M}}_2$ (cyan dash dotted).*

the sparse factor model. Two kinds of subspace sparsity constraints are imposed on $\mathbf{A}$. Specifically, for the row sparsity, we randomly select 80% rows of $\mathbf{A}$ to be zero vectors, while for the column sparsity, we randomly set 80% elements within each column of $\mathbf{A}$ as zeros. We focus on Scenario 1 with $\kappa_0 = 1$ and generate nonzero entries of $\mathbf{A}$ from Uniform$[-\sqrt{3}, \sqrt{3}]$.

Implementing TSPCA requires choosing the thresholding parameter η_k in (21) and the columnwise number of nonzero elements C_0 in (24). To select the optimal $\hat{\eta}_k$ for each $k = 1, \ldots, k_0$, we implement a G-fold *cross-validation* (CV) approach (Cai and Liu; 2011). To be specific, we first sequentially divide the set $\{1, \ldots, n\}$ into G blockwise groups $\mathcal{D}_1, \ldots, \mathcal{D}_G$ of approximately equal size. We then treat the g-th group as a validation set, compute the sample lag-k autocovariance functions $\widehat{\mathbf{\Sigma}}_{yy}^{(k),(g)}(u,v)$ and $\widehat{\mathbf{\Sigma}}_{yy}^{(k),(-g)}(u,v)$ for $u, v \in \mathcal{U}$ based on the validation set and the remaining $G - 1$ groups, respectively, and repeat the above procedure G times. We finally select $\hat{\eta}_k$ by minimizing $R_k(\eta_k) = G^{-1} \sum_{g=1}^{G} \big\| \mathcal{T}_{\eta_k}\big(\widehat{\mathbf{\Sigma}}_{yy}^{(k),(-g)}\big) - \widehat{\mathbf{\Sigma}}_{yy}^{(k),(g)} \big\|_{\mathcal{S},\mathrm{F}}^2$. We adopt a similar CV method to select the optimal $\widehat{C}_0$. Given the g-th group
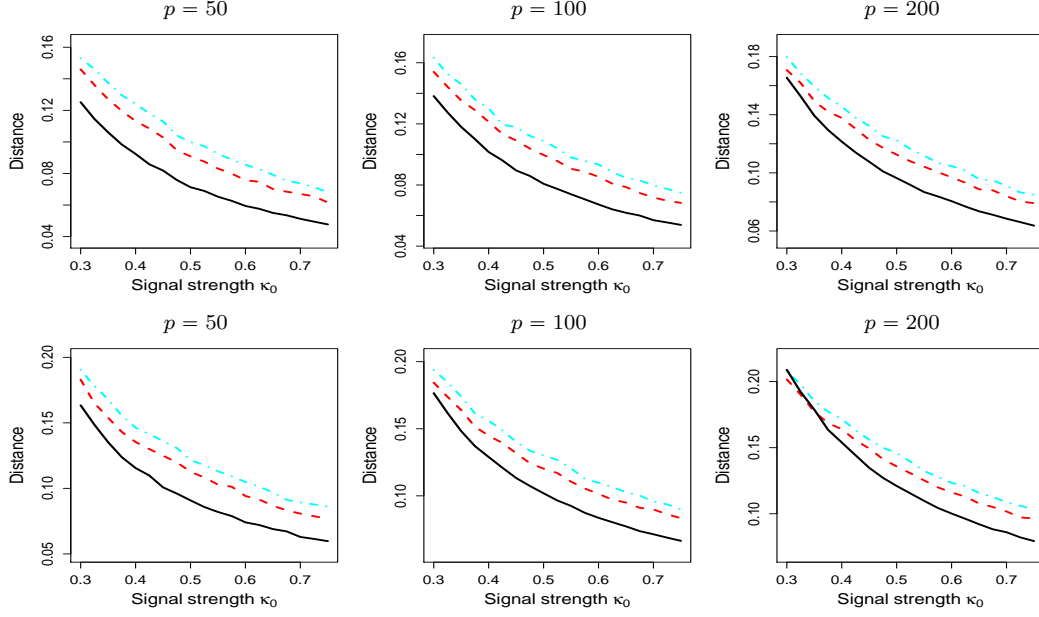
Figure 4: *Scenario 1 (top row) and Scenario 2 (bottom row) for $p = 50$, 100 and 200 with correct r when the factors are weak: Plots of average $\mathcal{D}\big(\mathcal{C}(\mathbf{K}),\mathcal{C}(\widehat{\mathbf{K}})\big)$ against κ_0 for three methods based on $\widehat{\mathbf{M}}$ (black solid), $\widehat{\mathbf{M}}_1$ (red dashed) and $\widehat{\mathbf{M}}_2$ (cyan dash dotted).*

as a validation set, we obtain the solution $\widetilde{\mathbf{K}}^{(-g)}$ to the constrained optimization problem in (24) with $\tau = 0$ based on the remaining $G - 1$ groups. To solve this problem, we apply sparse PCA (Moghaddam et al.; 2006; Mackey; 2009) to $\widetilde{\mathbf{M}}^{(-g)}$, which is formed by (23) using $\mathcal{T}_{\widehat{\eta}_k}\big(\widehat{\mathbf{\Sigma}}_{yy}^{(k),(-g)}\big)$ for $k = 1, \ldots, k_0$. We also obtain $\widehat{\mathbf{K}}^{(g)}$ by carrying out an eigenanalysis for $\widehat{\mathbf{M}}^{(g)}$ based on the validation set. The above procedure is repeated G times and $\widehat{C}_0$ is selected by minimizing $D(C_0) = G^{-1} \sum_{g=1}^{G} \mathcal{D}\big(\mathcal{C}(\widetilde{\mathbf{K}}^{(-g)}), \mathcal{C}(\widehat{\mathbf{K}}^{(g)})\big)$. Although the time break created by each leave-out validation set possibly jeopardize the autocovariance structure on the remaining $G - 1$ groups via k_0 mis-utilized lagged terms, its effect on the estimation of $\mathbf{\Sigma}_{yy}^{(k)(-g)}$'s is negligible especially when n is sufficiently large. Hence our proposed CV approach does not place a practical constraint.

We compare TSPCA with the ordinary PCA-based approach by performing an eigenanlysis on $\widehat{\mathbf{M}}$. The estimation quality is measured in terms of the distance between the estimated and true sparse factor loading spaces using the correct r. We report the numerical summaries

Table 1: The mean and standard error (in parentheses) of the distance between the estimated and true sparse factor loading spaces over 100 simulation runs. All entries have been multiplied by 10^2 for formatting reasons.

Sparsity	p	PCA	TSPCA	Sparsity	p	PCA	TSPCA
	100	3.86(0.53)	1.71(0.31)		100	4.27(0.83)	3.55(1.09)
Row	200	3.79(0.56)	1.74(0.26)	Column	200	4.25(0.65)	3.69(0.59)
	400	3.73(0.66)	1.76(0.29)		400	4.03(0.74)	3.17(0.88)

in Table 1 for $p = 100, 200, 400$. It is obvious to see that TSPCA uniformly improves the PCA-based estimation, demonstrating the advantage of our regularized estimation procedure to fit sparse factor models. The improvement for the row sparsity case is more significant than that for the column sparsity case. Compared with the functional sparsity patterns in Lemma 1 under the column sparsity constraint, it follows from the decomposition of $\mathbf{\Sigma}_{yy}^{(k)}$ in (3) that the row sparsity in $\mathbf{A}$ can lead to functional sparser $\mathbf{\Sigma}_{yy}^{(k)}$'s. In this sense, TSPCA benefits more from the row sparsity structure, thus resulting in enhanced improvement.

6 Real data analysis

6.1 U.K. temperature data

Our first dataset, available at https://www.metoffice.gov.uk/research/climate/maps-and-data/historic-station-data, consists of monthly average temperature collected at $p = 22$ measuring stations in the U.K. from 1959 to 2020 ($n = 62$). Let $Y_{ij}(u_k)$ ($i = 1, \ldots, n, j = 1, \ldots, p, k = 1, \ldots, 12$) denotes the average temperature during month $u_k = k \in \mathcal{U} = [1, 12]$ of year $1958 + i$ at the j-th measuring station. The observed temperature curves are smoothed based on 10-dimensional Fourier basis to capture the periodic structures over the annual cycle. The smoothed curve series exhibits very small autocorrelations beyond $k = 2$, so we use $k_0 = 2$ in computing $\widehat{\mathbf{M}}$. Although the ratio-based estimator suggests $\hat{r} = 1$, we illustrate further results using $\hat{r} = 3$.

Figure 5 displays spatial heatmaps of estimated factor loading matrix and the rotated matrix using the varimax procedure to maximize the sum of the variances of the squared loadings. Some interesting patterns can be observed from those heatmaps. Compared with the original results in Figure 5, the varimax rotation brings the factor loading matrix closer to a "simple structure", where (i) each component in $\mathbf{Y}_t(\cdot)$ has a high loading on one specific factor but near-zero loadings on other factors and (ii) each factor has high impact on a few components of $\mathbf{Y}_t(\cdot)$ with high loadings on this factor while the remaining variables have near-zero loadings on this factor. The varimax rotation leads to larger color variations in heatmaps for better interpretation. For example, it is apparent that the first factor via varimax rotation influences the dynamics in the southeast, while the original factor has uniform impact in all locations. Hence we focus on the interpretation of remaining factors after varimax rotation. Specifically, the second factor mainly impacts the dynamics of the northern region. The third factor can roughly be viewed as the main driving force for the dynamics in the middle north.
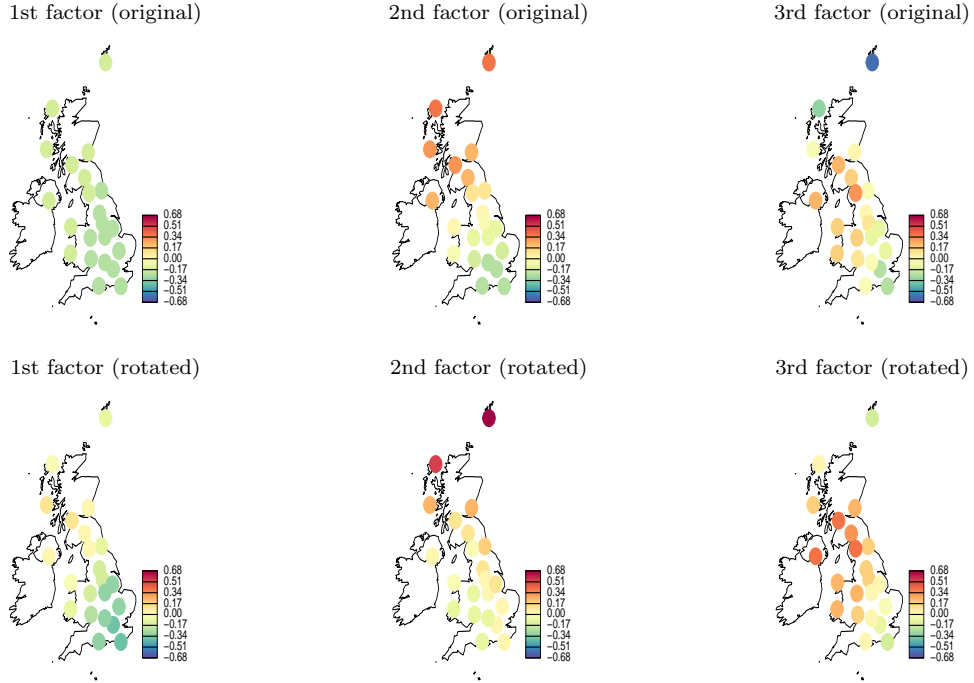

Figure 5: *Spatial heatmaps based on estimated factor loading matrix (top row) and varimax-rotated loading matrix (bottom row) of 22 U.K. locations.*

6.2 Japanese mortality data

Our second dataset contains age-specific and gender-specific mortality rates for $p = 47$ Japanese prefectures from 1975 to 2017 ($n = 43$). This dataset was also analyzed in Gao et al. (2019). Due to those sparse observations at old ages, we focus on data for ages below 96. We apply a log transformation to mortality rates and denote by $Y_{ij}(u_k)$ ($i = 1, \ldots, n, j = 1, \ldots, p, k = 1, \ldots, 96$) the log mortality rate of people aged $u_k = k - 1 \in \mathcal{U} = [0, 95]$ living in the j-th prefecture during the year $1974 + i$. We then perform smoothing for observed curves and replace the missing values via smoothing splines. The estimation of model (1) is done using $k_0 = 2$. The ratio-based estimator for r is $\hat{r} = 1$, though we use $\hat{r} = 2$ for the illustration. To enhance interpretability for identified factors under a high-dimensional $p > n$ regime, we also implement TSPCA to estimate the sparse factor loading matrix.

