Journal of Multivariate Analysis 186 (2021) 104786

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A semiparametric latent factor model for large scale temporal data with heteroscedasticity

Lyuou Zhang [a], Wen Zhou [b,*], Haonan Wang [b]

[a] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, PR China
[b] Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA

ARTICLE INFO

Article history:
Received 22 December 2020
Received in revised form 5 July 2021
Accepted 6 July 2021
Available online 17 July 2021

AMS 2020 subject classifications:
primary 62H25
secondary 62M10

Keywords:
Efficient estimation
Heteroscedasticity
Large scale temporal data
Latent semiparametric factor model
Projection

ABSTRACT

Large scale temporal data have flourished in a vast array of applications, and their sophisticated structures, especially the heteroscedasticity among subjects with inter- and intra-temporal dependence, have fueled a great demand for new statistical models. In this paper, with covariate information, we consider a flexible model for large scale temporal data with subject-specific heteroscedasticity. Formally, the model employs latent semiparametric factors to simultaneously account for the subject-specific heteroscedasticity and the contemporaneous and/or serial correlations. The subject-specific heteroscedasticity is modeled as the product of the unobserved factor process and subject's covariate effect, which is further characterized via additive models. For estimation, we propose a two-step procedure. First, the latent factor process and nonparametric loading are recovered through projection-based methods, and following, we estimate the regression components by approaches motivated from the generalized least squares. By scrupulously examining the non-asymptotic rates for recovering the factor process and its loading, we show the consistency and efficiency of estimated regression coefficients in the absence of prior knowledge of latent factor process and subject's covariate effect. The statistical guarantees remain valid even for finite time points that makes our method particularly appealing when the subjects significantly outnumber the observation time points. Using comprehensive simulations, we demonstrate the finite sample performance of our method, which corroborates the theoretical findings. Finally, we apply our method to a data set of air quality and energy consumption collected at 129 monitoring sites in the United States in 2015.

1. Introduction

Jointly modeling a large and possibly divergent number of temporally evolving subjects arise ubiquitously in genomics, proteomics, environmental science, econometrics, clinical studies, and neuroscience. An extensively used statistical model for explaining the interactions and co-movements among the temporally evolving subjects is $y_{it} = \boldsymbol{z}_{it}^\top \boldsymbol{\beta} + \varepsilon_{it}, \quad i \in \{1, \ldots, n\}, \ t \in \{1, \ldots, T\}$, where y_{it} is the observation for the ith subject at time point t, $\boldsymbol{\beta}$ is a p-dimensional regression coefficient, $\boldsymbol{z}_{it}$ is a p-dimensional covariate vector that might evolve in time, and $(\varepsilon_{1t}, \ldots, \varepsilon_{nt})^\top$ is a vector time series with possible contemporaneous correlations. Here, the number of subjects n is allowed to diverge much faster than the number of time points T. To name a few applications, y_{it} can model the expression level of the ith gene at time point t in

* Corresponding author.
E-mail addresses: zhanglvou@mail.shufe.edu.cn (L. Zhang), riczw@stat.colostate.edu (W. Zhou), wanghn@stat.colostate.edu (H. Wang).

https://doi.org/10.1016/j.jmva.2021.104786
0047-259X/

a time course sequencing experiment, see, e.g., [10,40], the concentration of certain air pollutant in county i at day t, see, e.g., [29], and the measurement from electroencephalograms at brain location i and time point t in a motor–visual task experiment, see, e.g., [33]. As n rapidly grows, heteroscedasticity across subjects becomes inevitable and brings substantial challenges to modeling, estimation, and inference [14,46]. First, ignoring the subject-specific heteroscedasticity leads to inefficient estimation and inference on the regression coefficient $\boldsymbol{\beta}$. In addition, in the presence of contemporaneous and serial correlations, when n rapidly outnumbers T the high dimensionality of data makes it even more difficult to accurately estimate the covariance of ϵ_{it}, which compromises the estimation efficiency of $\boldsymbol{\beta}$.

In this paper, inspired by the approximate factor structure and its variants [11,26,38,44], we introduce a flexible data-driven model, where the heteroscedasticity across subjects and serial dependence of ε_{it} are assumed to arise from a product of the subject-specific effect and some latent stationary process. Specifically, motivated by Connor and Linton [13], Daye et al. [14], and Fan et al. [19], with additional time invariant covariates $\boldsymbol{x}_i$, we model the subject-specific effect by $\boldsymbol{g}(\boldsymbol{x}_i) = (g_1(\boldsymbol{x}_i), \ldots, g_K(\boldsymbol{x}_i))^\top$ with nonparametric functions $g_1, \ldots, g_K$. For instance, $\boldsymbol{x}_i$ could be the genetic information in a clinical study or the market capitalization in the financial asset allocation. Then, consider a K-dimensional zero-mean latent process $\boldsymbol{f}_t$, our semiparametric latent factor model with subject-specific heteroscedasticity is

$$y_{it} = \boldsymbol{z}_{it}^\top \boldsymbol{\beta} + \boldsymbol{g}(\boldsymbol{x}_i)^\top \boldsymbol{f}_t + u_{it}, \tag{1}$$

where the residual process u_{it} is independent of $\boldsymbol{f}_t$. Analogous to the traditional factor models, $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ can be viewed as the loading and factor, respectively. Particularly, $\boldsymbol{g}(\boldsymbol{x}_i)$ models the heteroscedasticity across subjects and, together with $\boldsymbol{f}_t$, retains the cross-sectional dependence while $\boldsymbol{f}_t$ and u_{it} characterize the serial dependence. Model (1) features a large number of widely used models. For example, when $\boldsymbol{f}_t$ is degenerate, (1) reduces to the partially linear additive models [41]; when $\boldsymbol{g}(\cdot)$ is known and $\boldsymbol{f}_t$ is Gaussian, (1) is the traditional linear mixed model; with i replaced by one-dimensional spatial locations, (1) is analogous to the spatio-temporal model [31]; and when $\boldsymbol{g}(\cdot)$ reduces to constants, (1) is equivalent to the traditional factor models [2,11] or the panel data model with unobservable interactive effects [3,4]. It is worth mentioning that, focusing on the high-dimensional factor analysis, Connor and Linton [13] and Fan et al. [19] consider γ_{ik} in addition to $\boldsymbol{g}(\boldsymbol{x}_i)$ in the loading to account for that cannot be explained by $\boldsymbol{x}_i$. In contrast, one of our major goals is to efficiently estimate $\boldsymbol{\beta}$ in the presence of subject-specific heteroscedasticity and contemporaneous and serial correlations for any n and T. To that end, it requires recovering $\boldsymbol{f}_t$ and $\boldsymbol{g}(\boldsymbol{x}_i)$ with satisfactory rates as well as accurate estimation of the long run variance of residues, while from Theorem 4.1 in Fan et al. [19] and Theorem 1, the desired rate for estimating the long run variance of residues will not be satisfied in the presence of γ_{ik}, unless more stringent conditions on n, T and γ_{ik}'s are imposed.

Like the partially linear model and the linear mixed model, regardless of its consistency, the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ in (1) is inefficient without taking into account the dependence. Therefore, a careful estimation of the unobserved $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ are needed to guarantee some sort of efficiency in both estimation and inference on $\boldsymbol{\beta}$. In the literature, there are scatter approaches to estimate $\boldsymbol{f}_t$ and its loadings for models similar to (1). For instance, Connor and Linton [13] employed a kernel method to estimate $\boldsymbol{f}_t$ given $\boldsymbol{x}_i$ with finite values, and Connor et al. [12] extended such estimates for general $\boldsymbol{x}_i$. Additionally, the consistency on estimated loading and latent factor shed light on estimating the large covariance matrix under assumptions of factor structures [18]. Motivated from these pioneering works, we propose a two-stage projection-based estimator of $\boldsymbol{\beta}$, $\boldsymbol{g}(\boldsymbol{x}_i)$, and $\boldsymbol{f}_t$ in (1). Roughly speaking, adopting a projection-based principal component type estimator [2,19], we first estimate $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ from $y_{it} - \boldsymbol{z}_{it}^\top \widehat{\boldsymbol{\beta}}^0$ for some preliminary $\widehat{\boldsymbol{\beta}}^0$. Using the estimated $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ from the first-stage, we then estimate $\boldsymbol{\beta}$ with a generalized least squares (GLS) type approach.

Theoretically, the asymptotic properties such as consistency on estimating $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ are not sufficient to guarantee the consistency and, especially the efficiency, of the second-stage estimator of $\boldsymbol{\beta}$ due to the lack of the finite sample characterization of errors from the first-stage [6]. In fact, it is known that a naive plug-in GLS type estimator does not necessarily guarantee the efficiency. To circumvent these challenges, a major contribution of this paper is a careful non-asymptotic analysis on the projection-based estimator of $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$, by which we show that the consistency on estimating $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ is free from restrictions on the relationship between n and T. With the exponential type concentration inequalities on estimating $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$, we are able to obtain the finite sample deviation of the proposed two-stage estimator of $\boldsymbol{\beta}$ from the oracle GLS that enjoys full access to $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$. These nontrivial results show that our estimator of $\boldsymbol{\beta}$ is overwhelmingly close to the oracle GLS, which establishes the efficiency of our estimator of $\boldsymbol{\beta}$. In addition, we show the asymptotic normality of our estimator of $\boldsymbol{\beta}$ for drawing inference. The established concentration results for recovering $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ (Theorem 1) are of independent interest for extending the projection-based principal component analysis (PCA) to other high-dimensional problems such as modeling temporally evolving tensor data or the segmentation of high-dimensional time series.

After introducing our model with identification conditions in Section 2.1, we detail the two-stage projection-based estimation of the loading, latent factor process, and regression coefficients in Section 2.2. We carry out the non-asymptotic analysis of our estimator and explore its efficiency in Section 3. Section 4 presents the inference. Sections 5 and 6 report simulation studies and an application on air quality and energy consumption data in the United States to demonstrate our method. After discussing potential extensions of our method in Section 7, we conclude the paper with technical details in Section 8. Extra numerical results and a detailed discussion on the determination of unknown dimension K are deferred to the supplement.

2. Methodology

2.1. A semiparametric latent factor model

Consider an $n \times 1$ vector of temporally evolving subjects $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{nt})^\top$ along with p-dimensional predictors $\boldsymbol{z}_{it}$ and d-dimensional time invariant covariates $\boldsymbol{x}_i$ associated with the ith subject. Our objective is to study the long run movement of y_{it} with respect to $\boldsymbol{z}_{it}$ and model the dependence, over time, of each component of $\boldsymbol{y}_t$ and across components, where the heteroscedasticity across subjects is accounted by $\boldsymbol{x}_i$. In our baseline formulation, each subject is modeled by a multi-factor linear model $y_{it} = \boldsymbol{z}_{it}^\top \boldsymbol{\beta} + \varepsilon_{it}$ [8] for $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T\}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional vector of regression coefficients common across subjects. As discussed in Section 1, we adopt the semiparametric factor model $\varepsilon_{it} = \boldsymbol{g}(\boldsymbol{x}_i)^\top \boldsymbol{f}_t + u_{it}$, where the vector loading function $\boldsymbol{g}(\boldsymbol{x}_i) : \mathbb{R}^d \to \mathbb{R}^K$ accounts for the subject-specific heteroscedasticity and contemporaneous dependence and the K-dimensional latent factor process $\boldsymbol{f}_t$ models the serial dependence. We further model $g_k(\boldsymbol{x}_i)$ in an additive fashion, i.e., $g_k(\boldsymbol{x}_i) = \sum_{\ell=1}^d g_{k\ell}(x_{i\ell})$, without losing flexibility yet providing concision in techniques [21]. Function $\boldsymbol{g}$ offers structure that is flexible enough to allow dependence between $\{\boldsymbol{z}_{it}\}_{i\le n, t\le T}$ and $\{\boldsymbol{x}_i\}_{i\le n}$, which has also been noted for other high-dimensional heteroscedastic regression models [14].

For each t, denote $\mathbf{Z}_t = (\boldsymbol{z}_{1t}, \ldots, \boldsymbol{z}_{nt})^\top$, $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{nt})^\top$, and the $n \times K$ matrix of $g_k(\boldsymbol{x}_i)$ by $\mathbf{G} = (\boldsymbol{g}(\boldsymbol{x}_1), \ldots, \boldsymbol{g}(\boldsymbol{x}_n))^\top$. A more compact form of (1) reads

$$\boldsymbol{y}_t = \mathbf{Z}_t \boldsymbol{\beta} + \mathbf{G} \boldsymbol{f}_t + \boldsymbol{u}_t. \tag{2}$$

Similar to the traditional factor models, we impose the following conditions to control the rank and scale of latent loading function and factor process for model identification.

Condition 1. The rank of $\mathbf{G}$ is K. For each t, $f_{1t}, \ldots, f_{Kt}$ are uncorrelated with each other and have zero mean and unit variance; $u_{1t}, \ldots, u_{nt}$ are uncorrelated with each other and have zero mean and finite variances; $\mathbf{Z}_t$, $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$ are uncorrelated from each other.

In addition, we assume that $\boldsymbol{f}_t, u_{it}$ are independent from $\boldsymbol{x}_i$ so that $\mathrm{Cov}(y_{it}, y_{js} | \boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{g}(\boldsymbol{x}_i)^\top \mathrm{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_s) \boldsymbol{g}(\boldsymbol{x}_j) + \mathrm{Cov}(u_{it}, u_{js})$ for each i, j, t, s. Our model reaches beyond the existing literature [4,8] in the way that the inter- and intra-temporal dependence as well as the subject-specific heteroscedasticity are modeled simultaneously by $\boldsymbol{f}_t$ and $\boldsymbol{g}$. Condition 1 is similar to conditions following (1.1) in Chamberlain and Rothschild [11] and Condition (C1) in Lam and Yao [26] (with diagonal $\boldsymbol{\Sigma}_\varepsilon$ and integer k herein). It guarantees the identifiability of the column space of $\mathbf{G}$. To further identify $\mathbf{G}$ from its column space, consider T potentially dependent replicates $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ and $\mathbb{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_T)$. Let $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)^\top$ and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$, (2) reads

$$\mathbf{Y} = \mathbb{Z}(\mathbf{I}_T \otimes \boldsymbol{\beta}) + \mathbf{G}\mathbf{F}^\top + \mathbf{U} \tag{3}$$

where $\otimes$ is the Kronecker product. The following condition assures the identification of $\mathbf{G}$.

Condition 2. Almost surely, $T^{-1}\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k$ and $\mathbf{G}^\top\mathbf{G}$ is diagonal with distinct entries.

Same as the PC1 condition of Bai and Ng [5], the first part of Condition 2 has been commonly assumed in factor analysis [20] and it is compatible with Condition 1 as $T^{-1}\mathbf{F}^\top\mathbf{F}$ is an estimator of $\mathrm{Var}(\boldsymbol{f}_t)$. Under Condition 2, we can identify $\mathbf{GH}$ and $\mathbf{FH}$ for some $K \times K$ orthogonal matrix $\mathbf{H}$ with $\mathbf{H} = \mathbf{I} + o(\min(n, T)^{-1})$, while the distinction among entries of $\mathbf{G}^\top\mathbf{G}$ further prevents rotational indeterminacy.

In contrast to the approximate factor model that allows cross-sectional dependence in $\boldsymbol{u}_t$, the assumption on u_{it} over i in Condition 1 is designated for efficiently estimating $\boldsymbol{\beta}$ without any restrictions on n and T. In fact, in the absence of the modulating component in (3), mild cross-sectional dependence of $\boldsymbol{u}_t$ will not affect the estimation of $\mathbf{G}$ and $\mathbf{F}$. On the other hand, without Condition 1 on u_{it}, a consistent estimate of $\mathrm{Cov}(\boldsymbol{u}_t)$ is required for efficiently estimating $\boldsymbol{\beta}$. This will demand conditions on n and T, such as $\sqrt{n}\ln(n) = o(T)$ [18,43], which is more stringent compared to those in Section 3.

