On the Modelling and Prediction of High-Dimensional Functional Time Series

Jinyuan Chang[1], Qin Fang[2], Xinghao Qiao[2], and Qiwei Yao[2]

[1] *Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China*

[2] *Department of Statistics, London School of Economics, London, WC2A 2AE, UK*

Abstract

We propose a two-step procedure to model and predict high-dimensional functional time series, where the number of function-valued time series p is large in relation to the length of time series n. Our first step performs an eigenanalysis of a positive definite matrix, which leads to a one-to-one linear transformation for the original high-dimensional functional time series, and the transformed curve series can be segmented into several groups such that any two subseries from any two different groups are uncorrelated both contemporaneously and serially. Consequently in our second step those groups are handled separately without the information loss on the overall linear dynamic structure. The second step is devoted to establishing a finite-dimensional dynamical structure for all the transformed functional time series within each group. Furthermore the finite-dimensional structure is represented by that of a vector time series. Modelling and forecasting for the original high-dimensional functional time series are realized via those for the vector time series in all the groups. We investigate the theoretical properties of our proposed methods, and illustrate the finite-sample performance through both extensive simulation and three real datasets.

Keywords: Dimension reduction; Eigenanalysis; Functional thresholding; Hilbert–Schmidt norm; Permutation; Segmentation transformation.

1 Introduction

Functional time series typically refers to continuous-time records that are naturally divided into consecutive time intervals, such as days, months or years. With recent advances in data collection technology, multivariate or even high-dimensional functional time series arise ubiquitously in many applications, including daily pollution concentration curves over different locations, annual temperature curves at different stations, annual age-specific mortality rates for different prefectures, and intraday energy consumption trajectories from different households. Those data can be represented as a p-dimensional functional time series $\mathbf{Y}_t(u) = \{Y_{t1}(u), \ldots, Y_{tp}(u)\}^\top$ defined on a compact set $u \in \mathcal{U}$, and we observe $\mathbf{Y}_t(\cdot)$ for $t = 1, \ldots, n$. In this paper we tackle the high-dimensional settings when the dimension p is comparable to, or even greater than, the sample size n, which poses new challenges in modelling and forecasting $\mathbf{Y}_t(\cdot)$.

By assuming $\mathbf{Y}_t(\cdot)$ is stationary, a conventional approach is first to extract features by performing dimension reduction for each component series $Y_{tj}(\cdot)$ separately via, e.g., functional principal component analysis (FPCA) or dynamic FPCA (Hörmann et al., 2015; Bathia et al., 2010), and then to model p vector time series by, e.g., regularized vector autoregressions (Guo and Qiao, 2023; Chang et al., 2022) or factor model (Gao et al., 2019). However, more effective dimension-reduction can be achieved by pulling together the information from different component series in the first place. This is in the same spirit of multivariate FPCA (Chiou et al., 2014; Happ and Greven, 2018) (for fixed p) and sparse FPCA (Hu and Yao, 2022), though those approaches make no use of the information on the serial dependence which is the most relevant for future prediction.

To achieve a more effective dimension-reduction by incorporating the information on the serial dependence across different component series, we propose in this paper a two-step approach. Our first step is a segmentation step in which we seek for a linear transformation $\mathbf{Y}_t(\cdot) = \mathbf{A}\mathbf{Z}_t(\cdot)$, where $\mathbf{A}$ is a $p \times p$ invertible constant matrix, such that the transformed series $\mathbf{Z}_t(\cdot) = \{\mathbf{Z}_t^{(1)}(\cdot)^\top, \ldots, \mathbf{Z}_t^{(q)}(\cdot)^\top\}^\top$ can be segmented into q groups $\mathbf{Z}_t^{(1)}(\cdot), \ldots, \mathbf{Z}_t^{(q)}(\cdot)$, and curves subseries $\mathbf{Z}_t^{(i)}(\cdot)$ and $\mathbf{Z}_t^{(j)}(\cdot)$ are uncorrelated at all time lags for any $i \neq j$, i.e.

$$\mathrm{Cov}\{\mathbf{Z}_t^{(i)}(u), \mathbf{Z}_{t+k}^{(j)}(v)\} = \mathbf{0}\,, \quad (u,v) \in \mathcal{U}^2 \text{ and } k = 0, \pm 1, \pm 2, \ldots\,.$$

Hence each $\mathbf{Z}_t^{(i)}$ can be modelled and forecasted separately as far as the linear dynamics is concerned. Under the stationarity assumption, the estimation of the transformation matrix $\mathbf{A}$ boils down to the eigenanalysis of a positive definite matrix defined by the double integral of quadratic forms in the autocovariance function of $\mathbf{Y}_t(\cdot)$. An additional permutation on the components of $\mathbf{Z}_t(\cdot)$ will be specified in order to identify the latent group structure.

Our second step is to identify a finite-dimensional dynamic structure for each transformed subseries $\mathbf{Z}_t^{(l)}(\cdot)$ separately, which is based on a latent decomposition

$$\mathbf{Z}_t^{(l)}(u) = \mathbf{X}_t^{(l)}(u) + \boldsymbol{\varepsilon}_t^{(l)}(u)\,, \quad u \in \mathcal{U}\,, \tag{1}$$

where $\mathbf{X}_t^{(l)}(\cdot)$ represents the dynamics of $\mathbf{Z}_t^{(l)}(\cdot)$, $\boldsymbol{\varepsilon}_t^{(l)}(\cdot)$ is white noise in the sense that $\mathbb{E}\{\boldsymbol{\varepsilon}_t^{(l)}(u)\} = \mathbf{0}$ and $\mathbb{E}\{\boldsymbol{\varepsilon}_t^{(l)}(u)\boldsymbol{\varepsilon}_s^{(l)}(v)^\top\} = \mathbf{0}$ for any $(u,v) \in \mathcal{U}^2$ and $t \neq s$, and $\{\mathbf{X}_t^{(l)}(\cdot)\}_{t=1}^n$ are uncorrelated with $\{\boldsymbol{\varepsilon}_t^{(l)}(\cdot)\}_{t=1}^n$. Furthermore we assume that the dynamic structure of $\mathbf{X}_t^{(l)}(\cdot)$ admits a vector time series presentation via a variational multivariate FPCA. For given $\{\mathbf{Z}_t^{(l)}(\cdot)\}_{t=1}^n$, the standard multivariate FPCA performs dimension reduction based on the eigenanlysis of the sample covariance function of $\mathbf{Z}_t^{(l)}(\cdot)$, which cannot be used to identify the finite-dimensional dynamic structure of $\mathbf{X}_t^{(l)}(\cdot)$ due to the contamination of $\boldsymbol{\varepsilon}_t^{(l)}(\cdot)$. Inspired by the fact that the lag-k ($k \neq 0$) autocovariance function of $\mathbf{Z}_t^{(l)}(\cdot)$ automatically filters out the white noise, our variational multivariate FPCA is based on the eigenanalysis of a positive-definite matrix defined in terms of its nonzero lagged autocovariance functions; leading to a low-dimensional vector time series which bears all the dynamic structure of

$\mathbf{X}_t^{(l)}(\cdot)$, and consequently, also that of $\mathbf{Z}_t^{(l)}(\cdot)$. This is possible as the number of components in each $\mathbf{Z}_t^{(l)}(\cdot)$ is small. Finally, owing to the one-to-one linear transformation in the segmentation step, the good predictive performance of $\mathbf{Z}_t(\cdot)$ can be easily carried back to $\mathbf{Y}_t(\cdot)$.

The new contribution of this paper is threefold. First, the segmentation transformation in the first step transforms the serial correlations across different series into the autocorrelations within each of the identified q subseries. This not only avoids the direct modelling of the p functional time series together, but also makes each of those subseries more serially correlated and, hence, more predictable. As the serial correlations across different series are valuable for future prediction, the segmentation provides an effective way to use the information. Note that the prediction directly based on a multivariate ARMA-type model with even a moderately large dimension is not recommendable, as the gain from using the correlations across different component series is often cancelled off by the errors in estimating too many parameters. On the other hand, the assumed segmentation structure may not exist. The proposed estimation method will then lead to an approximate segmentation which neglects small and, therefore, practically intractable correlations. Such an approximation often yields more accurate future predictions than those based on the models without segmentation. See the empirical studies in Sections 5 and 6. Though the segmentation transformation is based on the same idea of the PCA for vector time series of Chang et al. (2018), our proposal relies on the double integral to take full advantage of the functional nature of the data by gathering the autocovariance information over $(u, v) \in \mathcal{U}^2$. Secondly, aided by the enforced sparsity, we propose a novel functional thresholding procedure, which guarantees the consistency of our estimation under the high-dimensional regime. Thirdly, the nonzero lagged autocovariance-based dimension reduction approach in the second step makes the good use of the serial dependence information in our estimation, which is most relevant in the context of time series. Our proposal extends the univariate method of Bathia et al. (2010) by taking into account the cross-autocovariance to accommodate multivariate functional time series.

Our work lies in the intersection of two strands of literature: functional time series and high-dimensional time series. In the context of functional time series, many standard univariate or vector time series theory and methods have been adapted to the functional domain, see, among others, Bathia et al. (2010), Panaretos and Tavakoli (2013), Aue et al. (2015), Hörmann et al. (2015), Li et al. (2020), Jiao et al. (2021), and references within. In the context of high-dimensional time series, the available methods to reduce the number of parameters can be loosely divided into three categories: (i) regularization (Han et al., 2015; Basu and Michailidis, 2015; Guo et al., 2016; Ghosh et al., 2019; Wilms et al., 2021), (ii) dimension reduction via factor model (Bai and Ng, 2002; Forni et al., 2005; Lam and Yao, 2012; Stock and Watson, 2012; Chang et al., 2015; Fan et al., 2016), and (iii) independent component analysis (Tiao and Tsay, 1989; Back and Weigend, 1997; Matteson and Tsay, 2011; Chang et al., 2018; Han et al., 2022). Since the literature in each category is large, we can only include a selection of references here.

The rest of the paper is organized as follows. In Section 2, we develop the methods employed in Step 1, i.e. the segmentation transformation, the permutation and the functional thresholding. Section 3 specifies the dimension reduction method used in Step 2. We investigate the associated theoretical properties of the proposed methods in Section 4. The finite-sample performance of our methods are examined through extensive simulations in Section 5. Section 6 applies our proposal to three real datasets, revealing its superior predictive performance over most frequently used competitors. All technical proofs are relegated to the supplementary material.

Notations. Denote by $I(\cdot)$ the indicator function. For a positive integer m, write $[m] = \{1, \ldots, m\}$ and denote by $\mathbf{I}_m$ the identity matrix of size $m \times m$. For $x, y \in \mathbb{R}$, we use $x \vee y = \max(x, y)$. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \ll b_n$ or $b_n \gg a_n$ if $\limsup_{n\to\infty} a_n/b_n = 0$. For a $p \times q$ real matrix $\mathbf{B}$, denote by $\mathbf{B}^\top$ its transpose, and write

$\mathbf{B}^{\otimes 2} = \mathbf{B}\mathbf{B}^\top$ and $\|\mathbf{B}\|_2 = \lambda_{\max}^{1/2}(\mathbf{B}^\top\mathbf{B})$, where $\lambda_{\max}(\mathbf{M})$ denotes the largest eigenvalue of the matrix $\mathbf{M}$. Let $L^2(\mathcal{U})$ be a Hilbert space of squared integrable functions defined on $\mathcal{U}$ and equipped with the inner product $\langle \mathcal{F}, \mathcal{G}\rangle = \int_{\mathcal{U}} \mathcal{F}(u)\mathcal{G}(u)\,\mathrm{d}u$ for $\mathcal{F}, \mathcal{G} \in L_2(\mathcal{U})$ and the induced norm $\|\cdot\| = \langle \cdot, \cdot\rangle^{1/2}$. For any $\mathcal{B}$ in the tensor product $\mathbb{S} \equiv L^2(\mathcal{U}) \otimes L^2(\mathcal{U})$, we denote the Hilbert–Schmidt norm by $\|\mathcal{B}\|_{\mathcal{S}} = \{\int_{\mathcal{U}}\int_{\mathcal{U}} \mathcal{B}^2(u, v)\,\mathrm{d}u\mathrm{d}v\}^{1/2}$.

2 Segmentation transformation

In this section, we seek for a $p \times p$ constant matrix $\mathbf{A}$ such that

$$\mathbf{Y}_t(u) = \mathbf{A}\mathbf{Z}_t(u) = \mathbf{A}\{\mathbf{Z}_t^{(1)}(u)^\top, \ldots, \mathbf{Z}_t^{(q)}(u)^\top\}^\top, \qquad u \in \mathcal{U}, \tag{2}$$

where $q \leqslant p$ is an unknown positive integer, $\mathbf{Z}_t^{(l)}(\cdot)$ is a p_l-dimensional functional time series with each $p_l \geqslant 1$ and $\sum_{l=1}^q p_l = p$, and $\mathrm{Cov}\{\mathbf{Z}_t^{(l)}(u), \mathbf{Z}_s^{(l')}(v)\} = \mathbf{0}$ for all $t, s \in [n]$, $l \neq l'$ and $(u, v) \in \mathcal{U}^2$. We also assume that $\max_{i\in[p]} \int_{\mathcal{U}} \mathbb{E}\{Z_{ti}^2(u)\}\,\mathrm{d}u = O(1)$. Then it holds that

$$\mathbf{\Sigma}_{y,k}(u, v) \equiv \mathrm{Cov}\{\mathbf{Y}_t(u), \mathbf{Y}_{t+k}(v)\} = \mathbf{A}\mathrm{Cov}\{\mathbf{Z}_t(u), \mathbf{Z}_{t+k}(v)\}\mathbf{A}^\top \equiv \mathbf{A}\mathbf{\Sigma}_{z,k}(u, v)\mathbf{A}^\top, \tag{3}$$

where $\mathbf{\Sigma}_{z,k}(u, v)$ is block-diagonal with the blocks on the main diagonal of sizes $p_1 \times p_1, \ldots, p_q \times p_q$.

We consider orthogonal transformations only, i.e., $\mathbf{A}^\top\mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}_p$, as we can replace $(\mathbf{Y}_t, \mathbf{Z}_t)$ in (2) by $(\mathbf{V}_y^{-1/2}\mathbf{Y}_t, \mathbf{V}_z^{-1/2}\mathbf{Z}_t)$, where $\mathbf{V}_y = \int_{\mathcal{U}} \mathbf{\Sigma}_{y,0}(u, u)\,\mathrm{d}u$ and $\mathbf{V}_z = \int_{\mathcal{U}} \mathbf{\Sigma}_{z,0}(u, u)\,\mathrm{d}u$. Then $\mathbf{A}$ is replaced by $\mathbf{V}_y^{-1/2}\mathbf{A}\mathbf{V}_z^{1/2}$ which is an orthogonal matrix as

$$\mathbf{I}_p = \int_{\mathcal{U}} \mathrm{Var}\{\mathbf{V}_y^{-1/2}\mathbf{Y}_t(u)\}\,\mathrm{d}u = \int_{\mathcal{U}} \mathrm{Var}\{\mathbf{V}_z^{-1/2}\mathbf{Z}_t(u)\}\,\mathrm{d}u. \tag{4}$$

Note we can take $\mathbf{V}_z^{-1/2}\mathbf{Z}_t$ as $\mathbf{Z}_t$ since they share the same block structure. In practice, we can replace observations $\mathbf{Y}_t$ by $\widehat{\mathbf{V}}_y^{-1/2}\mathbf{Y}_t$, where $\widehat{\mathbf{V}}_y$ is a consistent estimator of $\mathbf{V}_y$.

Write $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_q)$, where $\mathbf{A}_l$ has p_l columns. Then

$$\mathbf{Z}_t^{(l)}(u) = \mathbf{A}_l^\top\mathbf{Y}_t(u), \qquad l \in [q]. \tag{5}$$

Even with the orthogonality constraint, $\mathbf{A}$ cannot be identified uniquely in (2), as within-block rotations will not distort the uncorrelated block structure. In fact only the linear spaces spanned by the columns of $\mathbf{A}_l$, denoted by $\mathcal{C}(\mathbf{A}_l)$, $l \in [q]$, are uniquely defined by (2).

2.1 Estimation procedure

We consider how to estimate $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_q)$ in (2). For a given integer $k_0 \geqslant 1$, let

$$\mathbf{W}_z = \sum_{k=0}^{k_0} \int_{\mathcal{U}} \int_{\mathcal{U}} \boldsymbol{\Sigma}_{z,k}(u,v)^{\otimes 2} \, \mathrm{d}u \mathrm{d}v \quad \text{and} \quad \mathbf{W}_y = \sum_{k=0}^{k_0} \int_{\mathcal{U}} \int_{\mathcal{U}} \boldsymbol{\Sigma}_{y,k}(u,v)^{\otimes 2} \, \mathrm{d}u \mathrm{d}v \,. \tag{6}$$

Then both $\mathbf{W}_y$ and $\mathbf{W}_x$ are non-negative definite. Due to $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_p$, by (3) and (6),

$$\mathbf{W}_z = \mathbf{A}^\top \mathbf{W}_y \mathbf{A} \,. \tag{7}$$

As all $\boldsymbol{\Sigma}_{z,k}(u,v)$ for $k \geqslant 0$ and $(u,v) \in \mathcal{U}^2$ are block-diagonal matrices of the same sizes, so is $\mathbf{W}_z$. Perform the eigenanalysis for each of q blocks on the main diagonal of $\mathbf{W}_z$ separately, leading to q orthogonal matrices of sizes $p_l \times p_l$ for $l \in [q]$. The columns of each of those orthogonal matrices are the p_l orthonormal eigenvectors from the corresponding eigenanalysis. Line up those q orthogonal matrices as the blocks for a main diagonal, to form a $p \times p$ block diagonal orthogonal matrix $\boldsymbol{\Gamma}_z$. Then the columns of $\boldsymbol{\Gamma}_z$ are the orthonormal eigenvectors of $\mathbf{W}_z$, i.e.

$$\mathbf{W}_z \boldsymbol{\Gamma}_z = \boldsymbol{\Gamma}_z \mathbf{D} \,, \tag{8}$$

where $\mathbf{D}$ is a diagonal matrix consisting of the p eigenvalues. Then by (7) and (8), $\mathbf{W}_y \mathbf{A} \boldsymbol{\Gamma}_z = \mathbf{A} \mathbf{W}_z \boldsymbol{\Gamma}_z = \mathbf{A} \boldsymbol{\Gamma}_z \mathbf{D}$. Thus the columns of $\boldsymbol{\Gamma}_y \equiv \mathbf{A} \boldsymbol{\Gamma}_z$ are the orthonormal eigenvectors of $\mathbf{W}_y$. Combining this with (2) yields that

$$\boldsymbol{\Gamma}_y^\top \mathbf{Y}_t(\cdot) = \boldsymbol{\Gamma}_z^\top \mathbf{A}^\top \mathbf{Y}_t(\cdot) = \boldsymbol{\Gamma}_z^\top \mathbf{Z}_t(\cdot) \,. \tag{9}$$

Since $\boldsymbol{\Gamma}_z$ is a block-diagonal orthogonal matrix with q blocks, $\boldsymbol{\Gamma}_z^\top \mathbf{Z}_t(\cdot)$ effectively applies orthogonal transformation within each of the q groups of $\mathbf{Z}_t(\cdot)$. Thus $\boldsymbol{\Gamma}_z^\top \mathbf{Z}_t(\cdot)$ is of the same segmentation structure of $\mathbf{Z}_t(\cdot)$, i.e. knowing $\boldsymbol{\Gamma}_z^\top \mathbf{Z}_t(\cdot)$ is as good as knowing the latent segmentation of $\mathbf{Z}_t(\cdot)$. Hence $\mathbf{\Gamma}_y$ can be taken as the required transformation matrix $\mathbf{A}$; see (9). We summarize the above finding in Proposition 1 below.

Proposition 1. (i) *The orthogonal matrix* $\mathbf{\Gamma}_z$ *in* (8) *can be of the same block-diagonal form as* $\mathbf{W}_z$.

(ii) *An orthogonal matrix* $\mathbf{\Gamma}_z$ *satisfies* (8) *if and only if its columns are a permutation of the columns of a block-diagonal orthogonal matrix described in* (i)*, provided that any two different main diagonal blacks in* $\mathbf{W}_z$ *do not share the same eigenvalues.*

Let $\widehat{\mathbf{\Sigma}}_{y,k}(u,v)$ be some consistent estimator of $\mathbf{\Sigma}_{y,k}(u,v)$ for $k \in \{0\} \cup [k_0]$, to be specified in Section 2.2 below. We define an estimator of $\mathbf{W}_y$ as

$$\widehat{\mathbf{W}}_y = \sum_{k=0}^{k_0} \int_{\mathcal{U}} \int_{\mathcal{U}} \widehat{\mathbf{\Sigma}}_{y,k}(u,v)^{\otimes 2} \,\mathrm{d}u\mathrm{d}v \,, \tag{10}$$

and calculate its orthonormal eigenvectors $\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_p$. Let $\widehat{\mathbf{\Gamma}}_y = (\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_p)$. By (9) and Proposition 1, $\widehat{\mathbf{A}}$ should be a (latent) column-permutation of $\widehat{\mathbf{\Gamma}}_y$. More specifically, put

$$\widehat{\mathbf{Z}}_t(\cdot) \equiv \{\widehat{Z}_{t1}(\cdot), \ldots, \widehat{Z}_{tp}(\cdot)\}^\top = \widehat{\mathbf{\Gamma}}_y^\top \mathbf{Y}_t(\cdot)\,. \tag{11}$$

We propose below a data-driven procedure to divide the p components of $\widehat{\mathbf{Z}}_t(\cdot)$ into $\hat{q}$ uncorrelated groups.

Write $\mathbf{\Sigma}_{z,k}(\cdot,\cdot) = \{\Sigma_{z,k,ij}(\cdot,\cdot)\}_{i,j\in[p]}$. Recall $\mathbf{Z}_t(\cdot) = \{Z_{t1}(\cdot), \ldots, Z_{tp}(\cdot)\}^\top$. For two curve series $Z_{ti}(\cdot)$ and $Z_{tj}(\cdot)$ within the same group, one would expect that their lag-k cross-autocovariance function $\Sigma_{z,k,ij}(u,v)$ to be significantly different from zero for some integer k and $(u,v) \in \mathcal{U}^2$, thus leading to at least one large $\|\Sigma_{z,k,ij}\|_{\mathcal{S}}$ for some integer k. Based on $\widehat{\mathbf{Z}}_t(\cdot)$ defined as (11), we let $\widehat{\mathbf{\Sigma}}_{z,k}(u,v) \equiv \{\widehat{\Sigma}_{z,k,ij}(u,v)\}_{i,j\in[p]} = \widehat{\mathbf{\Gamma}}_y^\top \widehat{\mathbf{\Sigma}}_{y,k}(u,v)\widehat{\mathbf{\Gamma}}_y$ for any $(u,v) \in \mathcal{U}^2$. Given a fixed integer $m \geqslant 0$, we define the maximum cross-autocovariance over the lags between prespecified $-m$ and m as

$$\widehat{T}_{ij} = \max_{|k|\leqslant m} \|\widehat{\Sigma}_{z,k,ij}\|_{\mathcal{S}} \tag{12}$$

for any pair $(i,j) \in [p]^2$ such that $i < j$, and regard $\widehat{Z}_{ti}(\cdot)$ and $\widehat{Z}_{tj}(\cdot)$ from the same group if $\widehat{T}_{ij}$ takes some large value. To be specific, we rearrange $\aleph = p(p-1)/2$ values of $\widehat{T}_{ij}$ $(1 \leqslant i < j \leqslant p)$ in the descending order $\widehat{T}_{(1)} \geqslant \cdots \geqslant \widehat{T}_{(\aleph)}$ and compute

$$\hat{\varrho} = \arg\max_{j \in [\aleph]} \frac{\widehat{T}_{(j)} + \delta_n}{\widehat{T}_{(j+1)} + \delta_n} \tag{13}$$

for some $\delta_n > 0$. Corresponding to $\widehat{T}_{(1)}, \ldots, \widehat{T}_{(\hat{\varrho})}$, we identify $\hat{\varrho}$ pairs of cross-correlated curves. To divide the p components of $\widehat{\mathbf{Z}}_t(\cdot)$ into several uncorrelated groups, we can first start with p groups with each $\widehat{Z}_{tj}(\cdot)$ in one group and then repeatedly merge two groups if two cross-correlated curves are split over the two groups. The iteration is terminated until all the cross-correlated pairs are within one group. Hence we obtain the estimated group structure of $\widehat{\mathbf{Z}}_t(\cdot)$ with the number of the final groups $\hat{q}$ being the estimated value for q. Denote by $\widehat{\mathbf{Z}}_t^{(l)}(\cdot)$ the estimated l-th group for $l \in [\hat{q}]$. The transformation matrix $\widehat{\mathbf{A}} = (\widehat{\mathbf{A}}_1, \ldots, \widehat{\mathbf{A}}_{\hat{q}})$ can then be found by reorganizing the order of $(\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_p)$ such that

$$\widehat{\mathbf{Z}}_t^{(l)}(\cdot) = \widehat{\mathbf{A}}_l^\top \mathbf{Y}_t(\cdot), \qquad l \in [\hat{q}]. \tag{14}$$

Remark 1. (i) We include a small term $\delta_n > 0$ in (13) to stabilise the estimates for '0/0'. Given a suitable order of δ_n, we can establish the group recovery consistency. See Theorem 1 in Section 4. A common practice is to set $\delta_n = 0$ and replace $\aleph$ by $c_\varrho \aleph$ in (13) for some constant $c_\varrho \in (0,1)$, see Lam and Yao (2012); Ahn and Horenstein (2013); Chang et al. (2018).