Figures 6 and 7 in the Supplementary Material show spatial heatmaps of varimax-rotated loading matrix and sparse loading matrices with different sparsity levels for Japanese females and males, respectively. For TSPCA, we implement the CV method to select the optimal $\hat{\eta}_k$'s and set the columnwise sparsity in estimated loading matrix to $27/47$ or $32/47$ for better visualization of the results. A more systematic method to determine the sparsity level, e.g., via a significance testing, needs to be developed. Compared with the original results, both varimax rotation and TSPCA lead to enhanced interpretability in the sense of recovering the factor loading matrix with "simple structure". Among two competitors, TSPCA tends to reduce some near-zero loadings to exactly zero, thus providing a parsimonious model with more interpretable results than varimax rotation. In particular, we can observe clear regional patterns from those heatmaps for TSPCA, where each of two factors influences the dynamics of some complementary regions of Japan in all settings. Take the results under the sparsity $27/47$ in Figure 6 as an example, the first factor serves as the main driving force in regions of Hokkaido, Kanto, Chubu and Kansal, while Tohoku and Shikoku are heavily loaded regions on the second factor.

To further illustrate the developed methodology, we set upon the task of predicting high-dimensional functional time series $\mathbf{Y}_t(\cdot)$. Specifically, we incorporate the factor model
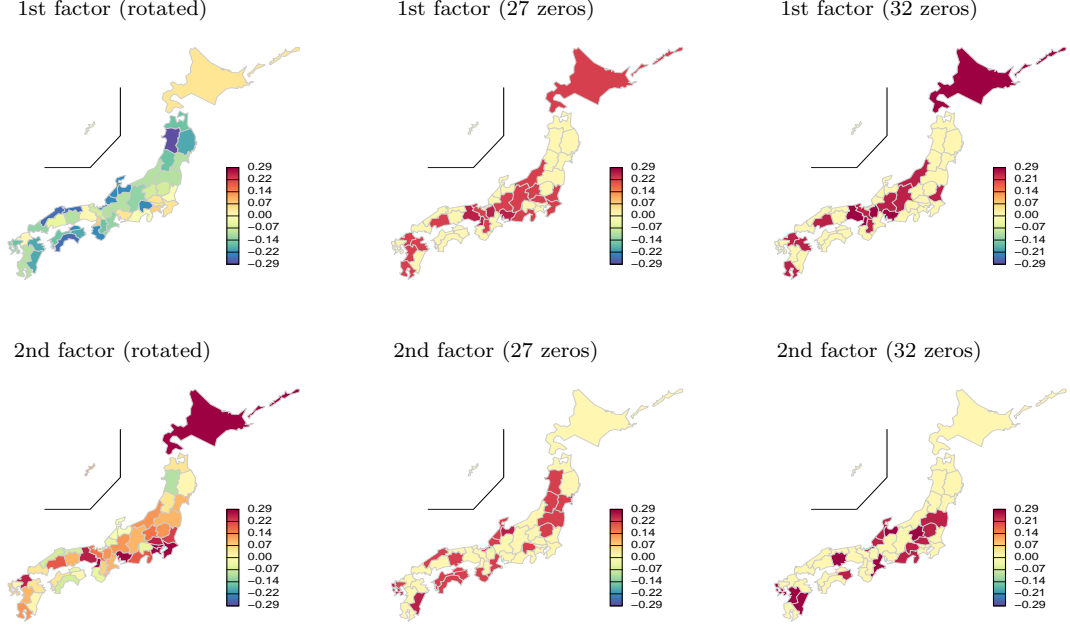
Figure 6: *Spatial heatmaps based on varimax-rotated loadings (left column) and sparse loadings with 27 zeros (middle column) and 32 zeros (right column) of 47 prefectures on two factors for Japanese females.*

framework into the h-step-ahead prediction (denoted as FMP) consisting of three steps below.

1. Apply our proposed method to estimate model (1) based on past observations $\{\mathbf{Y}_t(\cdot)\}_{t=1}^T$, thus obtaining the estimated factor loading matrix $\widehat{\mathbf{A}}$ and number of factors $\hat{r}$.

2. Compute $\hat{r}$ estimated factors by $\widehat{\mathbf{X}}_t(\cdot) = \widehat{\mathbf{A}}^{\mathrm{T}}\mathbf{Y}_t(\cdot)$. For each $k \in \{1, \ldots, \hat{r}\}$, predict $\widehat{X}_{(t+h)k}(\cdot)$ based on past fitted values for the k-th common factor, $\{\widehat{X}_{tk}(\cdot)\}_{t=1}^T$.

3. The h-step head predict for $\mathbf{Y}_{T+h}(\cdot)$ is $\widehat{\mathbf{Y}}_{T+h}(\cdot) = \widehat{\mathbf{A}}\widehat{\mathbf{X}}_{T+h}(\cdot)$.

The second step is on the prediction of univariate functional time series. For each common factor, we obtain the h-step-ahead prediction based on the best fitted functional ARMA model (Klepsch et al.; 2017) according to the BIC criterion. We then develop a sparse version of FMP (denoted as SFMP) by performing TSPCA in the first step to estimate sparse $\widehat{\mathbf{A}}$. For comparison, we also implement the method of Gao et al. (2019) (denoted as

GSY) to jointly predict p-vector of functional time series $\mathbf{Y}_t(\cdot)$, as well as the univariate prediction method (Aue et al.; 2015) (denoted as UP) on each $Y_{tj}(\cdot)$ separately.

To evaluate the predictive accuracy, we use the expanding window approach. The data is divided into a training set and a test set consisting of the first n_1 and the last n_2 observations, respectively. For any integer $h > 0$, we implement each fitting method on the training set, obtain h-step-ahead prediction on the test data based on the fitted model, increase the training size by one and repeat this procedure $n_2 - h + 1$ times to compute the h-step-ahead *mean absolute prediction error* (MAPE) and *mean squared prediction error* (MSPE) by

$$\mathrm{MAPE}(h) = \frac{1}{p \times (n_2 - h + 1) \times 96} \sum_{j=1}^{p} \sum_{t=n_1+h}^{n} \sum_{k=1}^{96} \left| \widehat{Y}_{tj}(u_k) - Y_{tj}(u_k) \right|,$$

$$\mathrm{MSPE}(h) = \frac{1}{p \times (n_2 - h + 1) \times 96} \sum_{j=1}^{p} \sum_{t=n_1+h}^{n} \sum_{k=1}^{96} \left\{ \widehat{Y}_{tj}(u_k) - Y_{tj}(u_k) \right\}^2.$$

The resulting MAPE and MSPE values are summarized in Figure 2. It is obvious that both FMP and SFMP significantly outperform two competitors in all settings, while SFMP is slightly inferior to FMP possibly due to the bias introduced by the enforced sparsity.

Table 2: MAPEs and MSPEs of FMP, SFMP and two competing methods for Japanese male and female mortality rates. All entries have been multiplied by 10 for formatting reasons.

		MAPE				MSPE			
	h	FMP	SFMP	GSY	UP	FMP	SFMP	GSY	UP
Male	1	1.33	1.36	1.75	1.42	0.48	0.49	0.84	0.59
	2	1.37	1.43	1.88	1.47	0.51	0.52	0.92	0.63
	3	1.43	1.49	1.87	1.53	0.54	0.55	0.86	0.67
	h	FMP	SFMP	GSY	UP	FMP	SFMP	GSY	UP
Female	1	1.40	1.44	1.77	1.52	0.61	0.64	0.87	0.82
	2	1.45	1.51	1.84	1.56	0.66	0.69	0.94	0.99
	3	1.50	1.56	1.77	1.66	0.71	0.74	0.89	1.14

7 Discussion

We conclude the paper by discussing two possible extensions. First, our estimation procedure is developed under the assumption that the idiosyncratic component is a white noise sequence. It is also interesting to estimate the factor model (1) following the framework in Fan et al. (2013), where the idiosyncratic noise is allowed to exhibit serial correlations. Under the uncorrelatedness assumption between $\mathbf{X}_t(\cdot)$ and $\boldsymbol{\varepsilon}_t(\cdot)$, we can decompose the integrated covariance function as

$$\int\int \boldsymbol{\Sigma}_{yy}^{(0)}(u,v)\,\mathrm{d}u\mathrm{d}v = \mathbf{A}\Big\{\int\int \boldsymbol{\Sigma}_{xx}^{(0)}(u,v)\,\mathrm{d}u\mathrm{d}v\Big\}\mathbf{A}^{\mathsf{T}} + \int\int \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{(0)}(u,v)\,\mathrm{d}u\mathrm{d}v. \qquad (27)$$

By imposing some eigenvalue conditions on two terms in the right-side of (27) similar to those in Fan et al. (2013), the common factors are asymptotically identifiable and hence the factor loading space can be recovered by carrying out an eigenanalysis for $\int\int \widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(u,v)\,\mathrm{d}u\mathrm{d}v$.

Second, it is conceivable to integrate the high-dimensional time series clustering approach (Zhang et al.; 2021) into our proposed factor model framework (1) for the purpose of clustering a large number of functional time series into different and unknown clusters such that curve dynamics within each cluster share similar dynamic structure. Specifically, we can model p-vector of functional time series $\mathbf{Y}_t(\cdot)$ based on a functional latent factor structure, where the dynamic structures are characterized by the latent common and cluster-specific functional factors with each factor represented in the form of (1). Whereas the common factors are strong, the cluster-specific factors are weak and only affect $\mathbf{Y}_t(\cdot)$ in a specific cluster. The clustering is then based on factor loadings on all weak factors.

Appendix

A Additional regularity conditions

To study theoretical properties in an ultra-high-dimensional regime, we need to rely on non-asymptotic error bounds on $\widehat{\Sigma}_{yy,ij}^{(k)}$'s and will use the functional stability measure of

$\{\mathbf{Y}_t(\cdot)\}_{t\in\mathbb{Z}}$ (Guo and Qiao; 2021; Fang et al.; 2021).