2.2. Two-stage projection-based estimation

2.2.1. First-stage: projection-based estimator of $\mathbf{G}$ and $\mathbf{F}$

Let $\widetilde{\mathbf{Y}} = \mathbf{Y} - \mathbb{Z}(\mathbf{I}_T \otimes \widehat{\boldsymbol{\beta}}^0)$ and $\widetilde{\mathbf{U}} = \mathbf{U} + \mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}$ for some preliminary $\widehat{\boldsymbol{\beta}}^0$, (3) can be written as

$$\widetilde{\mathbf{Y}} = \mathbf{G}\mathbf{F}^\top + \widetilde{\mathbf{U}}. \tag{4}$$

A naive approach is to estimate $\mathbf{G}$ and $\mathbf{F}$ using PCA. That is, $\mathbf{F}/\sqrt{T}$ are estimated using eigenvectors corresponding to the first K largest eigenvalues of the $T \times T$ matrix $\widetilde{\mathbf{Y}}^\top\widetilde{\mathbf{Y}}$, and $\mathbf{G}$ is estimated by right projecting $T^{-1}\widetilde{\mathbf{Y}}$ onto the estimated $\mathbf{F}$. This method, however, takes into no account for the functional structure of $\boldsymbol{g}$ or the smooth variation of $\{\widetilde{y}_{it}\}_{i=1}^n$ from (4) against $\boldsymbol{x}_i$ at each t. Fan et al. [19] proposed a projected principal component approach by smoothing $\{\widetilde{y}_{it}\}_{i=1}^n$ as a function of $\boldsymbol{x}_i$ at each t before implementing the aforementioned principal component estimation. Motivated by this, we replace $\widetilde{\mathbf{Y}}$ by

$\mathbf{P}\widetilde{\mathbf{Y}}$ for some projection $\mathbf{P}$ onto a linear space spanned by a set of basis functions. Not only leveraging the smoothness, but $\mathbf{P}$ can also be constructed to be orthogonal to errors $\widetilde{\mathbf{U}}$ so that the subsequent PCA procedure is approximately errorless.

To begin with, let $\mathbb{H}$ be a linear space spanned by a sequence of orthonormal basis functions $\{\phi_0(x) \equiv 1, \phi_1(x), \phi_2(x), \dots \phi_J(x)\}$, where the number of basis functions J diverges in n. For each $k \in \{1, \dots, K\}$, $i \in \{1, \dots, n\}$, and $\ell \in \{1, \dots, d\}$, we have $g_{k\ell}(x_{i\ell}) = b_{0,k\ell} + \sum_{j=1}^{J} b_{j,k\ell}\phi_j(x_{i\ell}) + R_{k\ell}(x_{i\ell})$, where $\{b_{j,k\ell}\}_{j\le J}$ are the coefficients and $R_{k\ell}$ is the approximation or projection error. Assume $Jd + 1 < n$ so that the coefficients are estimable. Denote $\boldsymbol{b}_k = (b_{0,k}, b_{1,k1}, \dots, b_{J,k1}, \dots, b_{1,kd}, \dots, b_{J,kd})^\top$ for each $k \in \{1, \dots, K\}$ and $i \in \{1, \dots, n\}$, where $b_{0,k} = \sqrt{J}\sum_{\ell=1}^{d} b_{0,k\ell}$, and $\boldsymbol{\varphi}_i = (1/\sqrt{J}, \phi_1(x_{i1}), \dots, \phi_J(x_{i1}), \dots, \phi_1(x_{id}), \dots, \phi_J(x_{id}))^\top$. Then, it admits $g_k(\boldsymbol{x}_i) = \boldsymbol{\varphi}_i^\top \boldsymbol{b}_k + \sum_{\ell=1}^{d} R_{k\ell}(x_{i\ell})$ and (4) can be rewritten as $\widetilde{\mathbf{Y}} = (\boldsymbol{\Phi}\mathbf{B} + \mathbf{R})\mathbf{F}^\top + \widetilde{\mathbf{U}}$, where $\boldsymbol{\Phi} = (\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_n)^\top$, $\mathbf{B} = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_K)$, and $\mathbf{R} = \{\sum_{\ell=1}^{d} R_{k\ell}(x_{i\ell})\}_{i=1,k=1}^{n,K}$. Then, we let $\mathbf{P} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top$ and apply the PCA procedure to $\mathbf{P}\widetilde{\mathbf{Y}}$. That is, we estimate $\widehat{\mathbf{F}}$ by letting $\widehat{\mathbf{F}}/\sqrt{T}$ be the eigenvectors corresponding to the first K largest eigenvalues of $\widetilde{\mathbf{Y}}^\top\mathbf{P}\widetilde{\mathbf{Y}}$ and estimate $\mathbf{G}$ by $\widehat{\mathbf{G}} = T^{-1}\mathbf{P}\widetilde{\mathbf{Y}}\widehat{\mathbf{F}}$. Moreover, we let $\widehat{\mathbf{B}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\widetilde{\mathbf{Y}}\widehat{\mathbf{F}}$.

2.2.2. Second-stage: GLS-type estimator of $\boldsymbol{\beta}$

First, from (3), consider $\bar{\boldsymbol{y}} = \mathbb{Z}_0^\top\boldsymbol{\beta} + \mathbf{G}T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t + T^{-1}\sum_{t=1}^{T}\boldsymbol{u}_t$, where $\mathbb{Z}_0 = T^{-1}\sum_{t=1}^{T}\mathbf{Z}_t$ and $\bar{\boldsymbol{y}} = T^{-1}\sum_{t=1}^{T}\boldsymbol{y}_t$. Conditional on $\mathbf{Z}_t$'s and $\boldsymbol{x}_i$'s, Condition 1 implies that the variance of $n \times 1$ vector $\bar{\boldsymbol{y}}$ is

$$\mathbf{V} = \mathbf{G}\,\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t\right)\mathbf{G}^\top + \boldsymbol{\mathcal{D}}, \tag{5}$$

where the $n \times n$ diagonal matrix $\boldsymbol{\mathcal{D}}$ has diagonal entries $\mathrm{Var}(T^{-1}\sum_{t=1}^{T}u_{1t}), \dots, \mathrm{Var}(T^{-1}\sum_{t=1}^{T}u_{nt})$. Then, (5) naturally leads to the oracle GLS-type estimate of $\boldsymbol{\beta}$,

$$\widetilde{\boldsymbol{\beta}} = (\mathbb{Z}_0^\top\mathbf{V}^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^\top\mathbf{V}^{-1}\bar{\boldsymbol{y}}. \tag{6}$$

With the full knowledge on $\mathbf{G}$ and $\mathbf{F}$ in (3), $\mathbf{V}$ in (6) can be estimated as follows. Let $\bar{\boldsymbol{f}} = T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t$, it is known that $\mathrm{Var}(T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t) = T^{-2}\sum_{t=-T+1}^{T-1}(T-|t|)\boldsymbol{\Sigma}_f(t)$, where $\boldsymbol{\Sigma}_f(s) = \mathrm{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_{t+s})$ and $\boldsymbol{\Sigma}_f(-s) = \mathrm{Cov}(\boldsymbol{f}_{t-s}, \boldsymbol{f}_t)$ can be estimated by $\widehat{\boldsymbol{\Sigma}}_f(s) = (T-s)^{-1}\sum_{t=1}^{T-s}(\boldsymbol{f}_t - \bar{\boldsymbol{f}})(\boldsymbol{f}_{t+s} - \bar{\boldsymbol{f}})^\top$ and $\widehat{\boldsymbol{\Sigma}}_f(-s) = (T-s)^{-1}\sum_{t=s}^{T}(\boldsymbol{f}_{t-s} - \bar{\boldsymbol{f}})(\boldsymbol{f}_t - \bar{\boldsymbol{f}})^\top$ for $s \ge 0$, respectively. Naturally, $\mathcal{V}(\boldsymbol{f}_t) = T^{-2}\sum_{t=-T+1}^{T-1}(T-|t|)\widehat{\boldsymbol{\Sigma}}_f(t)$ estimates $\mathrm{Var}(T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t)$ in (5), and similarly $\mathcal{V}(u_{it})$ estimates $\mathrm{Var}(T^{-1}\sum_{t=1}^{T}u_{it})$ for each $i \in \{1, \dots, n\}$ and the $n \times n$ diagonal matrix with diagonals $\mathcal{V}(u_{1t}), \dots, \mathcal{V}(u_{nt})$ estimates $\boldsymbol{\mathcal{D}}$.

Though the oracle GLS estimator $\widetilde{\boldsymbol{\beta}}$ is not accessible in practice as it depends on the full knowledge on $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$, it suggests a refined estimator of $\boldsymbol{\beta}$ by replacing $\mathbf{G}$ and $\mathbf{F}$ with $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{F}}$ in (6), respectively. With $\widehat{\mathbf{F}}$ from the first-stage, we can approximate $\mathcal{V}(\boldsymbol{f}_t)$ and $\mathcal{V}(u_{it})$ in (5) by $\mathcal{V}(\widehat{\boldsymbol{f}}_t)$ and $\mathcal{V}(\widehat{u}_{it})$ respectively, where $\widehat{\boldsymbol{f}}_t$ is the tth row of $\widehat{\mathbf{F}}$, and $\widehat{\boldsymbol{u}}_t$ is the tth column of corresponding $\widehat{\mathbf{U}} = \widetilde{\mathbf{Y}} - \widehat{\mathbf{G}}\widehat{\mathbf{F}}^\top$. In summary, we can estimate $\mathbf{V}$ by

$$\widehat{\mathbf{V}} = \widehat{\mathbf{G}}\mathcal{V}(\widehat{\boldsymbol{f}}_t)\widehat{\mathbf{G}}^\top + \widehat{\boldsymbol{\mathcal{D}}}, \tag{7}$$

where $\widehat{\boldsymbol{\mathcal{D}}}$ is the $n \times n$ diagonal matrix with diagonals $\mathcal{V}(\widehat{u}_{1t}), \dots, \mathcal{V}(\widehat{u}_{nt})$ and we arrive at the **TwO**-stage **P**rojection-based **E**stimator (**TOPE**) of $\boldsymbol{\beta}$, $\bar{\boldsymbol{\beta}} = (\mathbb{Z}_0^\top\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^\top\widehat{\mathbf{V}}^{-1}\bar{\boldsymbol{y}}$. The procedure is summarized in Algorithm 1 below.

Algorithm 1. TOPE (Two-stage projection-based estimator)

Input: Data $\{(y_{it}, \boldsymbol{x}_i, \mathbf{Z}_{it})\}_{i=1,t=1}^{n,T}$, predetermined K, and matrix of basis functions $\boldsymbol{\Phi}$.
Procedure:

1. For preliminary $\widehat{\boldsymbol{\beta}}^0$, compute $\widetilde{\mathbf{Y}} = \mathbf{Y} - \mathbb{Z}(\mathbf{I}_T \otimes \widehat{\boldsymbol{\beta}}^0)$.
2. First-stage: estimate $\mathbf{F}$ by letting the columns of $\widehat{\mathbf{F}}/\sqrt{T}$ be the eigenvectors corresponding to the first K largest eigenvalues of $\widetilde{\mathbf{Y}}^\top\mathbf{P}\widetilde{\mathbf{Y}}$ and estimate $\mathbf{G}$ by $\widehat{\mathbf{G}} = \mathbf{P}\widetilde{\mathbf{Y}}\widehat{\mathbf{F}}/T$.
3. Second-stage: compute $\widehat{\mathbf{V}} = \widehat{\mathbf{G}}\mathcal{V}(\widehat{\boldsymbol{f}}_t)\widehat{\mathbf{G}}^\top + \widehat{\boldsymbol{\mathcal{D}}}$ as in (7), where $\widehat{\boldsymbol{f}}_t$ is the tth row of $\widehat{\mathbf{F}}$ and $\widehat{\boldsymbol{u}}_t$ is the tth column of $\widehat{\mathbf{U}}$, and calculate **TOPE** $\bar{\boldsymbol{\beta}}$.

Output: $\widehat{\mathbf{F}}$, $\widehat{\mathbf{G}}$, $\bar{\boldsymbol{\beta}}$, and $\widehat{\mathbf{V}}$.

The choice of preliminary $\widehat{\boldsymbol{\beta}}^0$ for TOPE is quite flexible. Theoretically, the consistency and efficiency of TOPE are guaranteed whenever $\|\widehat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}\|_2 = O_P(n^{-1/2+\alpha}T^{-1/2})$ for $\alpha \in [0, 1/2)$, which is not difficult to acquire. Some concrete choices of $\widehat{\boldsymbol{\beta}}^0$ are discussed in depth in Section 8.3. Alternative to TOPE, one can first project $\mathbf{Y}$ using $(\mathbf{I}_n - \mathbf{P})$, where $\mathbf{P} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}$. This leads to $(\mathbf{I}_n - \mathbf{P})\mathbf{Y} = (\mathbf{I}_n - \mathbf{P})\mathbb{Z}(\mathbf{I}_T \otimes \boldsymbol{\beta}) + \mathbf{R}\mathbf{F}^\top + (\mathbf{I}_n - \mathbf{P})\mathbf{U}$, which is similar to the procedure of profile likelihood [16] or restricted maximum likelihood [24]. However, the validity of this approach relies on the assumption that $\mathbb{Z}$ and $\boldsymbol{\Phi}$ are linearly independent, which is more restricted than that of TOPE. Another seemingly straightforward approach is to project $\mathbf{Y}$ using $(\mathbf{I}_n - \widehat{\mathbf{P}}_{\mathbf{Z}})$ where $\widehat{\mathbf{P}}_{\mathbf{Z}} = \mathbb{Z}(\mathbb{Z}^\top\mathbb{Z})^{-1}\mathbb{Z}^\top$ and perform PCA on $(\mathbf{I}_n - \widehat{\mathbf{P}}_{\mathbf{Z}})\mathbf{Y} \approx (\mathbf{I}_n - \widehat{\mathbf{P}}_{\mathbf{Z}})\mathbf{G}\mathbf{F}^\top$ to estimate the loading and latent process. Though such an estimate of $\mathbf{F}$ remains consistent, as noted by Wang et al. [45],

this approach only identifies the part of the latent structure that is orthogonal to $\mathbb{Z}$. That is, one can only obtain a consistent estimate of $(\mathbf{I}_n - \widehat{\mathbf{P}}_{\mathbf{Z}})\mathbf{G}$, and in particular $\widehat{\mathbf{G}} + \widehat{\mathbf{P}}_{\mathbf{Z}}\mathbf{A}$ will also be a valid estimator for arbitrary $n \times K$ matrix $\mathbf{A}$.

3. Theoretical properties of TOPE

3.1. Preliminaries

We impose the following conditions on our model, in addition to Condition 1.

Condition 3. For any $\delta > 0$, $1 - n^{-1}\ln(1/\delta) \lesssim \lambda_{\min}(n^{-1}\mathbf{G}^\top\mathbf{G}) \le \lambda_{\max}(n^{-1}\mathbf{G}^\top\mathbf{G}) \lesssim 1 + n^{-1}\ln(1/\delta)$ with probability at least $1 - \delta$.

Condition 4. The density of $\boldsymbol{x}_i \in \mathcal{X}^d$, where $\mathcal{X} \subset \mathbb{R}$ is compact, is bounded away from zero and infinity.

Condition 5 (*Accuracy of the sieve approximation*)**.**

(i) For each $\ell \in \{1, \ldots, d\}$, $k \in \{1, \ldots, K\}$, the loading function $g_{k\ell}(\cdot)$ belongs to a Hölder class $\mathcal{G} = \{g : |g^{(r)}(s) - g^{(r)}(t)| \le L|s - t|^\gamma\}$ for some $L > 0$.
(ii) For $\kappa = 2(r + \gamma) \ge 4$, $\sup_{x\in\mathcal{X}} |g_{k\ell}(x) - \sum_{j=1}^J b_{k,j\ell}\phi_j(x)|^2 \lesssim J^{-\kappa}$.
(iii) It admits $\max_{k,j,\ell} b_{k,j\ell}^2 < \infty$.