(ii) All integrated terms in $\mathbf{W}_y$ are non-negative definite. Hence there is no information cancellation over different lags. Therefore the estimation is insensitive to the choice of k_0. In practice a small k_0 (such as $k_0 \leqslant 5$) is often sufficient, while further enlarging k_0 tends to add more noise to $\mathbf{W}_y$.

2.2 Selection of $\widehat{\mathbf{\Sigma}}_{y,k}(u,v)$

The estimate $\widehat{\mathbf{\Sigma}}_{y,k}(u,v)$ plays a key role in Section 2.1. A natural candidate for $\widehat{\mathbf{\Sigma}}_{y,k}(u,v)$ is the sample version of $\mathbf{\Sigma}_{y,k}(u,v)$ defined as

$$\widehat{\mathbf{\Sigma}}^{\mathrm{S}}_{y,k}(u,v) = \frac{1}{n-k}\sum_{t=1}^{n-k}\{\mathbf{Y}_t(u)-\bar{\mathbf{Y}}(u)\}\{\mathbf{Y}_{t+k}(v)-\bar{\mathbf{Y}}(v)\}^{\top}\,, \qquad k\in\{0\}\cup[k_0]\,. \tag{15}$$

When $p^2/n\to 0$, $\widehat{\mathbf{\Sigma}}^{\mathrm{S}}_{y,k}(u,v)$ is a valid estimator for $\mathbf{\Sigma}_{y,k}(u,v)$. However when p grows faster than $n^{1/2}$, it does not always hold that $\|\widehat{\mathbf{\Sigma}}^{\mathrm{S}}_{y,k}(u,v)-\mathbf{\Sigma}_{y,k}(u,v)\|_2\to 0$ in probability. We then impose some sparsity condition on the orthogonal transformation matrix $\mathbf{A}$, which will pass onto the autocovariance functions $\mathbf{\Sigma}_{y,k}(\cdot,\cdot)$, as $\mathbf{\Sigma}_{y,k}(\cdot,\cdot)=\mathbf{A}\mathbf{\Sigma}_{z,k}(\cdot,\cdot)\mathbf{A}^{\top}$.

Inspired by the spirit of threshold estimator for large covariance matrix (Bickel and Levina, 2008), we apply the functional thresholding rule, that combines the functional generalizations of hard thresholding and shrinkage with the aid of the Hilbert–Schmidt norm of functions, on the entries of the sample autocovariance function $\widehat{\mathbf{\Sigma}}^{\mathrm{S}}_{y,k}(u,v)=\{\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}(u,v)\}_{i,j\in[p]}$ in (15). This leads to the estimator

$$\mathcal{T}_{\omega_k}(\widehat{\mathbf{\Sigma}}^{\mathrm{S}}_{y,k})(u,v) = \big[\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}(u,v)I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}}\geqslant\omega_k\}\big]_{i,j\in[p]}\,, \qquad (u,v)\in\mathcal{U}^2\,, \tag{16}$$

where $\omega_k\geqslant 0$ is the thresholding parameter at lag k. Taking $\widehat{\mathbf{\Sigma}}_{y,k}(\cdot,\cdot)$ in (10) as $\mathcal{T}_{\omega_k}(\widehat{\mathbf{\Sigma}}^{\mathrm{S}}_{y,k})(\cdot,\cdot)$ yields

$$\widehat{\mathbf{W}}_y = \sum_{k=0}^{k_0}\int_{\mathcal{U}}\int_{\mathcal{U}}\mathcal{T}_{\omega_k}(\widehat{\mathbf{\Sigma}}^{\mathrm{S}}_{y,k})(u,v)^{\otimes 2}\,\mathrm{d}u\mathrm{d}v\,. \tag{17}$$

Remark 2. The thresholding parameter ω_k for each $k\in\{0\}\cup[k_0]$ can be selected using an L-fold cross-validation approach. Specifically, we sequentially divide the set $[n]$ into L validation sets $V_1,\ldots,V_L$ of approximately equal size. For each $l\in[L]$, let $\widehat{\mathbf{\Sigma}}^{\mathrm{S},(l)}_{y,k}(u,v)=\{\widehat{\Sigma}^{\mathrm{S},(l)}_{y,k,ij}(u,v)\}_{i,j\in[p]}$ and $\widehat{\mathbf{\Sigma}}^{\mathrm{S},(-l)}_{y,k}(u,v)=\{\widehat{\Sigma}^{\mathrm{S},(-l)}_{y,k,ij}(u,v)\}_{i,j\in[p]}$ be the sample lag-k autocovaraince functions based on the l-th validation set $\{\mathbf{Y}_t(\cdot):t\in V_l\}$ and the remaining $L-1$ sets $\{\mathbf{Y}_t(\cdot):t\in[n]\backslash V_l\}$, respectively. We select the optimal $\hat{\omega}_k$ by minimizing

$$\mathrm{Error}(\omega_k)=\frac{1}{L}\sum_{l=1}^{L}\sum_{i,j=1}^{p}\|\mathcal{T}_{\omega_k}(\widehat{\Sigma}^{\mathrm{S},(l)}_{y,k,ij})-\widehat{\Sigma}^{\mathrm{S},(-l)}_{y,k,ij}\|^2_{\mathcal{S}}\,,$$

where $\mathcal{T}_{\omega_k}(\widehat{\Sigma}^{\mathrm{s},(l)}_{y,k,ij})(u,v) = \widehat{\Sigma}^{\mathrm{s},(l)}_{y,k,ij}(u,v) I\{\|\widehat{\Sigma}^{\mathrm{s},(l)}_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k\}$. Given the time break from the leave-out validation set, the autocovariance estimation based on the remaining $L-1$ groups is affected by k_0 misutilized lagged terms. However, this effect is negligible for large n.

3 Identifying finite-dimensional structure of transformed curve subseries

Our Step 2 is to represent (linear) dynamic structure of each $\mathbf{Z}_t^{(l)}(\cdot)$, obtained in Step 1, in terms of a vector time series via representation (1). The key idea is to identify the finite decomposition for $\mathbf{X}_t^{(l)}(\cdot)$. For $(u,v) \in \mathcal{U}^2$ and $k \geqslant 0$, let $\boldsymbol{\mu}^{(l)}(u) = \mathbb{E}\{\mathbf{X}_t^{(l)}(u)\}$ and

$$\mathbf{M}_k^{(l)}(u,v) = \mathbb{E}[\{\mathbf{X}_t^{(l)}(u) - \boldsymbol{\mu}^{(l)}(u)\}\{\mathbf{X}_{t+k}^{(l)}(v) - \boldsymbol{\mu}^{(l)}(v)\}^\top]\,.$$

Then the multivariate Karhunen-Loève decomposition for $\mathbf{X}_t^{(l)}(\cdot)$ serving as the foundation of multivariate FPCA (Chiou et al., 2014; Happ and Greven, 2018) admits the form

$$\mathbf{M}_0^{(l)}(u,v) = \sum_{j=1}^{\infty} \lambda_j^{(l)} \boldsymbol{\varphi}_j^{(l)}(u)\boldsymbol{\varphi}_j^{(l)}(v)^\top\,, \qquad \mathbf{X}_t^{(l)}(u) - \boldsymbol{\mu}^{(l)}(u) = \sum_{j=1}^{\infty} \xi_{tj}^{(l)} \boldsymbol{\varphi}_j^{(l)}(u)\,, \tag{18}$$

where $\lambda_1^{(l)} \geqslant \lambda_2^{(l)} \geqslant \cdots \geqslant 0$ are the ordered eigenvalues of the covariance function $\mathbf{M}_0^{(l)}(\cdot,\cdot)$, $\boldsymbol{\varphi}_1^{(l)}(\cdot), \boldsymbol{\varphi}_2^{(l)}(\cdot), \ldots$ are the corresponding orthonormal eigenfunctions satisfying $\int_{\mathcal{U}} \boldsymbol{\varphi}_j^{(l)}(u)^\top \boldsymbol{\varphi}_k^{(l)}(u)\,\mathrm{d}u = I(j=k)$, and $\xi_{tj}^{(l)} = \int_{\mathcal{U}} \boldsymbol{\varphi}_j^{(l)}(u)^\top\{\mathbf{X}_t^{(l)}(u) - \boldsymbol{\mu}^{(l)}(u)\}\,\mathrm{d}u$ with $\mathbb{E}\{\xi_{tj}^{(l)}\} = 0$ and $\mathrm{Cov}\{\xi_{tj}^{(l)}, \xi_{tk}^{(l)}\} = \lambda_j^{(l)} I(j=k)$.

When $\mathbf{X}_t^{(l)}(\cdot)$ is r_l-dimensional in the sense that $\lambda_{r_l}^{(l)} > 0$ and $\lambda_{r_l+1}^{(l)} = 0$, the dynamics of $\mathbf{X}_t^{(l)}(\cdot)$ is entirely determined by that of r_l-vector time series $\boldsymbol{\xi}_t^{(l)} = \{\xi_{t1}^{(l)}, \ldots, \xi_{tr_l}^{(l)}\}^\top$. Unfortunately, under the latent decomposition (1), i.e.

$$\mathbf{Z}_t^{(l)}(u) = \mathbf{X}_t^{(l)}(u) + \boldsymbol{\varepsilon}_t^{(l)}(u) = \boldsymbol{\mu}^{(l)}(u) + \sum_{j=1}^{r_l} \xi_{tj}^{(l)} \boldsymbol{\varphi}_j^{(l)}(u) + \boldsymbol{\varepsilon}_t^{(l)}(u)\,, \qquad u \in \mathcal{U}\,, \tag{19}$$

the standard multivariate FPCA based on (18) is inappropriate as $\mathbf{X}_t^{(l)}(\cdot)$ is unobservable and we cannot provide a consistent estimator for $\mathbf{M}_0^{(l)}(u,v)$ based on $\mathbf{Z}_t^{(l)}(\cdot)$ due to the fact

$\text{Cov}\{\mathbf{Z}_t^{(l)}(u), \mathbf{Z}_t^{(l)}(v)\} = \mathbf{M}_0^{(l)}(u,v) + \text{Cov}\{\boldsymbol{\varepsilon}_t^{(l)}(u), \boldsymbol{\varepsilon}_t^{(l)}(v)\}$. A possible solution to deal with the difficulty is to delineate the dynamic structure of $\mathbf{X}_t^{(l)}(\cdot)$ by subtracting $\text{Cov}\{\boldsymbol{\varepsilon}_t^{(l)}(u), \boldsymbol{\varepsilon}_t^{(l)}(v)\}$ from $\text{Cov}\{\mathbf{Z}_t^{(l)}(u), \mathbf{Z}_t^{(l)}(v)\}$. To this end, a common practice is to assume that the covariance function of $\boldsymbol{\varepsilon}_t^{(l)}(\cdot)$ is diagonal (Yao et al., 2005) or banded (Descary and Panaretos, 2019), which is too restrictive in practice.

Now we introduce the variational multivariate FPCA based on a variational multivariate Karhunen-Loève decomposition for $\mathbf{X}_t^{(l)}(\cdot)$. Motivated from the fact $\text{Cov}\{\mathbf{Z}_t^{(l)}(u), \mathbf{Z}_{t+k}^{(l)}(v)\} = \mathbf{M}_k^{(l)}(u,v)$ for any $k \geqslant 1$, for a prespecified small integer $k_0 \geqslant 1$, we define

$$\mathbf{K}^{(l)}(u,v) = \sum_{k=1}^{k_0} \int_{\mathcal{U}} \mathbf{M}_k^{(l)}(u,w)\mathbf{M}_k^{(l)}(v,w)^\top \, \mathrm{d}w \,. \tag{20}$$

Note that $\mathbf{K}^{(l)}$ can be viewed as the kernel of the induced linear operator. For notational economy, we will use $\mathbf{K}^{(l)}$ to denote both the kernel and the operator. Similar to $\mathbf{M}_0^{(l)}$, $\mathbf{K}^{(l)}$ is also a non-negative definite operator which admits a spectral decomposition

$$\mathbf{K}^{(l)}(u,v) = \sum_{j=1}^{\infty} \theta_j^{(l)} \boldsymbol{\psi}_j^{(l)}(u)\boldsymbol{\psi}_j^{(l)}(v)^\top \,, \tag{21}$$

where $\theta_1^{(l)} \geqslant \theta_2^{(l)} \geqslant \cdots \geqslant 0$ are the eigenvalues of $\mathbf{K}^{(l)}$, and $\boldsymbol{\psi}_1^{(l)}(\cdot), \boldsymbol{\psi}_2^{(l)}(\cdot), \ldots$ are the corresponding orthonormal eigenfunctions. Let $\boldsymbol{\xi}_t^{(l)} = \{\xi_{t1}^{(l)}, \ldots, \xi_{tr_l}^{(l)}\}^\top$ and $\boldsymbol{\Omega}_k^{(l)} = \mathbb{E}[\boldsymbol{\xi}_t^{(l)}\{\boldsymbol{\xi}_{t+k}^{(l)}\}^\top]$. It then follows from Proposition 2 below that, under the expansion (19), the operator $\mathbf{K}^{(l)}$ has exactly r_l nonzero eigenvalues, and the dynamic space spanned by $\{\boldsymbol{\psi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\psi}_{r_l}^{(l)}(\cdot)\}$ remains the same as that spanned by $\{\boldsymbol{\varphi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\varphi}_{r_l}^{(l)}(\cdot)\}$.

Proposition 2. *Let $\boldsymbol{\Omega}_k^{(l)}$ be a full-ranked matrix for some $k \in [k_0]$. Then it holds that* (i) $\theta_{r_l}^{(l)} > 0$ *and* $\theta_{r_l+1}^{(l)} = 0$; (ii) $\text{span}\{\boldsymbol{\varphi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\varphi}_{r_l}^{(l)}(\cdot)\} = \text{span}\{\boldsymbol{\psi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\psi}_{r_l}^{(l)}(\cdot)\}$.

Therefore, $\mathbf{X}_t^{(l)}(\cdot)$ can be expanded using r_l basis functions $\boldsymbol{\psi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\psi}_{r_l}^{(l)}(\cdot)$, i.e.,

$$\mathbf{X}_t^{(l)}(u) - \boldsymbol{\mu}^{(l)}(u) = \sum_{j=1}^{r_l} \zeta_{tj}^{(l)} \boldsymbol{\psi}_j^{(l)}(u) \,, \qquad u \in \mathcal{U} \,, \tag{22}$$

where the basis coefficients $\zeta_{tj}^{(l)} = \int_{\mathcal{U}} \boldsymbol{\psi}_j^{(l)}(u)^\top \{\mathbf{X}_t^{(l)}(u) - \boldsymbol{\mu}^{(l)}(u)\} \, \mathrm{d}u$. Note that we take the sum in defining $\mathbf{K}^{(l)}(u,v)$ in (20) to accumulate the information from different lags, and there is no information cancellation as each term in the sum is non-negative definite. An additional advantage for using the lagged autocovariance-based decomposition is that the identified directions $\boldsymbol{\psi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\psi}_{r_l}^{(l)}(\cdot)$ catch the most significant serial dependence, which is advantageous for future prediction.

Noting that $\mathbf{Z}_t^{(l)}(\cdot)$ is not directly observable, we can only estimate $\mathbf{M}_k^{(l)}$ and $\mathbf{K}^{(l)}$ based on $\hat{p}_l$-vector of estimated transformed curve subseries $\widehat{\mathbf{Z}}_t^{(l)}(\cdot) = \{\widehat{Z}_{t1}^{(l)}(\cdot), \ldots, \widehat{Z}_{t\hat{p}_l}^{(l)}(\cdot)\}^\top$ obtained in Step 1. With the aid of (14), for $k \in \{0\} \cup [k_0]$, put

$$\widehat{\mathbf{M}}_k^{(l)}(u, v) = \widehat{\mathbf{A}}_l^\top \widehat{\boldsymbol{\Sigma}}_{y,k}(u, v) \widehat{\mathbf{A}}_l \,. \tag{23}$$

It is easy to see from (1) that $\widehat{\mathbf{M}}_k^{(l)}(u, v)$ is a reasonable estimator for $\mathbf{M}_k^{(l)}(u, v)$ when $k \geqslant 1$, as it filters out white noise $\boldsymbol{\varepsilon}_t^{(l)}(\cdot)$ automatically. However this is no longer the case when $k = 0$. This is the major reason why we employ the variational multivariate FPCA to estimate the finite-dimensional structure based on (21) and (22) instead of the multivariate FPCA based on (18). It is noteworthy that (23) requires the consistent estimators for $\boldsymbol{\Sigma}_{y,k}(u, v)$. Its implementation under the high-dimensional setting can thus be done by setting $\widehat{\boldsymbol{\Sigma}}_{y,k}(u, v) = \mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}_{y,k}^{\mathrm{s}})(u, v)$ defined in (16).

To estimate $\boldsymbol{\psi}_j^{(l)}(\cdot)$ and $\zeta_{tj}^{(l)}$ in (22), we perform the eigenanalysis for the estimated operator

$$\widehat{\mathbf{K}}^{(l)}(u, v) = \sum_{k=1}^{k_0} \int_{\mathcal{U}} \widehat{\mathbf{M}}_k^{(l)}(u, w) \widehat{\mathbf{M}}_k^{(l)}(v, w)^\top \, \mathrm{d}w \tag{24}$$

leading to the eigenvalues $\hat{\theta}_1^{(l)} \geqslant \hat{\theta}_2^{(l)} \geqslant \cdots \geqslant 0$, and the corresponding orthonormal eigenfunctions $\widehat{\boldsymbol{\psi}}_1^{(l)}(\cdot), \widehat{\boldsymbol{\psi}}_2^{(l)}(\cdot), \ldots$. To estimate r_l (i.e. the number of nonzero eigenvalues), we take the commonly-adopted ratio-based estimator for r_l as:

$$\hat{r}_l = \arg\max_{j \in [n-k_0]} \frac{\hat{\theta}_j^{(l)} + \tilde{\delta}_n}{\hat{\theta}_{j+1}^{(l)} + \tilde{\delta}_n} \tag{25}$$

for some $\tilde{\delta}_n > 0$. Under some regularity conditions, such defined $\hat{r}_l$ is a consistent estimator for r_l; see Theorem 3 in Section 4. In practice, since $\tilde{\delta}_n$ is usually unknown, we instead adopt

$\hat{r}_l = \arg\max_{j\in[c_r(n-k_0)]} \hat{\theta}_j^{(l)}/\hat{\theta}_{j+1}^{(l)}$, where $c_r \in (0,1)$ is a prescribed constant aiming to avoid fluctuations due to the ratios of extreme small values.

Let $\hat{\zeta}_{tj}^{(l)} = \int_{\mathcal{U}} \widehat{\boldsymbol{\psi}}_j^{(l)}(u)^\top\{\widehat{\mathbf{Z}}_t^{(l)}(u) - \bar{\mathbf{Z}}^{(l)}(u)\}\,\mathrm{d}u$ for $t \in [n]$, $j \in [\hat{r}_l]$ and $l \in [\hat{q}]$. We fit a model for the $\hat{r}_l$-dimensional vector time series $\widehat{\boldsymbol{\zeta}}_t^{(l)} = \{\hat{\zeta}_{t1}^{(l)}, \ldots, \hat{\zeta}_{t\hat{r}_l}^{(l)}\}^\top$ with $t \in [n]$ to obtain its h-step ahead prediction as $\mathring{\boldsymbol{\zeta}}_{n+h}^{(l)}$ and then recover the h-step ahead functional prediction as

$$\mathring{\mathbf{Z}}_{n+h}^{(l)}(u) = \bar{\mathbf{Z}}^{(l)}(u) + \sum_{j=1}^{\hat{r}_l} \mathring{\zeta}_{(n+h)j}^{(l)} \widehat{\boldsymbol{\psi}}_j^{(l)}(u)\,, \quad h \geqslant 1\,.$$

We finally obtain the h-step ahead prediction for original functional time series via $\widehat{\mathbf{Y}}_{n+h}(\cdot) = \widehat{\mathbf{A}}\mathring{\mathbf{Z}}_{n+h}(\cdot)$, where $\widehat{\mathbf{A}} = (\widehat{\mathbf{A}}_1, \ldots, \widehat{\mathbf{A}}_{\hat{q}})$ and $\mathring{\mathbf{Z}}_{n+h}(\cdot) = \{\mathring{\mathbf{Z}}_{n+h}^{(1)}(\cdot)^\top, \ldots, \mathring{\mathbf{Z}}_{n+h}^{(\hat{q})}(\cdot)^\top\}^\top$.

4 Theoretical properties

In this section, we present theoretical analysis of our two-step estimation procedure. To ease presentation, we focus on the high-dimensional scenario and develop the associated theoretical results based on the functional threshold estimator $\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}_{y,k}^{\mathrm{S}})(u,v)$ in (16). To simplify notation, we use $\mathcal{B}$ to denote the linear operator induced from the kernel function $\mathcal{B} \in \mathbb{S}$, i.e. for any $\mathcal{F} \in L_2(\mathcal{U})$, $\mathcal{B}(\mathcal{F})(\cdot) = \int_{\mathcal{U}} \mathcal{B}(\cdot\,,v)\mathcal{F}(v)\,\mathrm{d}v \in L_2(\mathcal{U})$. We denote the p-fold Cartesian product $\mathbb{H} = L^2(\mathcal{U}) \times \cdots \times L^2(\mathcal{U})$. For any $\boldsymbol{\mathcal{F}}, \boldsymbol{\mathcal{G}} \in \mathbb{H}$, we denote the inner product by $\langle \boldsymbol{\mathcal{F}}, \boldsymbol{\mathcal{G}} \rangle = \int_{\mathcal{U}} \boldsymbol{\mathcal{F}}(u)^\top \boldsymbol{\mathcal{G}}(u)\,\mathrm{d}u$ with the induced norm $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$, and use $\boldsymbol{\mathcal{B}}$ to denote the linear operator induced from the kernel matrix function $\boldsymbol{\mathcal{B}} = (\mathcal{B}_{ij})_{m_1\times m_2}$ with each $\mathcal{B}_{ij} \in \mathbb{S}$, i.e. for any $\boldsymbol{\mathcal{F}} \in \mathbb{H}$, $\boldsymbol{\mathcal{B}}(\boldsymbol{\mathcal{F}})(\cdot) = \int_{\mathcal{U}} \boldsymbol{\mathcal{B}}(\cdot\,,v)\boldsymbol{\mathcal{F}}(v)\,\mathrm{d}v \in \mathbb{H}$. We write $\|\boldsymbol{\mathcal{B}}\|_{\mathcal{S},\infty} = \max_{i\in[m_1]} \sum_{j=1}^{m_2} \|\mathcal{B}_{ij}\|_{\mathcal{S}}$. Before imposing the regularity conditions, we first define the functional version of sub-Gaussianity that facilities the development of non-asymptotic results for Hilbert space-valued random elements.

Definition 1. *Let $\Upsilon_t(\cdot)$ be a mean zero random variable in $L^2(\mathcal{U})$ and $\Sigma_0 : L^2(\mathcal{U}) \to L^2(\mathcal{U})$ be a covariance operator. We call $\Upsilon_t(\cdot)$ a sub-Gaussian process if there exists a constant*

$c > 0$ *such that* $\mathbb{E}[\exp\{\langle \mathcal{F}, \Upsilon_t - \mathbb{E}(\Upsilon_t)\rangle\}] \leqslant \exp\{2^{-1}c^2\langle \mathcal{F}, \Sigma_0(\mathcal{F})\rangle\}$ *for all* $\mathcal{F} \in L^2(\mathcal{U})$.