Before presenting regularity conditions, we first solidify some notation. We denote the p-fold Cartesian product of $\mathbb{H}$ by $\mathbb{H}^p = \mathbb{H} \times \cdots \times \mathbb{H}$. For $\mathbf{w} = (w_1, \ldots, w_g)^{\mathrm{T}}, \mathbf{g} = (g_1, \ldots, g_p)^{\mathrm{T}} \in \mathbb{H}^p$, we define $\langle \mathbf{w}, \mathbf{g} \rangle = \sum_{j=1}^p w_j g_j$. For an integral operator $\mathbf{K} : \mathbb{H}^p \to \mathbb{H}^p$ induced from the kernel function $\mathbf{K} = (K_{ij})_{p\times p}$, $\mathbf{K}(\mathbf{w})(u) = \left(\sum_{j=1}^p \langle K_{1j}(u, \cdot), w_j(\cdot) \rangle, \ldots, \sum_{j=1}^p \langle K_{pj}(u, \cdot), w_j(\cdot) \rangle \right)^{\mathrm{T}}$ for $\mathbf{w} \in \mathbb{H}^p$. To simplify notation, we use $\mathbf{K}$ to denote both the kernel function and the operator. We then define Hilbert space valued sub-Gaussian random variable, which corresponds to an infinite-dimensional analog of the sub-Gaussian random vector in $\mathbb{R}^p$.

Definition 1. *Let $W(\cdot)$ be a random variable in $\mathbb{H}$ and $\Sigma_W : \mathbb{H} \to \mathbb{H}$ be the covariance operator of $W(\cdot)$. Then $W(\cdot)$ is a sub-Gaussian process if there exists a constant $\alpha > 0$ such that for all $w \in \mathbb{H}$, $\mathbb{E}[\exp\{\langle w, W - \mathbb{E}(W) \rangle\}] \leqslant \exp\{2^{-1}\alpha^2 \langle w, \Sigma_W(w) \rangle\}$.*

Condition 7. *(i) $\{\mathbf{Y}_t(\cdot)\}_{t\in\mathbb{Z}}$ is a sequence of multivariate functional linear processes with sub-Gaussian errors, i.e., $\mathbf{Y}_t(\cdot) = \sum_{l=0}^{\infty} \mathbf{B}_l(\boldsymbol{\varepsilon}_{t-l})$ for any $t \in \mathbb{Z}$, where $\mathbf{B}_l = (B_{l,jk})_{p\times p}$ with each $B_{l,jk} \in \mathbb{H} \otimes \mathbb{H}$, $\boldsymbol{\varepsilon}_t = (\varepsilon_{t1}, \ldots, \varepsilon_{tp})^{\mathrm{T}} \in \mathbb{H}^p$ and the components in $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{Z}}$ are independent mean-zero sub-Gaussian processes satisfying Definition 1; (ii) The coefficient functions satisfy $\sum_{l=0}^{\infty} \|B_{l,ij}\|_{\mathcal{S},\infty} = O(1)$; (iii) $\max_j \int_{\mathcal{U}} Cov\{\varepsilon_{tj}(u), \varepsilon_{tj}(u)\} du = O(1)$.*

Condition 8. *For $\{\mathbf{Y}_t(\cdot)\}_{t\in\mathbb{Z}}$, its spectral density function at frequency $\theta \in [-\pi, \pi]$, $\mathbf{f}_{y,\theta} = (2\pi)^{-1} \sum_{k\in\mathbb{Z}} \mathbf{\Sigma}_{yy}^{(k)} e^{-ik\theta}$, exists, and its functional stability measure defined in (A.1) is finite,*

$$\mathcal{M}_y = 2\pi \cdot \underset{\theta \in [-\pi,\pi], \mathbf{w} \in \mathbb{H}_0^p}{esssup} \frac{\langle \mathbf{w}, \mathbf{f}_{y,\theta}(\mathbf{w}) \rangle}{\langle \mathbf{w}, \mathbf{\Sigma}_{yy}^{(0)}(\mathbf{w}) \rangle} < \infty, \tag{A.1}$$

where $\mathbb{H}_0^p = \{\mathbf{w} \in \mathbb{H} : \langle \mathbf{w}, \mathbf{\Sigma}_{yy}^{(0)}(\mathbf{w}) \rangle \in (0, \infty)\}$.

The multivariate functional linear process in Condition 7(i) generalizes the multivariate linear process to the functional domain. Under this setting, the functional stability measure $\mathcal{M}_y$ in (A.1) can be expressed based on $\mathbf{f}_{y,\theta}$ and $\mathbf{\Sigma}_{yy}^{(0)}$, both of which have explicit expressions (Fang et al.; 2021). Under Conditions 7 and 8, the rate in (22) can be implied from Lemma B3 in the Supplementary Material. In general, we can relax Conditions 7(ii) and (iii) by allowing both $\sum_{l=0}^{\infty} \|B_{l,ij}\|_{\mathcal{S},\infty}$ and $\max_j \int_{\mathcal{U}} Cov\{\varepsilon_{tj}(u), \varepsilon_{tj}(u)\} du$ to grow at very slow rates as p diverges, then the rate in (22) will depend on these two terms.

31

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors, *Econometrica* **81**: 1203–1227.

Aue, A., Norinho, D. D. and Hörmann, S. (2015). On the prediction of stationary functional time series, *Journal of the American Statistical Association* **110**: 378–392.

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

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

Bosq, D. (2000). *Linear Processes in Function Spaces*, Springer, New York.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation, *Journal of the American Statistical Association* **106**: 672–684.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics, *Journal of the American Statistical Association* **103**: 1438–1456.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets, *Econometrica* **51**: 1281–1304.

Chang, J., Guo, B. and Yao, Q. (2018). Principal component analysis for second-order stationary vector time series, *The Annals of Statistics* **46**: 2094–2124.

Chen, C., Guo, S. and Qiao, X. (2020). Functional linear regression: dependence and error contamination, *Journal of Business & Economic Statistics, in press* .

Chen, E. Y., Tsay, R. S. and Chen, R. (2020). Constrained factor models for high-dimensional matrix-variate time series, *Journal of the American Statistical Association* **115**: 775–793.

Chen, R., Yang, D. and Zhang, C.-H. (2021). Factor models for high-dimensional tensor time series, *Journal of the American Statistical Association, in press* .

d'Aspremont, A., Bach, F. and El Ghaoui, L. (2008). Optimal solutions for sparse principal component analysis., *Journal of Machine Learning Research* **9**: 1269–1294.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. (2007). A direct formulation for sparse pca using semidefinite programming, *SIAM review* **49**(3): 434–448.

Fan, J., Guo, J. and Zheng, S. (2020). Estimating number of factors by adjusted eigenvalues thresholding, *Journal of the American Statistical Association, in press* .

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements, *Journal of the Royal Statistical Society: Series B* **75**: 603–680.

Fan, J., Liao, Y. and Wang, W. (2016). Projected principal component analysis in factor models, *The Annals of Statistics* **44**: 219 – 254.

Fang, Q., Guo, S. and Qiao, X. (2021). Finite sample theory for high-dimensional functional/scalar time series with applications, *Electronic Journal of Statistics, in press* .

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation, *Review of Economics and statistics* **82**: 540–554.

Gao, Y., Shang, H. L. and Yang, Y. (2019). High-dimensional functional time series forecasting: An application to age-specific mortality rates, *Journal of Multivariate Analysis* **170**: 232–243.

Gataric, M., Wang, T. and Samworth, R. J. (2020). Sparse principal component analysis via axis-aligned random projections, *Journal of the Royal Statistical Society: Series B* **82**: 329–359.

Guo, S. and Qiao, X. (2021). On consistency and sparsity for high-dimensional functional time series with application to autoregressions, *arXiv:2003.11462v2* .

Hays, S., Shen, H. and Huang, J. Z. (2012). Functional dynamic factor models with application to yield curve forecasting, *The Annals of Applied Statistics* **6**: 870–894.

Hörmann, S., Kidzinski, L. and Hallin, M. (2015). Dynamic functional principal components, *Journal of the Royal Statistical Society: Series B* **77**: 319–348.

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data, *The Annals of Statistics* **38**: 1845–1884.

Klepsch, J., Kluppelberg, C. and Wei, T. (2017). Prediction of functional ARMA processes with an application to traffic data, *Econometrics and Statistics* **1**: 128–149.

Kokoszka, P., Miao, H. and Zhang, X. (2015). Functional dynamic factor model for intraday price curves, *Journal of Financial Econometrics* **13**: 456–477.

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 of latent factors for high-dimensional time series, *Biometrika* **98**: 901–918.

Li, D., Robinson, P. M. and Shang, H. L. (2020). Long-range dependent curve time series, *Journal of the American Statistical Association* **115**: 957–971.

Mackey, L. (2009). Deflation methods for sparse PCA, *Advances in Neural Information Processing Systems* **21**: 1017–1024.

Moghaddam, B., Weiss, Y. and Avidan, S. (2006). Generalized spectral bounds for sparse LDA, *Proceedings of the 23rd international conference on Machine learning*, pp. 641–648.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues, *The Review of Economics and Statistics* **92**: 1004–1016.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors, *Biometrika* **95**: 365–379.

Panaretos, V. and Tavakoli, S. (2013). Fourier analysis of stationary time series in function space, *The Annals of Statistics* **41**: 568–603.

Pena, D. and Box, G. E. (1987). Identifying a simplifying structure in time series, *Journal of the American statistical Association* **82**: 836–843.

Qiao, X., Qian, C., James, G. M. and Guo, S. (2020). Doubly functional graphical models in high dimensions, *Biometrika* **107**: 415–431.

Stock, J. and Watson, M. (2012). *Dynamic Factor Models*, Oxford University Press, Oxford.

Tavakoli, S., Nisol, G. and Hallin, M. (2021). Factor models for high-dimensional functional time series, *arXiv:1905.10325* .

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions, *The Annals of Statistics* **41**: 2905–2947.

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

Zhang, B., Pan, G., Yao, Q. and Zhou, W. (2021). Factor modelling for clustering high-dimensional time series, *arXiv:2101.01908* .

Zou, H. and Xue, L. (2018). A selective overview of sparse principal component analysis, *Proceedings of the IEEE* **106**: 1311–1320.

Supplementary material to "Factor Modelling for High-Dimensional Functional Time Series"

Shaojun Guo, Xinghao Qiao and Qingsong Wang

This supplementary material contains all technical proofs in Section B and additional empirical results in Section C.

B   Technical proofs

In addition to the notation defined in the main paper, we summarize here more notation to be used throughout the supplementary material. For a matrix $\mathbf{A} = (A_{ij})_{p \times q}$, we denote its Frobenius norm, matrix ℓ_1 norm and matrix ℓ_∞ norm by $\|\mathbf{A}\|_{\mathrm{F}} = (\sum_{i=1}^{p} \sum_{j=1}^{q} A_{ij}^2)^{1/2}$, $\|\mathbf{A}\|_1 = \max_{1 \leqslant j \leqslant q} \sum_{i=1}^{p} |A_{ij}|$ and $\|\mathbf{A}\|_\infty = \max_{1 \leqslant i \leqslant p} \sum_{j=1}^{q} |A_{ij}|$, respectively. For a matrix of bivariate functions $\mathbf{K} = (K_{ij}(\cdot, \cdot))_{p \times q}$ with each $K_{ij} \in \mathbb{H} \otimes \mathbb{H}$, we define functional versions of the Frobenius norm and matrix ℓ_1-norm by $\|\mathbf{K}\|_{\mathcal{S},\mathrm{F}} = (\sum_{i=1}^{p} \sum_{j=1}^{q} \|K_{ij}\|_{\mathcal{S}}^2)^{1/2}$ and $\|\mathbf{K}\|_{\mathcal{S},1} = \max_{1 \leqslant j \leqslant q} \sum_{i=1}^{p} \|K_{ij}\|_{\mathcal{S}}$, respectively.