Condition 3 is similar to the pervasive condition on loading matrix in the traditional factor model [38]. Since $\mathbf{G}\mathbf{G}^\top$ and $\mathbf{G}^\top\mathbf{G}$ have their first K largest eigenvalues in common, the K largest eigenvalues of $\mathbf{G}^\top\mathbf{G}$ also diverge in n. This condition ensures that $\boldsymbol{x}_i$ has non-vanishing explanatory power on loading so that $\mathbf{G}^\top\mathbf{G}$ has spiked eigenvalues. Condition 4 is standard in the literature of nonparametric and semiparametric statistics [22,23,39]. The accuracy of sieve approximation in Condition 5 can be obtained by common basis such as polynomial or B-splines [19,30].

Condition 6. For each $i \in \{1, \ldots, n\}$, z_{it} is weakly stationary. Almost surely, for each T we have,

(i) eigenvalues of $n^{-1}\mathbb{Z}_0^\top\mathbb{Z}_0$ are bounded away from 0 and infinity;
(ii) $\|\mathbf{P}_Z\mathbf{G}\|_{\mathbb{F}} = O(n^\alpha)$ for some $\alpha \in [0, 1/2)$, where $\mathbf{P}_Z$ is the projection matrix on $\mathbb{Z}_0$.

Condition 6(i) is similar to the standard condition on the design matrix in linear model that $\mathbb{Z}_0^\top\mathbb{Z}_0/n$ converges in n. Similar to conditions for semiparametric models in Robinson [35], (ii) guarantees identifications between the parametric and nonparametric parts in our model. Particularly, it allows consistent identification of the regression component without enforcing independence between $\boldsymbol{z}_{it}$ and $\boldsymbol{x}_i$. For instance, in Section 8.3, it is employed to show the existence of a legitimate preliminary estimator $\widehat{\boldsymbol{\beta}}^0$.

At last, we impose some widely-used conditions [2,38] regarding the serial dependence and stationarity on $\{\boldsymbol{f}_t, \boldsymbol{u}_t\}$ as well as their tail behavior. Denote $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ the σ-algebra generated by $\{(\boldsymbol{f}_t, \boldsymbol{u}_t) : t \le 0\}$ and $\{(\boldsymbol{f}_t, \boldsymbol{u}_t) : t \ge T\}$, and recall the α-mixing coefficient as $\alpha(T) = \sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_T^\infty} |\Pr(A)\Pr(B) - \Pr(A \cap B)|$.

Condition 7 (*Serial dependence, stationarity, and tail behavior*)**.**

(i) $\{\boldsymbol{u}_t, \boldsymbol{f}_t\}_{t\le T}$ are strictly stationary with zero mean and finite long run variances.
(ii) There exist $r_1, C_1 > 0$ such that for all $T > 0$, $\alpha(T) < \exp(-C_1T^{r_1})$.
(iii) There exist $r_2, r_3 > 1$ with $r_1^{-1} + r_2^{-1} + r_3^{-1} > 1$ and $b_1, b_2 > 0$ such that for each i, k, t and any $s > 0$, $\Pr(|u_{it}| > s) \le \exp\{-(s/b_1)^{r_2}\}$ and $\Pr(|f_{tk}| > s) \le \exp\{-(s/b_2)^{r_3}\}$.

3.2. Statistical guarantees

To establish the statistical guarantees of TOPE, we first perform a non-asymptotic analysis of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{G}}$, then obtain the deviation between $\bar{\boldsymbol{\beta}}$ and the GLS estimator $\widetilde{\boldsymbol{\beta}}$ to study the efficiency of TOPE.

Theorem 1. *Suppose that Conditions 1, 2, and 3–7 hold. Assume $Jd + 1 < n$ and $J = o(n^{1/2-\alpha})$ for $\alpha \in [0, 1/2)$. With probability at least $1 - \delta$ for any $\delta > 0$, we have*

$$\frac{1}{T}\|\widehat{\boldsymbol{F}} - \boldsymbol{F}\|_{\mathbb{F}}^2 \lesssim \left(\frac{1}{n} + \frac{p^2}{n^{1-2\alpha}T} + \frac{1}{J^\kappa}\right)\ln(1/\delta),$$
$$\frac{1}{n}\|\widehat{\boldsymbol{G}} - \boldsymbol{G}\|_{\mathbb{F}}^2 \lesssim \left(\frac{J}{n^2} + \frac{p^2J}{n^{1-2\alpha}T} + \frac{p^4J}{n^{2-4\alpha}T^2} + \frac{1}{J^{\kappa-1}}\right)\{\ln(1/\delta)\}^2,$$
$$\|\widehat{\boldsymbol{B}} - \boldsymbol{B}\|_{\mathbb{F}}^2 \lesssim \left(\frac{J}{n^2} + \frac{p^2J}{n^{1-2\alpha}T} + \frac{p^4J}{n^{2-4\alpha}T^2} + \frac{1}{J^{\kappa-1}}\right)\{\ln(1/\delta)\}^2.$$

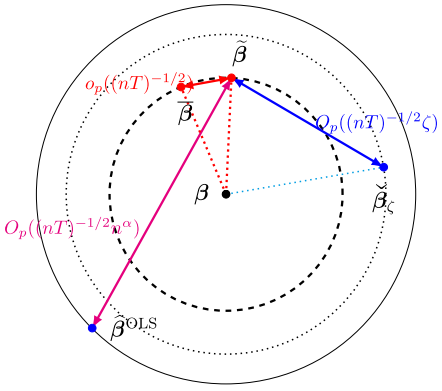


Fig. 1. A schematic illustration about different estimators to $\boldsymbol{\beta}$ in (1), where $\bar{\boldsymbol{\beta}}$ is the TOPE estimator and $\widetilde{\boldsymbol{\beta}}$ is the oracle GLS estimator with full knowledge on $\mathbf{G}$ and $\mathbf{F}$.

In contrast to the known asymptotic properties of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{G}}$ for traditional and semiparametric factor models with divergent n and T [5,19], Theorem 1 provides finite sample characterization of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{G}}$. Given a finite p, the rates obtained in Theorem 1 agree with the asymptotic results in Fan et al. [19]. Also, whenever $p = o(n^{1/2-\alpha}T^{1/2}J^{-1/2})$, $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{G}}$ are consistent in mean squared errors. Especially, for finite p, this consistency does not require T diverging to infinity which enables our method to be used for modeling a large number of short time series in practice. More importantly, results in Theorem 1 make it possible to establish the following finite sample results on both $\bar{\boldsymbol{\beta}}$ and its covariance with respect to the GLS estimator $\widetilde{\boldsymbol{\beta}}$ as defined in (6).

Theorem 2. *Under conditions in Theorem 1, with probability at least* $1-\delta$,

$$\|\bar{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2 \lesssim \frac{1}{\sqrt{nT}}\left\{\frac{\sqrt{J}}{n} + \frac{1}{\sqrt{n}} + \frac{1}{T} + \frac{p\sqrt{J}}{\sqrt{n^{1-2\alpha}T}} + \frac{1}{J^{(\kappa-1)/2}}\right\}\sqrt{\ln(1/\delta)},$$

where $\widetilde{\boldsymbol{\beta}}$ *is the oracle GLS estimator of* $\boldsymbol{\beta}$ *with full knowledge of* $\mathbf{G}$ *and* $\mathbf{F}$ *as in Section 2.2.2. In addition,*

$$\left\|\mathrm{Var}(\bar{\boldsymbol{\beta}}) - \mathrm{Var}(\widetilde{\boldsymbol{\beta}})\right\|_{\mathbb{F}} \lesssim \frac{p\vartheta_{n,T,J}}{nT} + \frac{p\vartheta_{n,T,J}^2}{(nT)^{3/2}},$$

where $\vartheta_{n,T,J} = n^{-1}J^{1/2} + n^{-1/2} + T^{-1} + pJ^{1/2}n^{-1/2+\alpha}T^{-1/2} + J^{-(\kappa-1)/2}$.

The nontrivial finite sample results in Theorem 2 imply that the deviation between $\bar{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ is due to: (i) the errors in estimating $\mathbf{G}$ with rate $n^{-1/2}T^{-1/2}(n^{-1}J^{1/2} + pJ^{1/2}n^{-1/2+\alpha}T^{-1/2} + J^{-(\kappa-1)/2})$, (ii) the errors in estimating $\mathbf{F}$ with rate $n^{-1/2}T^{-1/2}(n^{-1/2} + T^{-1} + pn^{-1/2+\alpha}T^{-1/2} + J^{-\kappa/2})$, and (iii) the deviation between $\mathcal{V}(\boldsymbol{f}_t)$ and $\mathrm{Var}(T^{-1}\sum_{t=1}^T \boldsymbol{f}_t)$ with rate $n^{-1/2}T^{-3/2}$.

Let $\|\mathbf{A}\|_{\mathbf{S}} := n^{-1/2}\|\mathbf{S}^{-1/2}\mathbf{A}\mathbf{S}^{-1/2}\|_2$ for any positive definite $\mathbf{S}$ and define the class of estimators to $\boldsymbol{\beta}$ with respect to working covariance $\mathbf{V}_\zeta$ by $\Theta_\zeta = \{\check{\boldsymbol{\beta}}_\zeta = (\mathbb{Z}_0^\top\mathbf{V}_\zeta^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^\top\mathbf{V}_\zeta^{-1}\bar{\boldsymbol{y}} : \|\mathbf{V}_\zeta - \mathbf{V}\|_{\mathbf{V}} \lesssim \zeta\}$, where the oracle GLS estimator $\widetilde{\boldsymbol{\beta}} \in \Theta_0$, TOPE $\bar{\boldsymbol{\beta}} \in \Theta_{\vartheta_{n,T,J}}$ by Theorem 1, and OLS estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} \in \Theta_{n^\alpha}$ by Proposition 5. From the proof of Theorem 2, $\|\check{\boldsymbol{\beta}}_\zeta - \widetilde{\boldsymbol{\beta}}\|_2 = O_p(n^{-1/2}T^{-1/2}\zeta)$ for any $\check{\boldsymbol{\beta}}_\zeta \in \Theta_\zeta$. Thus, $\|\check{\boldsymbol{\beta}}_\zeta - \widetilde{\boldsymbol{\beta}}\|_2 = O_p(n^{-1/2}T^{-1/2})$ if $\zeta = O(1)$ and $\|\check{\boldsymbol{\beta}}_\zeta - \widetilde{\boldsymbol{\beta}}\|_2 = o_p(n^{-1/2}T^{-1/2})$ if $\zeta = o(1)$. In the presence of heteroscedasticity across subjects and/or autocorrelation, the oracle GLS estimator is known to be efficient in general [6]. Particularly, the GLS estimator $\widetilde{\boldsymbol{\beta}}$ is unbiased and efficient in Θ_ζ given the full information on $\mathbf{G}$ and $\boldsymbol{\Sigma}_f(t)$ for each $t \in \{1-T, \ldots, T-1\}$. Therefore, Theorem 2 implies that TOPE $\bar{\boldsymbol{\beta}}$ is asymptotically unbiased, and given $p\vartheta_{n,T,J} = o(1)$, the non-asymptotic difference between the variances of $\bar{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ is bounded by a rate smaller than $(nT)^{-1}$, which is the rate of $\mathrm{Var}(\widetilde{\boldsymbol{\beta}})$. That is, the TOPE $\bar{\boldsymbol{\beta}}$ is asymptotically efficient in Θ_ζ. This discussion is visualized in Fig. 1.

Remark 1. Though Theorems 1 and 2 are implicitly related to a legitimate preliminary estimator $\widehat{\boldsymbol{\beta}}^0$ via α, the existence of such a $\widehat{\boldsymbol{\beta}}^0$ is easily guaranteed under Conditions 1, 2, and 6. In fact, as shown in Proposition 5, the OLS estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ automatically satisfies $\|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}\| = O_p(n^{-1/2+\alpha}T^{-1/2})$ and is therefore legitimate. Technically, $\widehat{\boldsymbol{\beta}}^0$ contributes to Theorem 2 through $\vartheta_{n,T,J}$, more specifically via $p\sqrt{n^{-1+2\alpha}T^{-1}J}$. For a better $\widehat{\boldsymbol{\beta}}^0$ with decreasing α, its contribution through $\vartheta_{n,T,J}$ diminishes and achieves $p\sqrt{n^{-1/2}T^{-1}}$ when $\alpha \approx 0$. On the contrary, when $\alpha = 1/2 - 2c$ for any small $c > 0$, $\vartheta_{n,T,J} = o(n^{-c(\kappa-1)} + pn^{-c}T^{-1/2})$ and does not alter the conclusion in Theorem 2. Hence, measured by α, $\widehat{\boldsymbol{\beta}}^0$ only affects on how $\vartheta_{n,T,J} = o(1)$ converges to zero as n and T increase, and it does not alter the conclusion about the efficiency of TOPE $\bar{\boldsymbol{\beta}}$ as long as $\alpha \in [0, 1/2)$. These are also demonstrated by numerical studies in Section A.4 in the supplement.

As a final remark, Theorem 3 establishes results analogous to Theorem 1 in the max norm and shares common observations with Wang and Fan [43] and Barigozzi et al. [7].

Theorem 3. *For model (3), under the same conditions of Theorem 1, with probability at least* $1-\delta$, $\|\widehat{\boldsymbol{F}}-\boldsymbol{F}\|_{\max} \lesssim \{n^{-1/2}+p^2(n^{1-\alpha}T)^{-1/2}+J^{-\kappa/2}\}\{\ln(T)\}^{2/r_3}\ln(1/\delta)$, $\|\widehat{\boldsymbol{G}}-\boldsymbol{G}\|_{\max} \lesssim [\{(T+p^2n^\alpha)\ln(n)T^{-2}\}^{1/2}+n^{-1/2}+p^2(n^{1-\alpha}T)^{-1/2}+J^{-\kappa/2}]\ln(1/\delta)$, *and* $\|\widehat{\boldsymbol{B}}-\boldsymbol{B}\|_{\max} \lesssim [\{(T+p^2n^\alpha)\ln(n)n^{-1}T^{-2}\}^{1/2}+n^{-1/2}+p^2(n^{1-\alpha}T)^{-1/2}+J^{-\kappa/2}]\ln(1/\delta)$.

4. TOPE-based inference

The following theorem provides inference on $\boldsymbol{\beta}$ based on TOPE. In general, the expectation with respect to $\{\mathbf{Z}_t, \boldsymbol{x}_i\}$ is unknown as their distribution is not accessible, therefore the asymptotic distribution of $\bar{\boldsymbol{\beta}}$ conditional on $\{\mathbf{Z}_t, \boldsymbol{x}_i\}$ in (2) below is more practical.

Theorem 4. *Under conditions in Theorem 1, we have*

(i) *with* $\boldsymbol{\Sigma} = \mathrm{E}_{\mathbf{Z}_t,\boldsymbol{X}}\{(\mathbb{Z}_0^\top \boldsymbol{V}^{-1}\mathbb{Z}_0)^{-1}\}$, $\boldsymbol{\Sigma}^{-1/2}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I}_p)$;

(ii) *conditional on* $\boldsymbol{Z}_t$ *and* $\boldsymbol{x}_i$, $(\mathbb{Z}_0^\top \boldsymbol{V}^{-1}\mathbb{Z}_0)^{1/2}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I}_p)$.