Condition 1. (i) $\{\mathbf{Y}_t(\cdot)\}$ is a sequence of multivariate functional linear processes with sub-Gaussian errors, i.e., $\mathbf{Y}_t(\cdot) = \sum_{l=0}^{\infty} \boldsymbol{\mathcal{D}}_l(\boldsymbol{\epsilon}_{t-l})$, where $\boldsymbol{\mathcal{D}}_l = (\mathcal{D}_{l,ij})_{p\times p}$ with each $\mathcal{D}_{l,ij} \in \mathbb{S}$ and $\boldsymbol{\epsilon}_t(\cdot) = \{\epsilon_{t1}(\cdot), \ldots, \epsilon_{tp}(\cdot)\}^\top$ with independent components of mean-zero sub-Gaussian processes satisfying Definition 1; (ii) The coefficient functions satisfy $\sum_{l=0}^{\infty} \|\boldsymbol{\mathcal{D}}_l\|_{\mathcal{S},\infty} = O(1)$; (iii) $\max_{j\in[p]} \int_{\mathcal{U}} \mathrm{Cov}\{\epsilon_{tj}(u), \epsilon_{tj}(u)\}\,\mathrm{d}u = O(1)$.

Condition 2. For $\{\mathbf{Y}_t(\cdot)\}$, its spectral density operator $\boldsymbol{f}_{y,\theta} = (2\pi)^{-1}\sum_{k\in\mathbb{Z}} \boldsymbol{\Sigma}_{y,k}\exp(-\mathrm{i}k\theta)$ for $\theta \in [-\pi, \pi]$ exists and the functional stability measure

$$\mathcal{M}_y = 2\pi \underset{\theta\in[-\pi,\pi],\boldsymbol{\Phi}\in\mathbb{H}_0}{\text{ess sup}} \frac{\langle \boldsymbol{\Phi}, \boldsymbol{f}_{y,\theta}(\boldsymbol{\Phi})\rangle}{\langle \boldsymbol{\Phi}, \boldsymbol{\Sigma}_{y,0}(\boldsymbol{\Phi})\rangle} < \infty\,, \tag{26}$$

where $\mathbb{H}_0 = \{\boldsymbol{\Phi} \in \mathbb{H} : \langle \boldsymbol{\Phi}, \boldsymbol{\Sigma}_{y,0}(\boldsymbol{\Phi})\rangle \in (0, \infty)\}$.

Condition 1(i) can be viewed as the functional (or multivariate) generalization of the multivariate (or functional) linear process. Conditions 1(ii) and 1(iii) guarantee the covariance-stationarity of $\{\mathbf{Y}_t(\cdot)\}$ and imply that $\max_{j\in[p]} \int_{\mathcal{U}} \Sigma_{y,0,jj}(u, u)\,\mathrm{d}u = O(1)$ (Fang et al., 2022). Both conditions are essential to derive the convergence rate for $\widehat{\boldsymbol{\Sigma}}^{\mathrm{S}}_{y,k}$ specified in (15), $\max_{i,j\in[p]} \|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij} - \Sigma_{y,k,ij}\|_{\mathcal{S}} = O_{\mathrm{p}}\{\mathcal{M}_y(n^{-1}\log p)^{1/2}\}$, which plays a crucial rule in our theoretical analysis. In general, we can relax Conditions 1(ii) and 1(iii) by allowing $\sum_{l=0}^{\infty} \|\boldsymbol{\mathcal{D}}_l\|_{\mathcal{S},\infty}$ and $\max_{j\in[p]} \int_{\mathcal{U}} \mathrm{Cov}\{\epsilon_{tj}(u), \epsilon_{tj}(u)\}\,\mathrm{d}u$ to diverge slowly with p, and our established rates below will depend on these two terms. Condition 2 places a finite upper bound on the functional stability measure, which characterizes the effect of small decaying eigenvalues of $\boldsymbol{\Sigma}_{y,0}$ on the numerator of (26), thus being able to handle infinite-dimensional functional objects $Y_{tj}(\cdot)$. See its detailed discussion in Guo and Qiao (2023).

Condition 3. For $\mathbf{A} = (A_{ij})_{p\times p}$, $\max_{i\in[p]} \sum_{j=1}^{p} |A_{ij}|^\alpha \leqslant s_1$ and $\max_{j\in[p]} \sum_{i=1}^{p} |A_{ij}|^\alpha \leqslant s_2$ for some constant $\alpha \in [0, 1)$.

The parameters s_1 and s_2 determine the row and column sparsity levels of $\mathbf{A}$, respectively. We may allow s_1 and s_2 to grow at some slow rates as p increases. The row sparsity with small s_1 entails that each component of $\mathbf{Y}_t(\cdot)$ is a linear combination of a small number of components in $\mathbf{Z}_t(\cdot)$, while the column sparsity with small s_2 corresponds to the case that each $Z_{tj}(\cdot)$ has impact on only a few components of $\mathbf{Y}_t(\cdot)$. The parameter α also controls the sparsity level of $\mathbf{A}$ with a smaller value yielding a sparser $\mathbf{A}$. Write

$$p_\dagger = \max_{l \in [q]} p_l \,. \tag{27}$$

Lemma A2 in the supplementary material reveals that the functional sparsity structures in columns/rows of $\boldsymbol{\Sigma}_{y,k}(\cdot,\cdot)$ are determined by $s_1 s_2 p_\dagger$ with smaller values of s_1, s_2 and $p_\dagger$ yielding functional sparser $\boldsymbol{\Sigma}_{y,k}(\cdot,\cdot)$.

Recall that $\mathbf{W}_z = \operatorname{diag}(\mathbf{W}_{z,1},\ldots,\mathbf{W}_{z,q})$ in (7) is a block-diagonal matrix, where $\mathbf{W}_{z,l}$ is a $p_l \times p_l$ matrix. We further define

$$\rho = \min_{j \neq l} \min_{\lambda \in \Lambda(\mathbf{W}_{z,l}),\, \tilde{\lambda} \in \Lambda(\mathbf{W}_{z,j})} |\lambda - \tilde{\lambda}| \,, \tag{28}$$

where $\Lambda(\mathbf{B})$ denotes the set of eigenvalues of the matrix $\mathbf{B}$, and assume $\rho > 0$.

We first establish the group recovery consistency of Step 1, i.e. the segmentation step. To do this, we reformulate the permutation in Section 2.1 in an equivalent graph representation way. For $\boldsymbol{\Gamma}_y$ specified in (9), it holds that $\boldsymbol{\Gamma}_y = \mathbf{A}\boldsymbol{\Gamma}_z$. As shown in Proposition 1(i), $\boldsymbol{\Gamma}_z$ can be taken as a block-diagonal orthogonal matrix with the main block sizes $p_1,\ldots,p_q$. Write $\boldsymbol{\Gamma}_z = \operatorname{diag}(\boldsymbol{\Gamma}_{z,1},\ldots,\boldsymbol{\Gamma}_{z,q})$. Since $\mathbf{A} = (\mathbf{A}_1,\ldots,\mathbf{A}_q)$, we have $\boldsymbol{\Gamma}_y \equiv (\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_p) = (\mathbf{A}_1\boldsymbol{\Gamma}_{z,1},\ldots,\mathbf{A}_q\boldsymbol{\Gamma}_{z,q})$. The columns of $\boldsymbol{\Gamma}_y$ are naturally partitioned in to q groups $G_1,\ldots,G_q$, where $G_l = \{\boldsymbol{\eta}_{\sum_{l'=0}^{l-1} p_{l'}+1},\ldots,\boldsymbol{\eta}_{\sum_{l'=0}^{l} p_{l'}}\}$ with $p_0 = 0$. To simplify the notation, we just write

$$G_l = \left\{ \sum_{l'=0}^{l-1} p_{l'} + 1, \ldots, \sum_{l'=0}^{l} p_{l'} \right\}, \qquad l \in [q] \,. \tag{29}$$

Recall that the columns of such defined $\boldsymbol{\Gamma}_y$ are the eigenvectors of $\mathbf{W}_y$. For ρ defined in (28), if $\|\widehat{\mathbf{W}}_y - \mathbf{W}_y\|_2 \leqslant \rho/5$, by Lemma A4 in the supplementary material, there exists an orthogonal matrix $\mathbf{H} = \operatorname{diag}(\mathbf{H}_1,\ldots,\mathbf{H}_q)$ with $\mathbf{H}_l \in \mathbb{R}^{p_l \times p_l}$ for each $l \in [q]$ and a column permutation matrix $\mathbf{R}$ for $\widehat{\boldsymbol{\Gamma}}_y$, such that $\widehat{\boldsymbol{\Gamma}}_y \mathbf{R} \equiv (\widehat{\boldsymbol{\Pi}}_1,\ldots,\widehat{\boldsymbol{\Pi}}_q)$ with $\widehat{\boldsymbol{\Pi}}_l \in \mathbb{R}^{p \times p_l}$, and

$$\|\widehat{\boldsymbol{\Pi}}_l - \mathbf{A}_l \boldsymbol{\Gamma}_{z,l} \mathbf{H}_l\|_2 \leqslant 8\rho^{-1} \|\widehat{\mathbf{W}}_y - \mathbf{W}_y\|_2 \,. \tag{30}$$

If the p eigenvalues of $\mathbf{W}_y$ are distinct, $\mathbf{H}$ is a diagonal matrix with elements in the diagonal being 1 or -1. Write $\boldsymbol{\Gamma}_y\mathbf{H} = (\mathbf{A}_1\boldsymbol{\Gamma}_{z,1}\mathbf{H}_1, \ldots, \mathbf{A}_q\boldsymbol{\Gamma}_{z,q}\mathbf{H}_q) \equiv (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$. For each $l \in [q]$, we can define a graph (G_l, E_l) such that $(i,j) \in E_l$ if and only if $\max_{|k| \leqslant m} \|\boldsymbol{\gamma}_i^\top \boldsymbol{\Sigma}_{y,k}\boldsymbol{\gamma}_j\|_{\mathcal{S}} \neq 0$.

Condition 4. There exists some $\varsigma > 0$ such that $\inf_{(i,j)\in E_l} \max_{|k|\leqslant m} \|\boldsymbol{\gamma}_i^\top \boldsymbol{\Sigma}_{y,k}\boldsymbol{\gamma}_j\|_{\mathcal{S}} \geqslant \varsigma$ for each $l \in [q]$, where m is specified in (12).

Condition 4 ensures that the group G_l is inseparable at the minimal signal level ς given the transformation $\mathbf{A}_l\boldsymbol{\Gamma}_{z,l}\mathbf{H}_l$ for each $l \in [q]$. Define $T_{ij} = \max_{|k|\leqslant m} \|\boldsymbol{\gamma}_i^\top \boldsymbol{\Sigma}_{y,k}\boldsymbol{\gamma}_j\|_{\mathcal{S}}$ and $\varrho = \sum_{l=1}^q |E_l|$. Rearrange $\aleph = p(p-1)/2$ values of T_{ij} $(1 \leqslant i < j \leqslant p)$ in the descending order, $T_{(1)} \geqslant \cdots \geqslant T_{(\aleph)}$. We then have $T_{(i)} \geqslant \varsigma$ for $i \in [\varrho]$ and $T_{(i)} = 0$ for $i \geqslant \varrho + 1$. Denote by $E = \{(i,j) : T_{ij} \geqslant T_{(\varrho)}, 1 \leqslant i < j \leqslant p\}$ the edge set of $G = [p]$ under the transformation $\boldsymbol{\Gamma}_y\mathbf{H}$. The true segmentation $\{G_1, \ldots, G_q\}$ in (29) can then be identified by splitting (G, E) into q isolated subgraphs $(G_1, E_1), \ldots, (G_q, E_q)$.

Recall that with the aid of $\widehat{\boldsymbol{\Gamma}}_y$, the estimated segmentation is obtained via the ratio-based estimator $\hat{\varrho}$ as defined in (13). To be specific, we build an estimated graph $(G, \widetilde{E})$ with vertex set $G = [p]$ and edge set $\widetilde{E} = \{(i,j) : \widehat{T}_{ij} \geqslant \widehat{T}_{(\hat{\varrho})}, 1 \leqslant i < j \leqslant p\}$, and split it into $\hat{q}$ isolated subgraphs $(\widetilde{G}_1, \widetilde{E}_1), \ldots, (\widetilde{G}_{\hat{q}}, \widetilde{E}_{\hat{q}})$. Note that p columns of $\widehat{\boldsymbol{\Gamma}}_y = (\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_p)$ correspond to the ordered eigenvalues $\lambda_1(\widehat{\mathbf{W}}_y) \geqslant \cdots \geqslant \lambda_p(\widehat{\mathbf{W}}_y)$. Write $\widehat{\boldsymbol{\Gamma}}_y\mathbf{R} \equiv (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_p)$ and let $\pi : [p] \to [p]$ denote the permutation associated with $\mathbf{R}$, i.e. $\hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{\eta}}_{\pi(i)}$. Based on the permutation mapping π, we let $\widehat{G}_l = \{\pi^{-1}(i) : i \in \widetilde{G}_l\}$ for $l \in [\hat{q}]$.

Theorem 1. *Let Conditions 1–4 hold. For each $|k| \leqslant k_0 \vee m$, select $\omega_k = c_k\mathcal{M}_y(n^{-1}\log p)^{1/2}$ in (16) for some sufficiently large constant $c_k > 0$. Assume $(\rho^{-1}s_1^2s_2^2p_\dagger^{3-\alpha})^{2/(1-\alpha)}\mathcal{M}_y^2 \log p = o(n)$ and δ_n in (13) satisfies $\rho^{-1}s_1^3s_2^3p_\dagger^{5-2\alpha}\mathcal{M}_y^{1-\alpha}(n^{-1}\log p)^{(1-\alpha)/2} \ll \delta_n \ll \varsigma^2T_{(1)}^{-1}$, where $p_\dagger$ and ρ are specified in (27) and (28), respectively. As $n \to \infty$, it holds that* (i) *$\mathbb{P}(\hat{q} = q) \to 1$ and* (ii) *there exists a permutation $\tilde{\pi} : [q] \to [q]$ such that $\mathbb{P}[\bigcap_{l=1}^q \{\widehat{G}_{\tilde{\pi}(l)} = G_l\} \,|\, \hat{q} = q] \to 1$.*

Theorem 1 gives the group recovery consistency of our Step 1. We next evaluate the errors in estimating $\mathcal{C}(\mathbf{A}_l)$ for $l \in [q]$. Based on the estimated groups $\{\widehat{G}_1, \ldots, \widehat{G}_{\hat{q}}\}$, we reorganize the order of $(\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_p) = (\widehat{\boldsymbol{\eta}}_{\pi(1)}, \ldots, \widehat{\boldsymbol{\eta}}_{\pi(p)})$ and define $\widehat{\mathbf{A}}_l$ in (14) as $\widehat{\mathbf{A}}_l = (\widehat{\boldsymbol{\gamma}}_i)_{i \in \widehat{G}_l}$ for $l \in [\hat{q}]$. We consider a general discrepancy measure (Chang et al., 2015) between two linear spaces $\mathcal{C}(\mathbf{B}_1)$ and $\mathcal{C}(\mathbf{B}_2)$ spanned by the columns of $\mathbf{B}_1 \in \mathbb{R}^{p \times \tilde{p}_1}$ and $\mathbf{B}_2 \in \mathbb{R}^{p \times \tilde{p}_2}$, respectively, with $\mathbf{B}_i^\top \mathbf{B}_i = \mathbf{I}_{\tilde{p}_i}$ for $i \in [2]$ as

$$D\{\mathcal{C}(\mathbf{B}_1), \mathcal{C}(\mathbf{B}_2)\} = \sqrt{1 - \frac{\operatorname{tr}(\mathbf{B}_1 \mathbf{B}_1^\top \mathbf{B}_2 \mathbf{B}_2^\top)}{\max(\tilde{p}_1, \tilde{p}_2)}} \in [0, 1]\,. \tag{31}$$

Then $D\{\mathcal{C}(\mathbf{B}_1), \mathcal{C}(\mathbf{B}_2)\}$ is equal to 0 if and only if $\mathcal{C}(\mathbf{B}_1) \subset \mathcal{C}(\mathbf{B}_2)$ or $\mathcal{C}(\mathbf{B}_2) \subset \mathcal{C}(\mathbf{B}_1)$, and to 1 if and only if the two spaces are orthogonal.

Theorem 2. *Let conditions for Theorem 1 hold. As $n \to \infty$, it holds that*

$$\max_{l \in [q]} \min_{j \in [\hat{q}]} D\{\mathcal{C}(\mathbf{A}_l), \mathcal{C}(\widehat{\mathbf{A}}_j)\} = O_{\mathrm{p}}\big\{\rho^{-1} s_1^2 s_2^2 p_\dagger^{3-\alpha} \mathcal{M}_y^{1-\alpha} (n^{-1} \log p)^{(1-\alpha)/2}\big\}\,.$$

Theorem 2 presents the uniform convergence rate for $\min_{j \in [\hat{q}]} D\{\mathcal{C}(\mathbf{A}_l), \mathcal{C}(\widehat{\mathbf{A}}_j)\}$ over $l \in [q]$, which is determined by both dimensionality parameters $(n, p, p_\dagger, s_1, s_2)$ and internal parameters $(\mathcal{M}_y, \alpha)$. The rate is faster for smaller values of $\{s_1, s_2, p_\dagger, \mathcal{M}_y, \alpha\}$, while enlarging the minimum eigen-gap between different blocks (i.e., larger ρ) reduces the difficulty of estimating each $\mathcal{C}(\mathbf{A}_l)$.

Supported by Theorems 1 and 2, our subsequent theoretical results are developed by assuming that the group structure of $\mathbf{Z}_t(\cdot)$ is correctly identified or known, i.e., $\hat{q} = q$ and $\widehat{G}_l = G_l$ for each l. We now turn to investigate the theoretical properties of Step 2, i.e. the dimension reduction step. Inherited from the segmentation step, $\mathbf{Z}_t^{(1)}(\cdot), \ldots, \mathbf{Z}_t^{(q)}(\cdot)$ rely on the specific form of $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_q)$, and thus is not uniquely defined. Yet intuitively, we only require a certain transformation matrix to make our subsequent analysis related to $\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_p$ mathematically tractable. Based on (30), we define $\boldsymbol{\Pi}_l = \mathbf{A}_l \boldsymbol{\Gamma}_{z,l} \mathbf{H}_l$ and it holds that $\mathcal{C}(\boldsymbol{\Pi}_l) = \mathcal{C}(\mathbf{A}_l)$ for each $l \in [q]$. Let $\mathbf{Z}_t^{(l)}(\cdot) = \boldsymbol{\Pi}_l^\top \mathbf{Y}_t(\cdot)$. Recall (1) and (22). The primary goal of Step 2 is to identify each r_l and to estimate the associated dynamic space $\mathcal{C}_l = \operatorname{span}\{\boldsymbol{\psi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\psi}_{r_l}^{(l)}(\cdot)\}$. Recall that $\{\hat{\theta}_j^{(l)}, \widehat{\boldsymbol{\psi}}_j^{(l)}(\cdot)\}_{j \geqslant 1}$ are the eigenvalue/eigenfunction pairs of $\widehat{\mathbf{K}}^{(l)}(\cdot,\cdot)$ defined in (24) with $\widehat{\mathbf{A}}_l = (\widehat{\gamma}_i)_{i\in G_l}$ and the dimension r_l is a fixed finite integer for all $l \in [q]$. Our asymptotic results are based on the following regularity condition:

Condition 5. For each $l \in [q]$, all r_l nonzero eigenvalues of $\mathbf{K}^{(l)}(\cdot,\cdot)$ are different, i.e., $\theta_1^{(l)} > \cdots > \theta_{r_l}^{(l)} > 0 = \theta_{r_l+1}^{(l)} = \cdots$.

Theorem 3. *Let Conditions 1–3 and 5 hold. Assume* $(\rho^{-1}s_1^3s_2^3p_\dagger^{5-2\alpha})^{2/(1-\alpha)}\mathcal{M}_y^2\log p = o(n)$ *and* $\tilde{\delta}_n$ *in (25) satisfies* $\rho^{-1}s_1^3s_2^3p_\dagger^{7-2\alpha}\mathcal{M}_y^{1-\alpha}(n^{-1}\log p)^{(1-\alpha)/2} \ll \tilde{\delta}_n \ll \min_{l\in[q]}\{\theta_{r_l}^{(l)}\}^2/\max_{l\in[q]}\theta_1^{(l)}$, *where* $p_\dagger$ *and* ρ *are specified in (27) and (28), respectively. As* $n \to \infty$, *it holds that* $\mathbb{P}[\bigcap_{l=1}^q\{\hat{r}_l = r_l\}] \to 1$.

Theorem 3 shows that r_l can be correctly identified with probability tending to one uniformly over $l \in [q]$. Let $\widehat{\mathcal{C}}_l = \operatorname{span}\{\widehat{\boldsymbol{\psi}}_1^{(l)}(\cdot),\ldots,\widehat{\boldsymbol{\psi}}_{\hat{r}_l}^{(l)}(\cdot)\}$ be the dynamic space spanned by $\hat{r}_l$ estimated eigenfunctions. To measure the discrepancy between $\mathcal{C}_l$ and $\widehat{\mathcal{C}}_l$, we introduce the following metric. For two subspaces $\mathcal{C}(\mathbf{b}_1) = \operatorname{span}\{\mathbf{b}_{11}(\cdot),\ldots,\mathbf{b}_{1\tilde{r}_1}(\cdot)\}$ and $\mathcal{C}(\mathbf{b}_2) = \operatorname{span}\{\mathbf{b}_{21}(\cdot),\ldots,\mathbf{b}_{2\tilde{r}_2}(\cdot)\}$ satisfying $\langle\mathbf{b}_{ij},\mathbf{b}_{ik}\rangle = I(j=k)$ for each $i \in [2]$, the discrepancy measure between $\mathcal{C}(\mathbf{b}_1)$ and $\mathcal{C}(\mathbf{b}_2)$ is defined as

$$\widetilde{D}\{\mathcal{C}(\mathbf{b}_1),\mathcal{C}(\mathbf{b}_2)\} = \sqrt{1 - \frac{1}{\max(\tilde{r}_1,\tilde{r}_2)}\sum_{j=1}^{\tilde{r}_1}\sum_{k=1}^{\tilde{r}_2}\langle\mathbf{b}_{1j},\mathbf{b}_{2k}\rangle^2} \in [0,1],$$

which equals 0 if and only if $\mathcal{C}(\mathbf{b}_1) \subset \mathcal{C}(\mathbf{b}_2)$ or $\mathcal{C}(\mathbf{b}_2) \subset \mathcal{C}(\mathbf{b}_1)$ and 1 if and only if two spaces are orthogonal.

Theorem 4. *Let conditions for Theorem 3 hold. Assume* $(\Delta^{-1}\rho^{-1}s_1^3s_2^3p_\dagger^{7-2\alpha})^{2/(1-\alpha)}\mathcal{M}_y^2\log p = o(n)$ *with* $\Delta = \min_{l\in[q],j\in[r_l]}\{\theta_j^{(l)} - \theta_{j+1}^{(l)}\}$. *As* $n \to \infty$, *it holds that*

$$\max_{l\in[q]}\widetilde{D}(\widehat{\mathcal{C}}_l,\mathcal{C}_l) = O_{\mathrm{p}}\{\Delta^{-1}\rho^{-1}s_1^3s_2^3p_\dagger^{7-2\alpha}\mathcal{M}_y^{1-\alpha}(n^{-1}\log p)^{(1-\alpha)/2}\}.$$

5 Simulation studies

We conduct a series of simulations to illustrate the finite sample performance of the proposed methods for cases with fixed p and large p in Sections 5.1 and 5.2, respectively.