B.1   Proof of Theorem 1

Recall that

$$\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \iint \widehat{\mathbf{\Sigma}}_{yy}^{(k)}(u, v) \widehat{\mathbf{W}}(u, v) \widehat{\mathbf{\Sigma}}_{yy}^{(k)}(u, v)^{\mathrm{T}} \, \mathrm{d}u \mathrm{d}v,$$

$$\widetilde{\mathbf{M}} = \sum_{k=1}^{k_0} \iint \mathbf{\Sigma}_{yy}^{(k)}(u, v) \widehat{\mathbf{W}}(u, v) \mathbf{\Sigma}_{yy}^{(k)}(u, v)^{\mathrm{T}} \, \mathrm{d}u \mathrm{d}v,$$

where $\widehat{\mathbf{W}}(u, v) = \mathbf{Q} \{ \mathbf{Q}^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_{yy}^{(0)}(v, v) \mathbf{Q} \}^{-1} \mathbf{Q}^{\mathrm{T}}$. Noting that $\widehat{\mathbf{W}}(u, v)$ is scale-invariant with respect to $\mathbf{Q}$, we assume here that the entries of $\mathbf{Q} = (Q_{ij})_{p \times q}$ are independent $N(0, p^{-1})$ instead of $N(0, 1)$ to ensure that the entries of $\mathbf{Q}^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_{yy}^{(0)}(v, v) \mathbf{Q}$ are $O_P(1)$ instead of $O_P(p^2)$. Provided that the maximum lag k_0 is fixed, without loss of generality we only need to consider the case $k_0 = 1$. We organize our proof in the following four steps.

Step 1. We will show that $\mathbf{Q}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v)\mathbf{Q}$ is positive definite with high probability uniformly over $\mathcal{U}$, i.e., there exists some positive constant $C > 0$ such that, with probability tending to one,

$$\inf_{v\in\mathcal{U}}\lambda_{\min}\{\mathbf{Q}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v)\mathbf{Q}\} > C. \tag{S.1}$$

Observe that $\mathbf{Q}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v)\mathbf{Q} = \mathbf{Q}^{\mathrm{T}}\boldsymbol{\Sigma}_{yy}^{(0)}(v,v)\mathbf{Q} + \mathbf{Q}^{\mathrm{T}}\left\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v) - \boldsymbol{\Sigma}_{yy}^{(0)}(v,v)\right\}\mathbf{Q}$. Then,

$$\left|\lambda_k\left\{\mathbf{Q}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v)\mathbf{Q}\right\} - \lambda_k\left\{\mathbf{Q}^{\mathrm{T}}\boldsymbol{\Sigma}_{yy}^{(0)}(v,v)\mathbf{Q}\right\}\right| \leqslant \left\|\mathbf{Q}^{\mathrm{T}}\left\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v) - \boldsymbol{\Sigma}_{yy}^{(0)}(v,v)\right\}\mathbf{Q}\right\|_{\mathrm{F}}, \tag{S.2}$$

where $\lambda_k(\mathbf{A})$ is the k-th largest eigenvalue of a symmetric matrix $\mathbf{A}$. By the Hanson–Wright inequality and union bound of probability, we obtain that there exists some constant $c > 0$ such that for any matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $t > 0$,

$$\mathbb{P}\left\{\frac{\|\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q} - \mathbb{E}\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}\|_{\mathrm{F}}}{\|\mathbf{A}\|_{\mathrm{F}}} \geqslant tqp^{-1}\right\} \leqslant 2q^2\exp\{-c\min(t^2,t)\}.$$

It follows from the concentration inequality above with $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v) - \boldsymbol{\Sigma}_{yy}^{(0)}(v,v)$, the independence between $\mathbf{Q}$ and $\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v)$, the Lipschitz-continuity of $\boldsymbol{\Sigma}_{yy}^{(0)}(v,v)$ and the uniform convergence rate in Condition 5 that

$$\begin{aligned}
\sup_{v\in\mathcal{U}}\left\|\mathbf{Q}^{\mathrm{T}}\left\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v) - \boldsymbol{\Sigma}_{yy}^{(0)}(v,v)\right\}\mathbf{Q}\right\|_{\mathrm{F}} &\leqslant O_P\left\{qp^{-1}\{\log(qn)\}^{1/2}\sup_{v\in\mathcal{U}}\|\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v) - \boldsymbol{\Sigma}_{yy}^{(0)}(v,v)\|_{\mathrm{F}}\right\} \\
&= O_P\{q\log(qn)n^{-1/2}\}.
\end{aligned}$$

Combining the above result with (S.2) and the fact that

$$\inf_{v\in\mathcal{U}}\lambda_q\{\mathbf{Q}^{\mathrm{T}}\boldsymbol{\Sigma}_{yy}^{(0)}(v,v)\mathbf{Q}\} \geqslant \inf_{v\in\mathcal{U}}\lambda_q\{\mathbf{Q}^{\mathrm{T}}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{(0)}(v,v)\mathbf{Q}\} > c > 0$$

under Condition 2 implies that (S.1) holds with high probability.

Step 2. The bounds of r largest eigenvalues of $\widecheck{\mathbf{M}}$ (i.e., $\nu_1 \geqslant \cdots \geqslant \nu_r$) satisfy

$$\nu_j \asymp p^{2-2\delta}, \quad j = 1,\cdots,r, \tag{S.3}$$

with probability tending to one.

First, note that

$$\nu_r \leqslant \nu_1 \leqslant \|\widecheck{\mathbf{M}}\| \leqslant \iint \|\boldsymbol{\Sigma}_{yy}^{(1)}(u,v)\|^2\|\widehat{\mathbf{W}}(u,v)\|\,\mathrm{d}u\mathrm{d}v. \tag{S.4}$$

2

Applying the fact $\|\mathbf{Q}\| \asymp 1 + o_P(1)$ and (S.1) from Step 1, we have that

$$\sup_{u,v\in\mathcal{U}} \|\widehat{\mathbf{W}}(u,v)\| \leqslant \|\mathbf{Q}\|^2 \big[\inf_{v\in\mathcal{U}} \lambda_{\min}\{\mathbf{Q}^T\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v)\mathbf{Q}\}\big]^{-1} \lesssim 1 + o_P(1). \tag{S.5}$$

It follows from Conditions 3–4, the decomposition (3) and Lemma B2 that

$$
\begin{aligned}
\iint \big\|\boldsymbol{\Sigma}_{yy}^{(1)}(u,v)\big\|^2 \,\mathrm{d}u\mathrm{d}v &\leqslant \iint \Big(\big\|\mathbf{A}\boldsymbol{\Sigma}_{xx}^{(1)}(u,v)\mathbf{A}^{\mathrm{T}}\big\| + \big\|\mathbf{A}\boldsymbol{\Sigma}_{x\varepsilon}^{(1)}(u,v)\big\|\Big)^2 \,\mathrm{d}u\mathrm{d}v \\
&\leqslant 2\Big\{\|\mathbf{A}\|^4 \iint \big\|\boldsymbol{\Sigma}_{xx}^{(1)}(u,v)\big\|^2 \,\mathrm{d}u\mathrm{d}v + \|\mathbf{A}\|^2 \iint \big\|\boldsymbol{\Sigma}_{x\varepsilon}^{(1)}(u,v)\big\|^2 \mathrm{d}u\mathrm{d}v\Big\} \\
&= O\Big\{p^{2-2\delta} + p^{1-\delta} \iint \big\|\boldsymbol{\Sigma}_{x\varepsilon}^{(1)}(u,v)\big\|^2 \,\mathrm{d}u\mathrm{d}v\Big\}, \tag{S.6}
\end{aligned}
$$

where, by Condition 4,

$$\iint \big\|\boldsymbol{\Sigma}_{x\varepsilon}^{(1)}(u,v)\big\|^2 \,\mathrm{d}u\mathrm{d}v \leqslant \big\|\boldsymbol{\Sigma}_{x\varepsilon}^{(1)}\big\|_{\mathcal{S},\mathrm{F}}^2 = O\Big\{\max_{1\leqslant i\leqslant r} \sum_{j=1}^p \big\|\Sigma_{x\epsilon,ij}^{(1)}\big\|_{\mathcal{S}}^2\Big\} \leqslant O\Big\{\big\|\boldsymbol{\Sigma}_{x\varepsilon}^{(1)}\big\|_{\mathcal{S},\infty}^2\Big\} = o(p^{1-\delta}). \tag{S.7}$$

Combining (S.4)–(S.7) yields that $\nu_r \lesssim p^{2-2\delta}\{1 + o_P(1)\}$.

We next give a lower bound on ν_r. By Conditions 1, 3, Lemma 3 of Wang and Xi (1997) and (S.1) from Step 1, we first obtain that

$$
\begin{aligned}
I_1 &= \lambda_r \left\{\iint \boldsymbol{\Sigma}_{xx}^{(1)}(u,v)\mathbf{A}^{\mathrm{T}}\widehat{\mathbf{W}}(v,v)\mathbf{A}\boldsymbol{\Sigma}_{xx}^{(1)}(u,v)^{\mathrm{T}} \,\mathrm{d}u\mathrm{d}v\right\} \\
&\geqslant \lambda_r \left\{\iint \boldsymbol{\Sigma}_{xx}^{(1)}(u,v)\boldsymbol{\Sigma}_{xx}^{(1)}(u,v)^{\mathrm{T}} \,\mathrm{d}u\mathrm{d}v\right\} \inf_{u,v\in\mathcal{U}} \lambda_r\left\{\mathbf{A}^{\mathrm{T}}\widehat{\mathbf{W}}(u,v)\mathbf{A}\right\} \\
&\gtrsim \lambda_r(\mathbf{A}^{\mathrm{T}}\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A}) \cdot \inf_{v\in\mathcal{U}} \lambda_{q-r+1}\left(\{\mathbf{Q}^T\widehat{\boldsymbol{\Sigma}}_{yy}^{(0)}(v,v)\mathbf{Q}\}^{-1}\right) \\
&\gtrsim p^{1-\delta}\{1 + o_P(1)\}.
\end{aligned}
$$

It follows from (S.5), (S.7), Lemma B2 and Condition 3 that

$$
\begin{aligned}
I_2 &= \left\|\iint \boldsymbol{\Sigma}_{x\varepsilon}^{(1)}(u,v)\widehat{\mathbf{W}}(u,v)\boldsymbol{\Sigma}_{xx}^{(1)}(u,v)\mathbf{A}^{\mathrm{T}} \,\mathrm{d}u\mathrm{d}v\right\| \\
&\leqslant \left\{\iint \|\boldsymbol{\Sigma}_{x\varepsilon}^{(1)}(u,v))\|^2 \,\mathrm{d}u\mathrm{d}v \iint \|\boldsymbol{\Sigma}_{xx}^{(1)}(u,v)\|^2 \,\mathrm{d}u\mathrm{d}v\right\}^{1/2} \sup_{u,v\in\mathcal{U}} \|\widehat{\mathbf{W}}(u,v)\|\|\mathbf{A}\| = o_P(p^{1-\delta}),
\end{aligned}
$$

By Weyl's inequality, the decomposition (3) and Condition 3, ν_r can be bounded from below

by

$$\nu_r = \lambda_r \Big\{ \iint \mathbf{\Sigma}_{yy}^{(1)}(u,v) \widehat{\mathbf{W}}(u,v) \mathbf{\Sigma}_{yy}^{(1)}(u,v)^{\mathrm{T}} \, \mathrm{d}u\mathrm{d}v \Big\}$$

$$\geqslant \lambda_r(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \cdot \lambda_r \Big[\iint \big\{ \mathbf{\Sigma}_{xx}^{(1)}(u,v)\mathbf{A}^{\mathrm{T}} + \mathbf{\Sigma}_{x\varepsilon}^{(1)}(u,v) \big\} \widehat{\mathbf{W}}(u,v) \big\{ \mathbf{\Sigma}_{xx}^{(1)}(u,v)\mathbf{A}^{\mathrm{T}} + \mathbf{\Sigma}_{x\varepsilon}^{(1)}(u,v) \big\}^{\mathrm{T}} \, \mathrm{d}u\mathrm{d}v \Big]$$

$$\geqslant \|\mathbf{A}\|_{\min}^2 \cdot \lambda_r \Big\{ \iint \mathbf{\Sigma}_{xx}^{(1)}(u,v)\mathbf{A}^{\mathrm{T}} \widehat{\mathbf{W}}(u,v) \mathbf{A}\mathbf{\Sigma}_{xx}^{(1)}(u,v)^{\mathrm{T}} \, \mathrm{d}u\mathrm{d}v \Big\}$$

$$+ 2\|\mathbf{A}\|_{\min}^2 \cdot \lambda_r \Big\{ \iint \mathbf{\Sigma}_{x\varepsilon}^{(1)}(u,v) \widehat{\mathbf{W}}(u,v) \mathbf{\Sigma}_{xx}^{(1)}(u,v)\mathbf{A}^{\mathrm{T}} \, \mathrm{d}u\mathrm{d}v \Big\}$$

$$\geqslant \|\mathbf{A}\|_{\min}^2 \cdot (I_1 - 2I_2) \gtrsim p^{2-2\delta}\{1 + o_P(1)\}.$$

Hence we obtain that $\nu_r \asymp p^{2-2\delta}$ with high probability.