Replacing $\mathbf{V}$ in Theorem 4 (ii) by $\widehat{\mathbf{V}} = \widehat{\mathbf{G}}\mathcal{V}(\widehat{\boldsymbol{f}}_t)\widehat{\mathbf{G}}^\top + \widehat{\mathcal{D}}$ from (7), for any estimable $\mathbf{C}\boldsymbol{\beta}$ with $q \times p$ matrix $\mathbf{C}$ and $q < p$, $\mathrm{CS}_{\mathbf{C}} = \{\mathbf{C}\boldsymbol{\beta} : (\mathbf{C}\boldsymbol{\beta}-\mathbf{C}\bar{\boldsymbol{\beta}})^\top\{\mathbf{C}(\mathbb{Z}_0^\top\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}\mathbf{C}^\top\}^{-1}(\mathbf{C}\boldsymbol{\beta}-\mathbf{C}\bar{\boldsymbol{\beta}}) < \chi^2_{q,1-\eta}\}$ defines a $100(1-\eta)\%$ confidence set, where $\chi^2_{q,1-\eta}$ is the $100(1-\eta)\%$ quantile of χ^2_q distribution. When rows of $\mathbf{C}$ are the natural basis of $\mathbb{R}^p$, $\mathrm{CS}_{\mathbf{C}}$ provides a confidence set of a subset of $\boldsymbol{\beta}$. Alternatively, denote $\widehat{\sigma}^2_\ell$ the ℓth diagonal entry of $(\mathbb{Z}_0^\top\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}$ for each $\ell \in \{1,\ldots,p\}$, a $100(1-\eta)\%$ confidence interval for the ℓth entry of $\boldsymbol{\beta}$ is $\mathrm{CI}_\ell = [\widehat{\beta}_\ell - \widehat{\sigma}_\ell\Phi^{-1}(1-\eta/2), \widehat{\beta}_\ell + \widehat{\sigma}_\ell\Phi^{-1}(1-\eta/2)]$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. Moreover, Theorem 4 implies that $\Pr(\|\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_\infty > \varepsilon) < p\exp(-\varepsilon^2 p^{-1}\sigma^{-2})$, where σ^2 can be estimated by the minimum diagonal of $(\mathbb{Z}_0^\top\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}$. Thus, it also leads to a uniform confidence set for $\boldsymbol{\beta}$ at level $100(1-\eta)\%$, denoted as $\mathrm{CI}' = \{\boldsymbol{\beta} : |\beta_\ell - \widehat{\beta}_\ell| \le \widehat{\sigma}\sqrt{p\ln(p/\eta)},\ \ell \in \{1,\ldots,p\}\}$.

To draw inference on the explaining power of covariates $\boldsymbol{x}_i$ on the dependence structure of data, Fan et al. [19] proposed a semiparametric specification testing statistic $\mathrm{S}_G = \mathrm{tr}\{(\widetilde{\mathbf{F}}^\top\widetilde{\mathbf{Y}}^\top\widetilde{\mathbf{Y}}\widetilde{\mathbf{F}})^{-1}\widetilde{\mathbf{F}}^\top\widetilde{\mathbf{Y}}^\top\mathbf{P}\widetilde{\mathbf{Y}}\widetilde{\mathbf{F}}\}$, where $\widetilde{\mathbf{F}}/\sqrt{T}$ are the eigenvectors corresponding to the K largest eigenvalues of $\widetilde{\mathbf{Y}}^\top\widetilde{\mathbf{Y}}$. In addition to Condition 3–7, assuming $T^{2/3} = o(n)$, $n\{\ln(n)\}^4 = o(T^2)$, $J = o(\min\{n^{1/2-\alpha}, \sqrt{T}\})$, and $\max\{T\sqrt{n}, n\} = o(J^\kappa)$, we have $(n\mathrm{S}_G - JdK)(2JdK)^{-1/2} \xrightarrow{d} N(0,1)$ whenever $\mathbf{G}(\mathbf{X}) = \mathbf{0}$. Thus, we can test $H_0 : \mathbf{G}(\mathbf{X}) = \mathbf{0}$ almost surely. Hence, S_G provides a diagnostic tool for the proposed model.

5. Numerical studies

5.1. Simulation settings

We demonstrate the finite sample performance of TOPE for both estimation and inference in comparison to three competing methods: the OLS estimator, which ignores heteroscedasticity across subjects and dependence; the GLS estimator, which naively utilizes the first K components of $T^{-1}\sum_{t=1}^T(\boldsymbol{y}_t-\bar{\boldsymbol{y}})(\boldsymbol{y}_t-\bar{\boldsymbol{y}})^\top$ as $\widehat{\mathbf{V}}$; and last, the oracle estimator, which is TOPE with known $\mathbf{G}$ without using approximation. To implement TOPE, we employ the OLS estimator as preliminary $\widehat{\boldsymbol{\beta}}^0$.

The mean squared error (MSE) and the empirical coverage probability (ECP) of the confidence region for $\boldsymbol{\beta}$ are employed to compare different methods. In addition, $\|\widehat{\mathbf{F}}-\mathbf{F}\|_{\mathrm{F}}/\sqrt{T}$ and $\|\widehat{\mathbf{G}}-\mathbf{G}\|_{\mathrm{F}}/\sqrt{n}$ are displayed to demonstrate estimations on $\mathbf{G}$ and $\mathbf{F}$ by TOPE. The maximum marginal length of the confidence set (MML) is used to demonstrate the efficiency. That is, the confidence set with ECP agreeing to the nominal level and small MML is preferable. For a clear presentation, we display MML of different methods normalized by the largest one (the MML of OLS, in general).

We consider $n \in \{50, 100, 200, 500, 1000, 2000\}$ and $T \in \{20, 50, 100, 200, 500\}$; also, we set $p = 4$ with $\boldsymbol{\beta} = (1,1,1,1)^\top$ and generate *i.i.d.* $z_{i\ell,t} \sim N(3\exp(t/30), 1)$ for each $i \in \{1,\ldots,n\}$, $\ell \in \{1,\ldots,p\}$, and $t \in \{1,\ldots,T\}$. A similar setting was used in Huang et al. [23]. For the loading, we set $d = 3$ and generate *i.i.d.* $\boldsymbol{x}_i \sim U([0,1]^d)$, then let $g_1(\boldsymbol{x}) = x_1$, $g_2(\boldsymbol{x}) = x_1^2 + x_2^2 - 1$, and $g_3(\boldsymbol{x}) = x_3^2 - 2x_1 + x_2$ for $K = 3$. As suggested by [19], with the initial realization $\mathbf{G}_0$ for g_1, g_2 and g_3, we further compute $\mathbf{H}_{\mathbf{G}} = \mathbf{G}_0^\top\mathbf{G}_0$ and set $\mathbf{G} = \mathbf{G}_0\mathbf{H}_{\mathbf{G}}$ in simulations so that Condition 2 is satisfied.

The latent process $\boldsymbol{f}_t = (f_{1t}, f_{2t}, f_{3t})^\top$ consists of $K = 3$ independent univariate time series governed by the same model with one of the following three settings: independent in t, AR(1) with autoregressive coefficient $\rho = 0.5$, or ARMA(1, 1) with autoregressive coefficient $\rho = 0.5$ and moving average coefficient $\theta = 0.5$. Also, three innovations are considered: the standard normal, centered χ^2_5, and t_8. After generating $\boldsymbol{f}_t$, $\mathbf{F} = (\boldsymbol{f}_1,\ldots,\boldsymbol{f}_T)^\top$ is further transformed so that $T^{-1}\mathbf{F}^\top\mathbf{F} = \mathbf{I}_K$ in Condition 2 is satisfied. Similar to $\boldsymbol{f}_t$, we generate n independent $\boldsymbol{u}_i$ from the same model, which includes two dependence structures: independent in t and AR(1) with autoregressive coefficient $\rho = 0.5$, as well as two innovations: $N(0, 0.01)$ and $(\chi^2_5 - 5)/10$. For each setting, 500 simulations are conducted.

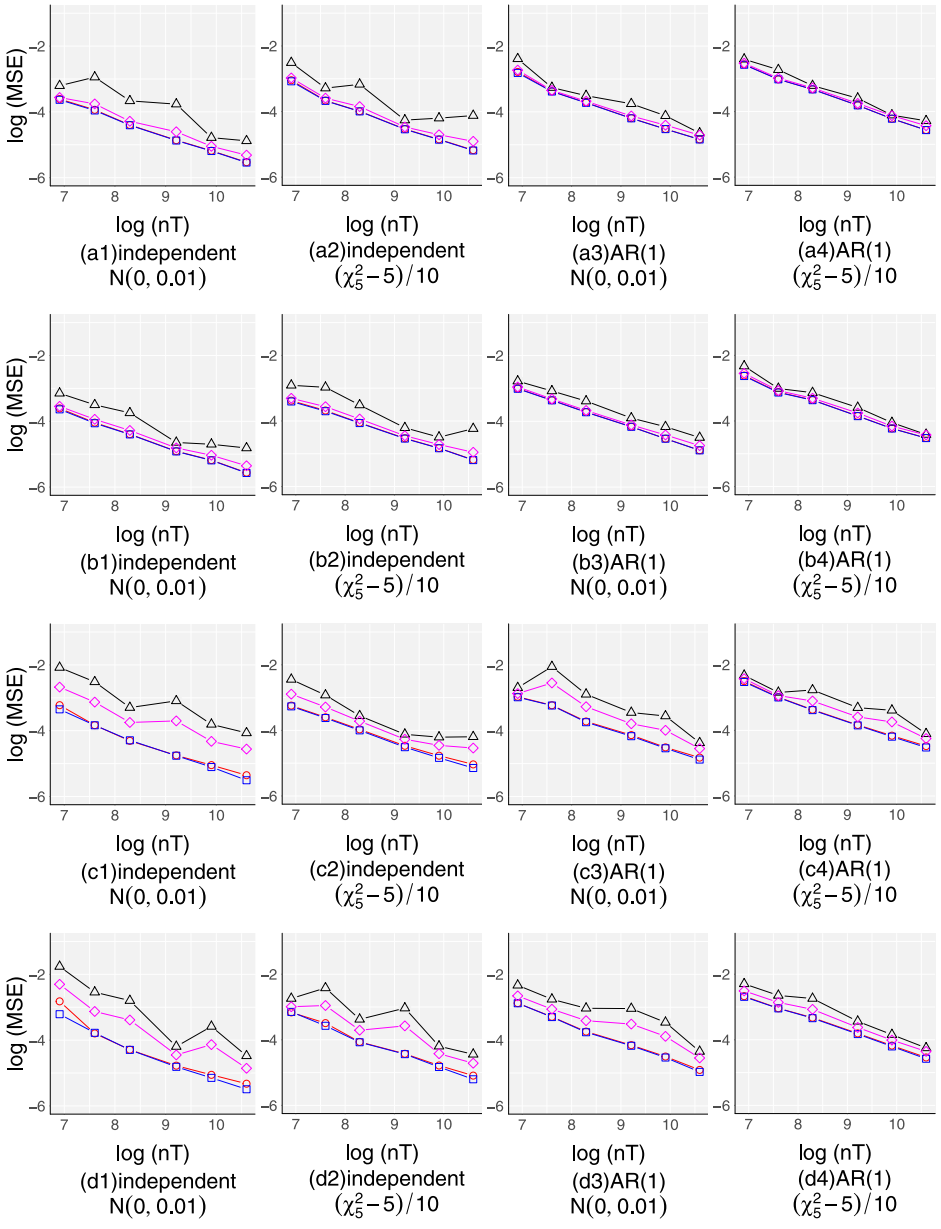


Fig. 2. Comparisons of the logarithm of MSE for estimating $\boldsymbol{\beta}$ by TOPE ("–○–") along those of the oracle estimator ("–□–"), the GLS estimator ("–◇–"), and the OLS ("–△–"). Results are about $T = 20$. In plots (a1)–(a4), $f_{kt} \sim N(0, 1)$ is independent in k, t. In plots (b1)–(b4), $f_{kt} \sim t_8$ is independent in k, t. In plots (c1)–(c4), f_{kt} follows the ARMA(1, 1) model with $N(0, 1)$ innovation for each $k = 1, 2, 3$. In plots (d1)–(d4), f_{kt} follows the ARMA(1, 1) model with t_8 innovation for each $k = 1, 2, 3$. Distributions and serial correlations of $\boldsymbol{u}_i$ are displayed in the plots.

5.2. Results

Fig. 2 displays the MSE with respect to $\ln(nT)$ on the logarithm scale when $T = 20$ and $\boldsymbol{f}_t$ is independent in t or follows the ARMA(1, 1) model with $\mathcal{N}(0, 1)$ or t_8 innovations. In Fig. 2, the MSEs of all estimators reduce as n increases. Both TOPE and GLS perform similarly as the oracle estimator when $\boldsymbol{f}_t$ is independent in t ((a1)–(a4) and (b1)–(b4)), and outperform OLS; on the other hand, temporal dependence in $\boldsymbol{u}_t$ slightly increases the MSE but does not alter the convergence rate ((c1)–(c4) and (d1)–(d4)). In the presence of temporal dependence in $\boldsymbol{f}_t$, GLS is outperformed while TOPE's performance remains comparable to the oracle estimator ((c1)–(c4) and (d1)–(d4)). In the supplementary files, additional results for settings similar to Fig. 2 but with $T = 100, 500$ are reported in Figs. S.1–S.4, and results for $\boldsymbol{f}_t$ following the AR(1) model

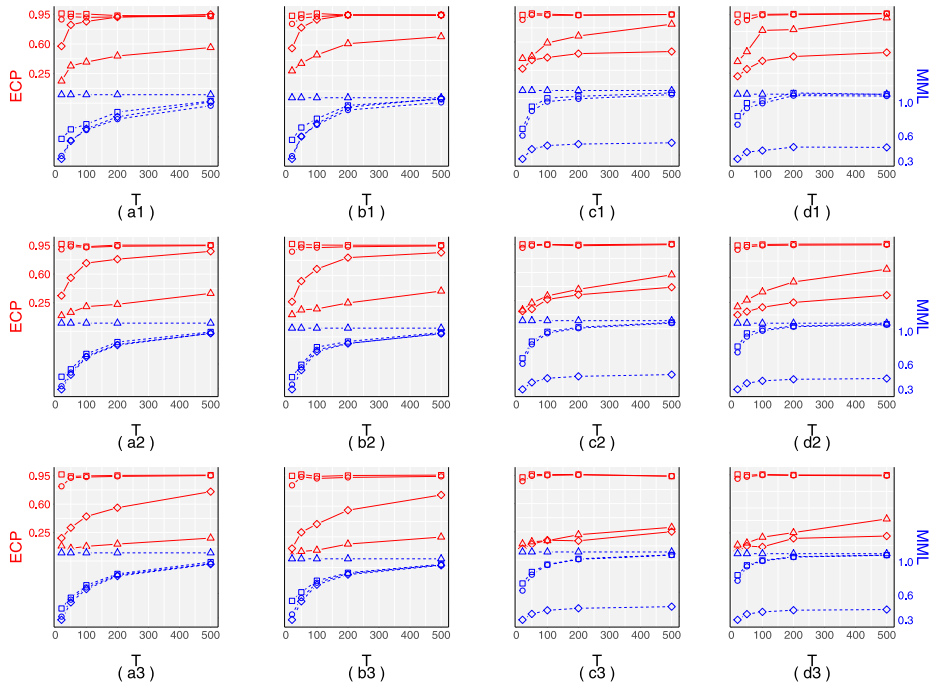


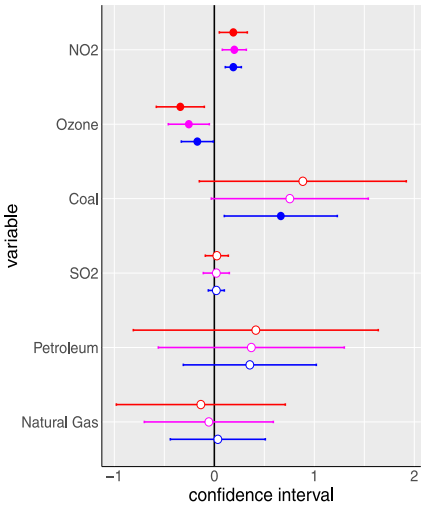
Fig. 3. Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator ("–□–" for ECP and "- -□- -" for MML), the GLS estimator ("–◇–" for ECP and "- -◇- -" for MML), and the OLS ("–△–" for ECP and "- -△- -" for MML). In simulations, $f_{kt} \sim N(0, 1)$ is independent in k, t; $n = 100, 500, 2000$ for the first, second, and third row, respectively. In plots (a1)–(a3) $u_{it} \sim N(0, 0.01)$ is independent in i, t. In plots (b1)–(b3), $u_{it} \sim (\chi^2_5 - 5)/10$ is independent in i, t. In plots (c1)–(c3) $\boldsymbol{u}_i$ follows the AR(1) model with $N(0, 0.01)$ innovation while the same model is used for $\boldsymbol{u}_i$ in plots (d1)–(d3) with $(\chi^2_5 - 5)/10$ innovation.

are displayed in Figs. S.5 and S.7. Similar observations are obtained for different settings of $\boldsymbol{f}_t$, and the differences among estimators decrease as T increases.