To simplify the data generating process, we consider a relaxed form of (2), written as

$$\breve{\mathbf{Y}}_t(u) = \breve{\mathbf{A}}\breve{\mathbf{Z}}_t(u) = \breve{\mathbf{A}}\{\breve{\mathbf{Z}}_t^{(1)}(u)^\top, \ldots, \breve{\mathbf{Z}}_t^{(q)}(u)^\top\}^\top, \quad u \in \mathcal{U} = [0,1], \tag{32}$$

with no orthogonality restriction on the transformation matrix $\breve{\mathbf{A}} = (\breve{\mathbf{A}}_1, \ldots, \breve{\mathbf{A}}_q)$. The p-dimensional transformed functional time series $\breve{\mathbf{Z}}_t(\cdot)$ is formed by q uncorrelated groups $\{\breve{\mathbf{Z}}_t^{(l)}(\cdot) : l \in [q]\}$, where each group $\breve{\mathbf{Z}}_t^{(l)}(\cdot)$ is decomposed as the sum of a dynamic element $\breve{\mathbf{X}}_t^{(l)}(\cdot)$ and a white noise element $\breve{\boldsymbol{\varepsilon}}_t^{(l)}(\cdot)$, i.e. $\breve{\mathbf{Z}}_t^{(l)}(\cdot) = \breve{\mathbf{X}}_t^{(l)}(\cdot) + \breve{\boldsymbol{\varepsilon}}_t^{(l)}(\cdot)$. Based on (4) in Section 2, model (32) can then be easily reformulated as (2) by setting

$$\mathbf{Y}_t(\cdot) = \mathbf{V}_{\breve{y}}^{-1/2}\breve{\mathbf{Y}}_t(\cdot), \quad \mathbf{A} = \mathbf{V}_{\breve{y}}^{-1/2}\breve{\mathbf{A}}\mathbf{V}_{\breve{z}}^{1/2} \text{ and } \mathbf{Z}_t(\cdot) = \mathbf{V}_{\breve{z}}^{-1/2}\breve{\mathbf{Z}}_t(\cdot), \tag{33}$$

where $\mathbf{V}_{\breve{y}} = \int_{\mathcal{U}} \mathrm{Cov}\{\breve{\mathbf{Y}}_t(u), \breve{\mathbf{Y}}_t(u)\}\,\mathrm{d}u$ and $\mathbf{V}_{\breve{z}} = \int_{\mathcal{U}} \mathrm{Cov}\{\breve{\mathbf{Z}}_t(u), \breve{\mathbf{Z}}_t(u)\}\,\mathrm{d}u$. Then the orthogonality of $\mathbf{A}$ is satisfied.

Write $\breve{\boldsymbol{\varepsilon}}_t(\cdot) = \{\breve{\boldsymbol{\varepsilon}}_t^{(1)}(\cdot)^\top, \ldots, \breve{\boldsymbol{\varepsilon}}_t^{(q)}(\cdot)^\top\}^\top \equiv \{\breve{\varepsilon}_{t1}(\cdot), \ldots, \breve{\varepsilon}_{tp}(\cdot)\}^\top$. We generate each curve component of $\breve{\boldsymbol{\varepsilon}}_t(\cdot)$ independently by $\breve{\varepsilon}_{tj}(u) = \sum_{l=1}^{10} 2^{-(l-1)} e_{tjl}\psi_l(u)$ for $j \in [p]$, where e_{tjl}'s are sampled independently from $\mathcal{N}(0,1)$ and $\{\psi_l(\cdot)\}_{l=1}^{10}$ is a 10-dimensional Fourier basis function. The finite-dimensional dynamics $\breve{\mathbf{X}}_t(\cdot) = \{\breve{\mathbf{X}}_t^{(1)}(\cdot)^\top, \ldots, \breve{\mathbf{X}}_t^{(q)}(\cdot)^\top\}^\top$ with prescribed group structure is generated based on some 5-dimensional curve dynamics $\vartheta_{tg}(u) = \sum_{l=1}^{5} \kappa_{tgl}\psi_l(u)$ for $g \in [20]$. The basis coefficients $\boldsymbol{\kappa}_{tg} = (\kappa_{tg1}, \ldots, \kappa_{tg5})^\top$ are generated from a stationary VAR model $\boldsymbol{\kappa}_{tg} = \mathbf{U}_g\boldsymbol{\kappa}_{(t-1)g} + \mathbf{e}_t$ for each g. To guarantee the stationarity of $\boldsymbol{\kappa}_{tg}$, we generate $\mathbf{U}_g = \iota\breve{\mathbf{U}}_g/\rho(\breve{\mathbf{U}}_g)$ with $\iota \sim \mathrm{Uniform}[0.5, 1]$ and $\rho(\breve{\mathbf{U}}_g)$ being the spectral radius of $\breve{\mathbf{U}}_g \in \mathbb{R}^{5\times 5}$, the entries of which are sampled independently from Uniform$[-3, 3]$. The components of the innovation $\mathbf{e}_t$ are sampled independently from $\mathcal{N}(0,1)$. We will specify the exact forms of $\breve{\mathbf{X}}_t(\cdot)$ under the fixed and large p scenarios in Sections 5.1 and 5.2, respectively. The white noise sequence $\breve{\boldsymbol{\varepsilon}}_t(\cdot)$ ensures that $\breve{\mathbf{Z}}_t(\cdot)$ as well as $\mathbf{Z}_t(\cdot)$ share the same group structure as $\breve{\mathbf{X}}_t(\cdot)$. Unless otherwise stated, we set $k_0 = m = 5$ and $c_r = c_\varrho = 0.75$ in our procedure, as our simulation results suggest that our procedure is robust to the choices of these parameters.

5.1 Cases with fixed p

We consider the following three examples of $\check{\mathbf{X}}_t(\cdot) = \{\check{X}_{t1}(\cdot), \ldots, \check{X}_{tp}(\cdot)\}^\top$ with different group structures for $p \in \{6, 10, 15\}$ based on independent $\vartheta_{t1}(\cdot), \ldots, \vartheta_{t5}(\cdot)$.

Example 1. $\check{X}_{t1}(\cdot) = \vartheta_{t1}(\cdot)$, $\check{X}_{tj}(\cdot) = \vartheta_{(t+j-2)2}(\cdot)$ for $j \in \{2, 3\}$ and $\check{X}_{tj}(\cdot) = \vartheta_{(t+j-4)3}(\cdot)$ for $j \in \{4, 5, 6\}$.

Example 2. $\check{X}_{tj}(\cdot)$ for $j \in [6]$ are the same as those in Example 1 and $\check{X}_{tj}(\cdot) = \vartheta_{(t+j-7)4}(\cdot)$ for $j \in \{7, \ldots, 10\}$.

Example 3. $\check{X}_{tj}(\cdot)$ for $j \in [10]$ are the same as those in Example 2 and $\check{X}_{tj}(\cdot) = \vartheta_{(t+j-11)5}(\cdot)$ for $j \in \{11, \ldots, 15\}$.

Therefore, $\check{\mathbf{X}}_t(\cdot)$ consists of $q = 3, 4$ and 5 uncorrelated groups of curve subseries in Examples 1, 2 and 3, respectively, where the number of component curves per group is $p_l = l$ for $l \in [q]$. The p-dimensional observed functional time series $\check{\mathbf{Y}}_t(\cdot) = \{\check{Y}_{t1}(\cdot), \ldots, \check{Y}_{tp}(\cdot)\}^\top$ for $t \in [n]$ is then generated by (32) with the entries of $\check{\mathbf{A}}$ sampled independently from Uniform$[-3, 3]$. To obtain h-step ahead prediction of $\check{\mathbf{Y}}_t(\cdot)$, we integrate the segmentation and dimension reduction steps respectively in Sections 2 and 3 into the VAR estimation as outlined in Algorithm 1. For each of the three examples introduced above, we select

$$\widehat{\mathbf{V}}_{\check{y}}^{(h)} = \frac{1}{n-h} \sum_{t=1}^{n-h} \int_{\mathcal{U}} \left\{ \check{\mathbf{Y}}_t(u) - \frac{1}{n-h} \sum_{t=1}^{n-h} \check{\mathbf{Y}}_t(u) \right\}^{\otimes 2} \mathrm{d}u \,, \tag{34}$$

$$\widehat{\boldsymbol{\Sigma}}_{\tilde{y},k}^{(h)}(u, v) = \frac{1}{n-h-k} \sum_{t=1}^{n-h-k} \{\tilde{\mathbf{Y}}_t(u) - \bar{\mathbf{Y}}_*(u)\}\{\tilde{\mathbf{Y}}_{t+k}(v) - \bar{\mathbf{Y}}_*(v)\}^\top \,, \tag{35}$$

with $\bar{\mathbf{Y}}_*(\cdot) = (n-h-k)^{-1} \sum_{t=1}^{n-h-k} \tilde{\mathbf{Y}}_t(\cdot)$, for the quantities involved in Step (i) of Algorithm 1. We refer to the **seg**mentation-and-**V**AR-based Algorithm 1 with selections of $\widehat{\mathbf{V}}_{\check{y}}^{(h)}$ in (34) and $\widehat{\boldsymbol{\Sigma}}_{\tilde{y},k}^{(h)}(u, v)$ in (35) as SegV hereafter.

Algorithm 1 General prediction procedure for multivariate functional time series

(i) Treat the first $n-h$ observations as training data, adopt the normalization step to obtain $\widetilde{\mathbf{Y}}_t(\cdot)=\{\widehat{\mathbf{V}}_{\breve{y}}^{(h)}\}^{-1/2}\breve{\mathbf{Y}}_t(\cdot)$, where $\widehat{\mathbf{V}}_{\breve{y}}^{(h)}$ is the consistent estimator of $\mathbf{V}_{\breve{y}}$ in (33), and implement the segmentation procedure on $\{\widetilde{\mathbf{Y}}_t(\cdot)\}_{t=1}^{n-h}$ as in Section 2.1 to estimate the transformation matrix $\widehat{\mathbf{A}}=(\widehat{\mathbf{A}}_1,\dots,\widehat{\mathbf{A}}_{\hat{q}})$ and the transformed curve subseries $\{\widehat{\mathbf{Z}}_t^{(l)}(\cdot):l\in[\hat{q}]\}$.

(ii) Apply the procedure in Section 3 on each $\{\widehat{\mathbf{Z}}_t^{(l)}(\cdot)\}_{t=1}^{n-h}$ to achieve the h-step ahead prediction denoted as $\mathring{\mathbf{Z}}_n^{(l)}(\cdot)$, for $l\in[\hat{q}]$. In particular, for each l, select the best VAR model that best fits each lower-dimensional vector process $\{\widehat{\boldsymbol{\zeta}}_t^{(l)}\}_{t=1}^{n-h}$ according to the AIC criterion.

(iii) Obtain the h-step ahead prediction $\widehat{\mathbf{A}}\mathring{\mathbf{Z}}_n(\cdot)$ for the normalized curves $\widetilde{\mathbf{Y}}_n(\cdot)$ with $\mathring{\mathbf{Z}}_n(\cdot)=\{\mathring{\mathbf{Z}}_n^{(1)}(\cdot)^\top,\dots,\mathring{\mathbf{Z}}_n^{(\hat{q})}(\cdot)^\top\}^\top$. Then the h-step ahead prediction for the original curves $\breve{\mathbf{Y}}_n(\cdot)$ is given by $\hat{\mathbf{Y}}_n(\cdot)\equiv\{\hat{Y}_{n1}(\cdot),\dots,\hat{Y}_{np}(\cdot)\}^\top=\{\widehat{\mathbf{V}}_{\breve{y}}^{(h)}\}^{1/2}\widehat{\mathbf{A}}\mathring{\mathbf{Z}}_n(\cdot)$.

The performance of our two-step proposal is examined in terms of linear space estimation, group identification and post-sample prediction. For $\mathbf{A}=(\mathbf{A}_1,\dots,\mathbf{A}_q)$ specified in (33), with the aid of (31), define $f(l)=\arg\min_{j\in[\hat{q}]}D^2\{\mathcal{C}(\mathbf{A}_l),\mathcal{C}(\widehat{\mathbf{A}}_j)\}$ for each $l\in[q]$. We then call $\widehat{\mathbf{A}}=(\widehat{\mathbf{A}}_1,\dots,\widehat{\mathbf{A}}_{\hat{q}})$ an effective segmentation of $\mathbf{A}$ if (i) $1<\hat{q}\leqslant q$, and (ii) $\operatorname{rank}(\widehat{\mathbf{A}}_{l'})=\sum_{l\in[q]:\,f(l)=l'}\operatorname{rank}(\mathbf{A}_l)$ for each $l'\in[\hat{q}]$. The intuition is as follows. The effective segmentation implies that each identified group in $\widehat{\mathbf{Z}}_t(\cdot)$ contains at least one, but not all, groups in $\mathbf{Z}_t(\cdot)$. Since our main target is to forecast $\breve{\mathbf{Y}}_t(\cdot)$ based on the cross-serial dependence in $\{\mathbf{Z}_t^{(l)}(\cdot):l\in[q]\}$, this segmentation result is effective in the sense that the linear dynamics in $\mathbf{Z}_t(\cdot)$ is well kept in $\{\widehat{\mathbf{Z}}_t^{(l)}(\cdot):l\in[\hat{q}]\}$ without any contamination or damage and a mild dimension reduction is achieved with $\hat{q}>1$. For the special case of complete segmentation ($\hat{q}=q$), we use the maximum and averaged estimation errors for $(\widehat{\mathbf{A}}_1,\dots,\widehat{\mathbf{A}}_{\hat{q}})$, respectively, defined as $\mathrm{MaxE}=\max_{l\in[q]}D^2\{\mathcal{C}(\mathbf{A}_l),\mathcal{C}(\widehat{\mathbf{A}}_{f(l)})\}$ and $\mathrm{AvgE}=q^{-1}\sum_{l=1}^q D^2\{\mathcal{C}(\mathbf{A}_l),\mathcal{C}(\widehat{\mathbf{A}}_{f(l)})\}$

to assess the ability of our method in fully recovering the spanned spaces $\mathcal{C}(\mathbf{A}_1), \ldots, \mathcal{C}(\mathbf{A}_q)$. Note that $\mathbf{A}$ in (33) can not be easily computed, as the true $\mathbf{V}_{\check{y}}$ and $\mathbf{V}_{\check{z}}$ are hard to find even for simulated examples. For $\check{\mathbf{A}}$ specified in (32), let $\widetilde{\mathbf{A}} = \mathbf{V}_{\check{y}}^{-1/2}\check{\mathbf{A}} \equiv (\widetilde{\mathbf{A}}_1, \ldots, \widetilde{\mathbf{A}}_q)$ with $\widetilde{\mathbf{A}}_l = \mathbf{V}_{\check{y}}^{-1/2}\check{\mathbf{A}}_l$. Since $\mathbf{V}_{\check{z}}$ is a block-diagonal matrix, then $\mathcal{C}(\widetilde{\mathbf{A}}_l) = \mathcal{C}(\mathbf{A}_l)$ for $l \in [q]$. Hence, we can replace $\mathcal{C}(\mathbf{A}_l)$ by $\mathcal{C}(\{\widehat{\mathbf{V}}_{\check{y}}^{(h)}\}^{-1/2}\check{\mathbf{A}}_l)$ to obtain the approximations of MaxE and AvgE in our simulations.

To evaluate the post-sample predictive accuracy, we define the mean squared prediction error (MSPE) as

$$\mathrm{MSPE} = \frac{1}{pN}\sum_{j=1}^{p}\sum_{i=1}^{N}\{\hat{Y}_{nj}(v_i) - \check{Y}_{nj}(v_i)\}^2 \tag{36}$$

with $v_1, \ldots, v_N$ being equally spaced time points in $[0,1]$, and compute the relative prediction error as the ratio of MSPE in (36) to that under the 'oracle' case. In the oracle case, we apply the procedure in Section 3 directly on each true $\{\check{\mathbf{Z}}_t^{(l)}(\cdot)\}_{t=1}^{n-h}$ to achieve the h-step ahead prediction for $\{\check{\mathbf{Z}}_n^{(l)}(\cdot) : l \in [q]\}$, denoted by $\{\check{\mathbf{Z}}_n^{(l)}(\cdot) : l \in [q]\}$, and further obtain the h-step ahead prediction $\check{\mathbf{A}}\{\check{\mathbf{Z}}_n^{(1)}(\cdot)^\top, \ldots, \check{\mathbf{Z}}_n^{(q)}(\cdot)^\top\}^\top$ for the original curves $\check{\mathbf{Y}}_t(\cdot)$. By comparison, we also implement an **uni**variate functional prediction method on each $\check{Y}_{tj}(\cdot)$ separately by performing univariate dimension reduction (Bathia et al., 2010), then predicting vector time series based on the best fitted **V**AR model and finally recovering functional prediction (denoted as UniV).

We generate $n \in \{200, 400, 800, 1600\}$ observations with $N = 30$ for each example and replicate each simulation 500 times. Table 1 provides numerical summaries, including the relative frequencies of the effective segmentation with $\hat{q} = q$ and $\hat{q} \geqslant q-1$, and the estimation errors for $\widehat{\mathbf{A}} = (\widehat{\mathbf{A}}_1, \ldots, \widehat{\mathbf{A}}_{\hat{q}})$ under the complete segmentation case. As one would expect, the proposed method provides higher proportions of effective segmentation and lower estimation errors as n increases, and performs fairly well for reasonably large n as p increases. For $(p, n) = (6, 200)$, we observe 62.6% complete segmentation with AvgE as low as 0.079. Furthermore, the proportions of effective segmentation with $\hat{q} \geqslant q - 1$ are above 93% for

$n \geqslant 200$. Similar results can be found for cases of $(p, n) = (10, 800+)$ and $(15, 1600)$, whose proportions of effective segmentation with $\hat{q} \geqslant q - 1$ remain higher than 87.4% and 83.2%, respectively. Table 1 also reports the relative one-step ahead prediction errors. It is evident that SegV significantly outperforms UniV in all settings, demonstrating the effectiveness of our proposed segmentation transformation and dimension reduction in predicting future values. Although the proportions of complete segmentation are not high especially when $p = 15$, the corresponding proportions of $\hat{q} \geqslant q - 1$ become substantially higher, and SegV performs very similarly to the oracle case with its relative prediction errors being close to 1.

5.2 Cases with large p

Under a large p scenario, a natural question to ask is whether the segmentation method based on the classical estimation for autocovariance functions of $\widetilde{\mathbf{Y}}_t(\cdot)$ (denoted as NonT) as (35) in Section 5.1 still performs well, and if not, whether a satisfactory improvement is attainable via the functional-thresholding estimation (denoted as FunT) developed in Section 2.2. To this end, we generate $\check{\mathbf{Y}}_t(\cdot)$ from model (32) with $p \in \{30, 60\}$ and $n \in \{200, 400\}$. Specifically, we let $\check{X}_{t(3l-2)}(\cdot) = \vartheta_{tl}(\cdot)$, $\check{X}_{t(3l-1)}(\cdot) = \vartheta_{(t+1)l}(\cdot)$, $\check{X}_{t(3l)}(\cdot) = \vartheta_{(t+2)l}(\cdot)$ for $l \in [q]$. This setting ensures q uncorrelated groups of curve subseries in $\check{\mathbf{X}}_t(\cdot)$ with $p_l = 3$ component curves per group and hence $q = 10$ and 20 correspond to $p = 30$ and 60, respectively. Let the $p \times p$ transformation matrix $\check{\mathbf{A}} = \mathbf{\Delta}_1 + \delta \mathbf{\Delta}_2$. Here $\mathbf{\Delta}_1 = \text{diag}\{\mathbf{\Delta}_{11}, \ldots, \mathbf{\Delta}_{1(p/6)}\}$ with elements of each $\mathbf{\Delta}_{1i} \in \mathbb{R}^{6\times 6}$ being sampled independently from Uniform$[-3, 3]$ for $i \in [p/6]$, and $\mathbf{\Delta}_2$ is a matrix with two randomly selected nonzero elements from Uniform$[-1, 1]$ each row. We set $\delta \in \{0.1, 0.5\}$. It is notable that our setting results in a very high-dimensional learning task in the sense that the intrinsic dimension $30 \times 5 = 150$ or $60 \times 5 = 300$ is large relative to the sample size $n = 200$ or 400.

We assess the performance of NonT and FunT in discovering the group structure. The optimal thresholding parameters $\hat{\omega}_k$ in FunT are selected by the five-fold cross-validation

Table 1: The relative frequencies of effective segmentation with respect to $\hat{q} = q$ and $\hat{q} \geqslant q-1$, and the means (standard deviations) of MaxE, AvgE, and relative MSPEs over 500 simulation runs.

		$n = 200$	$n = 400$	$n = 800$	$n = 1600$
	$\hat{q} = q$	0.626	0.722	0.772	0.880
	$\hat{q} \geqslant q-1$	0.930	0.988	0.998	1.000
Example 1	MaxE	0.128(0.088)	0.089(0.066)	0.053(0.048)	0.035(0.037)
($p = 6$)	AvgE	0.079(0.052)	0.053(0.038)	0.030(0.025)	0.019(0.019)
	SegV	1.081(0.172)	1.048(0.105)	1.026(0.065)	1.014(0.048)
	UniV	1.584(0.453)	1.598(0.423)	1.596(0.379)	1.651(0.443)
	$\hat{q} = q$	0.324	0.444	0.644	0.806
	$\hat{q} \geqslant q-1$	0.490	0.688	0.874	0.972
Example 2	MaxE	0.301(0.108)	0.193(0.090)	0.117(0.064)	0.072(0.049)
($p = 10$)	AvgE	0.183(0.059)	0.115(0.047)	0.069(0.035)	0.041(0.024)
	SegV	1.291(0.271)	1.174(0.215)	1.089(0.143)	1.059(0.091)
	UniV	1.708(0.404)	1.836(0.410)	1.841(0.436)	1.862(0.392)
	$\hat{q} = q$	0.032	0.178	0.410	0.622
	$\hat{q} \geqslant q-1$	0.086	0.344	0.616	0.832
Example 3	MaxE	0.426(0.091)	0.347(0.121)	0.241(0.113)	0.157(0.091)
($p = 15$)	AvgE	0.273(0.054)	0.195(0.050)	0.128(0.042)	0.077(0.033)
	SegV	1.477(0.313)	1.363(0.277)	1.166(0.156)	1.091(0.098)
	UniV	1.805(0.370)	1.967(0.394)	2.033(0.394)	2.001(0.384)

as discussed in Remark 2 in Section 2.2, and $\mathbf{V}_{\breve{y}}$ in the normalization step is estimated by $\widehat{\mathbf{V}}_{\breve{y}}^{(0)}$ given in (34), as the threshold version of $\widehat{\mathbf{V}}_{\breve{y}}^{(0)}$ might not be positive definite. In practice, when p is large, FunT may lead to segmentation with a small $\hat{q}$, indicating that some groups of $\{\widehat{\mathbf{Z}}_t^{(l)}(\cdot) : l \in [\hat{q}]\}$ contain multiple groups in $\{\mathbf{Z}_t^{(l)}(\cdot) : l \in [q]\}$. To ease the modelling burden of complex VAR process, we may consider performing further segmentation transformation on the estimated groups by repeating FunT R times. To be precise, the i-th round of segmentation transformation via FunT is performed within each group discovered in the $(i-1)$-th round with $c_{\varrho} = 1$ for $i \in [R]$, and hence $(\widehat{\mathbf{A}}_1, \ldots, \widehat{\mathbf{A}}_{\hat{q}})$ is updated after each iteration. Table 2 reports the relative frequencies of the effective segmentation for NonT and FunT with $R \in \{1, 5, 10\}$. Finally, we apply FunT-based SegV (denoted as FTSegV) in conjunction with the R-round segmentation transformation for $R \in \{1, 5, 10\}$ in Step (i) of Algorithm 1, and compare their one-step ahead predictive performance with UniV and SegV. Table 3 summarizes the relative prediction errors for all five comparison methods.

Several conclusions can be drawn from Tables 2 and 3. First, the performance of SegV severely deteriorates under the high-dimensional setting. Specifically, this procedure does not detect any effective segmentation, thus resulting in elevated prediction errors. By comparison, FTSegV exhibits superior predictive ability over SegV and UniV. In particular, for large n, e.g. $n = 400$, FTSegV does a reasonably good job in recovering the group structure of $\mathbf{Z}_t(\cdot)$ and performs comparably well to the oracle method with the relative prediction errors lower than 1.149 in all scenarios. Second, comparing the results for $n = 200$ among different R, we observe an interesting phenomenon that even though the relative frequencies of effective segmentation for FunT drop as R increases, implying that some groups in $\{\widehat{\mathbf{Z}}_t^{(l)}(\cdot) : l \in [\hat{q}]\}$ are split incorrectly before forecasting, the prediction errors stay low and slightly decrease as shown in Table 3. This is not surprising, since further segmentation based on FunT yields fewer parameters to be estimated in VAR models and thus benefits the forecasting accuracy even if a few small but significant cross-covariances of $\mathbf{Z}_t(\cdot)$ are ignored. Such finding highlights the success of FTSegV and its R-round segmentation in the sense that although FTSegV may not be able to accurately recover the group structure in $\mathbf{Z}_t(\cdot)$ for a small n, it achieves an appropriate dimension reduction to provide significant improvement in high-dimensional functional prediction.

Table 2: The relative frequencies of effective segmentation over 500 simulation runs.

(p, δ)	NonT		FunT					
			$R = 1$		$R = 5$		$R = 10$	
	$n = 200$	$n = 400$	$n = 200$	$n = 400$	$n = 200$	$n = 400$	$n = 200$	$n = 400$
(30, 0.1)	0	0	0.706	1.000	0.556	1.000	0.546	1.000
(30, 0.5)	0	0	0.588	1.000	0.436	1.000	0.420	1.000
(60, 0.1)	0	0	0.298	1.000	0.148	1.000	0.144	1.000
(60, 0.5)	0	0	0.194	0.996	0.078	0.990	0.072	0.990

Table 3: Means (standard deviations) of relative MSPEs over 500 simulation runs.