Step 3. We will show that

$$\|\widehat{\mathbf{M}} - \widetilde{\mathbf{M}}\| = O_P(p^{2-\delta}n^{-1/2}). \tag{S.8}$$

Observe that

$$\widehat{\mathbf{\Sigma}}_{yy}^{(1)}(u,v) \widehat{\mathbf{W}}(u,v) \widehat{\mathbf{\Sigma}}_{yy}^{(1)}(u,v)^{\mathrm{T}} - \mathbf{\Sigma}_{yy}^{(1)}(u,v) \widehat{\mathbf{W}}(u,v) \mathbf{\Sigma}_{yy}^{(1)}(u,v)^{\mathrm{T}}$$

$$= \big\{ \widehat{\mathbf{\Sigma}}_{yy}^{(1)}(u,v) - \mathbf{\Sigma}_{yy}^{(1)}(u,v) \big\} \widehat{\mathbf{W}}(u,v) \big\{ \widehat{\mathbf{\Sigma}}_{yy}^{(1)}(u,v) - \mathbf{\Sigma}_{yy}^{(1)}(u,v) \big\}^{\mathrm{T}}$$

$$+ \big\{ \widehat{\mathbf{\Sigma}}_{yy}^{(1)}(u,v) - \mathbf{\Sigma}_{yy}^{(1)}(u,v) \big\} \widehat{\mathbf{W}}(u,v) \mathbf{\Sigma}_{yy}^{(1)}(u,v)^{\mathrm{T}}$$

$$+ \mathbf{\Sigma}_{yy}^{(1)}(u,v) \widehat{\mathbf{W}}(u,v) \big\{ \widehat{\mathbf{\Sigma}}_{yy}^{(1)}(u,v) - \mathbf{\Sigma}_{yy}^{(1)}(u,v) \big\}^{\mathrm{T}}.$$

By $\|\mathbf{Q}\| \asymp 1 + o_P(1)$, Condition 5 and Markov's inequality, we have that

$$\iint \Big\| \big\{ \widehat{\mathbf{\Sigma}}_{yy}^{(1)}(u,v) - \mathbf{\Sigma}_{yy}^{(1)}(u,v) \big\} \mathbf{Q} \Big\|^2 \, \mathrm{d}u\mathrm{d}v \ \leqslant \sum_{i=1}^{p} \sum_{j=1}^{p} \Big\| \widehat{\Sigma}_{yy,ij}^{(1)} - \Sigma_{yy,ij}^{(1)} \Big\|_{\mathcal{S}}^2 \|\mathbf{Q}\|^2 = O_P(p^2 n^{-1}).$$

Furthermore, (S.6) and (S.7) imply that

$$\iint \Big\| \mathbf{\Sigma}_{yy}^{(1)}(u,v)\mathbf{Q} \Big\|^2 \, \mathrm{d}u\mathrm{d}v \leqslant \iint \Big\| \mathbf{\Sigma}_{yy}^{(1)}(u,v) \Big\|^2 \, \mathrm{d}u\mathrm{d}v \, \|\mathbf{Q}\|^2 = p^{2-2\delta}\{1 + o_P(1)\}.$$

It follows from the above decomposition, two upper bound results, Cauchy–Schwartz inequality and $\sup_{v\in\mathcal{U}} \lambda_{\max}\big(\{\mathbf{Q}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{yy}^{(0)}(v,v)\mathbf{Q}\}^{-1} \big) = O_P(1)$ from Step 1 that (S.8) holds.

Step 4. Combining (S.3) and (S.8) from Steps 2 and 3 with Lemma 3 and the proof of Theorem 1 in Lam et al. (2011), we have that there exists $\widetilde{\mathbf{K}}$ and an orthogonal matrix $\mathbf{U}$ such that $\widehat{\mathbf{K}} = \widetilde{\mathbf{K}}\mathbf{U}$ and

$$\|\widetilde{\mathbf{K}} - \mathbf{K}\| = O_P(p^{\delta}n^{-1/2}).$$

It then follows from the orthogonality of $\mathbf{K}$ that

$$
\begin{aligned}
\mathcal{D}^2\Big(\mathcal{C}(\mathbf{K}),\mathcal{C}(\widehat{\mathbf{K}})\Big) &= r^{-1}\mathrm{tr}\Big(\mathbf{I} - \mathbf{KK}^{\mathrm{T}}\widehat{\mathbf{K}}\widehat{\mathbf{K}}^{\mathrm{T}}\Big) = r^{-1}\mathrm{tr}(\mathbf{K}^{\mathrm{T}}(\mathbf{I} - \widehat{\mathbf{K}}\widehat{\mathbf{K}}^{\mathrm{T}})\mathbf{K}) \\
&= r^{-1}\Big\{\mathrm{tr}(\mathbf{K}^{\mathrm{T}}(\mathbf{I} - \widehat{\mathbf{K}}\widehat{\mathbf{K}}^{\mathrm{T}})\mathbf{K}) - \mathrm{tr}(\mathbf{K}^{\mathrm{T}}(\mathbf{I} - \mathbf{KK}^{\mathrm{T}})\mathbf{K})\Big\} \\
&= r^{-1}\mathrm{tr}\Big(\mathbf{K}^{\mathrm{T}}(\mathbf{KK}^{\mathrm{T}} - \widehat{\mathbf{K}}\widehat{\mathbf{K}}^{\mathrm{T}})\mathbf{K}\Big) \leqslant \Big\|\mathbf{K}^{\mathrm{T}}(\mathbf{KK}^{\mathrm{T}} - \widetilde{\mathbf{K}}\widetilde{\mathbf{K}}^{\mathrm{T}})\mathbf{K}\Big\|,
\end{aligned}
$$

where the last term is further bounded by

$$
\| -\mathbf{K}^{\mathrm{T}}(\mathbf{K} - \widetilde{\mathbf{K}})(\mathbf{K} - \widetilde{\mathbf{K}})^{\mathrm{T}}\mathbf{K} + (\mathbf{K} - \widetilde{\mathbf{K}})^{\mathrm{T}}(\mathbf{K} - \widetilde{\mathbf{K}})\| \leqslant 2\|\widetilde{\mathbf{K}} - \mathbf{K}\|^2 = O_P(p^{2\delta}n^{-1}).
$$

The result in Theorem 1 follows immediately, which completes the proof. □

B.2   Proof of Theorem 2

Denote $(\nu_j, \boldsymbol{\gamma}_j)$ and $(\hat{\nu}_j, \widehat{\boldsymbol{\gamma}}_j)$ by the eigenpairs of $\widetilde{\mathbf{M}}$ and $\widehat{\mathbf{M}}$, respectively. For simplicity we assume $\boldsymbol{\gamma}_j^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_j \geqslant 0$ here. First note that for $j = 1, 2, \ldots, r$,

$$
\hat{\nu}_j - \nu_j = \widehat{\boldsymbol{\gamma}}_j^{\mathrm{T}}\widehat{\mathbf{M}}\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^{\mathrm{T}}\widetilde{\mathbf{M}}\boldsymbol{\gamma}_j = I_1 + I_2 + I_3 + I_4 + I_5,
$$

where

$$
\begin{aligned}
I_1 &= (\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)^{\mathrm{T}}(\widehat{\mathbf{M}} - \widetilde{\mathbf{M}})\widehat{\boldsymbol{\gamma}}_j, \qquad I_2 = (\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)^{\mathrm{T}}\widetilde{\mathbf{M}}(\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j) \\
I_3 &= (\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)^{\mathrm{T}}\widetilde{\mathbf{M}}\boldsymbol{\gamma}_j, \qquad I_4 = \boldsymbol{\gamma}_j^{\mathrm{T}}(\widehat{\mathbf{M}} - \widetilde{\mathbf{M}})\widehat{\boldsymbol{\gamma}}_j, \qquad I_5 = \boldsymbol{\gamma}_j^{\mathrm{T}}\widetilde{\mathbf{M}}(\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j).
\end{aligned}
$$

Since $\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\| \leqslant \|\widetilde{\mathbf{K}} - \mathbf{K}\| = O_P(p^{\delta}n^{-1/2})$, $\|\widehat{\mathbf{M}} - \widetilde{\mathbf{M}}\| = O_P(p^{2-\delta}n^{-1/2})$ and $\|\widetilde{\mathbf{M}}\| = O_P(p^{2-2\delta})$, we have that

$$
|I_1| + |I_2| = O_P(p^2 n^{-1}), \qquad |I_3| + |I_4| + |I_5| = O_P(p^{2-\delta}n^{-1/2}),
$$

which means that $|\hat{\nu}_j - \nu_j| = O_P(p^{2-\delta}n^{-1/2})$ under $p^{\delta}n^{-1/2} = o(1)$.

To prove the result for $j = r+1, \ldots, p$, we introduce

$$
\overline{\mathbf{M}} = \sum_{k=1}^{k_0} \iint \widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}(u, v)\widehat{\mathbf{W}}(v, v)\boldsymbol{\Sigma}_{yy}^{(k)}(u, v)^{\mathrm{T}}\,\mathrm{d}u\mathrm{d}v.
$$

For each $j = r+1, \ldots, p$, consider the following decomposition

$$
\hat{\nu}_j = \widehat{\boldsymbol{\gamma}}_j^{\mathrm{T}}\widehat{\mathbf{M}}\widehat{\boldsymbol{\gamma}}_j = I_6 + I_7 + I_8,
$$

5

where

$$I_6 = \widehat{\boldsymbol{\gamma}}_j^{\mathrm{T}}(\widehat{\mathbf{M}} - \overline{\mathbf{M}} - \overline{\mathbf{M}}^{\mathrm{T}} + \widetilde{\mathbf{M}})\widehat{\boldsymbol{\gamma}}_j, \qquad I_7 = 2\widehat{\boldsymbol{\gamma}}_j^{\mathrm{T}}(\overline{\mathbf{M}} - \widetilde{\mathbf{M}})(\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j),$$

$$I_8 = (\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)^{\mathrm{T}}\widetilde{\mathbf{M}}(\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j).$$

Note that, under Condition 5,

$$|I_6| \leqslant \|\widehat{\mathbf{M}} - \overline{\mathbf{M}} - \overline{\mathbf{M}}^{\mathrm{T}} + \widetilde{\mathbf{M}}\| \lesssim \sum_{k=1}^{k_0} \|\widehat{\boldsymbol{\Sigma}}_{yy}^{(k)} - \boldsymbol{\Sigma}_{yy}^{(k)}\|_{\mathcal{S}}^2 = O_P(p^2 n^{-1}).$$

By the similar arguments to prove Theorem 1, we can show that $\|\overline{\mathbf{M}} - \widetilde{\mathbf{M}}\| = O_P(p^{2-\delta}n^{-1/2})$ and $\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\| = O_P(p^\delta n^{-1/2})$, which together imply that $|I_7| + |I_8| = O_P(p^2 n^{-1})$. Combining the above results yields that $\hat{\nu}_j = O_p(p^2 n^{-1})$ for $j = r+1, \ldots, p$, which completes the proof of the first part.