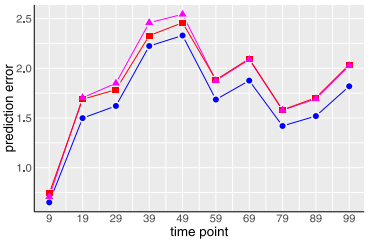
Fig. 3 displays the ECP and MML with respect to different T and n. The nominal level is 0.95. In Fig. 3, the confidence region of TOPE has ECP close to the nominal level with a small MML. Meanwhile, the coverage probabilities of OLS and GLS are both deviated from the nominal level and the deviations are substantial when n increases. In the presence of temporal dependence in $\boldsymbol{u}_t$, TOPE still outperforms GLS and OLS. The MML of TOPE substantially improves when n increases, particularly for large T, which reflects the fact that the estimation of $\mathbf{F}$ in TOPE prefers large n (see (c1) and (c2), (d1)–(d2) in Fig. 3 for example). In the presence of the dependence of $\boldsymbol{f}_t$ in t, TOPE performs remarkably well in terms of maintaining small MML and its ECP quickly converges to the nominal level in T (for example, Figs. S.17 and S.18 in the supplementary file). Meanwhile, due to the heteroscedasticity across subjects and the serial/cross-sectional correlations, both GLS and OLS fail to maintain the nominal coverage probability. As MML reflects the largest marginal variance of an estimator, OLS has large marginal variance in the presence of serial correlations in $\boldsymbol{u}_t$ (Fig. 3 (d1)–(d3)). However, the ECP of OLS substantially deviates from the nominal level, which reflects the inconsistent covariance estimate of OLS. Also, it is interesting to notice that both the ECP and MML of GLS are smaller than those of others (Fig. 3 (c1)–(c3) and (d1)–(d3)), which shows that the naive GLS tends to ignore the serial correlations and greatly underestimate the variance that results in the poor confidence sets with low ECP. More simulation results are retained in the supplementary files and provide similar observations. Specifically, Figs. S.9 and S.12 in the supplementary files display results for f_{kt} independent in k, t with either independent u_{it} in i, t or $\boldsymbol{u}_i$ following the AR(1) model with different innovations. Results for $\boldsymbol{f}_t$ following the AR(1) model or the ARMA(1, 1) model with different innovations are included in Figs. S.13–S.24 in the supplementary files.

6. Study on air quality and energy consumption data using TOPE

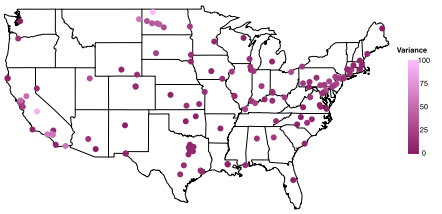
In this section, we apply our method to analyze air quality data collected in the United States in 2015. The data consists of the mean PM2.5 concentration (in $\mu g/m^3$) from 129 monitoring sites on each Tuesday and Thursday in 2015 (https://www.epa.gov/outdoor-air-quality-data). We also include daily max 1-hour concentration of three common air pollutants, including NO_2, SO_2, and ozone, and the latitude and longitude of each monitoring site in our analysis. Sources of energy consumption are known as a potential factor to affect concentration of air pollutants. For this study, as covariates, we include the annual state-level energy consumption proportions of three major sources out of all possible resources,



(a) The 95% confidence intervals (the OLS estimator in red, the estimator from traditional factor model in magenta, and TOPE in blue) of the effects of energy consumption proportions of coal, natural gas, and petroleum and daily max 1-hour concentration of NO_2, SO_2, and ozone on the PM2.5 concentration.



(b) The prediction errors for TOPE ("– ● –"), the OLS estimator ("– ■ –"), and the estimator from traditional factor model ("– ▲ –") at the testing points.



(c) Variance of the mean PM2.5 concentration (over all time points) at 129 monitoring sites across the United States.

Fig. 4. Data display, resulting confidence intervals, and prediction comparison for the real data analysis. Panel (a), (b), and (c) are the 95% confidence intervals of the effects of different explanatory variables on the PM2.5 concentrations, the prediction errors for different methods, and the variance of the mean PM2.5 concentration across different monitoring sites, respectively.

namely coal, natural gas, and petroleum, in 2015 (https://www.eia.gov/electricity/data/browser/). For analysis, we log-transform the air pollutant data and remove potential seasonality. Also, we transform the latitude and longitude to keep their values within [0, 1].

From Figs. 4(c) and S.29 in the supplement, it is observed that both geographical variables and energy consumption proportions help explaining the observed heteroscedasticity across monitoring sites so that we consider them as $\boldsymbol{x}_i$ in (1). In this analysis, the daily max 1-hour concentration of NO_2, SO_2, and ozone, as well as the energy consumption proportions of coal, natural gas, and petroleum are considered as $\boldsymbol{z}_{it}$ in (1).

To determine the dimension K of latent factor process, we apply both the eigenvalue-ratio procedure and the HDWN testing-based procedure proposed in the supplementary files. Ratios of the first ten adjacent eigenvalues of $\widetilde{\mathbf{Y}}^\top \mathbf{P}\widetilde{\mathbf{Y}}$ are 4.13, 5.26, 6.58, 1.27, 1.68, 1.17, 1.29, 1.21, 1.10 such that the ratio between the third and fourth eigenvalues are the largest. On the other hand, for the HDWN testing-based procedure, the p-values for testing (B.1) with $K_0 = 1, 2$ and 3 are 0.026, 0.040 and 0.104, respectively. That is, we reject $H_0(1)$ and $H_0(2)$ but fail to reject $H_0(3)$ for (B.1). Thus, both eigenvalue-ratio procedure and the proposed HDWN testing-based procedure suggest $\widehat{K} = 3$. Also, by the procedure discussed at the end in Section 4, we test $H_0 : \mathbf{G}(\mathbf{X}) = \mathbf{0}$ to further explore the statistical evidence to include geographical variables and energy consumption proportions to explain the heteroscedasticity across monitoring sites. We obtain $S_G = 2.34$ with p-value 4.84×10^{-14}; thus, these covariates are included for modeling. Then, the complete model in the form of (1) for performing analysis on this data is

$$\begin{aligned}\ln(\mathrm{PM2.5}_{it}) = {} & \beta_1 \ln(\mathrm{NO}_{2,it}) + \beta_2 \ln(\mathrm{SO}_{2,it}) + \beta_3 \ln(\mathrm{Oz}_{it}) + \beta_4 \mathrm{Cl}_i + \beta_5 \mathrm{Ng}_i + \beta_6 \mathrm{Pe}_i \\ & + \sum_{k=1}^{3} \left\{ g_{k1}(\mathrm{La}_i) + g_{k2}(\mathrm{Lo}_i) + g_{k3}(\mathrm{Cl}_i) + g_{k4}(\mathrm{Ng}_i) + g_{k5}(\mathrm{Pe}_i) \right\} f_{kt} + u_{it},\end{aligned}$$

where $\ln(\mathrm{PM2.5}_{it})$ is the log concentration of PM2.5 from the monitoring site i at time t; $\ln(\mathrm{NO}_{2,it})$, $\ln(\mathrm{SO}_{2,it})$, and $\ln(\mathrm{Oz}_{it})$ are the log daily max 1-hour concentration of NO_2, SO_2, and ozone, respectively, from the same monitoring site i at time t; Cl_i, Ng_i and Pe_i are the state-level energy consumption proportions of coal, natural gas, and petroleum out of all possible energy resources for the monitoring site i, respectively; and La_i and Lo_i are the latitude and longitude of the monitor site i, respectively.

For $g_{k\ell}$ in the above model, $\ell \in \{1, \ldots, 5\}$, we use cubic spline with 11 knots to construct $\boldsymbol{\Phi}$ for projection. We fit the above model using TOPE and draw inference as proposed in Section 4 to inspect the effects of covariates on the PM2.5 concentration. As an expected advantage, no further restrictions need to be imposed to model (1) and TOPE. In Fig. 4(a), the 95% confidence intervals for estimated coefficients using TOPE, the OLS estimator (by ignoring the variance

components), and the traditional factor model (using $\mathbf{A}$ instead of $\boldsymbol{g}(\boldsymbol{x}_i)$ in the above model) are displayed for comparison. It reflects the efficiency of TOPE in the presence of heteroscedasticity across monitoring sites and serial/contemporaneous correlations discussed in Section 3. Specifically, the confidence intervals constructed by TOPE are the shortest among all three methods for all the six covariates. All methods suggest significant positive correlation between daily max 1-hour concentration of NO_2 and PM2.5 concentration and significant negative correlation between ozone concentration and PM2.5 concentration. TOPE reveals a significant positive correlation between coal consumption and PM2.5 concentration, which agrees with Liang et al. [28] that coal consumption positively contributes to PM2.5 concentration. However, this is missed by both the OLS estimator and the traditional factor model. In Fig. S.30 in the supplement, the recovered $\boldsymbol{g}_k$ for $k \in \{1, 2, 3\}$ displays clear non-linearity.

In addition, we examine the prediction performance of TOPE, the OLS estimator, and the traditional factor model. For the 104 time points, we select 10 (from the 9th to the 99th, apart by 10 points) as the testing set and train the aforementioned three models using the remaining data points. With the estimated $\boldsymbol{\beta}$ from each method, the squared prediction errors at the testing points are displayed in Fig. 4(b). Compared with OLS and the traditional factor model, our model, alone with TOPE, has smaller prediction errors across all testing points, which demonstrates its superior prediction accuracy.

7. Discussions

Methodologically, we propose a flexible subject-specific heteroscedasticity model with latent semiparametric factor structures for analyzing large scale data with both intertemporal and intratemporal dependence. The model simultaneously accounts for the heteroscedasticity across subjects as well as the contemporaneous and serial correlations. We advocate a two-stage projection-based estimator for both the modulating and dependence components of the model, and establish an inference procedure for regression coefficients. We study the non-asymptotic rates for recovering the latent factor process and estimating the nonparametric loading function, which leads to the non-asymptotic properties of the estimated regression coefficients. As a result, we show that our proposed TOPE is asymptotically efficient within a fairly broad class of estimators including both the OLS and naive GLS estimators.

The widely-used Condition 2 essentially restricts $\mathbf{F}$ to subspace $\{\mathbf{F} \in \mathbb{R}^{T\times K} : T^{-1}\mathbf{F}^\top\mathbf{F} = \mathbf{I}_K\}$, which might be stringent for some applications. In fact, we notice that it can be greatly relaxed by a concentration assumption of $T^{-1}\mathbf{F}^\top\mathbf{F}$ to $\mathbf{I}_K$, which can be derived from Condition 7 with the help of the so-called τ-mixing coefficient. As a result, this will alter the convergence rate of $\widehat{\mathbf{F}} - \mathbf{F}$ in Theorem 1. Furthermore, as noted after Condition 2, we assume that the residual process u_{it} is uncorrelated over i to establish the statistical guarantee of TOPE on estimating $\boldsymbol{\beta}$. This condition is similar to that of the traditional PCA that assumes uncorrelated samples. It can be further relaxed to, for example, $\max_{j\le n}\sum_{i=1}^n |\mathrm{E}(u_{it}u_{jt})| < C_2$, $\max_{i\le n}\sum_{k=1}^n\sum_{m=1}^n\sum_{t=1}^T\sum_{s=1}^T |\mathrm{cov}(u_{it}u_{kt}, u_{is}u_{ms})| < C_2$, and $(nT)^{-1}\sum_{i=1}^n\sum_{j=1}^n\sum_{t=1}^T\sum_{s=1}^T |\mathrm{E}(u_{it}u_{js})| < C_2$ for some $C_2 > 0$. However, as a result, the $n \times n$ covariance matrix $\mathrm{Cov}(\boldsymbol{u}_t)$ must be used in place of $\boldsymbol{\mathcal{D}}$ in (5) to retain the efficiency of TOPE. For that purpose, both the weighted PCA [25] and the estimator using thresholding principal orthogonal complements [18] can be employed in conjunction with TOPE. Then, in addition to some more stringent conditions on n and T, the non-asymptotic results must be re-established to obtain the similar conclusions in Section 3. Finally, from its construction, TOPE also paves a potentially effective way to model high-dimensional temporal data with multiple responses and simultaneously draw inference on the heteroscedasticity. We will explore these questions in future efforts.

8. Proofs of main theorems and technical results

We begin by presenting some notation. For a matrix $\mathbf{M} = (m_{ij})_{1\le i,j\le p} \in \mathbb{R}^{p\times p}$, denote $\|\mathbf{M}\|_{\mathbb{F}} = (\sum_{i=1}^p\sum_{j=1}^p m_{ij}^2)^{1/2}$ the Frobenius norm, $\|\mathbf{M}\|_{\max} = \max_{i,j}|m_{ij}|$ the maximum norm, and $\|\mathbf{M}\|_\infty = \max_i\sum_j|m_{ij}|$ the induced ℓ_∞ norm. The spectral norm of $\mathbf{M}$ corresponds to its largest singular value, defined as $\|\mathbf{M}\|_2 = \sup_{\boldsymbol{a}\in S}\|\mathbf{M}\boldsymbol{a}\|_2$, where $S = \{\boldsymbol{a} \in \mathbb{R}^p : \|\boldsymbol{a}\|_2 = 1\}$ and the ℓ_q-norm of p-dimensional vector $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$ is defined by $\|\boldsymbol{a}\|_q = (\sum_{j=1}^p |a_j|^q)^{1/q}$ with $1 \le q < \infty$. Denote the minimum and maximum eigenvalues of $\mathbf{M}$ by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$, respectively. Let $\mathrm{tr}(\mathbf{M}) = \sum_{j=1}^p m_{jj}$ and $\mathrm{vec}(\mathbf{M})$ be the trace and vectorization of $\mathbf{M}$, and $\otimes$ be the Kronecker product. We write $\mathbf{I}$ for an identity matrix. For sequences $\{a_n\}$ and $\{b_n\}$, $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$ and $a_n = O(b_n)$ if $\limsup_{n\to\infty}|a_n|/b_n < \infty$; $X_n = o_p(a_n)$ and $X_n = O_p(a_n)$ are similarly defined for a sequence of random variables X_n; $a_n \lesssim b_n$ if and only if $a_n \le Cb_n$ for some C independent of n; and $a_n \asymp b_n$ if and only if there exist C, D independent on n such that $C|b_n| \le |a_n| \le D|b_n|$. Denote $\xrightarrow{p}$ and $\xrightarrow{d}$ the convergence in probability and in distribution, respectively. Unless specified otherwise, $\delta > 0$ and $C > 0$ denote absolute constants independent of n, T, p.