Method	(p, δ)	n =200	n =400	(p, δ)	n =200	n =400
FTSegV ($R = 1$)	(30, 0.1)	1.243(0.162)	1.095(0.105)	(60, 0.1)	1.249(0.122)	1.110(0.073)
FTSegV ($R = 5$)		1.225(0.153)	1.091(0.101)		1.250(0.123)	1.104(0.071)
FTSegV ($R = 10$)		1.222(0.151)	1.087(0.099)		1.249(0.122)	1.099(0.071)
SegV		1.814(0.376)	1.901(0.368)		1.813(0.271)	1.907(0.265)
UniV		1.631(0.313)	1.735(0.317)		1.599(0.214)	1.682(0.210)
FTSegV ($R = 1$)	(30, 0.5)	1.268(0.176)	1.134(0.134)	(60, 0.5)	1.285(0.134)	1.149(0.101)
FTSegV ($R = 5$)		1.255(0.171)	1.128(0.130)		1.282(0.136)	1.142(0.098)
FTSegV ($R = 10$)		1.250(0.168)	1.128(0.127)		1.281(0.136)	1.141(0.099)
SegV		1.815(0.377)	1.903(0.369)		1.813(0.271)	1.905(0.264)
UniV		1.635(0.315)	1.740(0.317)		1.603(0.215)	1.684(0.209)

6 Real data analysis

In this section, we apply our proposed SegV and FTSegV to three real data examples arising from different fields. Our main goal is to evaluate the post-sample predictive accuracy of both methods. By comparison, we also implement componentwise univariate prediction method (UniV) and the multivariate prediction method of Gao et al. (2019) (denoted as GSY) to jointly predict p component series by fitting a factor model to estimated scores obtained via eigenanalysis of the long-run covariance function (Hörmann et al., 2015). To evaluate the effectiveness of the segmentation transformation and its impact on prediction, we forge two other segmentation cases, namely **under**-segmentation and **over**-segmentation, for both SegV and FTSegV (denoted as Under.SegV, Over.SegV, Under.FTSegV and Over.FTSegV, respectively). Denote by $\{\widehat{G}_l : l \in [\hat{q}]\}$ the segmented groups of $\{\widehat{\mathbf{Z}}_t^{(l)}(\cdot) : l \in [\hat{q}]\}$ discovered in Step (i) of Algorithm 1 (seen also as correct-segmentation). The under-segmentation updates $\{\widehat{G}_l : l \in [\hat{q}]\}$ by merging two groups $\widehat{G}_{l_1}$ and $\widehat{G}_{l_1'}$ together before subsequent analysis, where $\arg\max_{(i,j):\, i\in\widehat{G}_l,\, j\in\widehat{G}_{l'},\, 1\leqslant l\neq l'\leqslant\hat{q}} \widehat{T}_{ij} \in \widehat{G}_{l_1} \times \widehat{G}_{l_1'}$ with $\widehat{T}_{ij}$ defined in (12). The over-segmentation, on the other hand, regards each curve component of $\{\widehat{\mathbf{Z}}_t^{(l)}(\cdot) : l \in [\hat{q}]\}$ as an individual group and then applies UniV componentwisely. For a fair comparison, the orders of VAR models adopted in all SegV/FTSegV-related methods and UniV are determined by the AIC criterion without any fine-tuning being applied, while GSY is implemented using the R package `ftsa`.

To examine the predictive performance, we apply an expanding window approach to the observed data $\check{Y}_{tj}(v_i)$ for $t \in [n], j \in [p], i \in [N]$. We first split the dataset into a training set and a test set respectively consisting of the first n_1 and the remaining n_2 observations. For any positive integer h, we implement each comparison method on the training set $\{\check{Y}_{tj}(v_i) : t \in [n_1], j \in [p], i \in [N]\}$ and obtain its h-step ahead prediction, denoted as $\hat{Y}^{(h)}_{(n_1+h)j}(v_i)$, based on the fitted model. We then increase the training size by one, i.e. $\{\check{Y}_{tj}(v_i) : t \in [n_1 + 1], j \in [p], i \in [N]\}$, refit the model and compute the next h-step ahead prediction $\hat{Y}^{(h)}_{(n_1+1+h)j}(v_i)$ for $j \in [p], i \in [N]$. Repeat the above procedure until the last h-step ahead prediction $\hat{Y}^{(h)}_{nj}(v_i)$ is produced. Finally, we compute the h-step ahead mean absolute prediction error (MAPE) and mean squared prediction error (MSPE), respectively, defined as

$$
\begin{aligned}
\mathrm{MAPE}(h) &= \frac{1}{(n_2+1-h)pN} \sum_{t=n_1+h}^{n} \sum_{j=1}^{p} \sum_{i=1}^{N} |\hat{Y}^{(h)}_{tj}(v_i) - \check{Y}_{tj}(v_i)|\,, \\
\mathrm{MSPE}(h) &= \frac{1}{(n_2+1-h)pN} \sum_{t=n_1+h}^{n} \sum_{j=1}^{p} \sum_{i=1}^{N} \{\hat{Y}^{(h)}_{tj}(v_i) - \check{Y}_{tj}(v_i)\}^2\,.
\end{aligned}
\tag{37}
$$

6.1 UK annual temperature data

The first dataset, which is available at `https://www.metoffice.gov.uk/research/climate/maps-and-data/historic-station-data`, consists of monthly mean temperature collected at $p = 22$ measuring stations across Britain from 1959 to 2020 ($n = 62$). Let $\check{Y}_{tj}(v_i)$ ($t \in [62]$, $j \in [22]$, $i \in [12]$) be the mean temperature during month $v_i = i$ of year $1958 + t$ measured at the j-th station. The observed temperature curves are smoothed using a 10-dimensional Fourier basis that characterize the periodic pattern over the annual cycle. We divide the smoothed dataset into the training set of size $n_1 = 41$ and the test set of size $n_2 = 21$. Since the smoothed curve series exhibit very weak autocorrelations when lags are beyond 3 and the training size is relatively small, we use $k_0 = m = 3$ in our procedure for this example.

The values of MAPE and MSPE for $h \in \{1, 2, 3\}$ defined in (37) are summarized in Table 4. Several obvious patterns are observable. First, our proposed SegV and FTSegV perform similarly well and both provide the highest predictive accuracies among all comparison methods for all h. This demonstrates the effectiveness of reducing the number of parameters via the segmentation in predicting high-dimensional functional time series, while the latent transformation matrix may not be approximately sparse in practice. Second, although the cases of under- and over-segmentation are slightly inferior to the correct-segmentation case, they significantly outperform UniV and GSY in one- and two-step-ahead predictions. It is worth noting that the over segmentation ignores all the correlations among different components of transformed curves, whereas UniV neglects those of original curves. This observation reveals that the transformation can also improve the prediction efficiently.

6.2 Japanese mortality data

The second dataset, which can be downloaded from `https://www.ipss.go.jp/p-toukei/JMD/index-en.html`, contains age-specific and gender-specific mortality rates for $p = 47$ prefectures in Japan during 1975 to 2017 ($n = 43$). Following the recent proposal of Gao et al.

Table 4: MAFEs and MSFEs for eight competing methods on the UK temperature curves for $h \in \{1, 2, 3\}$.

Method	MAFE			MSFE		
	$h = 1$	$h = 2$	$h = 3$	$h = 1$	$h = 2$	$h = 3$
SegV	**0.786**	0.806	**0.827**	**1.073**	1.075	**1.155**
Under.SegV	0.805	0.826	0.883	1.152	1.135	1.266
Over.SegV	0.797	0.821	0.845	1.101	1.126	1.174
FTSegV	0.789	**0.806**	0.828	1.077	**1.073**	1.158
Under.FTSegV	0.791	0.820	0.872	1.105	1.112	1.250
Over.FTSegV	0.797	0.821	0.845	1.101	1.126	1.174
UniV	0.936	0.951	0.976	1.450	1.450	1.458
GSY	0.894	0.884	0.854	1.346	1.338	1.219

(2019), we model the log transformation of the mortality rate of people aged $v_i = i-1$ living in the j-th prefecture during year $1974+t$ as a random curve $\check{Y}_{tj}(v_i)$ ($t \in [43]$, $j \in [47]$, $i \in [96]$) and perform smoothing for observed mortality curves via smoothing splines. The post-sample prediction are carried out in an identical way to Section 6.1. We choose $k_0 = m = 3$ in our estimation procedure and treat the smoothed curves in the first $n_1 = 33$ years and the last $n_2 = 10$ years as the training sample and the test sample, respectively.

Table 5 reports the MAPEs and MSPEs for Japanese females and males. Again it is obvious that SegV and FTSegV provide the best predictive performance uniformly for both females and males, and all h. One may also notice that, compared with SegV and Under.SegV, Over.SegV does not perform well for males. In most cases, the transformed curve series for males admits $\hat{q} = 44$ groups with 43 groups of size 1 and one large group of size 4. However, Over.SegV fails to account for the cross serial dependence within the large group of 4, thus leading to less accurate predictions. On the other hand, the transformed curves for females reveal a common structure with one group of size 2 and the remaining groups of size 1. As expected, Over.SegV performs slightly better in this case. This finding again confirms the effectiveness of our procedure, in particular, the within group cross dependence information is also valuable in the post-sample prediction.

Table 5: MAFEs and MSFEs for eight competing methods on the Japanese female and male mortality curves for $h \in \{1, 2, 3\}$. All numbers are multiplied by 10.

Method		MAFE			MSFE				MAFE			MSFE		
		$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$		$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
SegV	Female	1.393	**1.414**	1.468	**0.482**	**0.470**	0.486	Male	**1.374**	1.461	1.543	0.436	0.453	0.481
Under.SegV		1.537	1.661	1.853	0.528	0.560	0.642		1.394	1.491	1.608	0.443	0.464	0.503
Over.SegV		1.427	1.610	1.814	0.482	0.520	0.588		1.506	1.678	1.897	0.468	0.514	0.603
FTSegV		**1.392**	1.417	**1.468**	0.484	0.471	**0.484**		1.376	**1.444**	**1.521**	**0.435**	**0.446**	**0.473**
Under.FTSegV		1.542	1.661	1.846	0.533	0.560	0.638		1.389	1.482	1.596	0.440	0.460	0.499
Over.FTSegV		1.433	1.617	1.816	0.484	0.523	0.588		1.512	1.673	1.894	0.470	0.512	0.604
UniV		1.602	1.858	2.136	0.523	0.618	0.737		1.568	1.855	2.167	0.485	0.595	0.743
GSY		1.618	1.691	1.682	0.678	0.733	0.706		1.550	1.581	1.576	0.669	0.663	0.628

6.3 Energy consumption data

Our third dataset contains energy consumption readings (in kWh) taken at half hourly intervals for thousands of London households, and is available at https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households. In our study, we select households with flat energy prices during the period between December 2012 and May 2013 ($n = 182$) after removing samples with too many missing records, and hence construct 4000 samples of daily energy consumption curves observed at $N = 48$ equally spaced time points. To alleviate the impact of randomness from individual curves, we randomly split the data into p groups of equal size, then take the sample average of curves within each group and finally smooth the averaged curves based on a 15-dimensional Fourier basis. We target to evaluate the h-day ahead predictive accuracy for the p-dimensional intraday energy consumption averaged curves in May 2013 based on the training data from December 2012 to the previous day. The eight comparison methods are built in the same manner as Section 6.1 with $k_0 = m = 5$.

Table 6 presents the mean prediction errors for $h \in \{1, 2, 3\}$ and $p \in \{40, 80\}$. A few trends are apparent. First, the prediction errors for $p = 80$ are higher than those for $p = 40$ as higher dimensionality poses more challenges in prediction. Second, likewise in previous examples, SegV and FTSegV attain the lowest prediction errors in comparison to five competing methods under all scenarios. All segmentation-based methods consistently outperform UniV and GSY by a large margin. Third, despite being developed for high-dimensional functional time series prediction, GSY provides the worst result in this example.

Table 6: MAFEs and MSFEs for eight competing methods on the energy consumption curves for $h \in \{1, 2, 3\}$ and $p \in \{40, 80\}$. All numbers are multiplied by 10^2.

Method		MAFE			MSFE				MAFE			MSFE		
		$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$		$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$
SegV	$p=40$	1.639	1.748	1.793	0.047	0.053	0.054	$p=80$	**1.996**	2.058	2.071	**0.070**	0.075	0.075
Under.SegV		1.669	1.766	1.794	0.048	0.054	0.054		2.025	2.092	2.104	0.072	0.077	0.077
Over.SegV		1.709	1.873	1.964	0.049	0.058	0.062		2.022	2.132	2.187	0.070	0.078	0.081
FTSegV		**1.637**	**1.747**	**1.791**	**0.047**	**0.053**	**0.054**		2.012	**2.055**	**2.070**	0.071	**0.074**	**0.074**
Under.FTSegV		1.669	1.766	1.793	0.048	0.054	0.054		2.040	2.087	2.104	0.073	0.076	0.077
Over.FTSegV		1.708	1.872	1.963	0.049	0.058	0.062		2.045	2.138	2.190	0.072	0.078	0.081
UniV		1.867	2.009	2.109	0.058	0.067	0.072		2.221	2.362	2.463	0.083	0.093	0.100
GSY		2.142	2.264	2.320	0.099	0.110	0.119		2.833	2.826	2.781	0.159	0.159	0.159

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.

Back, A. D. and Weigend, A. S. (1997). A first application of independent component analysis to extracting structure from stock returns, *International Journal of Neural Systems* **8**: 473–484.

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

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models, *The Annals of Statistics* **43**: 1535–1567.

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

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding, *The Annals of Statistics* **36**: 2577–2604.

Chang, J., Chen, C., Qiao, X. and Yao, Q. (2022). An autocovariance-based learning framework for high-dimensional functional time series, *arXiv:2008.12885v3* .

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

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.

Chiou, J.-M., Chen, Y.-T. and Yang, Y.-F. (2014). Multivariate functional principal component analysis: a normalization approach, *Statistica Sinica* **24**: 1571–1596.

Descary, M.-H. and Panaretos, V. M. (2019). Functional data analysis by matrix completion, *The Annals of Statistics* **47**: 1–38.

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. (2022). Finite sample theory for high-dimensional functional/scalar time series with applications, *Electronic Journal of Statistics* **16**: 527–591.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting, *Journal of the American Statistical Association* **100**: 830–840.

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.

Ghosh, S., Khare, K. and Michailidis, G. (2019). High-dimensional posterior consistency in Bayesian vector autoregressive models, *Journal of the American Statistical Association* **114**: 735–748.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, fourth edn, Johns Hopkins University Press, Baltimore, MD.

Guo, S. and Qiao, X. (2023). On consistency and sparsity for high-dimensional functional time series with application to autoregressions, *Bernoulli* **29**: 451–472.

Guo, S., Wang, Y. and Yao, Q. (2016). High-dimensional and banded vector autoregressions, *Biometrika* **103**: 889–903.

Han, F., Lu, H. and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions, *Journal of Machine Learning Research* **16**: 3115–3150.

Han, Y., Chen, R., Zhang, C.-H. and Yao, Q. (2022). Simultaneous decorrelation of matrix time series, *Journal of the American Statistical Association, in press* .

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains, *Journal of the American Statistical Association* **113**: 649–659.

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

Hu, X. and Yao, F. (2022). Sparse functional principal component analysis in high dimensions, *Statistica Sinica* **32**: 1939–1960.

Jiao, S., Aue, A. and Ombao, H. (2021). Functional time series prediction under partial observation of the future curve, *Journal of the American Statistical Association, in press* .

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

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.

Matteson, D. S. and Tsay, R. S. (2011). Dynamic orthogonal components for multivariate time series, *Journal of the American Statistical Association* **106**: 1450–1463.

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

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors, *Journal of Business & Economic Statistics* **30**: 481–493.

Tiao, G. C. and Tsay, R. S. (1989). Model specification in multivariate time series, *Journal of the Royal Statistical Society: Series B* **51**: 157–195.

Wilms, I., Basu, S., Bien, J. and Matteson, D. S. (2021). Sparse identification and estimation of large-scale vector autoregressive moving averages, *Journal of the American Statistical Association, in press* .

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association* **100**: 577–590.

Supplementary material to "On the modelling and prediction of high-dimensional functional time series"

Jinyuan Chang, Qin Fang, Xinghao Qiao and Qiwei Yao

This supplementary material contains all technical proofs supporting Section 4. We begin by introducing some notation. For $x, y \in \mathbb{R}$, we use $x \vee y = \max(x, y)$. For a vector $\mathbf{b} \in \mathbb{R}^p$, we denote its ℓ_2 norm by $\|\mathbf{b}\|_2 = (\sum_{j=1}^p |b_j|^2)^{1/2}$. For any $\boldsymbol{\mathcal{F}} = (\mathcal{F}_1, \ldots, \mathcal{F}_p)^\top$ and $\boldsymbol{\mathcal{G}} = (\mathcal{G}_1, \ldots, \mathcal{G}_p)^\top \in \mathbb{H}$, we define the inner product as $\langle \boldsymbol{\mathcal{F}}, \boldsymbol{\mathcal{G}} \rangle = \int_{\mathcal{U}} \boldsymbol{\mathcal{F}}(u)^\top \boldsymbol{\mathcal{G}}(u) \, \mathrm{d}u = \sum_{j=1}^p \int_{\mathcal{U}} \mathcal{F}_j(u) \mathcal{G}_j(u) \, \mathrm{d}u$ with the induced norm $\| \cdot \| = \langle \cdot, \cdot \rangle^{1/2}$, and denote by $\boldsymbol{\mathcal{F}} \otimes \boldsymbol{\mathcal{G}}^\top = (\mathcal{F}_i \otimes \mathcal{G}_j)_{i,j \in [p]}$. We further denote by $\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{L}}(\mathbb{H}, \mathbb{H})$ the space of continuous linear operators from $\mathbb{H}$ to $\mathbb{H}$. For $\boldsymbol{\mathcal{B}} = (\mathcal{B}_{ij})_{p \times p}$ with each $\mathcal{B}_{ij} \in \mathbb{S}$, we write $\|\boldsymbol{\mathcal{B}}\|_{\mathcal{S},\mathrm{F}} = (\sum_{i=1}^p \sum_{j=1}^p \|\mathcal{B}_{ij}\|_{\mathcal{S}}^2)^{1/2}$, $\|\boldsymbol{\mathcal{B}}\|_{\mathcal{S},1} = \max_{j \in [p]} \sum_{i=1}^p \|\mathcal{B}_{ij}\|_{\mathcal{S}}$, $\|\boldsymbol{\mathcal{B}}\|_{\mathcal{S},\infty} = \max_{i \in [p]} \sum_{j=1}^p \|\mathcal{B}_{ij}\|_{\mathcal{S}}$ and $\|\boldsymbol{\mathcal{B}}\|_{\mathcal{L}} = \sup_{\|\boldsymbol{\mathcal{F}}\| \leqslant 1, \boldsymbol{\mathcal{F}} \in \mathbb{H}} \|\boldsymbol{\mathcal{B}}(\boldsymbol{\mathcal{F}})\|$. We define the image space of $\boldsymbol{\mathcal{B}}$ as $\mathrm{Im}(\boldsymbol{\mathcal{B}}) = \{\boldsymbol{\mathcal{G}} \in \mathbb{H} : \boldsymbol{\mathcal{G}} = \boldsymbol{\mathcal{B}}(\boldsymbol{\mathcal{F}}), \boldsymbol{\mathcal{F}} \in \mathbb{H}\}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exist a positive constant c such that $a_n / b_n \leqslant c$ and write $a_n \asymp b_n$ if and only if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously. We further write $a_n \ll b_n$ or $b_n \gg a_n$ if $\limsup_{n \to \infty} a_n / b_n = 0$. Throughout, we use c, c_0 to denote generic positive finite constants that may be different in different uses.

A Auxiliary lemmas

To prove Theorems 1–4, we need the following inequalities, equality and auxiliary lemmas, the proofs of which are deferred to Section G.

Inequality 1. *Let $\boldsymbol{\mathcal{B}}_1 = (\mathcal{B}_{1,ij})_{p \times p}$ and $\boldsymbol{\mathcal{B}}_2 = (\mathcal{B}_{2,ij})_{p \times p}$ with $\mathcal{B}_{1,ij}, \mathcal{B}_{2,ij} \in \mathbb{S}$ for any $i, j \in [p]$. It holds that* (i) $\| \int\int \boldsymbol{\mathcal{B}}_1(u,v) \boldsymbol{\mathcal{B}}_2(u,v)^\top \, \mathrm{d}u\mathrm{d}v \|_2 \leqslant \|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},\infty}^{1/2} \|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},1}^{1/2} \|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\infty}^{1/2} \|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},1}^{1/2}$, (ii) $\| \int \boldsymbol{\mathcal{B}}_1(\cdot, w) \boldsymbol{\mathcal{B}}_2(\cdot, w)^\top \, \mathrm{d}w \|_{\mathcal{S},\mathrm{F}} \leqslant \|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},\mathrm{F}} \|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\mathrm{F}}$, *and* (iii) $\|\boldsymbol{\mathcal{B}}_1 + \boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\mathrm{F}} \leqslant \|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},\mathrm{F}} + \|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\mathrm{F}}$.

Inequality 2. *Let* $\mathcal{B} = (B_{ij})_{p\times p}$ *with each* $B_{ij} \in \mathbb{S}$, $\mathbf{b}_1 \in \mathbb{R}^p$ *and* $\mathbf{b}_2 \in \mathbb{R}^p$. *Then* $\|\mathbf{b}_1^\top \mathcal{B}\mathbf{b}_2\|_{\mathcal{S}} \leqslant \|\mathbf{b}_1\|_2\|\mathbf{b}_2\|_2\|\mathcal{B}\|_{\mathcal{S},\infty}^{1/2}\|\mathcal{B}\|_{\mathcal{S},1}^{1/2}$.

Inequality 3. *Let* $\mathcal{B} = (B_{ij})_{p\times p}$ *with each* $B_{ij} \in \mathbb{S}$ *and* $\mathcal{F} \in \mathbb{H}$. *Then* $\|\int \mathcal{B}(\cdot, v)\mathcal{F}(v)\,\mathrm{d}v\| \leqslant \|\mathcal{B}\|_{\mathcal{S},\mathrm{F}}\|\mathcal{F}\|$ *and* $\|\mathcal{B}\|_{\mathcal{L}} \leqslant \|\mathcal{B}\|_{\mathcal{S},\mathrm{F}}$.

Equality 1. *For any* $\mathcal{F}$ *and* $\mathcal{G} \in \mathbb{H}$, *it holds that* $\|\mathcal{F} \otimes \mathcal{G}^\top\|_{\mathcal{S},\mathrm{F}} = \|\mathcal{F}\|\|\mathcal{G}\|$.

Lemma A1. *Let* $\{\mathbf{Y}_t(\cdot)\}$ *satisfy Conditions 1 and 2. There exists some universal constant* $\tilde{c} > 0$ *such that* $\mathbb{P}\{\|\widehat{\Sigma}_{y,k,ij}^{\mathrm{S}} - \Sigma_{y,k,ij}\|_{\mathcal{S}} > \mathcal{M}_y\eta\} \leqslant 8\exp\{-\tilde{c}n\min(\eta^2, \eta)\}$ *for any* $\eta > 0$, $|k| \leqslant k_0 \vee m$ *and* $i, j \in [p]$.

Lemma A2. *Suppose Condition 3 holds. Then* $\max_{|k|\leqslant k_0\vee m}\sum_{i=1}^p \|\Sigma_{y,k,ij}\|_{\mathcal{S}}^{\alpha} = O(\Xi) = \max_{|k|\leqslant k_0\vee m}\sum_{j=1}^p \|\Sigma_{y,k,ij}\|_{\mathcal{S}}^{\alpha}$ *and* $\max_{|k|\leqslant k_0\vee m}\|\mathbf{\Sigma}_{y,k}\|_{\mathcal{S},1} = O(\Xi p_\dagger^{1-\alpha}) = \max_{|k|\leqslant k_0\vee m}\|\mathbf{\Sigma}_{y,k}\|_{\mathcal{S},\infty}$, *where* $\Xi = s_1s_2(2p_\dagger + 1)$ *with* $p_\dagger = \max_{l\in[q]} p_l$.