We next turn to prove the second part. It follows from $\nu_j \asymp p^{2-2\delta}$ for $j = 1, \ldots, r$ in (S.3) and $p^\delta n^{-1/2} = o(1)$ that

$$\frac{|\hat{\nu}_j - \nu_j|}{\nu_j} = O_P(p^{2-\delta}n^{-1/2}p^{2\delta-2}) = O_P(p^\delta n^{-1/2}) = o_P(1), \quad j = 1, \ldots, r.$$

As a result, for $j = 1, \ldots, r-1$,

$$\frac{\hat{\nu}_{j+1}}{\hat{\nu}_j} = \frac{\nu_{j+1}}{\nu_j}\frac{1 + \nu_{j+1}^{-1}(\hat{\nu}_{j+1} - \nu_{j+1})}{1 + \nu_j^{-1}(\hat{\nu}_j - \nu_j)} = \frac{\nu_{j+1}}{\nu_j}\frac{1 + o_P(1)}{1 + o_P(1)} \asymp 1,$$

and

$$\frac{\hat{\nu}_{r+1}}{\hat{\nu}_r} = \frac{1}{\nu_r}\frac{\hat{\nu}_{r+1}}{1 + \nu_r^{-1}(\hat{\nu}_r - \nu_r)} = O_P(p^{2\delta-2}p^2 n^{-1}) = O_P(p^{2\delta}n^{-1}).$$

We complete the proof of the second part. □

B.3 Proof of Lemma 1

Denote by $\mathbf{A} = (A_{ij})_{p \times r}$. By the decomposition (3) and the fact $(x + y)^\tau \leqslant x^\tau + y^\tau$ for $x, y \geqslant 0$ and $\tau \in [0, 1)$, we have that

$$\|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}^\tau \leqslant \left\| \sum_{1 \leqslant m,m' \leqslant r} A_{im}\Sigma_{xx,mm'}^{(k)}A_{jm'} \right\|_{\mathcal{S}}^\tau + \left\| \sum_{m=1}^r A_{im}\Sigma_{x\epsilon,mj}^{(k)} \right\|_{\mathcal{S}}^\tau$$

$$\leqslant \sum_{1 \leqslant m,m' \leqslant r} \left\| A_{im}\Sigma_{xx,mm'}^{(k)}A_{jm'} \right\|_{\mathcal{S}}^\tau + \sum_{m=1}^r \left\| A_{im}\Sigma_{x\epsilon,mj}^{(k)} \right\|_{\mathcal{S}}^\tau$$

$$= \sum_{1 \leqslant m,m' \leqslant r} \left| A_{im}A_{jm'} \right|^\tau \left\| \Sigma_{xx,mm'}^{(k)} \right\|_{\mathcal{S}}^\tau + \sum_{m=1}^r \left| A_{im} \right|^\tau \left\| \Sigma_{x\epsilon,mj}^{(k)} \right\|_{\mathcal{S}}^\tau,$$

6

where, under Conditions 1 and 2, it holds that

$$\big\|\Sigma^{(k)}_{xx,mm'}\big\|^2_{\mathcal{S}} = \iint [\mathbb{E}X_{(t+k)m}(u)X_{t,m'}(v)]^2 \,\mathrm{d}u\mathrm{d}v \leqslant \mathbb{E}\|X_{(t+k)m}\|^2\,\mathbb{E}\|X_{tm'}\|^2 = O(1),$$

$$\max_{1\leqslant j\leqslant p}\big\|\Sigma^{(k)}_{x\epsilon,mj}\big\|^2_{\mathcal{S}} \leqslant \mathbb{E}\|X_{(t+k)m}\|^2 \max_{1\leqslant j\leqslant p}\mathbb{E}\|\epsilon_{tj}\|^2 = O(1). \tag{S.9}$$

For each j, summing over i and applying (S.9) under Condition 6 with fixed L and r, we obtain that

$$\max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\big\|\Sigma^{(k)}_{yy,ij}\big\|^\tau_{\mathcal{S}} \leqslant \max_{1\leqslant j\leqslant p}\sum_{1\leqslant m,m'\leqslant r}\sum_{i=1}^{p}\Big\{|A_{im}A_{jm'}|^\tau\|\Sigma^{(k)}_{xx,mm'}\|^\tau_{\mathcal{S}}\Big\} + \max_{1\leqslant j\leqslant p}\sum_{m=1}^{r}\sum_{i=1}^{p}\Big\{|A_{im}|^\tau\|\Sigma^{(k)}_{x\epsilon,mj}\|^\tau_{\mathcal{S}}\Big\}$$

$$\leqslant c_1(p)\Big\{r^2\cdot\max_{\substack{1\leqslant j\leqslant p,\\ 1\leqslant m,m'\leqslant r}}|A_{jm'}|^\tau\|\Sigma^{(k)}_{xx,mm'}\|^\tau_{\mathcal{S}} + r\cdot\max_{\substack{1\leqslant j\leqslant p,\\ 1\leqslant m\leqslant r}}\|\Sigma^{(k)}_{x\epsilon,mj}\|^\tau_{\mathcal{S}}\Big\} \lesssim c_1(p).$$

By the similar arguments above and Condition $4'$,

$$\max_{1\leqslant i\leqslant p}\sum_{j=1}^{p}\big\|\Sigma^{(k)}_{yy,ij}\big\|^\tau_{\mathcal{S}} \leqslant \max_{1\leqslant i\leqslant p}\sum_{1\leqslant m,m'\leqslant r}\sum_{j=1}^{p}\Big\{|A_{im}A_{jm'}|^\tau\|\Sigma^{(k)}_{xx,mm'}\|^\tau_{\mathcal{S}}\Big\} + \max_{1\leqslant i\leqslant p}\sum_{m=1}^{r}\sum_{j=1}^{p}\Big\{|A_{im}|^\tau\|\Sigma^{(k)}_{x\epsilon,mj}\|^\tau_{\mathcal{S}}\Big\}$$

$$\leqslant c_1(p)r^2\cdot\max_{\substack{1\leqslant i\leqslant p,\\ 1\leqslant m,m'\leqslant r}}|A_{im}|^\tau\|\Sigma^{(k)}_{xx,mm'}\|^\tau_{\mathcal{S}} + c_2(p)r\cdot\max_{\substack{1\leqslant i\leqslant p,\\ 1\leqslant m\leqslant r}}|A_{im}|^\tau$$

$$\lesssim c_1(p) + c_2(p).$$

Hence we complete the proof of this lemma. □

B.4 Proof of Theorem 3

We organize our proof in the following three steps.

Step 1. With the choice of $\eta_k \asymp (\log p/n)^{1/2}$, we will show that

$$\big\|\mathcal{T}_{\eta_k}\{\widehat{\boldsymbol{\Sigma}}^{(k)}_{yy}\} - \boldsymbol{\Sigma}^{(k)}_{yy}\big\|_{\mathcal{S},1} = O_P\left\{c_1(p)\mathcal{M}_y^{1-\tau}\Big(\frac{\log p}{n}\Big)^{\frac{1-\tau}{2}}\right\},$$

$$\big\|\mathcal{T}_{\eta_k}\{\widehat{\boldsymbol{\Sigma}}^{(k)}_{yy}\} - \boldsymbol{\Sigma}^{(k)}_{yy}\big\|_{\mathcal{S},\infty} = O_P\left[\{c_1(p)+c_2(p)\}\mathcal{M}_y^{1-\tau}\Big(\frac{\log p}{n}\Big)^{\frac{1-\tau}{2}}\right]. \tag{S.10}$$

By the definition of functional matrix ℓ_1 norm $\|\cdot\|_{\mathcal{S},1}$ and the triangle inequality,

$$\big\|\mathcal{T}_{\eta_k}\{\widehat{\boldsymbol{\Sigma}}^{(k)}_{yy}\} - \boldsymbol{\Sigma}^{(k)}_{yy}\big\|_{\mathcal{S},1} \leqslant \big\|\mathcal{T}_{\eta_k}\{\boldsymbol{\Sigma}^{(k)}_{yy}\} - \boldsymbol{\Sigma}^{(k)}_{yy}\big\|_{\mathcal{S},1} + \big\|\mathcal{T}_{\eta_k}\{\widehat{\boldsymbol{\Sigma}}^{(k)}_{yy}\} - \mathcal{T}_{\eta_k}\{\boldsymbol{\Sigma}^{(k)}_{yy}\}\big\|_{\mathcal{S},1}.$$

For the first term, it follows from Lemma 1 that

$$\big\|\mathcal{T}_{\eta_k}\{\boldsymbol{\Sigma}^{(k)}_{yy}\} - \boldsymbol{\Sigma}^{(k)}_{yy}\big\|_{\mathcal{S},1} = \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\|\Sigma^{(k)}_{yy,ij}\|_{\mathcal{S}}I\{\|\Sigma^{(k)}_{yy,ij}\|_{\mathcal{S}} < \eta_k\}$$

$$< \eta_k^{1-\tau}\max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\|\Sigma^{(k)}_{yy,ij}\|^\tau_{\mathcal{S}} \lesssim \eta_k^{1-\tau}c_1(p). \tag{S.11}$$

For the second term, observe that

$$\big\|\mathcal{T}_{\eta_k}\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}\} - \mathcal{T}_{\eta_k}\{\boldsymbol{\Sigma}_{yy}^{(k)}\}\big\|_{\mathcal{S},1} = \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\big\|\widehat{\Sigma}_{yy,ij}^{(k)}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k\} - \Sigma_{yy,ij}^{(k)}I\{\|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k\}\big\|_{\mathcal{S}}$$

$$\leqslant \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\big\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\big\|_{\mathcal{S}}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k, \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k\}$$

$$+ \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\big\|\widehat{\Sigma}_{yy,ij}^{(k)}\big\|_{\mathcal{S}}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k, \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}<\eta_k\}$$

$$+ \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\big\|\Sigma_{yy,ij}^{(k)}\big\|_{\mathcal{S}}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}}<\eta_k, \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k\}$$

$$:= I_1 + I_2 + I_3.$$

Denote $Z = \max\limits_{1\leqslant i,j\leqslant p}\big\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\big\|_{\mathcal{S}}$. We first bound I_1. By Lemma 1,

$$I_1 \leqslant \max_{1\leqslant i,j\leqslant p}\big\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\big\|_{\mathcal{S}} \cdot \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k, \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k\}$$

$$\leqslant Z \cdot \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}I\{\|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k\} \lesssim Z\eta_k^{-\tau}c_1(p).$$