Remark 2. The techniques in this section primarily depend on the derivation of a series of nontrivial exponential type concentration inequalities for preliminary estimators (such as $\widehat{\boldsymbol{\beta}}^0$ or $\widehat{\mathbf{F}}$) and their approximations (such as $\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H} = \sum_{i=1}^8(\mathbf{A}_i)\mathbf{K}^{-1}$ before Lemma 6). Together with the union bounds, it avoids entangling with the correlations between any preliminary estimators and the data.

8.1. Proof of the main results

8.1.1. Invertibility of the projection matrix

Without loss of generality, we take $\mathcal{X}^d = [0,1]^d$. Consider coefficients $\boldsymbol{a}_k = (a_0^{(k)}, a_{11}^{(k)}, \ldots, a_{J1}^{(k)}, \ldots, a_{1d}^{(k)}, \ldots, a_{Jd}^{(k)})^\top \in \mathbb{R}^{Jd+1}$ for $k \geq 1$, and define

$$\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle_n = \frac{1}{n} \sum_i \{a_0^{(1)} + \sum_j \sum_\ell a_{j\ell}^{(1)} \phi_j(X_{i\ell})\} \{a_0^{(2)} + \sum_j \sum_\ell a_{j\ell}^{(2)} \phi_j(X_{i\ell})\}. \tag{8}$$

In the literature, conditions on the largest and smallest eigenvalues of $n^{-1}\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ are usually stated as key assumptions for theoretical guarantees, see, e.g., [19]. Under standard nonparametric settings, we can establish it as follows.

Lemma 1. *Under Condition 4, whenever $J = o(\sqrt{n})$ and $d < J^{-1}n$, with probability at least $1-\delta$,*

$$n\left\{1 - \frac{J}{n}\ln(J^2/\delta)\right\} \lesssim \lambda_{\min}\left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right) < \lambda_{\max}\left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right) \lesssim n\left\{1 + \frac{J}{n}\ln(J^2/\delta)\right\},$$

where, as defined in Section 2.2,

$$\boldsymbol{\Phi} = \begin{bmatrix} 1/\sqrt{J} & \phi_1(x_{11}) & \ldots & \phi_J(x_{11}) & \ldots & \phi_1(x_{1d}) & \ldots & \phi_J(x_{1d}) \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 1/\sqrt{J} & \phi_1(x_{n1}) & \ldots & \phi_J(x_{n1}) & \ldots & \phi_1(x_{nd}) & \ldots & \phi_J(x_{nd}) \end{bmatrix}.$$

Proof. From (8), $\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n = \boldsymbol{a}^\top \left(n^{-1}\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right) \boldsymbol{a}$ for any $\boldsymbol{a} \in \mathbb{R}^{Jd+1}$. For any $\delta > 0$, let $\mathcal{A}_\delta = \{|\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n - \mathrm{E}(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n)| \gtrsim n^{-1}J\ln(J^2/\delta)\,\mathrm{E}(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n)\}$. On $\mathcal{A}_\delta^c$, we have $\{1 - n^{-1}J\ln(J^2/\delta)\}\,\mathrm{E}(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n) \lesssim \langle \boldsymbol{a}, \boldsymbol{a} \rangle_n \lesssim \{1 + n^{-1}J\ln(J^2/\delta)\}\,\mathrm{E}(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n)$. By Lemma 2, $\mathrm{E}(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n) \asymp \|\boldsymbol{a}\|_2^2$. Thus, $\{1 - n^{-1}J\ln(J^2/\delta)\}\|\boldsymbol{a}\|_2^2 \lesssim \boldsymbol{a}^\top(n^{-1}\boldsymbol{\Phi}^\top \boldsymbol{\Phi})\boldsymbol{a} \lesssim \{1 + n^{-1}J\ln(J^2/\delta)\}\|\boldsymbol{a}\|_2^2$. The conclusion follows Lemma 3, which implies $\Pr\{\mathcal{A}_\delta\} < \delta$. □

8.1.2. Proof of main theorems

Proof of Theorems 1 and 3. Theorems 1 and 3 readily follow from Propositions 1–4. □

Proof of Theorem 2. Recall that $\widehat{\mathbf{V}} = \widehat{\mathbf{G}}\mathcal{V}(\widehat{\boldsymbol{f}}_t)\widehat{\mathbf{G}}^\top + \widehat{\mathcal{D}}$, similarly to the proof of Lemma 14, we have

$$\lambda_{\min}(\widehat{\mathbf{V}}) \gtrsim \frac{1}{T}\left[1 + \frac{1}{\sqrt{nT}} + \sqrt{\frac{T + p^2n^{2\alpha}}{n^3T}} + \frac{\{(T + p^2n^{2\alpha})\ln(n)\}^{1/4}}{\sqrt{n^2T}} + \frac{1}{nJ^{\kappa/2}}\right]\{1 + \sqrt{\ln(1/\delta)}\},$$

with probability at least $1-\delta$. Then, by Lemma 14, with probability at least $1-\delta$,

$$\begin{aligned} \left\|\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}\right\|_2 &= \left\|\widehat{\mathbf{V}}^{-1}(\mathbf{V} - \widehat{\mathbf{V}})\mathbf{V}^{-1}\right\|_2 \leq \left\|\widehat{\mathbf{V}}^{-1}\right\|_2 \left\|\mathbf{V}^{-1}(\widehat{\mathbf{V}} - \mathbf{V})\right\|_2 \\ &\lesssim T\left\{\frac{\sqrt{J}}{n} + \frac{1}{\sqrt{n}} + \frac{1}{T} + \frac{p\sqrt{J}}{\sqrt{n^{1-2\alpha}T}} + \frac{1}{J^{(\kappa-1)/2}}\right\}\{1 + \sqrt{\ln(1/\delta)}\}. \end{aligned}$$

By Lemmas 16 and 17, $\|\mathbb{Z}_0^\top(\mathbf{G}\bar{\boldsymbol{f}} + \bar{\boldsymbol{u}})\|_2 \lesssim \|\mathbb{Z}_0\|_{\mathbb{F}} T^{-1/2}\sqrt{\ln(1/\delta)}$ with probability at least $1-\delta$. Thus, with probability at least $1-\delta$,

$$\begin{aligned} \left\|\bar{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right\|_2 \leq& \|\{(\mathbb{Z}_0^\top\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1} - (\mathbb{Z}_0^\top\mathbf{V}^{-1}\mathbb{Z}_0)^{-1}\}\mathbb{Z}_0^\top\mathbf{V}^{-1}(\mathbf{G}\bar{\boldsymbol{f}} + \bar{\boldsymbol{u}})\|_2 \\ &+ \|(\mathbb{Z}_0^\top\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^\top(\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1})(\mathbf{G}\bar{\boldsymbol{f}} + \bar{\boldsymbol{u}})\|_2 \\ \lesssim& \frac{1}{\sqrt{nT}}\left\{\frac{\sqrt{J}}{n} + \frac{1}{\sqrt{n}} + \frac{1}{T} + \frac{p\sqrt{J}}{\sqrt{n^{1-2\alpha}T}} + \frac{1}{J^{(\kappa-1)/2}}\right\}\{1 + \sqrt{\ln(1/\delta)}\}. \end{aligned}$$

Therefore, for any $a > 0$,

$$\begin{aligned} \mathrm{E}(\|\bar{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2^2) &= \int_0^\infty \Pr(\|\bar{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2^2 > s)ds = \int_0^a \Pr(\|\bar{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2^2 > s)ds + \int_a^\infty \Pr(\|\bar{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2^2 > s)ds \\ &\leq a + e\int_a^\infty \exp\{-snT(C\vartheta_{n,T,J}^2)^{-1}\}ds \\ &= a + Ce\vartheta_{n,T,J}^2(nT)^{-1}\exp\{-anT(C\vartheta_{n,T,J}^2)^{-1}\}, \end{aligned}$$

with $\vartheta_{n,T,J} = J^{1/2}n^{-1} + n^{-1/2} + T^{-1} + pJ^{1/2}n^{-1/2+\alpha}T^{-1/2} + J^{-(\kappa-1)/2}$ and constant $C > 0$. Letting $a = (nT)^{-1}C\vartheta_{n,T,J}^2$ gives $\mathrm{E}(\|\bar{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2^2) \leq 2(nT)^{-1}C\vartheta_{n,T,J}^2$. For TOPE $\bar{\boldsymbol{\beta}}$ and the oracle GLS estimator $\widetilde{\boldsymbol{\beta}}$ whose jth components are denoted by $\bar{\beta}_j$ and

$\widetilde{\beta}_j$ respectively, repeatedly employing Cauchy–Schwarz inequality to each of the (i,j) pair with $i,j\in\{1,\ldots,p\}$ leads to

$$\begin{aligned}|\mathrm{Cov}(\bar{\beta}_i,\bar{\beta}_j)-\mathrm{Cov}(\widetilde{\beta}_i,\widetilde{\beta}_j)| &= |\mathrm{E}\{(\bar{\beta}_i-\beta_i)(\bar{\beta}_j-\widetilde{\beta}_j)\}+\mathrm{E}\{(\bar{\beta}_i-\widetilde{\beta}_i)(\widetilde{\beta}_j-\beta_j)\}|\\ &\le [\mathrm{E}\{(\bar{\beta}_i-\beta_i)^2\}]^{1/2}[\mathrm{E}\{(\bar{\beta}_j-\widetilde{\beta}_j)^2\}]^{1/2}+[\mathrm{E}\{(\widetilde{\beta}_j-\beta_j)^2\}]^{1/2}[\mathrm{E}\{(\bar{\beta}_i-\widetilde{\beta}_i)^2\}]^{1/2}\\ &\lesssim \frac{\vartheta_{n,T,J}}{nT}+\frac{\vartheta_{n,T,J}^2}{(nT)^{3/2}},\end{aligned}$$

which yields

$$\|\mathrm{Var}(\bar{\boldsymbol{\beta}})-\mathrm{Var}(\widetilde{\boldsymbol{\beta}})\|_{\mathbb{F}}=[\sum_{i,j=1}^{p}\{\mathrm{Cov}(\bar{\beta}_i,\bar{\beta}_j)-\mathrm{Cov}(\widetilde{\beta}_i,\widetilde{\beta}_j)\}^2]^{1/2}\lesssim\frac{p\vartheta_{n,T,J}}{nT}+\frac{p\vartheta_{n,T,J}^2}{(nT)^{3/2}}.\quad\square$$

Proof of Theorem 4. (i) For the oracle GLS estimator $\widetilde{\boldsymbol{\beta}}$, it holds

$$\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}=(\mathbb{Z}_0^\top\mathbf{V}^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^\top\mathbf{V}^{-1}\left(\mathbf{G}\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t+\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\right):=\mathbf{A}\left(\mathbf{G}\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t+\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\right),$$

where $\mathbf{A}=(\mathbb{Z}_0^\top\mathbf{V}^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0\mathbf{V}^{-1}$ and $\mathbf{V}=\mathbf{G}\,\mathrm{Var}(T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t)\mathbf{G}^\top+\mathrm{Var}(T^{-1}\sum_{t=1}^{T}u_{1t})\mathbf{I}_n$ as defined in (5). For any p-vector $\boldsymbol{c}$, $(nT)^{1/2}\boldsymbol{c}^\top(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}):=\sum_{t=1}^{T}(W_{nt}+\widetilde{W}_{nt})$, where $W_{nt}=n^{1/2}T^{-1/2}\boldsymbol{c}^\top\mathbf{A}\boldsymbol{u}_t$ and $\widetilde{W}_{nt}=n^{1/2}T^{-1/2}\boldsymbol{c}^\top\mathbf{A}\mathbf{G}\boldsymbol{f}_t$. Then $\sum_{t=1}^{T}\mathrm{E}[|W_{nt}|^3]=n^{3/2}T^{-1/2}\|\boldsymbol{c}\|_2^3\,\mathrm{E}[\|\mathbf{A}\|_2^3\,\|]\,\mathrm{E}[\|\boldsymbol{u}_1\|_2^3]<\infty$ for any n, and since $\|\mathbf{A}\|_2\le\|\mathbf{A}\|_{\mathbb{F}}\le p\|\mathbf{A}\|_2$ we have

$$\frac{\sum_{t=1}^{T}\mathrm{E}[|W_{nt}|^3]}{(\sum_{t=1}^{T}\mathrm{E}[W_{nt}^2])^{3/2}}\le\frac{T^{-1/2}\|\boldsymbol{c}\|_2^3\,\mathrm{E}\left[\|\mathbf{A}\|_2^3\right]\mathrm{E}\left[\|\boldsymbol{u}_1\|_2^3\right]}{\left\{\boldsymbol{c}^\top\mathrm{E}\left[\mathbf{A}\,\mathrm{Var}(\boldsymbol{u}_1)\mathbf{A}^\top\right]\boldsymbol{c}\right\}^{3/2}}\le\frac{T^{-1/2}\,\mathrm{E}\left[\|\mathbf{A}\|_2^3\right]\mathrm{E}\left[\|\boldsymbol{u}_1\|^3\right]}{\{\max_i\mathrm{Var}(u_{i1})\}^{3/2}\left\{\mathrm{E}\left[\|\mathbf{A}\|_{\mathbb{F}}^2\right]\right\}^{3/2}}\to 0$$

as T diverges to infinity. By the Lyapunov central limit theorem [Theorem 27.3 in [9]], $\sum_{t=1}^{T}W_{nt}$ is hence asymptotically normal. Similarly, under Condition 6, we can show that $\widetilde{W}_{nt}$ is asymptotically normal. In addition, $\sum_{t=1}^{T}W_{nt}$ and $\sum_{t=1}^{T}\widetilde{W}_{nt}$ are uncorrelated since $\{\boldsymbol{u}_t\}$ and $\{\boldsymbol{f}_t\}$ are uncorrelated mean zero processes. Therefore, $n^{1/2}T^{1/2}\boldsymbol{c}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically normal for any $\boldsymbol{c}$, and we have $\boldsymbol{\Sigma}^{-1/2}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}N(\mathbf{0},\mathbf{I}_p)$, where $\boldsymbol{\Sigma}=\mathrm{E}[(\mathbb{Z}_0^\top\mathbf{V}^{-1}\mathbb{Z}_0)^{-1}]$. By Theorem 2, we have $\sqrt{nT}(\bar{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\xrightarrow{p}\mathbf{0}$. Notice that $\|\boldsymbol{\Sigma}\|_{\mathbb{F}}^2=O_P(nT)$, Slutsky's theorem yields $\boldsymbol{\Sigma}^{-1/2}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}N(\mathbf{0},\mathbf{I}_p)$.

(ii) Similar to (i) and conditional on $\mathbf{Z}_t$ and $\mathbf{X}$, the Lyapunov central limit theorem yields $(\mathbb{Z}_0^\top\mathbf{V}^{-1}\mathbb{Z}_0)^{1/2}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}N(\mathbf{0},\mathbf{I}_p)$, and Slusky's theorem leads to $(\mathbb{Z}_0^\top\mathbf{V}^{-1}\mathbb{Z}_0)^{1/2}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}N(\mathbf{0},\mathbf{I}_p)$. $\square$

8.2. Technical results

We first collect some preliminary results for spline estimators in Lemmas 2 and 3.

Lemma 2. *Under Condition 4, there exist constants c_1, c_2 such that $c_1\|\boldsymbol{a}\|_2^2\le\mathrm{E}(\langle\boldsymbol{a},\boldsymbol{a}\rangle_n)\le c_2\|\boldsymbol{a}\|_2^2$.*

Proof. It follows from Condition 4 that, for any $\ell=1,\ldots,d$, the marginal density of X_ℓ on its support is bounded away from 0 and ∞. Without loss of generality, we assume that, the support of $\mathbf{X}$ is $[0,1]^d$ and density $h(\mathbf{X})$ is bounded from below and above by m_1 and m_2 with $0<m_1\le m_2<\infty$.