Lemma A3. *Let Conditions 1–3 hold. For each* $|k| \leqslant k_0 \vee m$, *select* $\omega_k = c_k\mathcal{M}_y(n^{-1}\log p)^{1/2}$ *for some sufficiently large constant* $c_k > 0$. *If* $\log p = o(n)$, *then* $\max_{|k|\leqslant k_0\vee m}\|\mathcal{T}_{\omega_k}(\widehat{\mathbf{\Sigma}}_{y,k}^{\mathrm{S}}) - \mathbf{\Sigma}_{y,k}\|_{\mathcal{S},1} = O_{\mathrm{p}}\{\Xi\mathcal{M}_y^{1-\alpha}(n^{-1}\log p)^{(1-\alpha)/2}\} = \max_{|k|\leqslant k_0\vee m}\|\mathcal{T}_{\omega_k}(\widehat{\mathbf{\Sigma}}_{y,k}^{\mathrm{S}}) - \mathbf{\Sigma}_{y,k}\|_{\mathcal{S},\infty}$, *where* $\Xi = s_1s_2(2p_\dagger + 1)$ *with* $p_\dagger = \max_{l\in[q]} p_l$. *Moreover, if* $p_\dagger^{-2}\mathcal{M}_y^2\log p = o(n)$ *is also satisfied, then* $\|\int\int\{\mathcal{T}_{\omega_k}(\widehat{\mathbf{\Sigma}}_{y,k}^{\mathrm{S}})(u,v)^{\otimes 2} - \mathbf{\Sigma}_{y,k}(u,v)^{\otimes 2}\}\,\mathrm{d}u\mathrm{d}v\|_2 = O_{\mathrm{p}}\{\Xi^2 p_\dagger^{1-\alpha}\mathcal{M}_y^{1-\alpha}(n^{-1}\log p)^{(1-\alpha)/2}\}$ *for each* $|k| \leqslant k_0 \vee m$.

Lemma A4 (Theorem 8.1.10 of Golub and Van Loan (1996))**.** *Suppose* $\mathbf{B}$ *and* $\mathbf{B} + \mathbf{E}$ *are* $m \times m$ *symmetric matrices and* $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$, *with* $\mathbf{Q}_1 \in \mathbb{R}^{m\times l}$ *and* $\mathbf{Q}_2 \in \mathbb{R}^{m\times(m-l)}$, *is an orthogonal matrix such that* $\mathcal{C}(\mathbf{Q}_1)$ *is an invariant subspace for* $\mathbf{B}$, *that is,* $\mathbf{B}\cdot\mathcal{C}(\mathbf{Q}_1) \subset \mathcal{C}(\mathbf{Q}_1)$. *Partition the matrices* $\mathbf{Q}^\top\mathbf{B}\mathbf{Q}$ *and* $\mathbf{Q}^\top\mathbf{E}\mathbf{Q}$ *as follows:*

$$\mathbf{Q}^\top\mathbf{B}\mathbf{Q} = \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{pmatrix} \quad \textit{and} \quad \mathbf{Q}^\top\mathbf{E}\mathbf{Q} = \begin{pmatrix} \mathbf{E}_{11} & \mathbf{E}_{21}^\top \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{pmatrix}.$$

If $\mathrm{sep}(\mathbf{D}_1, \mathbf{D}_2) = \min_{\mu_1\in\Lambda(\mathbf{D}_1),\mu_2\in\Lambda(\mathbf{D}_2)}|\mu_1 - \mu_2| > 0$, *where* $\Lambda(\mathbf{M})$ *denotes the set of eigenvalues of the matrix* $\mathbf{M}$, *and* $\|\mathbf{E}\|_2 \leqslant \mathrm{sep}(\mathbf{D}_1, \mathbf{D}_2)/5$, *then there exists a matrix* $\mathbf{P} \in \mathbb{R}^{(m-l)\times l}$ *with*

$\|\mathbf{P}\|_2 \leqslant 4\|\mathbf{E}_{21}\|_2/\mathrm{sep}(\mathbf{D}_1, \mathbf{D}_2)$ *such that the columns of* $\mathbf{Q}_1^\star = (\mathbf{Q}_1 + \mathbf{Q}_2\mathbf{P})(\mathbf{I} + \mathbf{P}^\top\mathbf{P})^{-1/2}$ *define an orthonormal basis for a subspace that is invariant for* $\mathbf{B} + \mathbf{E}$.

From Lemma A4, we have

$$\begin{aligned}\|\mathbf{Q}_1^\star - \mathbf{Q}_1\|_2 &= \|\{\mathbf{Q}_1 + \mathbf{Q}_2\mathbf{P} - \mathbf{Q}_1(\mathbf{I} + \mathbf{P}^\top\mathbf{P})^{1/2}\}(\mathbf{I} + \mathbf{P}^\top\mathbf{P})^{-1/2}\|_2 \\ &\leqslant \|\mathbf{Q}_1\{\mathbf{I} - (\mathbf{I} + \mathbf{P}^\top\mathbf{P})^{1/2}\}\|_2 + \|\mathbf{Q}_2\mathbf{P}\|_2 \\ &\leqslant 2\|\mathbf{P}\|_2 \leqslant \frac{8}{\mathrm{sep}(\mathbf{D}_1, \mathbf{D}_2)}\|\mathbf{E}_{21}\|_2 \leqslant \frac{8}{\mathrm{sep}(\mathbf{D}_1, \mathbf{D}_2)}\|\mathbf{E}\|_2\,.\end{aligned}$$

Lemma A5. *Let* $\{\theta_j, \boldsymbol{\phi}_j(\cdot)\}_{j\geqslant 1}$ *and* $\{\hat{\theta}_j, \widehat{\boldsymbol{\phi}}_j(\cdot)\}_{j\geqslant 1}$ *be the eigenvalue/eigenfunction pairs of* $\mathbf{Q}(\cdot,\cdot)$ *and* $\widehat{\mathbf{Q}}(\cdot,\cdot)$ *respectively, with the corresponding nonzero eigenvalues sorted in decreasing order. Then we have* (i) $\sup_{j\geqslant 1}|\hat{\theta}_j - \theta_j| \leqslant \|\widehat{\mathbf{Q}} - \mathbf{Q}\|_{\mathcal{S},\mathrm{F}}$, *and* (ii) $\sup_{j\geqslant 1}\Delta_j\|\widehat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j\| \leqslant 2\sqrt{2}\|\widehat{\mathbf{Q}} - \mathbf{Q}\|_{\mathcal{S},\mathrm{F}}$, *where* $\Delta_j = \min_{k\in[j]}(\theta_k - \theta_{k+1})$.

B Proof of Proposition 2

Let $\boldsymbol{\Omega}_k^{(l)} = (\Omega_{k,ij}^{(l)})_{r_l\times r_l}$. By the decomposition (18), we write

$$\mathbf{M}_k^{(l)} = \sum_{i=1}^{r_l}\sum_{j=1}^{r_l}\Omega_{k,ij}^{(l)}\boldsymbol{\varphi}_i^{(l)} \otimes \{\boldsymbol{\varphi}_j^{(l)}\}^\top. \tag{S.1}$$

Hence, $\mathrm{Im}(\mathbf{M}_k^{(l)}) \subset \mathrm{span}\{\boldsymbol{\varphi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\varphi}_{r_l}^{(l)}(\cdot)\}$. Define $\tilde{\lambda}_{k,i}^{(l)} = \|\sum_{j=1}^{r_l}\Omega_{k,ij}^{(l)}\boldsymbol{\varphi}_j^{(l)}\|$ and $\boldsymbol{\phi}_{k,i}^{(l)}(\cdot) = \sum_{j=1}^{r_l}\Omega_{k,ij}^{(l)}\boldsymbol{\varphi}_j^{(l)}(\cdot)/\|\sum_{j=1}^{r_l}\Omega_{k,ij}^{(l)}\boldsymbol{\varphi}_j^{(l)}\|$, we then rewrite (S.1) as

$$\mathbf{M}_k^{(l)} = \sum_{i=1}^{r_l}\tilde{\lambda}_{k,i}^{(l)}\boldsymbol{\varphi}_i^{(l)} \otimes \{\boldsymbol{\phi}_{k,i}^{(l)}\}^\top. \tag{S.2}$$

We next show that the set $\{\boldsymbol{\phi}_{k,1}^{(l)}(\cdot), \ldots, \boldsymbol{\phi}_{k,r_l}^{(l)}(\cdot)\}$ is linearly independent for some $k \in [k_0]$. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{r_l})^\top$ denote an arbitrary vector in $\mathbb{R}^{r_l}$ and $\boldsymbol{\varphi}^{(l)}(\cdot) = \{\boldsymbol{\varphi}_1^{(l)}(\cdot)^\top, \ldots, \boldsymbol{\varphi}_{r_l}^{(l)}(\cdot)^\top\}^\top$. Since the set $\{\boldsymbol{\varphi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\varphi}_{r_l}^{(l)}(\cdot)\}$ is linearly independent and $\boldsymbol{\Omega}_k^{(l)}$ is of full rank for some $k \in [k_0]$, the only solution of

$$\sum_{i=1}^{r_l}\beta_i\Big\|\sum_{j=1}^{r_l}\Omega_{k,ij}^{(l)}\boldsymbol{\varphi}_j^{(l)}\Big\|\boldsymbol{\phi}_{k,i}^{(l)}(\cdot) = \boldsymbol{\beta}^\top\boldsymbol{\Omega}_k^{(l)}\boldsymbol{\varphi}^{(l)}(\cdot) = 0$$

is $\boldsymbol{\beta} = 0$ for such k. Hence, the set $\{\boldsymbol{\phi}_{k,1}^{(l)}(\cdot), \ldots, \boldsymbol{\phi}_{k,r_l}^{(l)}(\cdot)\}$ in (S.2) is linearly independent for some $k \in [k_0]$. Together with the decomposition (S.2) and the fact that any linearly independent set of r_l elements in a r_l-dimensional space forms a basis for that space, it implies that $\mathrm{Im}(\mathbf{M}_k^{(l)}) = \mathrm{span}\{\boldsymbol{\varphi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\varphi}_{r_l}^{(l)}(\cdot)\}$ for some $k \in [k_0]$.

By the definition of the image space, we further have

$$
\begin{aligned}
&\mathrm{Im}\Big\{ \int_{\mathcal{U}} \mathbf{M}_k^{(l)}(\cdot, w)\mathbf{M}_k^{(l)}(\cdot, w)^\top \,\mathrm{d}w \Big\} \\
&\quad = \Big\{ \mathfrak{G} \in \mathbb{H} : \mathfrak{G} = \int_{\mathcal{U}}\int_{\mathcal{U}} \mathbf{M}_k^{(l)}(u, w)\mathbf{M}_k^{(l)}(v, w)^\top \,\mathrm{d}w \boldsymbol{\mathfrak{F}}(v)\,\mathrm{d}v,\ \boldsymbol{\mathfrak{F}} \in \mathbb{H} \Big\} \\
&\quad = \Big\{ \mathfrak{G} \in \mathbb{H} : \mathfrak{G} = \int_{\mathcal{U}} \mathbf{M}_k^{(l)}(u, w) \int_{\mathcal{U}} \mathbf{M}_k^{(l)}(v, w)^\top \boldsymbol{\mathfrak{F}}(v)\,\mathrm{d}v\,\mathrm{d}w,\ \boldsymbol{\mathfrak{F}} \in \mathbb{H} \Big\} \\
&\quad = \Big\{ \mathfrak{G} \in \mathbb{H} : \mathfrak{G} = \int_{\mathcal{U}} \mathbf{M}_k^{(l)}(u, w)\widetilde{\boldsymbol{\mathfrak{F}}}(w)\,\mathrm{d}w,\ \widetilde{\boldsymbol{\mathfrak{F}}} \in \mathbb{H} \Big\} = \mathrm{Im}\{\mathbf{M}_k^{(l)}\}.
\end{aligned}
$$

Due to the nonnegativity of $\mathbf{K}^{(l)}(\cdot,\cdot)$, we have that $\int_{\mathcal{U}} \mathbf{K}^{(l)}(u, v)\boldsymbol{\vartheta}(v)\mathrm{d}v = 0$ if and only if $\int_{\mathcal{U}}\int_{\mathcal{U}} \mathbf{M}_k^{(l)}(u, w)\mathbf{M}_k^{(l)}(v, w)^\top \,\mathrm{d}w\boldsymbol{\vartheta}(v)\mathrm{d}v = 0$ for all $k \in [k_0]$. This further leads to $\mathrm{Im}(\mathbf{K}^{(l)}) = \bigcup_{k\in[k_0]} \mathrm{Im}\{\int_{\mathcal{U}} \mathbf{M}_k^{(l)}(\cdot, w)\mathbf{M}_k^{(l)}(\cdot, w)^\top \,\mathrm{d}w\} = \mathrm{span}\{\boldsymbol{\varphi}_1^{(l)}(\cdot), \ldots, \boldsymbol{\varphi}_{r_l}^{(l)}(\cdot)\}$. Hence, we complete the proof of part (ii). Furthermore, since $\dim[\mathrm{Im}\{\mathbf{K}^{(l)}\}] = r_l$, part (i) follows. □

C Proof of Theorem 1

Let $\nu_n = \Xi^2 p_\dagger^{1-\alpha}\mathcal{M}_y^{1-\alpha}(n^{-1}\log p)^{(1-\alpha)/2}$, where $\Xi = s_1 s_2(2p_\dagger + 1)$ with $p_\dagger = \max_{l\in[q]} p_l$. Recall $\mathbf{W}_y$ in (6) and $\widehat{\mathbf{W}}_y$ in (17). Since $\rho^{-1}\nu_n \to 0$ implies that $p_\dagger^{-2}\mathcal{M}_y^2 \log p = o(n)$, it follows from Lemma A3 and fixed k_0 that

$$
\|\widehat{\mathbf{W}}_y - \mathbf{W}_y\|_2 \leqslant \sum_{k=0}^{k_0} \Big\| \int\int \big\{ \mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}_{y,k}^{\mathrm{S}})(u, v)^{\otimes 2} - \boldsymbol{\Sigma}_{y,k}(u, v)^{\otimes 2} \big\} \,\mathrm{d}u\mathrm{d}v \Big\|_2 = O_{\mathrm{p}}(\nu_n)\,. \tag{S.3}
$$

Recall that $\widehat{\boldsymbol{\Gamma}}_y\mathbf{R} = (\widehat{\boldsymbol{\Pi}}_1, \ldots, \widehat{\boldsymbol{\Pi}}_q) = (\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_p)$ and $\boldsymbol{\Gamma}_y\mathbf{H} = (\mathbf{A}_1\boldsymbol{\Gamma}_{z,1}\mathbf{H}_1, \ldots, \mathbf{A}_q\boldsymbol{\Gamma}_{z,q}\mathbf{H}_q) = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$. By Lemma A4, for each $l \in [q]$, we have that

$$
\|\widehat{\boldsymbol{\Pi}}_l - \mathbf{A}_l\boldsymbol{\Gamma}_{z,l}\mathbf{H}_l\|_2 \leqslant 8\rho^{-1}\|\widehat{\mathbf{W}}_y - \mathbf{W}_y\|_2\,. \tag{S.4}
$$

Combining (S.3) and (S.4), it is immediate to see that

$$\max_{j\in[p]}\|\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\|_2=O_{\mathrm{p}}(\rho^{-1}\nu_n)\,. \tag{S.5}$$

Recall that $\widehat{\boldsymbol{\gamma}}_i=\widehat{\boldsymbol{\eta}}_{\pi(i)}$, $\widehat{T}_{ij}=\max_{|k|\leqslant m}\|\widehat{\boldsymbol{\eta}}_i^{\top}\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}_{y,k}^{\mathrm{S}})\widehat{\boldsymbol{\eta}}_j\|_{\mathcal{S}}$ and $T_{ij}=\max_{|k|\leqslant m}\|\boldsymbol{\gamma}_i^{\top}\boldsymbol{\Sigma}_{y,k}\boldsymbol{\gamma}_j\|_{\mathcal{S}}$. Notice that

$$\widehat{\boldsymbol{\eta}}_{\pi(i)}^{\top}\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}_{y,k}^{\mathrm{S}})\widehat{\boldsymbol{\eta}}_{\pi(j)}-\boldsymbol{\gamma}_i^{\top}\boldsymbol{\Sigma}_{y,k}\boldsymbol{\gamma}_j=I_1+I_2+I_3+I_4+I_5\,, \tag{S.6}$$

where $I_1=(\widehat{\boldsymbol{\gamma}}_i-\boldsymbol{\gamma}_i)^{\top}\{\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}_{y,k}^{\mathrm{S}})-\boldsymbol{\Sigma}_{y,k}\}\widehat{\boldsymbol{\gamma}}_j$, $I_2=(\widehat{\boldsymbol{\gamma}}_i-\boldsymbol{\gamma}_i)^{\top}\boldsymbol{\Sigma}_{y,k}(\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j)$, $I_3=(\widehat{\boldsymbol{\gamma}}_i-\boldsymbol{\gamma}_i)^{\top}\boldsymbol{\Sigma}_{y,k}\boldsymbol{\gamma}_j$, $I_4=\boldsymbol{\gamma}_i^{\top}\{\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}_{y,k}^{\mathrm{S}})-\boldsymbol{\Sigma}_{y,k}\}\widehat{\boldsymbol{\gamma}}_j$ and $I_5=\boldsymbol{\gamma}_i^{\top}\boldsymbol{\Sigma}_{y,k}(\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j)$. Let $\omega_n=\Xi\mathcal{M}_y^{1-\alpha}(n^{-1}\log p)^{(1-\alpha)/2}$. Hence $\omega_n\to 0$ as implied by $\rho^{-1}\nu_n\to 0$. By (S.5), the orthonormality of $\widehat{\boldsymbol{\gamma}}_j$ and $\boldsymbol{\gamma}_j$, Lemma A3 and Inequality 2, we obtain that

$$\max_{i,j\in[p]}\|I_1\|_{\mathcal{S}}=O_{\mathrm{p}}(\rho^{-1}\nu_n\omega_n)\,,\quad \max_{i,j\in[p]}\|I_2\|_{\mathcal{S}}=O_{\mathrm{p}}(\rho^{-2}\Xi p_{\dagger}^{1-\alpha}\nu_n^2)\,,$$
$$\max_{i,j\in[p]}\|I_4\|_{\mathcal{S}}=O_{\mathrm{p}}(\omega_n)\,,\quad \max_{i,j\in[p]}\big(\|I_3\|_{\mathcal{S}}+\|I_5\|_{\mathcal{S}}\big)=O_{\mathrm{p}}(\rho^{-1}\Xi p_{\dagger}^{1-\alpha}\nu_n)\,.$$

Together with $\rho^{-1}\nu_n\to 0$, $\omega_n\to 0$ and $\omega_n=o(\nu_n)$, it holds that

$$\begin{aligned}\max_{i,j\in[p]}|\widehat{T}_{\pi(i)\pi(j)}-T_{ij}|&\leqslant\max_{i,j\in[p],|k|\leqslant m}\big|\|\widehat{\boldsymbol{\eta}}_{\pi(i)}^{\top}\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}_{y,k}^{\mathrm{S}})\widehat{\boldsymbol{\eta}}_{\pi(j)}\|_{\mathcal{S}}-\|\boldsymbol{\gamma}_i^{\top}\boldsymbol{\Sigma}_{y,k}\boldsymbol{\gamma}_j\|_{\mathcal{S}}\big| \\ &\leqslant\max_{i,j\in[p],|k|\leqslant m}\|\widehat{\boldsymbol{\eta}}_{\pi(i)}^{\top}\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}_{y,k}^{\mathrm{S}})\widehat{\boldsymbol{\eta}}_{\pi(j)}-\boldsymbol{\gamma}_i^{\top}\boldsymbol{\Sigma}_{y,k}\boldsymbol{\gamma}_j\|_{\mathcal{S}}=O_{\mathrm{p}}(\rho^{-1}\Xi p_{\dagger}^{1-\alpha}\nu_n)\,.\end{aligned} \tag{S.7}$$

We now show that $\hat{\varrho}$ in (13) is a consistent estimate for ϱ. For $k\in[\aleph]$, without loss of generality, we write $T_{(k)}=T_{i_kj_k}$ with $i_k,j_k\in[p]$. Since $\rho^{-1}\Xi p_{\dagger}^{1-\alpha}\nu_n/\delta_n\to 0$ and $\delta_n/\varsigma\to 0$, we can find some h_n such that $\rho^{-1}\Xi p_{\dagger}^{1-\alpha}\nu_n\ll h_n\ll\delta_n\ll\varsigma$. Let $\widetilde{\Omega}=\{\max_{i,j\in[p]}|\widehat{T}_{\pi(i)\pi(j)}-T_{ij}|\leqslant h_n\leqslant\varsigma/2\}$. It is immediate to see that under the event $\widetilde{\Omega}$ we have $\varsigma/2\leqslant\widehat{T}_{\pi(i_k)\pi(j_k)}\leqslant\varsigma/2+T_{(1)}$ for $k\in[\varrho]$ and $0\leqslant\widehat{T}_{\pi(i_k)\pi(j_k)}\leqslant h_n$ for $k>\varrho$. Due to the definition of $\hat{\varrho}$, we have that

$$\frac{\widehat{T}_{(\hat{\varrho})}+\delta_n}{\widehat{T}_{(\hat{\varrho}+1)}+\delta_n}\geqslant\frac{\widehat{T}_{(\varrho)}+\delta_n}{\widehat{T}_{(\varrho+1)}+\delta_n}\geqslant\frac{\varsigma/2+\delta_n}{h_n+\delta_n}\asymp\frac{\varsigma}{\delta_n} \tag{S.8}$$

under $\widetilde{\Omega}$. If $\hat{\varrho}<\varrho$, under $\widetilde{\Omega}$,

$$\frac{\widehat{T}_{(\hat{\varrho})}+\delta_n}{\widehat{T}_{(\hat{\varrho}+1)}+\delta_n}\leqslant\frac{\varsigma/2+T_{(1)}+\delta_n}{\varsigma/2+\delta_n}\asymp\frac{T_{(1)}}{\varsigma}\,. \tag{S.9}$$

Since $\delta_n T_{(1)}/\varsigma^2 \to 0$, (S.8) and (S.9) imply $\mathbb{P}(\hat{\varrho} < \varrho \,|\, \widetilde{\Omega}) \to 0$. Similarly, if $\hat{\varrho} > \varrho$, under $\widetilde{\Omega}$,

$$\frac{\widehat{T}_{(\hat{\varrho})} + \delta_n}{\widehat{T}_{(\hat{\varrho}+1)} + \delta_n} \leqslant \frac{h_n + \delta_n}{0 + \delta_n} \to 1\,.$$

This together with (S.8) and $\delta_n/\varsigma \to 0$ yields that $\mathbb{P}(\hat{\varrho} > \varrho \,|\, \widetilde{\Omega}) \to 0$. Hence, $\mathbb{P}(\hat{\varrho} = \varrho \,|\, \widetilde{\Omega}) \to 1$. By (S.7), $\mathbb{P}(\widetilde{\Omega}) \to 1$. Combining the above results, we have $\mathbb{P}(\hat{\varrho} = \varrho) \to 1$. Recall $E = \{(i,j) : T_{ij} \geqslant T_{(\varrho)}, 1 \leqslant i < j \leqslant p\}$ and $\widetilde{E} = \{(i,j) : \widehat{T}_{ij} \geqslant \widehat{T}_{(\hat{\varrho})}, 1 \leqslant i < j \leqslant p\}$. Under the event $\{\hat{\varrho} = \varrho\}$, the permutation $\pi : [p] \to [p]$ actually provides a bijective mapping from the graph $([p], E)$ to $([p], \widetilde{E})$ in the sense that $\{k, (i,j)\} \in [p] \times E \to \{\pi(k), (\pi(i), \pi(j))\} \in [p] \times \widetilde{E}$. Hence we complete the proof of Theorem 1. $\square$

D Proof of Theorem 2

Let $\nu_n = \Xi^2 p_\dagger^{1-\alpha} \mathcal{M}_y^{1-\alpha} (n^{-1} \log p)^{(1-\alpha)/2}$, where $\Xi = s_1 s_2 (2p_\dagger + 1)$ with $p_\dagger = \max_{l \in [q]} p_l$. Since $\rho^{-1}\nu_n \to 0$, the result in (S.3) holds. This, together with (S.4) and the remark for Lemma 1 of Chang et al. (2018), yields that

$$\max_{l \in [q]} D\{\mathcal{C}(\widehat{\mathbf{\Pi}}_l), \mathcal{C}(\mathbf{A}_l)\} \lesssim \rho^{-1} \|\widehat{\mathbf{W}}_y - \mathbf{W}_y\|_2 = O_{\mathrm{p}}(\rho^{-1}\nu_n)\,. \tag{S.10}$$