We next bound I_2. By the triangle inequality and Lemma 1,

$$I_2 \leqslant \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\big\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\big\|_{\mathcal{S}}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k, \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}<\eta_k\}$$

$$+ \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\big\|\Sigma_{yy,ij}^{(k)}\big\|_{\mathcal{S}}I\{\|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}<\eta_k\}$$

$$\lesssim \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\big\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\big\|_{\mathcal{S}}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k, \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}<\eta_k\} + \eta_k^{1-\tau}c_1(p)$$

$$:= I_4 + \eta_k^{1-\tau}c_1(p).$$

We take certain $\theta \in (0,1)$. Let $N(t) = \max\limits_{1\leqslant j\leqslant p}\sum_{i=1}^{p}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant t\}$. By the triangle inequality and Lemma 1,

$$I_4 \leqslant \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\big\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\big\|_{\mathcal{S}}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k, \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}\leqslant\theta\eta_k\}$$

$$+ \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}\big\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\big\|_{\mathcal{S}}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant\eta_k, \theta\eta_k<\|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}<\eta_k\}$$

$$\leqslant Z \cdot \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}I\{\|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}\geqslant(1-\theta)\eta_k\} + Z \cdot \max_{1\leqslant j\leqslant p}\sum_{i=1}^{p}I\{\|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}>\theta\eta_k\}$$

$$\lesssim Z \cdot N\{(1-\theta)\eta_k\} + Z(\theta\eta_k)^{-\tau}c_1(p).$$

8

The above bounds imply that

$$I_2 \lesssim Z \cdot N\{(1-\theta)\eta_k\} + \eta_k^{1-\tau} c_1(p) + Z(\theta\eta_k)^{-\tau} c_1(p).$$

We finally bound I_3. By the triangle inequality and Lemma 1,

$$I_3 \leqslant \max_{1\leqslant j\leqslant p} \sum_{i=1}^{p} \|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}} < \eta_k, \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}} \geqslant \eta_k\}$$

$$+ \max_{1\leqslant j\leqslant p} \sum_{i=1}^{p} \|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}} < \eta_k, \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}} \geqslant \eta_k\}$$

$$\leqslant Z \cdot \max_{1\leqslant j\leqslant p} \sum_{i=1}^{p} I\{\|\widehat{\Sigma}_{yy,ij}^{(k)}\|_{\mathcal{S}} \geqslant \eta_k\} + \eta_k \max_{1\leqslant j\leqslant p} \sum_{i=1}^{p} I\{\|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}} \geqslant \eta_k\}$$

$$\lesssim Z\eta_k^{-\tau} c_1(p) + \eta_k^{1-\tau} c_1(p).$$

Combining the bounds for I_1, I_2 and I_3 yields that

$$\left\|\mathcal{T}_{\eta_n}\{\widehat{\mathbf{\Sigma}}_{yy}^{(k)}\} - \mathcal{T}_{\eta_k}\{\mathbf{\Sigma}_{yy}^{(k)}\}\right\|_{\mathcal{S},1} \lesssim Z\eta_k^{-\tau} c_1(p) + \eta_k^{1-\tau} c_1(p) + Z(\theta\eta_k)^{-\tau} c_1(p) + Z\cdot N\{(1-\theta)\eta_k\}. \quad \text{(S.12)}$$

By Lemma B3 and the choice of $\eta_k = c_3 \mathcal{M}_y (\log p/n)^{1/2} = o(1)$ for sufficiently large c_3,

$$\mathbb{P}\Big(N\{(1-\theta)\eta_k\} > 0\Big) = \mathbb{P}\Big(\max_{1\leqslant i,j\leqslant p} \|\widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}} \geqslant (1-\theta)\eta_k\Big)$$

$$\leqslant c_1 p^2 \exp\{-c_2(1-\theta)^2 c_3^2 \log p\} \to 0.$$

Hence $N\{(1-\theta)\eta_k\} = o_P(1)$. Applying Lemma B3 again, we obtain that $Z = O_P(\eta_k)$, which together with (S.11) and (S.12) imply that

$$\left\|\mathcal{T}_{\eta_n}\{\widehat{\mathbf{\Sigma}}_{yy}^{(k)}\} - \mathbf{\Sigma}_{yy}^{(k)}\right\|_{\mathcal{S},1} = O_P\left\{c_1(p)\mathcal{M}_y^{1-\tau}\Big(\frac{\log p}{n}\Big)^{\frac{1-\tau}{2}}\right\}.$$

By the similar procedure, the second result of (S.10) can be derived and the proof is omitted.

Step 2. We will show that

$$\left\|\widetilde{\mathbf{M}} - \overset{\smile}{\mathbf{M}}\right\| = O_P\left[c_1(p)\{c_1(p) + c_2(p)\}\mathcal{M}_y^{1-\tau}\Big(\frac{\log p}{n}\Big)^{\frac{1-\tau}{2}}\right]. \quad \text{(S.13)}$$

Note that

$$\left\|\widetilde{\mathbf{M}} - \overset{\smile}{\mathbf{M}}\right\|$$

$$= O_P\Bigg(\left\|\iint \Big[\mathcal{T}_{\eta_k}\{\widehat{\mathbf{\Sigma}}_{yy}^{(k)}\}(u,v) - \mathbf{\Sigma}_{yy}^{(k)}(u,v)\Big]\widehat{\mathbf{W}}(u,v)\Big[\mathcal{T}_{\eta_k}\{\widehat{\mathbf{\Sigma}}_{yy}^{(k)}\}(u,v) - \mathbf{\Sigma}_{yy}^{(k)}(u,v)\Big]^{\mathrm{T}} \mathrm{d}u\mathrm{d}v\right\|$$

$$+ \left\|\iint \Big[\mathcal{T}_{\eta_k}\{\widehat{\mathbf{\Sigma}}_{yy}^{(k)}\}(u,v) - \mathbf{\Sigma}_{yy}^{(k)}(u,v)\Big]\widehat{\mathbf{W}}(u,v)\mathbf{\Sigma}_{yy}^{(k)}(u,v)^{\mathrm{T}} \mathrm{d}u\mathrm{d}v\right\|\Bigg).$$

9

By Lemma B4 and (S.5), (S.10), we obtain that

$$\left\| \iint \left[\mathcal{T}_{\eta_k}\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}\}(u,v) - \boldsymbol{\Sigma}_{yy}^{(k)}(u,v) \right] \widehat{\mathbf{W}}(u,v) \left[\mathcal{T}_{\eta_k}\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}\}(u,v) - \boldsymbol{\Sigma}_{yy}^{(k)}(u,v) \right]^{\mathrm{T}} \mathrm{d}u\mathrm{d}v \right\|$$

$$\lesssim \left\| \mathcal{T}_{\eta_k}\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}\} - \boldsymbol{\Sigma}_{yy}^{(k)} \right\|_{\mathcal{S},1} \left\| \mathcal{T}_{\eta_k}\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}\} - \boldsymbol{\Sigma}_{yy}^{(k)} \right\|_{\mathcal{S},\infty} \{1 + o_P(1)\}$$

$$= O_P\left[c_1(p)\{c_1(p) + c_2(p)\} \mathcal{M}_y^{2(1-\tau)} \left(\frac{\log p}{n} \right)^{1-\tau} \right]$$

and

$$\left\| \iint \left[\mathcal{T}_{\eta_k}\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}\}(u,v) - \boldsymbol{\Sigma}_{yy}^{(k)}(u,v) \right] \widehat{\mathbf{W}}(u,v) \boldsymbol{\Sigma}_{yy}^{(k)}(u,v)^{\mathrm{T}} \mathrm{d}u\mathrm{d}v \right\|$$

$$\lesssim \left[\left\| \mathcal{T}_{\eta_n}\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}\} - \boldsymbol{\Sigma}_{yy}^{(k)} \right\|_{\mathcal{S},1} \left\| \mathcal{T}_{\eta_n}\{\widehat{\boldsymbol{\Sigma}}_{yy}^{(k)}\} - \boldsymbol{\Sigma}_{yy}^{(k)} \right\|_{\mathcal{S},\infty} \right]^{1/2} \left[\|\boldsymbol{\Sigma}_{yy}^{(k)}\|_{\mathcal{S},1} \|\boldsymbol{\Sigma}_{yy}^{(k)}\|_{\mathcal{S},\infty} \right]^{1/2} \{1 + o_P(1)\}.$$

For the second term above, it follows that

$$\|\boldsymbol{\Sigma}_{yy}^{(k)}\|_{\mathcal{S},1} = \max_{1\leqslant j\leqslant p} \sum_{i=1}^{p} \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}} \leqslant \max_{1\leqslant j\leqslant p} \sum_{i=1}^{p} \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}^{\tau} \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}^{1-\tau} \lesssim c_1(p),$$

where $\tau \in (0,1)$ and $\|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}}$ is uniformly bounded under Condition 7. Following the similar procedure, we can show that $\|\boldsymbol{\Sigma}_{yy}^{(k)}\|_{\mathcal{S},\infty} \lesssim c_1(p) + c_2(p)$. For the truly sparse case with $\tau = 0$, it follows from the derivation in Section B.3 under Conditions 1, 2, 6 with fixed r and L that

$$\max_{1\leqslant j\leqslant p} \sum_{i=1}^{p} \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}} \leqslant \max_{1\leqslant j\leqslant p} \sum_{1\leqslant m,m'\leqslant r} \sum_{i=1}^{p} |A_{im}A_{jm'}| \|\Sigma_{xx,mm'}^{(k)}\|_{\mathcal{S}} + \max_{1\leqslant j\leqslant p} \sum_{m=1}^{r} \sum_{i=1}^{p} |A_{im}| \|\Sigma_{x\epsilon,mj}^{(k)}\|_{\mathcal{S}}$$

$$\lesssim r^2 L c_1(p) + r c_1(p) \asymp c_1(p).$$

Similarly, under Condition 4′, we can show that

$$\max_{1\leqslant i\leqslant p} \sum_{j=1}^{p} \|\Sigma_{yy,ij}^{(k)}\|_{\mathcal{S}} \lesssim r^2 L c_1(p) + r L c_2(p) \asymp c_1(p) + c_2(p).$$

The above results together with (S.10) from Step 1 implies (S.13).

Step 3. Denote $\hat{\epsilon} = \mathcal{D}\big(\mathcal{C}(\mathbf{K}), \mathcal{C}(\widehat{\mathbf{K}})\big)$. By the fact that $\hat{\epsilon}^2 = (2r)^{-1}\|\mathbf{K}\mathbf{K}^{\mathrm{T}} - \widehat{\mathbf{K}}\widehat{\mathbf{K}}^{\mathrm{T}}\|_{\mathrm{F}}^2$ and Corollary 4.1 in Vu and Lei (2013), we obtain that

$$2r\nu_r\hat{\epsilon}^2 \leqslant |\langle \widetilde{\mathbf{M}} - \widecheck{\mathbf{M}}, \widehat{\mathbf{K}}\widehat{\mathbf{K}}^{\mathrm{T}} - \mathbf{K}\mathbf{K}^{\mathrm{T}}\rangle|$$

$$\leqslant \left| \left\langle \widetilde{\mathbf{M}} - \widecheck{\mathbf{M}}, \frac{\widehat{\mathbf{K}}\widehat{\mathbf{K}}^{\mathrm{T}} - \mathbf{K}\mathbf{K}^{\mathrm{T}}}{\|\widehat{\mathbf{K}}\widehat{\mathbf{K}}^{\mathrm{T}} - \mathbf{K}\mathbf{K}^{\mathrm{T}}\|_{\mathrm{F}}} \right\rangle \right| \sqrt{2r}\hat{\epsilon},$$

which further implies that

$$\nu_r\hat{\epsilon} \leqslant (2r)^{-1/2} \left| \langle \widetilde{\mathbf{M}} - \widecheck{\mathbf{M}}, \widehat{\boldsymbol{\Delta}}\rangle \right|,$$

10

where $\widehat{\mathbf{\Delta}} = \frac{\hat{\mathbf{K}}\hat{\mathbf{K}}^{\mathrm{T}} - \mathbf{K}\mathbf{K}^{\mathrm{T}}}{\|\hat{\mathbf{K}}\hat{\mathbf{K}}^{\mathrm{T}} - \mathbf{K}\mathbf{K}^{\mathrm{T}}\|_{\mathrm{F}}}$ and $\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} = 1$.