Denote $f_\ell(X_\ell)=\sum_j a_{j\ell}\phi_j(X_\ell)$, $\ell\in\{1,\ldots,d\}$ and $f_0\equiv a_0$. Then, we have $\mathrm{E}(\langle\boldsymbol{a},\boldsymbol{a}\rangle_n)=\mathrm{E}[\{a_0+\sum_j\sum_l a_{j\ell}\phi_j(X_\ell)\}^2]=\mathrm{E}[\{a_0+\sum_{\ell=1}^{d}f_\ell(X_\ell)\}^2]$. Since the basis functions are centralized,

$$\mathrm{E}(\langle\boldsymbol{a},\boldsymbol{a}\rangle_n)=\int_{\mathcal{X}}\{a_0+\sum_{\ell=1}^{d}f_\ell(X_\ell)\}^2h(\mathbf{X})d\mathbf{X}\asymp\int_{\mathcal{X}}\{a_0+\sum_{\ell=1}^{d}f_\ell(X_\ell)\}^2d\mathbf{X}=a_0^2+\int_{\mathcal{X}}\{\sum_{\ell=1}^{d}f_\ell(X_\ell)\}^2d\mathbf{X}$$

By Lemma 1 of Stone [39], we obtain

$$\int_{\mathcal{X}}\{\sum_{\ell=1}^{d}f_\ell(X_\ell)\}^2d\mathbf{X}\ge\left(\frac{C_0}{2}\right)^{d-1}\sum_{\ell=1}^{d}\int_0^1f_\ell^2(x)dx=\left(\frac{C_0}{2}\right)^{d-1}(\sum_\ell\sum_j a_{j\ell}^2),$$

where $C_0=1-(1-m_1/m_2)^{1/2}$. Consequently, we have $\mathrm{E}(\langle\boldsymbol{a},\boldsymbol{a}\rangle_n)\ge a_0^2+(C_0/2)^{d-1}(\sum_\ell\sum_j a_{j\ell}^2)\ge\min\{1,(C_0/2)^d\}\|\boldsymbol{a}\|^2$. Similarly, we can establish that $\int_0^1\{\sum_{\ell=1}^{d}f_\ell(X_\ell)\}^2d\mathbf{X}\le d^2\sum_{\ell=1}^{d}\int_0^1f_\ell^2(x)dx=d^2(\sum_\ell\sum_j a_{j\ell}^2)$, and consequently, $\mathrm{E}(\langle\boldsymbol{a},\boldsymbol{a}\rangle_n)\le a_0^2+d^2(\sum_\ell\sum_j a_{j\ell}^2)\le(1+d^2)\|\boldsymbol{a}\|^2$. $\square$

Lemma 3. *Under Condition 4, for some constant $C_1, C_2 > 0$, we have*

$$\Pr\left\{\sup_{\boldsymbol{a}_1,\boldsymbol{a}_2\in\mathbb{R}^{Jd+1}}\frac{|\langle\boldsymbol{a}_1,\boldsymbol{a}_2\rangle_n-\mathrm{E}(\langle\boldsymbol{a}_1,\boldsymbol{a}_2\rangle_n|)}{\sqrt{\mathrm{E}(\langle\boldsymbol{a}_1,\boldsymbol{a}_1\rangle_n)\,\mathrm{E}(\langle\boldsymbol{a}_2,\boldsymbol{a}_2\rangle_n)}}>s\right\}\le C_1J^2\exp\left\{-C_2\frac{n}{J}\frac{s^2}{1+s}\right\}.$$

Proof. The proof is similar to that of Lemma A.2 in Huang et al. [23]. First, notice that $\langle\boldsymbol{a}_1,\boldsymbol{a}_2\rangle_n-\mathrm{E}(\langle\boldsymbol{a}_1,\boldsymbol{a}_2\rangle_n)=n^{-1}\sum_{\ell,\ell'}\sum_{j,j'}a_{j\ell}^{(1)}a_{j'\ell'}^{(2)}\{\boldsymbol{e}_{j\ell}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{e}_{j'\ell'}-\mathrm{E}(\boldsymbol{e}_{j\ell}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{e}_{j'\ell'})\}$, where $\boldsymbol{e}_{j\ell}$ is the $(J\ell+j+1)$th natural basis of $\mathbb{R}^{Jd+1}$. Hence, we have $\boldsymbol{\Phi}\boldsymbol{e}_{j\ell}=\{\phi_j(X_{1\ell}),\ldots,\phi_j(X_{n\ell})\}^\top$. For any j,j',ℓ,ℓ', $\mathrm{Var}(n^{-1}\boldsymbol{e}_{j\ell}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{e}_{j'\ell'})\le n^{-2}\sum_i\mathrm{E}\{\phi_j^2(X_{i\ell})\phi_{j'}^2(X_{i\ell'})\}\lesssim n^{-1}$. As $|\phi_j(X_\ell)|\le M$ for each j,ℓ for some $M>0$,

$$\Pr\left\{\left|\frac{1}{n}\boldsymbol{e}_{j\ell}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{e}_{j'\ell'}-\mathrm{E}\left(\frac{1}{n}\boldsymbol{e}_{j\ell}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{e}_{j'\ell'}\right)\right|>c_2s\right\}\le\exp\left\{-\frac{ns^2}{M_1+M_2s}\right\},\quad s>0$$

by Bernstein's inequality with constants $M_1,M_2>0$. By the union bound,

$$\Pr\left[\bigcup_{j,j',\ell,\ell'}\left\{\left|\frac{1}{n}\boldsymbol{e}_{j\ell}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{e}_{j'\ell'}-\mathrm{E}\left(\frac{1}{n}\boldsymbol{e}_{j\ell}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{e}_{j'\ell'}\right)\right|>\frac{c_2s}{Jd}\right\}\right]\le C_1J^2\exp\left\{-\frac{C_2ns^2}{J^2+sJ}\right\},\quad s>0$$

for constants $C_1,C_2>0$. Denote $\mathcal{B}=\bigcup_{j,j',\ell,\ell'}\{|n^{-1}\boldsymbol{e}_{j\ell}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{e}_{j'\ell'}-\mathrm{E}(n^{-1}\boldsymbol{e}_{j\ell}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{e}_{j'\ell'})|>c_2s(Jd)^{-1}\}$, so that $\Pr(\mathcal{B})<C_1J^2\exp\{-C_2ns^2(J^2+sJ)^{-1}\}$. For each $s>0$, on $\mathcal{B}^c$, $|\langle\boldsymbol{a}_1,\boldsymbol{a}_2\rangle_n-\mathrm{E}(\langle\boldsymbol{a}_1,\boldsymbol{a}_2\rangle_n)|\le\sum_{j,\ell}\sum_{j',\ell'}|a_{j\ell}^{(1)}||a_{j'\ell'}^{(2)}|c_2s/(Jd)\lesssim s\sqrt{\mathrm{E}(\langle\boldsymbol{a}_1,\boldsymbol{a}_1\rangle_n)\,\mathrm{E}(\langle\boldsymbol{a}_2,\boldsymbol{a}_2\rangle_n)}$, where the last inequality is due to Lemma 2. The conclusion follows. □

Next, we document technical results for the proof of Theorem 1 in Lemmas 4–10 and Propositions 1–4.

Lemma 4. *Under Conditions 3 and 5, for each n, with probability at least $1-\delta$,*

$$1-n^{-1}\ln(1/\delta)\lesssim\lambda_{\min}\left(\frac{1}{n}\boldsymbol{G}^\top\boldsymbol{P}\boldsymbol{G}\right)<\lambda_{\max}\left(\frac{1}{n}\boldsymbol{G}^\top\boldsymbol{P}\boldsymbol{G}\right)\lesssim1+n^{-1}\ln(1/\delta).$$

Proof. Denote $\mathbf{R}=\mathbf{G}-\mathbf{P}\mathbf{G}$, and we have $\mathbf{G}^\top\mathbf{P}\mathbf{G}=\mathbf{G}^\top\mathbf{G}-\mathbf{G}^\top\mathbf{R}$. Thus, $\lambda_{\min}\left(n^{-1}\mathbf{G}^\top\mathbf{P}\mathbf{G}\right)\ge\lambda_{\min}\left(n^{-1}\mathbf{G}^\top\mathbf{G}\right)+\lambda_{\min}\left(-n^{-1}\mathbf{G}^\top\mathbf{R}\right)$, and $\lambda_{\max}\left(n^{-1}\mathbf{G}^\top\mathbf{P}\mathbf{G}\right)\le\lambda_{\max}\left(n^{-1}\mathbf{G}^\top\mathbf{G}\right)+\lambda_{\max}\left(-n^{-1}\mathbf{G}^\top\mathbf{R}\right)$. Note that $\|\mathbf{R}\|_{\mathbb{F}}^2\lesssim nJ^{-\kappa}$ by Condition 5. Thus, combining Condition 3, it holds that, with probability at least $1-\delta$,

$$\|n^{-1}\mathbf{G}^\top\mathbf{R}\|_{\mathbb{F}}^2=\frac{1}{n^2}\mathrm{tr}(\mathbf{R}^\top\mathbf{G}\mathbf{G}^\top\mathbf{R})\le\lambda_{\max}\left(\frac{1}{n}\mathbf{G}\mathbf{G}^\top\right)\frac{1}{n}\mathrm{tr}(\mathbf{R}^\top\mathbf{R})\lesssim J^{-\kappa}\{1+n^{-1}\ln(1/\delta)\}$$

and $|\lambda(\mathbf{G}^\top\mathbf{R}/n)|\lesssim J^{-\kappa}\{1+n^{-1}\ln(1/\delta)\}$. By Condition 3, with probability at least $1-\delta$, $1-n^{-1}\ln(1/\delta)\lesssim\lambda_{\min}\left(n^{-1}\mathbf{G}^\top\mathbf{G}\right)<\lambda_{\max}\left(n^{-1}\mathbf{G}^\top\mathbf{G}\right)\lesssim1+n^{-1}\ln(1/\delta)$. The conclusion follows. □

Lemma 5. *Consider $\widehat{\boldsymbol{\beta}}^0$ satisfying $\|\widehat{\boldsymbol{\beta}}^0-\boldsymbol{\beta}\|_2=O_P(n^{-1/2+\alpha}T^{-1/2})$ for $\alpha\in[0,1/2)$, such as the estimator in (14) in Section 8.3. Under Conditions 1 and 4–7, for $\widetilde{\boldsymbol{U}}=\boldsymbol{U}+\mathbb{Z}\{\boldsymbol{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}^0)\}$ defined in Section 2.2,*

(i) $\mathrm{E}(\|\boldsymbol{F}^\top\widetilde{\boldsymbol{U}}^\top\|_{\mathbb{F}}^2)=O((n+p^2n^{2\alpha})T)$, $\mathrm{E}(\|\widetilde{\boldsymbol{U}}^\top\boldsymbol{\Phi}\|_{\mathbb{F}}^2)=O(nJ(T+p^2n^{2\alpha}))$, $\mathrm{E}(\|\boldsymbol{\Phi}^\top\widetilde{\boldsymbol{U}}\boldsymbol{F}\|_{\mathbb{F}}^2)=O(p^2n^{1+2\alpha}TJ)$, $\mathrm{E}(\|\widetilde{\boldsymbol{U}}^\top\boldsymbol{\Phi}\boldsymbol{B}\|_{\mathbb{F}}^2)\lesssim n(T+p^2n^{2\alpha})$, *and* $\mathrm{E}(\|\boldsymbol{B}^\top\boldsymbol{\Phi}^\top\widetilde{\boldsymbol{U}}\boldsymbol{F}\|_{\mathbb{F}}^2)\lesssim p^2n^{1+2\alpha}T$.

(ii) With probability at least $1-5\delta$, $\|\boldsymbol{F}^\top\widetilde{\boldsymbol{U}}^\top\|_{\mathbb{F}}\lesssim\{(n+p^2)T\}^{1/2}\{1+\sqrt{\ln(1/\delta)}\}$, $\|\widetilde{\boldsymbol{U}}^\top\boldsymbol{\Phi}\|_{\mathbb{F}}\lesssim\{nJ(T+p^2)\}^{1/2}\{1+\sqrt{\ln(1/\delta)}\}$, $\|\boldsymbol{\Phi}^\top\widetilde{\boldsymbol{U}}\boldsymbol{F}\|_{\mathbb{F}}\lesssim(p^2nTJ)^{1/2}\{1+\sqrt{\ln(1/\delta)}\}$, $\|\widetilde{\boldsymbol{U}}^\top\boldsymbol{\Phi}\boldsymbol{B}\|_{\mathbb{F}}\lesssim\sqrt{n(T+p^2n^{2\alpha})}\{1+\sqrt{\ln(1/\delta)}\}$, *and* $\|\boldsymbol{B}^\top\boldsymbol{\Phi}^\top\widetilde{\boldsymbol{U}}\boldsymbol{F}\|_{\mathbb{F}}\lesssim p\sqrt{n^{1+2\alpha}T}\{1+\sqrt{\ln(1/\delta)}\}^2\{1+\sqrt{\ln(1/\delta)}\}$.

(iii) With probability at least $1-4\delta$, $\|\boldsymbol{P}\widetilde{\boldsymbol{U}}\|_{\mathbb{F}}\lesssim\sqrt{J(T+p^2n^{2\alpha})}\{1+n^{-1}J\ln(J^2/\delta)\}^{3/2}\{1+\sqrt{\ln(1/\delta)}\}$.

Proof.

(i) By Lemma B.1 of [19], $\mathrm{E}(\|\mathbf{F}^\top\mathbf{U}^\top\|_{\mathbb{F}}^2)=O(nT)$, $\mathrm{E}(\|\mathbf{U}^\top\boldsymbol{\Phi}\|_{\mathbb{F}}^2)=O(nJT)$, $\mathrm{E}(\|\boldsymbol{\Phi}^\top\mathbf{U}\mathbf{F}\|_{\mathbb{F}}^2)=O(nTJ)$, and $\mathrm{E}(\|\mathbf{P}\mathbf{U}\|_{\mathbb{F}}^2)=O(JT)$, and by Lemma C.6 in Fan et al. [19], $\mathrm{E}\left(\|\mathbf{U}\boldsymbol{\Phi}\mathbf{B}\|_{\mathbb{F}}^2\right)=O(nT)$ and $\mathrm{E}\left(\|\mathbf{B}\boldsymbol{\Phi}^\top\mathbf{U}\mathbf{F}\|_{\mathbb{F}}^2\right)=O(nT)$. Thus, it suffices to show

$$\mathrm{E}[\|\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}^0)\}\mathbf{F}\|_{\mathbb{F}}^2]=O(p^2T),\tag{9}$$

$$\mathrm{E}[\|\boldsymbol{\Phi}^\top\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}^0)\}\|_{\mathbb{F}}^2]=O(p^2nJ),\tag{10}$$

$$\mathrm{E}[\|\boldsymbol{\Phi}^\top\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}^0)\}\mathbf{F}\|_{\mathbb{F}}^2]=O(p^2nTJ),\tag{11}$$

$$\mathrm{E}[\|\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}^0)\}^\top\boldsymbol{\Phi}\mathbf{B}\|_{\mathbb{F}}^2]=O(p^2n^{1+2\alpha}),\tag{12}$$

$$\mathrm{E}[\|\mathbf{B}^\top \boldsymbol{\Phi}^\top \mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\mathbf{F}\|_{\mathbb{F}}^2] = O(p^2 n^{1+2\alpha} T). \tag{13}$$

By Proposition 5, $\mathrm{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\|_{\mathbb{F}}^2] \le \mathrm{E}(\|\mathbb{Z}\|_{\mathbb{F}}^2)\|\mathbf{I}_T\|_2^2\,\mathrm{E}(\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0\|_{\mathbb{F}}^2) = O(p^2 n^{2\alpha})$. Then (9) follows from Cauchy–Schwarz inequality that $\mathrm{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\mathbf{F}\|_{\mathbb{F}}^2] \le \mathrm{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\|_{\mathbb{F}}^2]\,\mathrm{E}(\|\mathbf{F}\|_{\mathbb{F}}^2) = O(p^2 n^{2\alpha} T)$. As a consequence of Lemma 2, we have $\mathrm{E}(\|\boldsymbol{\Phi}\|_2^2) = O(n)$, and consequently $\mathrm{E}(\|\boldsymbol{\Phi}\|_{\mathbb{F}}^2) \le (Jd+1)\,\mathrm{E}(\|\boldsymbol{\Phi}\|_2^2) = O(nJ)$, and (10) holds since $\mathrm{E}[\|\boldsymbol{\Phi}^\top \mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\|_{\mathbb{F}}^2] \le \mathrm{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\|_{\mathbb{F}}^2]\,\mathrm{E}(\|\boldsymbol{\Phi}\|_{\mathbb{F}}^2) = O(p^2 n^{1+2\alpha} J)$. Applying Cauchy–Schwarz inequality, (11) follows $\mathrm{E}[\|\boldsymbol{\Phi}^\top \mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\mathbf{F}\|_{\mathbb{F}}^2] \le \mathrm{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\|_{\mathbb{F}}^2]\,\mathrm{E}(\|\boldsymbol{\Phi}\|_{\mathbb{F}}^2)\,\mathrm{E}(\|\mathbf{F}\|_{\mathbb{F}}^2) = O(p^2 n^{1+2\alpha} TJ)$. Also, (12) and (13) follow $\mathrm{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}^\top \boldsymbol{\Phi}\mathbf{B}\|_{\mathbb{F}}^2] \le \mathrm{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\mathbf{G}\|_{\mathbb{F}}^2] + \mathrm{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\mathbf{R}\|_{\mathbb{F}}^2] = O(p^2 n^{1+2\alpha})$, and $\mathrm{E}[\|\mathbf{B}^\top \boldsymbol{\Phi}^\top \mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\mathbf{F}\|_{\mathbb{F}}^2] \le \mathrm{E}[\|\mathbf{G}^\top \mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\mathbf{F}\|_{\mathbb{F}}^2] + \mathrm{E}[\|\mathbf{R}^\top \mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}\mathbf{F}\|_{\mathbb{F}}^2] = O(p^2 n^{1+2\alpha} T)$, respectively.