Recall that $\widehat{\mathbf{\Pi}}_l = (\widehat{\boldsymbol{\gamma}}_i)_{i \in G_l}$ and $\widehat{\mathbf{A}}_l = (\widehat{\boldsymbol{\gamma}}_i)_{i \in \widehat{G}_l}$. Theorem 1 implies that there exists a permutation $\tilde{\pi} : [q] \to [q]$ such that $\mathbb{P}[\bigcap_{l=1}^q \{\widehat{G}_{\tilde{\pi}(l)} = G_l\}\,,\, \hat{q} = q] \to 1$. Let $\Omega_l = \{\widehat{G}_{\tilde{\pi}(l)} = G_l\,,\, \hat{q} = q\}$ and $d_n = \rho^{-1}\nu_n$. For any $\epsilon > 0$, by (S.10), there exists a constant $C > 0$ such that

$$\mathbb{P}\bigg[\max_{l \in [q]} d_n^{-1} D\{\mathcal{C}(\widehat{\mathbf{\Pi}}_l), \mathcal{C}(\mathbf{A}_l)\} > C\bigg] < \epsilon\,,$$

which implies

$$\begin{aligned}
&\mathbb{P}\bigg[\max_{l \in [q]} \min_{j \in [\hat{q}]} d_n^{-1} D\{\mathcal{C}(\widehat{\mathbf{A}}_j), \mathcal{C}(\mathbf{A}_l)\} > C\bigg] \\
&\quad \leqslant \mathbb{P}\bigg[\max_{l \in [q]} \min_{j \in [\hat{q}]} d_n^{-1} D\{\mathcal{C}(\widehat{\mathbf{A}}_j), \mathcal{C}(\mathbf{A}_l)\} > C, \bigcap_{l=1}^q \Omega_l\bigg] + \mathbb{P}\bigg(\bigcup_{l=1}^q \Omega_l^c\bigg) \\
&\quad \leqslant \mathbb{P}\bigg[\max_{l \in [q]} d_n^{-1} D\{\mathcal{C}(\widehat{\mathbf{\Pi}}_l), \mathcal{C}(\mathbf{A}_l)\} > C\bigg] + o(1) < \epsilon + o(1)\,.
\end{aligned}$$

Hence, $\max_{l \in [q]} \min_{j \in [\hat{q}]} D\{\mathcal{C}(\widehat{\mathbf{A}}_j), \mathcal{C}(\mathbf{A}_l)\} = O_{\mathrm{p}}(d_n)$. We complete the proof of Theorem 2. $\square$

E Proof of Theorem 3

Recall that $\mathbf{\Pi}_l = \mathbf{A}_l\mathbf{\Gamma}_{z,l}\mathbf{H}_l = (\boldsymbol{\gamma}_i)_{i\in G_l}$ and $\widehat{\mathbf{A}}_l = (\widehat{\boldsymbol{\gamma}}_i)_{i\in G_l}$. Write $\mathbf{M}_k^{(l)}(u,v) = \mathbf{\Pi}_l^\top\mathbf{\Sigma}_{y,k}(u,v)\mathbf{\Pi}_l \equiv \{M_{k,ij}^{(l)}(u,v)\}_{i,j\in[p_l]}$ and $\widehat{\mathbf{M}}_k^{(l)}(u,v) = \widehat{\mathbf{A}}_l^\top\mathcal{T}_{\omega_k}(\widehat{\mathbf{\Sigma}}_{y,k}^{\mathrm{s}})(u,v)\widehat{\mathbf{A}}_l \equiv \{\widehat{M}_{k,ij}^{(l)}(u,v)\}_{i,j\in[p_l]}$. Let $\nu_n = \Xi^2 p_\dagger^{1-\alpha}\mathcal{M}_y^{1-\alpha}(n^{-1}\log p)^{(1-\alpha)/2}$, where $\Xi = s_1s_2(2p_\dagger+1)$ with $p_\dagger = \max_{l\in[q]}p_l$. By a similar decomposition to (S.6) and $\rho^{-1}\Xi p_\dagger^{1-\alpha}\nu_n\to 0$, we obtain that $\max_{i,j\in[p_l],l\in[q]}\|\widehat{M}_{k,ij}^{(l)} - M_{k,ij}^{(l)}\|_{\mathcal{S}} = O_{\mathrm{p}}(\rho^{-1}\Xi p_\dagger^{1-\alpha}\nu_n)$. Hence,

$$\max_{l\in[q]}\|\widehat{\mathbf{M}}_k^{(l)} - \mathbf{M}_k^{(l)}\|_{\mathcal{S},\mathrm{F}} = \max_{l\in[q]}\bigg(\sum_{i,j\in[p_l]}\|\widehat{M}_{k,ij}^{(l)} - M_{k,ij}^{(l)}\|_{\mathcal{S}}^2\bigg)^{1/2} = O_{\mathrm{p}}(\rho^{-1}\Xi p_\dagger^{2-\alpha}\nu_n)\,.$$

Write $\mathbf{Z}_t^{(l)}(\cdot) = \{Z_{t,1}^{(l)}(\cdot),\ldots,Z_{t,p_l}^{(l)}(\cdot)\}^\top$. It follows from Cauchy–Schwartz inequality that

$$\begin{aligned}\max_{l\in[q]}\|\mathbf{M}_k^{(l)}\|_{\mathcal{S},\mathrm{F}}^2 &= \max_{l\in[q]}\sum_{i,j=1}^{p_l}\int\int\{M_{k,ij}^{(l)}(u,v)\}^2\,\mathrm{d}u\mathrm{d}v\\ &\leqslant \max_{l\in[q]}\sum_{i=1}^{p_l}\int\mathbb{E}[\{Z_{t,i}^{(l)}(u)\}^2]\,\mathrm{d}u\cdot\max_{l\in[q]}\sum_{j=1}^{p_l}\int\mathbb{E}[\{Z_{t+k,j}^{(l)}(u)\}^2]\,\mathrm{d}u = O(p_\dagger^2)\,.\end{aligned}$$

Observe that $\widehat{\mathbf{K}}^{(l)}(u,v) - \mathbf{K}^{(l)}(u,v) = \sum_{k=1}^{k_0}\int_{\mathcal{U}}\mathbf{M}_k^{(l)}(u,w)\{\widehat{\mathbf{M}}_k^{(l)}(v,w) - \mathbf{M}_k^{(l)}(v,w)\}^\top\,\mathrm{d}w + \sum_{k=1}^{k_0}\int_{\mathcal{U}}\{\widehat{\mathbf{M}}_k^{(l)}(u,w) - \mathbf{M}_k^{(l)}(u,w)\}\mathbf{M}_k^{(l)}(v,w)^\top\,\mathrm{d}w + \sum_{k=1}^{k_0}\int_{\mathcal{U}}\{\widehat{\mathbf{M}}_k^{(l)}(u,w) - \mathbf{M}_k^{(l)}(u,w)\}\{\widehat{\mathbf{M}}_k^{(l)}(v,w) - \mathbf{M}_k^{(l)}(v,w)\}^\top\,\mathrm{d}w$. Together with Inequality 1, $\rho^{-1}\Xi p_\dagger^{1-\alpha}\nu_n\to 0$ and fixed k_0, it holds that

$$\begin{aligned}\max_{l\in[q]}\|\widehat{\mathbf{K}}^{(l)} - \mathbf{K}^{(l)}\|_{\mathcal{S},\mathrm{F}} &\leqslant \max_{l\in[q]}\sum_{k=1}^{k_0}\|\widehat{\mathbf{M}}_k^{(l)} - \mathbf{M}_k^{(l)}\|_{\mathcal{S},\mathrm{F}}^2 + 2\max_{l\in[q]}\sum_{k=1}^{k_0}\|\mathbf{M}_k^{(l)}\|_{\mathcal{S},\mathrm{F}}\|\widehat{\mathbf{M}}_k^{(l)} - \mathbf{M}_k^{(l)}\|_{\mathcal{S},\mathrm{F}}\\ &= O_{\mathrm{p}}(\rho^{-2}\Xi^2p_\dagger^{4-2\alpha}\nu_n^2) + O_{\mathrm{p}}(\rho^{-1}\Xi p_\dagger^{3-\alpha}\nu_n) = O_{\mathrm{p}}(\rho^{-1}\Xi p_\dagger^{3-\alpha}\nu_n)\,.\end{aligned}$$

This, together with Lemma A5, implies that

$$\max_{l\in[q],j\in[r_l]}|\hat{\theta}_j^{(l)} - \theta_j^{(l)}| = O_{\mathrm{p}}(\rho^{-1}\Xi p_\dagger^{3-\alpha}\nu_n)\,,\quad \max_{l\in[q],j\in[r_l]}\|\widehat{\boldsymbol{\psi}}_j^{(l)} - \boldsymbol{\psi}_j^{(l)}\| = O_{\mathrm{p}}(\Delta^{-1}\rho^{-1}\Xi p_\dagger^{3-\alpha}\nu_n)\,,\tag{S.11}$$

where $\Delta = \min_{l\in[q],j\in[r_l]}\{\theta_j^{(l)} - \theta_{j+1}^{(l)}\}$.

Recall that

$$\hat{r}_l = \arg\max_{j\in[n-k_0]}\frac{\hat{\theta}_j^{(l)} + \tilde{\delta}_n}{\hat{\theta}_{j+1}^{(l)} + \tilde{\delta}_n}\,.$$

Note that the condition $\tilde{\delta}_n \max_{l\in[q]} \theta_1^{(l)}/\min_{l\in[q]}\{\theta_{r_l}^{(l)}\}^2 \to 0$ implies that $\tilde{\delta}_n = o(\min_{l\in[q]} \theta_{r_l}^{(l)})$. By $\rho^{-1}\Xi p_\dagger^{3-\alpha}\nu_n/\tilde{\delta}_n \to 0$ and $\tilde{\delta}_n/\min_{l\in[q]} \theta_{r_l}^{(l)} \to 0$, we can find some $\tilde{h}_n$ such that $\rho^{-1}\Xi p_\dagger^{3-\alpha}\nu_n \ll \tilde{h}_n \ll \tilde{\delta}_n \ll \min_{l\in[q]} \theta_{r_l}^{(l)}$. Let $\breve{\Omega} = \{\max_{l\in[q],j\in[r_l]} |\hat{\theta}_j^{(l)} - \theta_j^{(l)}| \leqslant \tilde{h}_n \leqslant \min_{l\in[q]} \theta_{r_l}^{(l)}/2\}$. Under the event $\breve{\Omega}$, we thus have $\min_{l\in[q]} \theta_{r_l}^{(l)}/2 \leqslant \hat{\theta}_j^{(l)} \leqslant \min_{l\in[q]} \theta_{r_l}^{(l)}/2 + \max_{l\in[q]} \theta_1^{(l)}$ if $j \in [r_l]$ and $0 \leqslant \hat{\theta}_j^{(l)} \leqslant \tilde{h}_n$ if $j > r_l$, for each $l \in [q]$. Due to the definition of $\hat{r}_l$, for each $l \in [q]$, we have that

$$\frac{\hat{\theta}_{\hat{r}_l}^{(l)} + \tilde{\delta}_n}{\hat{\theta}_{\hat{r}_l+1}^{(l)} + \tilde{\delta}_n} \geqslant \frac{\hat{\theta}_{r_l}^{(l)} + \tilde{\delta}_n}{\hat{\theta}_{r_l+1}^{(l)} + \tilde{\delta}_n} \geqslant \frac{\min_{l\in[q]} \theta_{r_l}^{(l)}/2 + \tilde{\delta}_n}{\tilde{h}_n + \tilde{\delta}_n} \asymp \frac{\min_{l\in[q]} \theta_{r_l}^{(l)}}{\tilde{\delta}_n} \tag{S.12}$$

under $\breve{\Omega}$. For each $l \in [q]$, if $\hat{r}_l < r_l$, under $\breve{\Omega}$,

$$\frac{\hat{\theta}_{\hat{r}_l}^{(l)} + \tilde{\delta}_n}{\hat{\theta}_{\hat{r}_l+1}^{(l)} + \tilde{\delta}_n} \leqslant \frac{\min_{l\in[q]} \theta_{r_l}^{(l)}/2 + \max_{l\in[q]} \theta_1^{(l)} + \tilde{\delta}_n}{\min_{l\in[q]} \theta_{r_l}^{(l)}/2 + \tilde{\delta}_n} \asymp \frac{\max_{l\in[q]} \theta_1^{(l)}}{\min_{l\in[q]} \theta_{r_l}^{(l)}}\,. \tag{S.13}$$

Since $\tilde{\delta}_n \max_{l\in[q]} \theta_1^{(l)}/\min_{l\in[q]}\{\theta_{r_l}^{(l)}\}^2 \to 0$, (S.12) and (S.13) imply $\mathbb{P}[\bigcup_{i=1}^q\{\hat{r}_l < r_l\} \,|\, \breve{\Omega}] \to 0$. Similarly, for each $l \in [q]$, if $\hat{r}_l > r_l$, under $\breve{\Omega}$,

$$\frac{\hat{\theta}_{\hat{r}_l}^{(l)} + \tilde{\delta}_n}{\hat{\theta}_{\hat{r}_l+1}^{(l)} + \tilde{\delta}_n} \leqslant \frac{\tilde{h}_n + \tilde{\delta}_n}{0 + \tilde{\delta}_n} \to 1\,.$$

This together with (S.12) and $\tilde{\delta}_n/\min_{l\in[q]} \theta_{r_l}^{(l)} \to 0$ yields that $\mathbb{P}[\bigcup_{i=1}^q\{\hat{r}_l > r_l\} \,|\, \breve{\Omega}] \to 0$. Thus $\mathbb{P}[\bigcap_{i=1}^q\{\hat{r}_l = r_l\} \,|\, \breve{\Omega}] \to 1$. By (S.11), $\mathbb{P}(\breve{\Omega}) \to 1$. We complete the proof of Theorem 3. □

F Proof of Theorem 4

Let $\boldsymbol{\psi}_j^{(l)}(\cdot) = \{\psi_{j1}^{(l)}(\cdot), \ldots, \psi_{jp_l}^{(l)}(\cdot)\}^\top$ and $\widehat{\boldsymbol{\psi}}_j^{(l)}(\cdot) = \{\hat{\psi}_{j1}^{(l)}(\cdot), \ldots, \hat{\psi}_{jp_l}^{(l)}(\cdot)\}^\top$. Due to the orthonormality of $\boldsymbol{\psi}_j^{(l)}(\cdot)$ and $\widehat{\boldsymbol{\psi}}_j^{(l)}(\cdot)$, we obtain that

$$\begin{aligned}
&\bigg\| \sum_{j=1}^{r_l} [\widehat{\boldsymbol{\psi}}_j^{(l)} \otimes \{\widehat{\boldsymbol{\psi}}_j^{(l)}\}^\top - \boldsymbol{\psi}_j^{(l)} \otimes \{\boldsymbol{\psi}_j^{(l)}\}^\top] \bigg\|_{\mathcal{S},\mathrm{F}}^2 \\
&\quad = \int\int \sum_{k=1}^{p_l}\sum_{m=1}^{p_l} \bigg[\sum_{j=1}^{r_l} \{\hat{\psi}_{jk}^{(l)}(u)\hat{\psi}_{jm}^{(l)}(v) - \psi_{jk}^{(l)}(u)\psi_{jm}^{(l)}(v)\} \bigg]^2 \mathrm{d}u\mathrm{d}v \\
&\quad = \sum_{i=1}^{r_l}\sum_{j=1}^{r_l} \int\int \sum_{k=1}^{p_l}\sum_{m=1}^{p_l} \hat{\psi}_{ik}^{(l)}(u)\hat{\psi}_{im}^{(l)}(v)\hat{\psi}_{jk}^{(l)}(u)\hat{\psi}_{jm}^{(l)}(v)\,\mathrm{d}u\mathrm{d}v
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{r_l}\sum_{j=1}^{r_l}\int\int\sum_{k=1}^{p_l}\sum_{m=1}^{p_l}\psi_{ik}^{(l)}(u)\psi_{im}^{(l)}(v)\psi_{jk}^{(l)}(u)\psi_{jm}^{(l)}(v)\,\mathrm{d}u\mathrm{d}v\\
&-2\sum_{i=1}^{r_l}\sum_{j=1}^{r_l}\int\int\sum_{k=1}^{p_l}\sum_{m=1}^{p_l}\psi_{ik}^{(l)}(u)\psi_{im}^{(l)}(v)\hat{\psi}_{jk}^{(l)}(u)\hat{\psi}_{jm}^{(l)}(v)\,\mathrm{d}u\mathrm{d}v\\
=&\sum_{i=1}^{r_l}\sum_{j=1}^{r_l}\langle\widehat{\boldsymbol{\psi}}_i^{(l)},\widehat{\boldsymbol{\psi}}_j^{(l)}\rangle^2+\sum_{i=1}^{r_l}\sum_{j=1}^{r_l}\langle\boldsymbol{\psi}_i^{(l)},\boldsymbol{\psi}_j^{(l)}\rangle^2-2\sum_{i=1}^{r_l}\sum_{j=1}^{r_l}\langle\widehat{\boldsymbol{\psi}}_i^{(l)},\boldsymbol{\psi}_j^{(l)}\rangle^2\\
=&\,2r_l-2\sum_{i=1}^{r_l}\sum_{j=1}^{r_l}\langle\widehat{\boldsymbol{\psi}}_i^{(l)},\boldsymbol{\psi}_j^{(l)}\rangle^2\,.
\end{aligned}
$$

Denote by $\widetilde{\mathcal{C}}_l=\mathrm{span}\{\widehat{\boldsymbol{\psi}}_1^{(l)}(\cdot),\ldots,\widehat{\boldsymbol{\psi}}_{r_l}^{(l)}(\cdot)\}$ the dynamic space spanned by r_l estimated eigenfunctions. By the definition of $\widetilde{D}(\widetilde{\mathcal{C}}_l,\mathcal{C}_l)$, we thus have $\sqrt{2r_l}\widetilde{D}(\widetilde{\mathcal{C}}_l,\mathcal{C}_l)=\|\sum_{j=1}^{r_l}[\widehat{\boldsymbol{\psi}}_j^{(l)}\otimes\{\widehat{\boldsymbol{\psi}}_j^{(l)}\}^\top-\boldsymbol{\psi}_j^{(l)}\otimes\{\boldsymbol{\psi}_j^{(l)}\}^\top]\|_{\mathcal{S},\mathrm{F}}$. Let $\nu_n=\Xi^2p_\dagger^{1-\alpha}\mathcal{M}_y^{1-\alpha}(n^{-1}\log p)^{(1-\alpha)/2}$, where $\Xi=s_1s_2(2p_\dagger+1)$ with $p_\dagger=\max_{l\in[q]}p_l$. Since $\Delta^{-1}\rho^{-1}\Xi p_\dagger^{3-\alpha}\nu_n\to0$, the result in (S.11) holds. This, together with Equality 1, fixed r_l and the orthonormality of $\boldsymbol{\psi}_j^{(l)}(\cdot)$, leads to

$$
\begin{aligned}
\max_{l\in[q]}\sqrt{2r_l}\widetilde{D}(\widetilde{\mathcal{C}}_l,\mathcal{C}_l)&=\max_{l\in[q]}\bigg\|\sum_{j=1}^{r_l}\Big[\widehat{\boldsymbol{\psi}}_j^{(l)}\otimes\{\widehat{\boldsymbol{\psi}}_j^{(l)}\}^\top-\boldsymbol{\psi}_j^{(l)}\otimes\{\boldsymbol{\psi}_j^{(l)}\}^\top\Big]\bigg\|_{\mathcal{S},\mathrm{F}}\\
&\leqslant\max_{l\in[q]}\sum_{j=1}^{r_l}\|\widehat{\boldsymbol{\psi}}_j^{(l)}-\boldsymbol{\psi}_j^{(l)}\|^2+2\max_{l\in[q]}\sum_{j=1}^{r_l}\|\boldsymbol{\psi}_j^{(l)}\|\|\widehat{\boldsymbol{\psi}}_j^{(l)}-\boldsymbol{\psi}_j^{(l)}\|\\
&=O_{\mathrm{p}}(\Delta^{-2}\rho^{-2}\Xi^2p_\dagger^{6-2\alpha}\nu_n^2)+O_{\mathrm{p}}(\Delta^{-1}\rho^{-1}\Xi p_\dagger^{3-\alpha}\nu_n)\\
&=O_{\mathrm{p}}(\Delta^{-1}\rho^{-1}\Xi p_\dagger^{3-\alpha}\nu_n)\,.
\end{aligned}
$$

Let $e_n=\Delta^{-1}\rho^{-1}\Xi p_\dagger^{3-\alpha}\nu_n$. Theorem 3 implies that $\mathbb{P}[\bigcap_{l=1}^q\{\hat{r}_l=r_l\}]\to1$. Thus, for any $\epsilon>0$, there exists a constant $C>0$ such that

$$
\mathbb{P}\bigg\{\max_{l\in[q]}e_n^{-1}\widetilde{D}(\widetilde{\mathcal{C}}_l,\mathcal{C}_l)>C\bigg\}<\epsilon\,,
$$

which implies

$$
\begin{aligned}
&\mathbb{P}\bigg\{\max_{l\in[q]}e_n^{-1}\widetilde{D}(\widehat{\mathcal{C}}_l,\mathcal{C}_l)>C\bigg\}\\
&\qquad\leqslant\mathbb{P}\bigg[\max_{l\in[q]}e_n^{-1}\widetilde{D}(\widehat{\mathcal{C}}_l,\mathcal{C}_l)>C,\,\bigcap_{l=1}^q\{\hat{r}_l=r_l\}\bigg]+\mathbb{P}\bigg\{\bigcup_{l=1}^q[\hat{r}_l\neq r_l\}\bigg]\\
&\qquad\leqslant\mathbb{P}\bigg\{\max_{l\in[q]}e_n^{-1}\widetilde{D}(\widetilde{\mathcal{C}}_l,\mathcal{C}_l)>C\bigg\}+o(1)<\epsilon+o(1)\,.
\end{aligned}
$$

Hence, $\max_{l\in[q]}\widetilde{D}(\widehat{\mathcal{C}}_l,\mathcal{C}_l)=O_{\mathrm{p}}(e_n)$. We complete the proof of Theorem 4. □

G Proofs of auxiliary lemmas

G.1 Proof of Inequality 1

By Cauchy–Schwartz inequality, we notice that

$$\begin{aligned}\left\|\int\int \boldsymbol{\mathcal{B}}_1(u,v)\boldsymbol{\mathcal{B}}_2(u,v)^\top \,\mathrm{d}u\mathrm{d}v\right\|_1 &= \max_{j\in[p]}\sum_{i=1}^p\left|\int\int\sum_{k=1}^p \mathcal{B}_{1,ik}(u,v)\mathcal{B}_{2,jk}(u,v)\,\mathrm{d}u\mathrm{d}v\right|\\ &\leqslant \max_{j\in[p]}\sum_{i=1}^p\sum_{k=1}^p \|\mathcal{B}_{1,ik}\|_{\mathcal{S}}\|\mathcal{B}_{2,jk}\|_{\mathcal{S}} \qquad \text{(S.14)}\\ &\leqslant \max_{k\in[p]}\sum_{i=1}^p\|\mathcal{B}_{1,ik}\|_{\mathcal{S}}\cdot\max_{j\in[p]}\sum_{k=1}^p\|\mathcal{B}_{2,jk}\|_{\mathcal{S}} = \|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},1}\|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\infty}\,.\end{aligned}$$

By similar argument, we obtain that

$$\left\|\int\int \boldsymbol{\mathcal{B}}_1(u,v)\boldsymbol{\mathcal{B}}_2(u,v)^\top \mathrm{d}u\mathrm{d}v\right\|_\infty \leqslant \|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},\infty}\|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},1}\,. \qquad \text{(S.15)}$$

Combining (S.14) and (S.15) and applying the inequality $\|\mathbf{E}\|^2 \leqslant \|\mathbf{E}\|_\infty\|\mathbf{E}\|_1$ for any matrix $\mathbf{E}\in\mathbb{R}^{p\times p}$, we complete the proof of part (i).