Note that $\widehat{\mathbf{\Delta}}$ is a matrix with rank at most $2r$ and the associated singular values $\hat{\sigma}_1 \geqslant$ $\hat{\sigma}_2 \geqslant \cdots \geqslant \hat{\sigma}_{2r} \geqslant 0$. Then by the von-Neumann's trace inequality, we obtain that

$$
\nu_r \hat{\epsilon} \leqslant (2r)^{-1/2} \|\widetilde{\mathbf{M}} - \widecheck{\mathbf{M}}\| \sum_{i=1}^{2r} \hat{\sigma}_i
$$
$$
\leqslant (2r)^{-1/2} \|\widetilde{\mathbf{M}} - \widecheck{\mathbf{M}}\| \sqrt{2r \sum_{i=1}^{2r} \hat{\sigma}_i^2}
$$
$$
= \|\widetilde{\mathbf{M}} - \widecheck{\mathbf{M}}\| \|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} = \|\widetilde{\mathbf{M}} - \widecheck{\mathbf{M}}\|.
$$

By (S.13) from Step 2, we complete the proof of Theorem 3. ▫

B.5 Auxiliary lemmas

In this section, we present some auxiliary lemmas that are used in the proofs of theoretical results in Sections 3 and 4. The first lemma is a direct corollary of Hahn-Banach theorem.

Lemma B1. *For a function $f : \mathcal{D} \to \mathcal{B}$, where $\mathcal{D} \in \mathbb{R}^p$ is compact set and $\mathcal{B}$ is Banach space with norm $\|\cdot\|$, then $\|\int_{\mathcal{D}} f(x)\mathrm{d}x\| \leqslant \int_{\mathcal{D}} \|f(x)\|\mathrm{d}x$.*

The second lemma specifies the order of $\mathbf{\Sigma}_{xx}^{(k)}(u, v)$.

Lemma B2. *Suppose that Condition 1 holds, then for $k = 1, \ldots, k_0$, $\iint \|\mathbf{\Sigma}_{xx}^{(k)}(u, v)\|^2 \, \mathrm{d}u\mathrm{d}v = O(1)$ and $\|\iint \mathbf{\Sigma}_{xx}^{(k)}(u, v) \, \mathrm{d}u\mathrm{d}v\| = O(1)$.*

Proof of Lemma B2. By Cauchy-Schwartz inequality, Condition 1 and Fubini Theorem,

$$
\iint \|\mathbf{\Sigma}_{xx}^{(k)}(u, v)\|^2 \, \mathrm{d}u\mathrm{d}v \leqslant \iint \|E\{\mathbf{X}_{t+k}(u)\mathbf{X}_t(v)^{\mathrm{T}}\}\|_{\mathrm{F}}^2 \, \mathrm{d}u\mathrm{d}v = \sum_{i=1}^{r}\sum_{j=1}^{r} \iint \left[\mathbb{E}\{X_{(t+k)i}(u)X_{tj}(v)\} \right]^2 \mathrm{d}u\mathrm{d}v
$$
$$
\leqslant \left\{ \sum_{i=1}^{r} \mathbb{E}\big(\|X_{(t+k)i}\|^2\big) \right\} \left\{ \sum_{j=1}^{r} \mathbb{E}\big(\|X_{tj}\|^2\big) \right\} = O(1).
$$

On the other hand, by Lemma B1, we have that

$$
\left\| \iint \mathbf{\Sigma}_{xx}^{(k)}(u, v) \, \mathrm{d}u\mathrm{d}v \right\|^2 \leqslant \left(\iint \|\mathbf{\Sigma}_{xx}^{(k)}(u, v)\| \, \mathrm{d}u\mathrm{d}v \right)^2 \leqslant \mathrm{Vol}(\mathcal{U}^2) \iint \|\mathbf{\Sigma}_{xx}^{(k)}(u, v)\|^2 \, \mathrm{d}u\mathrm{d}v = O(1),
$$

where $\mathrm{Vol}(\cdot)$ represents the volume. □

11

The third lemma provides the non-asymptotic error bound on $\widehat{\Sigma}_{yy,ij}^{(k)}(u,v)$.

Lemma B3. *Suppose that Conditions 7 and 8 hold. Then there exists some positive constants c_1 and c_2 such that for $k = 1, \ldots, k_0$ and $t > 0$,*

$$\mathbb{P}\left(\max_{1 \leqslant i,j \leqslant p} \big\| \widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)} \big\|_{\mathcal{S}} > \mathcal{M}_y t \right) \leqslant c_1 p^2 \exp\big\{ - c_2 n \min(t, t^2) \big\},$$

If $n \geqslant \rho^2 \log p$ with $\rho > \sqrt{2} c_2^{-1/2}$, then

$$\max_{1 \leqslant i,j \leqslant p} \big\| \widehat{\Sigma}_{yy,ij}^{(k)} - \Sigma_{yy,ij}^{(k)} \big\|_{\mathcal{S}} \leqslant \rho \mathcal{M}_y \sqrt{\frac{\log p}{n}}$$

holds with probability greater than $1 - c_1 p^{2 - c_2 \rho^2}$.

Proof of Lemma B3. This lemma follows directly from Guo and Qiao (2021); Fang et al. (2021) and the choice of $t = \rho(\log p/n)^{1/2} \leqslant 1$, and hence the proof is omitted here. □

The following lemma is used in the proof of Theorem 3.

Lemma B4. *For $\mathbf{S}(u,v) = \{S_{ij}(u,v)\}_{p \times p}$ and $\mathbf{T}(u,v) = \{T_{ij}(u,v)\}_{p \times p}$ with each S_{ij} and $T_{ij} \in \mathbb{H} \otimes \mathbb{H}$, it holds that*

$$\left\| \iint \mathbf{S}(u,v) \mathbf{T}^{\mathrm{T}}(u,v) \, \mathrm{d}u \mathrm{d}v \right\| \leqslant \big(\|\mathbf{S}\|_{\mathcal{S},\infty} \|\mathbf{S}\|_{\mathcal{S},1} \big)^{1/2} \big(\|\mathbf{T}\|_{\mathcal{S},\infty} \|\mathbf{T}\|_{\mathcal{S},1} \big)^{1/2}.$$

Proof of Lemma B4. Notice that

$$\left\| \iint \mathbf{S}(u,v) \mathbf{T}(u,v)^{\mathrm{T}} \, \mathrm{d}u \mathrm{d}v \right\|_1 = \max_{1 \leqslant j \leqslant p} \sum_{i=1}^{p} \left| \iint \sum_{k=1}^{p} S_{ik}(u,v) T_{jk}(u,v) \, \mathrm{d}u \mathrm{d}v \right|$$

$$\leqslant \max_{1 \leqslant j \leqslant p} \sum_{i=1}^{p} \sum_{k=1}^{p} \|S_{ik}\|_{\mathcal{S}} \|T_{jk}\|_{\mathcal{S}} \qquad (\text{S.14})$$

$$\leqslant \left\{ \max_{1 \leqslant j \leqslant p} \sum_{k=1}^{p} \|T_{jk}\|_{\mathcal{S}} \right\} \left\{ \max_{1 \leqslant k \leqslant p} \sum_{i=1}^{p} \|S_{ik}\|_{\mathcal{S}} \right\} = \|\mathbf{S}\|_{\mathcal{S},1} \|\mathbf{T}\|_{\mathcal{S},\infty}.$$

By the similar arguments, we obtain that

$$\left\| \iint \mathbf{S}(u,v) \mathbf{T}(u,v)^{\mathrm{T}} \, \mathrm{d}u \mathrm{d}v \right\|_\infty \leqslant \|\mathbf{S}\|_{\mathcal{S},\infty} \|\mathbf{T}\|_{\mathcal{S},1}. \qquad (\text{S.15})$$

Lemma B4 follows immediately from (S.14), (S.15), the matrix norm inequality $\|\mathbf{E}\|^2 \leqslant \|\mathbf{E}\|_\infty \|\mathbf{E}\|_1$ for any $p \times p$ matrix $\mathbf{E}$ and the choice of $\mathbf{E} = \iint \mathbf{S}(u,v) \mathbf{T}(u,v)^{\mathrm{T}} \, \mathrm{d}u \mathrm{d}v$. □

12

C   Additional empirical results

Figure 7 provides spatial heatmaps of varimax-rotated loading matrix and sparse loading matrices with different sparsity levels for Japanese males.
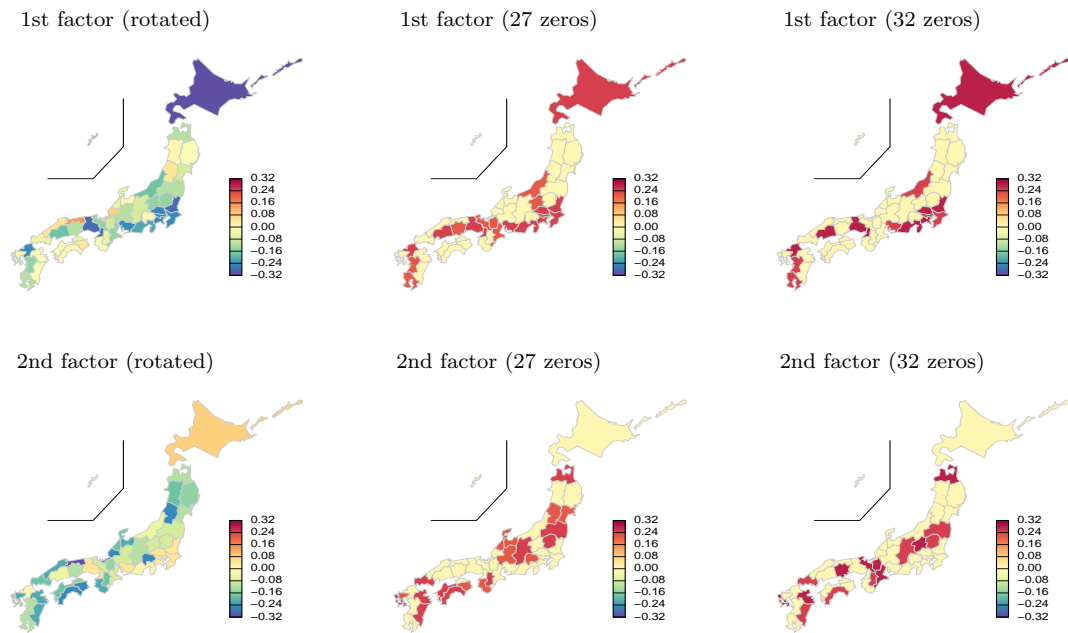

Figure 7: *Spatial heatmaps based on varimax-rotated loadings (left column) and sparse loadings with 27 zeros (middle column) and 32 zeros (right column) of 47 prefectures on two factors for Japanese males.*

References

Fang, Q., Guo, S. and Qiao, X. (2021). Finite sample theory for high-dimensional functional/scalar time series with applications, *Electronic Journal of Statistics, in press.*

Guo, S. and Qiao, X. (2021). On consistency and sparsity for high-dimensional functional time series with application to autoregressions, *arXiv:2003.11462v2.*

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

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions, *The Annals of Statistics* **41**: 2905–2947.

Wang, B.-Y. and Xi, B.-Y. (1997). Some inequalities for singular values of matrix products, *Linear Algebra and Its Applications* **264**: 109–115.