Let $\boldsymbol{\mathcal{C}}(u,v) = \int \boldsymbol{\mathcal{B}}_1(u,w)\boldsymbol{\mathcal{B}}_2(v,w)^\top \mathrm{d}w = \{\mathcal{C}_{ij}(u,v)\}_{i,j\in[p]}$. It then follows from Cauchy–Schwartz inequality that

$$\begin{aligned}\|\boldsymbol{\mathcal{C}}\|_{\mathcal{S},\mathrm{F}}^2 &= \sum_{i,j=1}^p\int\int \mathcal{C}_{ij}^2(u,v)\,\mathrm{d}u\mathrm{d}v = \sum_{i,j=1}^p\int\int\left\{\sum_{k=1}^p\int \mathcal{B}_{1,ik}(u,w)\mathcal{B}_{2,jk}(v,w)\,\mathrm{d}w\right\}^2\mathrm{d}u\mathrm{d}v\\ &\leqslant \sum_{i,j=1}^p\int\int\left\{\sum_{k=1}^p\int\mathcal{B}_{1,ik}^2(u,w)\,\mathrm{d}w\cdot\sum_{k=1}^p\int\mathcal{B}_{2,jk}^2(u,w)\,\mathrm{d}w\right\}\mathrm{d}u\mathrm{d}v = \|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},\mathrm{F}}^2\|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\mathrm{F}}^2\,.\end{aligned}$$

Hence, we complete the proof of part (ii).

By Cauchy–Schwartz inequality, we further obtain that

$$\begin{aligned}\|\boldsymbol{\mathcal{B}}_1+\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\mathrm{F}}^2 &= \sum_{i,j=1}^p\int\int\{\mathcal{B}_{1,ij}(u,v)+\mathcal{B}_{2,ij}(u,v)\}^2\,\mathrm{d}u\mathrm{d}v\\ &= 2\sum_{i,j=1}^p\int\int\mathcal{B}_{1,ij}^2(u,v)\mathcal{B}_{2,ij}^2(u,v)\,\mathrm{d}u\mathrm{d}v + \|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},\mathrm{F}}^2+\|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\mathrm{F}}^2\\ &\leqslant 2\|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},\mathrm{F}}\|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\mathrm{F}}+\|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},\mathrm{F}}^2+\|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\mathrm{F}}^2 = (\|\boldsymbol{\mathcal{B}}_1\|_{\mathcal{S},\mathrm{F}}+\|\boldsymbol{\mathcal{B}}_2\|_{\mathcal{S},\mathrm{F}})^2\,.\end{aligned}$$

Hence, we complete the proof of part (iii). $\square$

G.2 Proof of Inequality 2

By elementary calculations and Inequality 1, we obtain that

$$
\begin{aligned}
\|\mathbf{b}_1^\top \boldsymbol{\mathcal{B}} \mathbf{b}_2\|_{\mathcal{S}}^2 &= \int\int \mathbf{b}_1^\top \boldsymbol{\mathcal{B}}(u,v)\mathbf{b}_2\mathbf{b}_2^\top \boldsymbol{\mathcal{B}}(u,v)^\top \mathbf{b}_1 \,\mathrm{d}u\mathrm{d}v \\
&\leqslant \int\int \|\mathbf{b}_2\mathbf{b}_2^\top\|_2 \|\boldsymbol{\mathcal{B}}(u,v)^\top \mathbf{b}_1\|_2^2 \,\mathrm{d}u\mathrm{d}v \\
&\leqslant \|\mathbf{b}_2\|_2^2 \int\int \mathbf{b}_1^\top \boldsymbol{\mathcal{B}}(u,v)\boldsymbol{\mathcal{B}}(u,v)^\top \mathbf{b}_1 \,\mathrm{d}u\mathrm{d}v \leqslant \|\mathbf{b}_1\|_2^2 \|\mathbf{b}_2\|_2^2 \|\boldsymbol{\mathcal{B}}\|_{\mathcal{S},\infty} \|\boldsymbol{\mathcal{B}}\|_{\mathcal{S},1}\,,
\end{aligned}
$$

which completes our proof. □

G.3 Proof of Inequality 3

Let $\boldsymbol{\mathcal{F}}(u) = \{\mathcal{F}_1(u),\ldots,\mathcal{F}_p(u)\}^\top$ and $\boldsymbol{\mathcal{G}}(u) = \int \boldsymbol{\mathcal{B}}(u,v)\boldsymbol{\mathcal{F}}(v)\,\mathrm{d}v = \{\mathcal{G}_1(u),\ldots,\mathcal{G}_p(u)\}^\top$. By Cauchy–Schwartz inequality, we obtain that

$$
\begin{aligned}
\|\boldsymbol{\mathcal{G}}\|^2 &= \sum_{i=1}^p \int \mathcal{G}_i(u)^2 \,\mathrm{d}u = \sum_{i=1}^p \int \bigg\{ \sum_{k=1}^p \int \mathcal{B}_{ik}(u,v)\mathcal{F}_k(v)\,\mathrm{d}v \bigg\}^2 \mathrm{d}u \\
&\leqslant \sum_{i=1}^p \int \bigg\{ \sum_{k=1}^p \int \mathcal{B}_{ik}^2(u,v)\,\mathrm{d}v \sum_{k=1}^p \int \mathcal{F}_k^2(v)\,\mathrm{d}v \bigg\} \mathrm{d}u = \|\boldsymbol{\mathcal{B}}\|_{\mathcal{S},\mathrm{F}}^2 \|\boldsymbol{\mathcal{F}}\|^2\,,
\end{aligned}
$$

This further leads to $\|\boldsymbol{\mathcal{B}}\|_{\mathcal{L}} = \sup_{\|\boldsymbol{\mathcal{F}}\|\leqslant 1, \boldsymbol{\mathcal{F}}\in\mathbb{H}} \|\boldsymbol{\mathcal{B}}(\boldsymbol{\mathcal{F}})\| \leqslant \sup_{\|\boldsymbol{\mathcal{F}}\|\leqslant 1, \boldsymbol{\mathcal{F}}\in\mathbb{H}} \|\boldsymbol{\mathcal{B}}\|_{\mathcal{S},\mathrm{F}} \|\boldsymbol{\mathcal{F}}\| \leqslant \|\boldsymbol{\mathcal{B}}\|_{\mathcal{S},\mathrm{F}}$, which completes our proof. □

G.4 Proof of Equality 1

For $\boldsymbol{\mathcal{F}} = (\mathcal{F}_1,\ldots,\mathcal{F}_p)^\top$ and $\boldsymbol{\mathcal{G}} = (\mathcal{G}_1,\ldots,\mathcal{G}_p)^\top \in \mathbb{H}$, it holds that

$$
\|\boldsymbol{\mathcal{F}} \otimes \boldsymbol{\mathcal{G}}^\top\|_{\mathcal{S},\mathrm{F}}^2 = \sum_{i,j=1}^p \int\int \mathcal{F}_i^2(u)\mathcal{G}_j^2(v)\,\mathrm{d}u\mathrm{d}v = \bigg\{ \sum_{i=1}^p \int \mathcal{F}_i^2(u)\,\mathrm{d}u \bigg\} \bigg\{ \sum_{j=1}^p \int \mathcal{G}_j^2(v)\,\mathrm{d}v \bigg\} = \|\boldsymbol{\mathcal{F}}\|^2 \|\boldsymbol{\mathcal{G}}\|^2\,.
$$

Hence, we complete our proof. □

G.5 Proof of Lemma A1

This lemma follows directly from Theorem 1 of Fang et al. (2022) and Theorem 2 of Guo and Qiao (2023) and hence the proof is omitted here. □

G.6 Proof of Lemma A2

Recall that $\max_{i\in[p]}\int \mathbb{E}\{Z_{ti}^2(u)\}\,\mathrm{d}u = O(1)$. Hence,

$$
\begin{aligned}
\max_{i,j\in[p],|k|\leqslant k_0\vee m}\|\Sigma_{z,k,ij}\|_{\mathcal{S}}^2 &= \max_{i,j\in[p],|k|\leqslant k_0\vee m}\int\int[\mathbb{E}\{Z_{ti}(u)Z_{(t+k)j}(v)\}]^2\,\mathrm{d}u\mathrm{d}v \\
&\leqslant \max_{i\in[p]}\int \mathbb{E}\{Z_{ti}^2(u)\}\,\mathrm{d}u\cdot\max_{j\in[p]}\int\mathbb{E}\{Z_{(t+k)j}^2(v)\}\,\mathrm{d}v = O(1)\,. \qquad \text{(S.16)}
\end{aligned}
$$

Let $p_\dagger = \max_{l\in[q]} p_l$. Since $\boldsymbol{\Sigma}_{y,k}(u,v) = \mathbf{A}\boldsymbol{\Sigma}_{z,k}(u,v)\mathbf{A}^\top$ and $\Sigma_{z,k,lm} = 0$ for $|l-m|>p_\dagger$, then $\Sigma_{y,k,ij}(u,v) = \sum_{l,m=1}^p A_{il}\Sigma_{z,k,lm}(u,v)A_{jm} = \sum_{|l-m|\leqslant p_\dagger} A_{il}\Sigma_{z,k,lm}(u,v)A_{jm}$. By the inequality $(a+b)^\alpha \leqslant a^\alpha + b^\alpha$ for $a,b\geqslant 0$ and $\alpha\in[0,1)$, we obtain that

$$
\begin{aligned}
\sum_{i=1}^p\|\Sigma_{y,k,ij}\|_{\mathcal{S}}^\alpha &= \sum_{i=1}^p\bigg\|\sum_{|l-m|\leqslant p_\dagger}A_{il}\Sigma_{z,k,lm}A_{jm}\bigg\|_{\mathcal{S}}^\alpha \leqslant \sum_{i=1}^p\bigg(\sum_{|l-m|\leqslant p_\dagger}\|A_{il}\Sigma_{z,k,lm}A_{jm}\|_{\mathcal{S}}\bigg)^\alpha \\
&\leqslant \sum_{i=1}^p\sum_{|l-m|\leqslant p_\dagger}|A_{il}|^\alpha|A_{jm}|^\alpha\|\Sigma_{z,k,lm}\|_{\mathcal{S}}^\alpha \\
&\leqslant \max_{l,m\in[p],|k|\leqslant k_0\vee m}\|\Sigma_{z,k,lm}\|_{\mathcal{S}}^\alpha\cdot\max_{l\in[p]}\sum_{i=1}^p|A_{il}|^\alpha\cdot\sum_{|l-m|\leqslant p_\dagger}|A_{jm}|^\alpha \\
&\leqslant \max_{l,m\in[p],|k|\leqslant k_0\vee m}\|\Sigma_{z,k,lm}\|_{\mathcal{S}}^\alpha\cdot s_2\cdot(2p_\dagger+1)\sum_{m=1}^p|A_{jm}|^\alpha \\
&= O(\Xi)\,,
\end{aligned}
$$

where $\Xi = s_1s_2(2p_\dagger+1)$. By (S.16) and the block structure of $\boldsymbol{\Sigma}_{z,k}(u,v)$, we further obtain that $\|\boldsymbol{\Sigma}_{z,k}\|_{\mathcal{S},1} = \max_{j\in[p]}\sum_{i=1}^p\|\Sigma_{z,k,ij}\|_{\mathcal{S}} = O(p_\dagger)$. Similarly, we also have $\|\boldsymbol{\Sigma}_{z,k}\|_{\mathcal{S},\infty} = O(p_\dagger)$. Together with Inequality 2 and the orthonormality of the rows in $\mathbf{A}$, it holds that

$$
\|\boldsymbol{\Sigma}_{y,k}\|_{\mathcal{S},1} = \max_{j\in[p]}\sum_{i=1}^p\|\Sigma_{y,k,ij}\|_{\mathcal{S}} \leqslant \max_{i,j\in[p]}\|\Sigma_{y,k,ij}\|_{\mathcal{S}}^{1-\alpha}\cdot\max_{j\in[p]}\sum_{i=1}^p\|\Sigma_{y,k,ij}\|_{\mathcal{S}}^\alpha = O(\Xi p_\dagger^{1-\alpha})\,.
$$

Recall k_0 and m are fixed integers. Similarly, we can prove the rest of this lemma. □

G.7 Proof of Lemma A3

Denote by $\mathcal{T}_{\omega_k}(\widehat{\Sigma}_{y,k,ij}^{\mathrm{S}})$ the (i,j)-th component of $\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}_{y,k}^{\mathrm{S}})$. Due to the fact that $\mathcal{T}_{\omega_k}(\widehat{\Sigma}_{y,k,ij}^{\mathrm{S}}) = \widehat{\Sigma}_{y,k,ij}^{\mathrm{S}}(u,v)I\{\|\widehat{\Sigma}_{y,k,ij}^{\mathrm{S}}\|_{\mathcal{S}}\geqslant\omega_k\}$, we have $\|\mathcal{T}_{\omega_k}(\widehat{\Sigma}_{y,k,ij}^{\mathrm{S}}) - \widehat{\Sigma}_{y,k,ij}^{\mathrm{S}}\|_{\mathcal{S}}\leqslant\omega_k$. Under the event $\Omega =$

$\{\max_{i,j\in[p]} \|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij} - \Sigma_{y,k,ij}\|_{\mathcal{S}} \leqslant \tilde{\theta}\omega_k\}$ for $\tilde{\theta} \in (0,1)$ and $\omega_k = c_k \mathcal{M}_y (n^{-1}\log p)^{1/2}$, we have

$$
\begin{aligned}
&\max_{j\in[p]} \sum_{i=1}^{p} \|\mathcal{T}_{\omega_k}(\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}) - \Sigma_{y,k,ij}\|_{\mathcal{S}} \\
&\quad = \max_{j\in[p]} \sum_{i=1}^{p} \|\mathcal{T}_{\omega_k}(\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}) - \Sigma_{y,k,ij}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k\} \\
&\qquad + \max_{j\in[p]} \sum_{i=1}^{p} \|\mathcal{T}_{\omega_k}(\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}) - \Sigma_{y,k,ij}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}} < \omega_k\} \\
&\quad \leqslant \max_{j\in[p]} \sum_{i=1}^{p} \|\mathcal{T}_{\omega_k}(\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}) - \widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k, \|\Sigma_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k\} \\
&\qquad + \max_{j\in[p]} \sum_{i=1}^{p} \|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij} - \Sigma_{y,k,ij}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k, \|\Sigma_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k\} \\
&\qquad + \max_{j\in[p]} \sum_{i=1}^{p} \|\mathcal{T}_{\omega_k}(\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}) - \Sigma_{y,k,ij}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k, \|\Sigma_{y,k,ij}\|_{\mathcal{S}} < \omega_k\} \\
&\qquad + \max_{j\in[p]} \sum_{i=1}^{p} \|\Sigma_{y,k,ij}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}} < \omega_k\} \\
&\quad \leqslant \underbrace{\omega_k \sum_{i=1}^{p} I\{\|\Sigma_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k\}}_{Q_1} + \underbrace{\max_{j\in[p]} \sum_{i=1}^{p} \|\Sigma_{y,k,ij}\|_{\mathcal{S}} I\{\|\Sigma_{y,k,ij}\|_{\mathcal{S}} < 2\omega_k\}}_{Q_2} \\
&\qquad + \underbrace{\max_{j\in[p]} \sum_{i=1}^{p} \|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij} - \Sigma_{y,k,ij}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k, \|\Sigma_{y,k,ij}\|_{\mathcal{S}} < \omega_k\}}_{Q_3} .
\end{aligned}
$$

By Lemma A2, we have $Q_1 + Q_2 \lesssim \omega_k^{1-\alpha} \sum_{i=1}^{p} \|\Sigma_{y,k,ij}\|_{\mathcal{S}}^{\alpha} \lesssim \omega_k^{1-\alpha}\Xi$ under the event Ω. Also,

$$
\begin{aligned}
Q_3 &\leqslant \max_{j\in[p]} \sum_{i=1}^{p} \|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij} - \Sigma_{y,k,ij}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k, \|\Sigma_{y,k,ij}\|_{\mathcal{S}} < (1-\tilde{\theta})\omega_k\} \\
&\quad + \max_{j\in[p]} \sum_{i=1}^{p} \|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij} - \Sigma_{y,k,ij}\|_{\mathcal{S}} I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}\|_{\mathcal{S}} \geqslant \omega_k, (1-\tilde{\theta})\omega_k \leqslant \|\Sigma_{y,k,ij}\|_{\mathcal{S}} < \omega_k\} \\
&\leqslant \omega_k \max_{j\in[p]} \sum_{i=1}^{p} I\{\|\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij} - \Sigma_{y,k,ij}\|_{\mathcal{S}} > \tilde{\theta}\omega_k\} + \omega_k \max_{j\in[p]} \sum_{i=1}^{p} I\{\|\Sigma_{y,k,ij}\|_{\mathcal{S}} \geqslant (1-\tilde{\theta})\omega_k\} \\
&= \omega_k \max_{j\in[p]} \sum_{i=1}^{p} I\{\|\Sigma_{y,k,ij}\|_{\mathcal{S}} \geqslant (1-\tilde{\theta})\omega_k\} \lesssim \omega_k^{1-\alpha}\Xi
\end{aligned}
$$

under the event Ω. By Lemma A1, if $n \gtrsim \log p$ and $\tilde{c}\tilde{\theta}^2 c_k^2 > 2$, then $\mathbb{P}(\Omega^c) \leqslant 8p^{2-\tilde{c}\tilde{\theta}^2 c_k^2} \to 0$. Combining the above results, we thus have

$$\max_{j\in[p]} \sum_{i=1}^{p} \|\mathcal{T}_{\omega_k}(\widehat{\Sigma}^{\mathrm{S}}_{y,k,ij}) - \Sigma_{y,k,ij}\|_{\mathcal{S}} = O_{\mathrm{p}}\left\{\Xi \mathcal{M}_y^{1-\alpha}\left(\frac{\log p}{n}\right)^{(1-\alpha)/2}\right\}.$$

Recall k_0 and m are fixed integers. We have the first result. The second result can be proved in the similar manner. Due to $\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{S}}_{y,k})(u,v)^{\otimes 2} - \boldsymbol{\Sigma}_{y,k}(u,v)^{\otimes 2} = \{\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{S}}_{y,k})(u,v) - \boldsymbol{\Sigma}_{y,k}(u,v)\}^{\otimes 2} + \boldsymbol{\Sigma}_{y,k}(u,v)\{\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{S}}_{y,k})(u,v) - \boldsymbol{\Sigma}_{y,k}(u,v)\}^{\top} + \{\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{S}}_{y,k})(u,v) - \boldsymbol{\Sigma}_{y,k}(u,v)\}\boldsymbol{\Sigma}_{y,k}(u,v)^{\top}$, it follows from Inequality 1 and Lemma A2 that

$$\begin{aligned}
&\left\|\int\int \{\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{S}}_{y,k})(u,v)^{\otimes 2} - \boldsymbol{\Sigma}_{y,k}(u,v)^{\otimes 2}\}\,\mathrm{d}u\mathrm{d}v\right\|_2 \\
&\leqslant 2\|\boldsymbol{\Sigma}_{y,k}\|_{\mathcal{S},1}^{1/2}\|\boldsymbol{\Sigma}_{y,k}\|_{\mathcal{S},\infty}^{1/2}\|\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{S}}_{y,k}) - \boldsymbol{\Sigma}_{y,k}\|_{\mathcal{S},1}^{1/2}\|\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{S}}_{y,k}) - \boldsymbol{\Sigma}_{y,k}\|_{\mathcal{S},\infty}^{1/2} \\
&\quad + \|\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{S}}_{y,k}) - \boldsymbol{\Sigma}_{y,k}\|_{\mathcal{S},1}\|\mathcal{T}_{\omega_k}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{S}}_{y,k}) - \boldsymbol{\Sigma}_{y,k}\|_{\mathcal{S},\infty} \\
&= O_{\mathrm{p}}\left\{\Xi^2 p_{\dagger}^{1-\alpha}\mathcal{M}_y^{1-\alpha}\left(\frac{\log p}{n}\right)^{(1-\alpha)/2}\right\} + O_{\mathrm{p}}\left\{\Xi^2\mathcal{M}_y^{2-2\alpha}\left(\frac{\log p}{n}\right)^{1-\alpha}\right\}.
\end{aligned}$$

Since $p_{\dagger}^{-2}\mathcal{M}_y^2 \log p = o(n)$, we have the third result. □

G.8 Proof of Lemma A5

By the definition of spectral decomposition, we have that $\theta_j = \min_{\mathcal{B}\in\mathcal{L}_{j-1}} \|\mathbf{Q} - \mathcal{B}\|_{\mathcal{L}}$ and $\hat{\theta}_j = \min_{\mathcal{B}\in\mathcal{L}_{j-1}} \|\widehat{\mathbf{Q}} - \mathcal{B}\|_{\mathcal{L}}$, where $\mathcal{L}_{j-1} = \{\mathcal{B} : \mathcal{B} \in \mathcal{L}, \dim(\mathrm{Im}(\mathcal{B})) \leqslant j-1\}$. Thus, $\theta_j = \min_{\mathcal{B}\in\mathcal{L}_{j-1}} \|\mathbf{Q} - \mathcal{B}\|_{\mathcal{L}} \leqslant \|\mathbf{Q} - \widehat{\mathbf{Q}}\|_{\mathcal{L}} + \min_{\mathcal{B}\in\mathcal{L}_{j-1}} \|\widehat{\mathbf{Q}} - \mathcal{B}\|_{\mathcal{L}} = \|\widehat{\mathbf{Q}} - \mathbf{Q}\|_{\mathcal{L}} + \hat{\theta}_j$. Similarly, we have $\hat{\theta}_j = \min_{\mathcal{B}\in\mathcal{L}_{j-1}} \|\widehat{\mathbf{Q}} - \mathcal{B}\|_{\mathcal{L}} \leqslant \|\widehat{\mathbf{Q}} - \mathbf{Q}\|_{\mathcal{L}} + \min_{\mathcal{B}\in\mathcal{L}_{j-1}} \|\mathbf{Q} - \mathcal{B}\|_{\mathcal{L}} = \|\widehat{\mathbf{Q}} - \mathbf{Q}\|_{\mathcal{L}} + \theta_j$. Combining the two above results with Inequality 3, we obtain that

$$|\hat{\theta}_j - \theta_j| \leqslant \|\widehat{\mathbf{Q}} - \mathbf{Q}\|_{\mathcal{L}} \leqslant \|\widehat{\mathbf{Q}} - \mathbf{Q}\|_{\mathcal{S},\mathrm{F}} \tag{S.17}$$

holds for all $j \geqslant 1$, which completes our proof of part (i).

Without loss of generality, we assume that $\langle\widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_j\rangle \geqslant 0$. Since $\sum_{l=1}^{\infty}\langle\widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_l\rangle^2 = \|\widehat{\boldsymbol{\phi}}_j\|^2 = 1$ and $0 \leqslant \langle\widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_j\rangle \leqslant 1$, it holds that

$$\|\widehat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j\|^2 = \sum_{l=1}^{\infty}\left(\langle\widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_l\rangle - \langle\boldsymbol{\phi}_j, \boldsymbol{\phi}_l\rangle\right)^2 = \left(\langle\widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_j\rangle - 1\right)^2 + \sum_{l\neq j}\langle\widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_l\rangle^2$$

$$= \langle \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_j \rangle^2 - 2\langle \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_j \rangle + \sum_{l=1}^{\infty} \langle \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_l \rangle^2 + \sum_{l \neq j} \langle \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_l \rangle^2$$

$$= 2\sum_{l \neq j} \langle \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_l \rangle^2 + 2(\langle \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_j \rangle^2 - \langle \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_j \rangle) \leqslant 2\sum_{l \neq j} \langle \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_l \rangle^2 . \tag{S.18}$$

Observe that $\mathbf{Q}(\widehat{\boldsymbol{\phi}}_j)(u) - \theta_j \widehat{\boldsymbol{\phi}}_j(u) = (\mathbf{Q} - \widehat{\mathbf{Q}})(\widehat{\boldsymbol{\phi}}_j)(u) + (\hat{\theta}_j - \theta_j)\widehat{\boldsymbol{\phi}}_j(u)$. This together with Inequality 3, (S.17) and the orthonormality of $\widehat{\boldsymbol{\phi}}_j$ implies that

$$\|\mathbf{Q}(\widehat{\boldsymbol{\phi}}_j) - \theta_j \widehat{\boldsymbol{\phi}}_j\| \leqslant 2\|\widehat{\mathbf{Q}} - \mathbf{Q}\|_{\mathcal{S},\mathrm{F}} . \tag{S.19}$$

We further write

$$\|\mathbf{Q}(\widehat{\boldsymbol{\phi}}_j) - \theta_j \widehat{\boldsymbol{\phi}}_j\|^2 = \sum_{l=1}^{\infty} \big(\langle \widehat{\boldsymbol{\phi}}_j, \mathbf{Q}(\boldsymbol{\phi}_l) \rangle - \langle \theta_j \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_l \rangle\big)^2$$

$$= \sum_{l \neq j} (\theta_l - \theta_j)^2 \langle \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_l \rangle^2 \geqslant \Delta_j^2 \sum_{l \neq j} \langle \widehat{\boldsymbol{\phi}}_j, \boldsymbol{\phi}_l \rangle^2 . \tag{S.20}$$

Combining (S.18)–(S.20), we complete the proof of part (ii). □