(ii) By Condition 7 (iii), for $r_6 \in (0,1)$ and any $x \ge 1$,

$$\tau(x) \le 4\max(b_1, b_2)^2 r_6 \{\max(r_2, r_3)(1-r_6)/2\}^{2/\min(r_2,r_3)} \exp[2/\{\max(r_2, r_3)(1-r_6)\}]\{2\alpha(x)\}^{r_6},$$

which implies that (f_{tk}, u_{it}) is τ-mixing [32] by Condition 7 (ii). Then, following Theorem 1 in Merlevède et al. [32] and Davydov's inequality [Corollary 16.2.4 in [1]], for each $k \in \{1,\dots,K\}$ and $i \in \{1,\dots,n\}$ and any $x > 0$, $\Pr(|\sum_{t=1}^T f_{tk}\widetilde{u}_{it}| > x) \le T\exp(-Cx^{r_{fu}}) + \exp\{-Cx^2/\mathrm{E}(\sum_{t,s=1}^T f_{tk}\widetilde{u}_{it}f_{sk}\widetilde{u}_{is})\}$, where $r_{fu} = \{r_1^{-1} + \min(r_2, r_3)^{-1}\}^{-1}$. Let e^{-s} be the maximum of $T\exp(-Cx^{r_{fu}})$ and $\exp\{-Cx^2/\mathrm{E}(\sum_{t,s=1}^T f_{tk}\widetilde{u}_{it}f_{sk}\widetilde{u}_{is})\}$. By Bonferroni inequality, with probability at least $1-2e^{-s}$, $|\sum_{t=1}^T f_{tk}\widetilde{u}_{it}| \lesssim \max[\{s+\ln(pT)\}^{1/r_{fu}}, \{\mathrm{E}(\sum_{t,s=1}^T f_{tk}\widetilde{u}_{it}f_{sk}\widetilde{u}_{is})\}^{1/2}\sqrt{s+\ln p}]$ uniformly for each $k \in \{1,\dots,K\}$ and $i \in \{1,\dots,n\}$. Then, by (i), with probability at least $1-\delta$, $\|\mathbf{F}^\top\widetilde{\mathbf{U}}^\top\|_{\mathbb{F}} \lesssim \sqrt{(n+p^2n^{2\alpha})T(s+\ln p)}$. The remaining two bounds follows similarly.

(iii) By Lemma 1, with probability at least $1-\delta$, $\|\boldsymbol{\Phi}\|_2^2 = \lambda_{\max}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}) \lesssim n\{1+n^{-1}J\ln(J^2/\delta)\}$ and $\|(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\|_2 = \lambda_{\min}^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}) \lesssim n^{-1}\{1+n^{-1}J\ln(J^2/\delta)\}$. Hence, with probability at least $1-4\delta$,

$$\|\mathbf{P}\widetilde{\mathbf{U}}\|_{\mathbb{F}} \le \|\boldsymbol{\Phi}\|_2\|(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\|_2\|\boldsymbol{\Phi}^\top\widetilde{\mathbf{U}}\|_{\mathbb{F}} \lesssim \sqrt{J(T+p^2n^{2\alpha})}\left\{1+Jn^{-1}\ln(J^2/\delta)\right\}^{3/2}\{1+\sqrt{\ln(1/\delta)}\}. \quad \square$$

Denote $\mathbf{K}$ a $K\times K$ diagonal matrix whose diagonals are the first K eigenvalues of $(nT)^{-1}\widetilde{\mathbf{Y}}^\top\mathbf{P}\widetilde{\mathbf{Y}}$. Then $(nT)^{-1}\widetilde{\mathbf{Y}}^\top\mathbf{P}\widetilde{\mathbf{Y}}\widehat{\mathbf{F}} = \widehat{\mathbf{F}}\mathbf{K}$. Let $\mathbf{H} = (nT)^{-1}\mathbf{B}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\mathbf{B}\mathbf{F}^\top\widehat{\mathbf{F}}\mathbf{K}^{-1}$. Using $\widetilde{\mathbf{Y}} = (\boldsymbol{\Phi}\mathbf{B}+\mathbf{R})\mathbf{F}^\top + \widetilde{\mathbf{U}}$ from Section 2.2.1, we have $\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H} = (\sum_{i=1}^8 \mathbf{A}_i)\mathbf{K}^{-1}$ where $\mathbf{A}_1 = (nT)^{-1}\mathbf{F}\mathbf{B}^\top\boldsymbol{\Phi}^\top\widetilde{\mathbf{U}}\widehat{\mathbf{F}}$, $\mathbf{A}_2 = (nT)^{-1}\widetilde{\mathbf{U}}^\top\boldsymbol{\Phi}\mathbf{B}\mathbf{F}^\top\widehat{\mathbf{F}}$, $\mathbf{A}_3 = (nT)^{-1}\widetilde{\mathbf{U}}^\top\mathbf{P}\widetilde{\mathbf{U}}\widehat{\mathbf{F}}$, $\mathbf{A}_4 = (nT)^{-1}\mathbf{F}\mathbf{B}^\top\boldsymbol{\Phi}^\top\mathbf{R}\mathbf{F}^\top\widehat{\mathbf{F}}$, $\mathbf{A}_5 = (nT)^{-1}\mathbf{F}\mathbf{R}^\top\boldsymbol{\Phi}\mathbf{B}\mathbf{F}^\top\widehat{\mathbf{F}}$, $\mathbf{A}_6 = (nT)^{-1}\mathbf{F}\mathbf{R}^\top\mathbf{P}\mathbf{R}\mathbf{F}^\top\widehat{\mathbf{F}}$, $\mathbf{A}_7 = (nT)^{-1}\mathbf{F}\mathbf{R}^\top\mathbf{P}\widetilde{\mathbf{U}}\widehat{\mathbf{F}}$, and $\mathbf{A}_8 = (nT)^{-1}\widetilde{\mathbf{U}}^\top\mathbf{P}\mathbf{R}\mathbf{F}^\top\widehat{\mathbf{F}}$. Next, in Lemmas 6–10, we will provide a bound on $\|\mathbf{H}-\mathbf{I}\|_{\mathbb{F}}$ in probability.

Lemma 6. *With probability at least* $1-5\delta$, $\|\boldsymbol{K}^{-1}\|_2 \lesssim 1+n^{-1}\ln(1/\delta)$.

Proof. The K largest eigenvalues of $(nT)^{-1}\widetilde{\mathbf{Y}}^\top\mathbf{P}\widetilde{\mathbf{Y}}$ are the same as those of $\mathbf{W} = (nT)^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1/2}\boldsymbol{\Phi}^\top\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^\top\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1/2}$. Substituting $\widetilde{\mathbf{Y}} = \mathbf{G}\mathbf{F}^\top + \widetilde{\mathbf{U}}$ and $T^{-1}\mathbf{F}^\top\mathbf{F} = \mathbf{I}_K$, we have $\mathbf{W} = \sum_{i=1}^4 \mathbf{W}_i$ where $\mathbf{W}_1 = n^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1/2}\boldsymbol{\Phi}^\top\mathbf{G}\mathbf{G}^\top\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1/2}$, $\mathbf{W}_2 = (nT)^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1/2}\boldsymbol{\Phi}^\top\mathbf{G}\mathbf{F}^\top\widetilde{\mathbf{U}}^\top\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1/2}$, $\mathbf{W}_3 = \mathbf{W}_2^\top$, and $\mathbf{W}_4 = (nT)^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1/2}\boldsymbol{\Phi}^\top\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1/2}$. By Lemma 1, with probability at least $1-\delta$, $\|\boldsymbol{\Phi}\|_2^2 = \lambda_{\max}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}) \lesssim n\left\{1+n^{-1}J\ln(J^2/\delta)\right\}$ and $\|(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\|_2 = \lambda_{\min}^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}) \lesssim n^{-1}\{1+n^{-1}J\ln(J^2/\delta)\}$. By Lemma 4, with probability at least $1-\delta$, $\|\mathbf{P}\mathbf{G}\|_2^2 = \lambda_{\max}\left(\mathbf{G}^\top\mathbf{P}\mathbf{G}\right) \lesssim n(1+J^{-\kappa})\{1+n^{-1}\ln(1/\delta)\}$. Hence, with probability at least $1-5\delta$,

$$\begin{aligned}\|\mathbf{W}_2\|_2 &\le \frac{1}{n}\|(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1/2}\|_2^2\|\boldsymbol{\Phi}\|_2\|\mathbf{P}\mathbf{G}\|_2\left\|\frac{1}{T}\mathbf{F}^\top\widetilde{\mathbf{U}}^\top\boldsymbol{\Phi}\right\|_{\mathbb{F}} \\ &\lesssim p\sqrt{Jn^{2\alpha-1}T^{-1}}(1+J^{-\kappa})\{1+n^{-1}J\ln(J^2/\delta)\}^{3/2}\{1+\sqrt{\ln(1/\delta)}\}\{1+n^{-1}\ln(1/\delta)\},\end{aligned}$$

and by Lemma 5, with probability at least $1-4\delta$,

$$\|\mathbf{W}_4\|_2 \le \frac{1}{nT}\|(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1/2}\|_2^2\|\boldsymbol{\Phi}^\top\widetilde{\mathbf{U}}\|_{\mathbb{F}}^2 \lesssim \frac{J(T+p^2n^{2\alpha})}{nT}\{1+J\ln(J^2/\delta)/n\}\{1+\sqrt{\ln(1/\delta)}\}.$$

By Weyl's Theorem, $|\lambda_k(\mathbf{W}) - \lambda_k(\mathbf{W}_1)| \le \|\mathbf{W}-\mathbf{W}_1\|_2$ for each $k \in \{1,\dots,K\}$. Hence, with probability at least $1-5\delta$,

$$|\lambda_k(\mathbf{W}) - \lambda_k(\mathbf{W}_1)| \lesssim [p\sqrt{J}(n^{1-2\alpha}T)^{-1/2} + \{J(T+p^2n^{2\alpha})\}/(nT)]\{1+J\ln(J^2/\delta)/n\}^{3/2}\{1+\sqrt{\ln(1/\delta)}\}.$$

Note that the K largest eigenvalues of $\mathbf{W}_1$ is also the K largest eigenvalues of $n^{-1}\mathbf{G}^\top\mathbf{P}\mathbf{G}$. Thus, by Lemma 4, with probability at least $1-5\delta$, $\|\mathbf{K}^{-1}\|_2 \lesssim 1+n^{-1}\ln(1/\delta)$. $\square$

Lemma 7. *With probability at least* $1-7\delta$, *(i)* $\|\boldsymbol{A}_1\|_{\mathbb{F}}, \|\boldsymbol{A}_2\|_{\mathbb{F}} \lesssim \sqrt{n^{-1}(T+p^2n^{2\alpha})}\{1+\sqrt{\ln(1/\delta)}\}$, *(ii)* $\|\boldsymbol{A}_3\|_{\mathbb{F}} \lesssim n^{-1}T^{-1/2}J(T+p^2n^{2\alpha})\{1+\sqrt{\ln(1/\delta)}\}$, *(iii)* $\|\boldsymbol{A}_4\|_{\mathbb{F}}, \|\boldsymbol{A}_5\|_{\mathbb{F}} \lesssim (J^{-\kappa/2}\sqrt{T})\{1+\sqrt{\ln(1/\delta)}\}$, *(iv)* $\|\boldsymbol{A}_7\|_{\mathbb{F}}, \|\boldsymbol{A}_8\|_{\mathbb{F}} \lesssim \sqrt{(T+p^2n^{2\alpha})(nJ^{\kappa-1})^{-1}}\{1+\sqrt{\ln(1/\delta)}\}$; *and* $\|\boldsymbol{A}_6\|_{\mathbb{F}} \lesssim J^{-\kappa}\sqrt{T}$.

Proof. Notice that $\|\mathbf{F}\|_{\mathbb{F}} = \sqrt{KT}$ with probability 1 and $\|\widehat{\mathbf{F}}\|_{\mathbb{F}} = \sqrt{KT}$. Then, both (i) and (ii) follow from Lemma 5. By Condition 5 that $\|\mathbf{R}\|_{\mathbb{F}}^2 \lesssim nJ^{-\kappa}$, (iii) follows from Lemma 4 and $\boldsymbol{\Phi}\mathbf{B} = \mathbf{PG}$. Part (iv) follows from Lemma 5 and $\|\mathbf{R}\|_{\mathbb{F}}^2 \lesssim nJ^{-\kappa}$. Result on $\mathbf{A}_6$ follows similarly to (iii) given $\|\mathbf{P}\|_2 = 1$. □

Lemma 8. *With probability at least* $1 - 3\delta$, *(i)* $\|\boldsymbol{A}_1\|_{\max}, \|\boldsymbol{A}_2\|_{\max} \lesssim n^{-1/2}T^{-1}\sqrt{T + p^2n^{2\alpha}}\{\ln(T)\}^{2/r_2}\{1 + \ln(1/\delta)\}$, *(ii)* $\|\boldsymbol{A}_3\|_{\max} \lesssim n^{-1/2}T^{-1}\sqrt{T + p^2n^{2\alpha}}\{\ln(T)\}^{1/r_2}\{1 + \ln(1/\delta)\}$, *(iii)* $\|\boldsymbol{A}_4\|_{\max}, \|\boldsymbol{A}_5\|_{\max} \lesssim n^{-1}T^{-1}\{\ln(T)\}^{3/r_2}J^{-\kappa}\{1 + \ln(1/\delta)\}$, *(iv)* $\|\boldsymbol{A}_7\|_{\max}, \|\boldsymbol{A}_8\|_{\max} \lesssim (nT)^{-1}J^{-\kappa}\sqrt{J(T + p^2n^{2\alpha})}\{\ln(T)\}^{2/r_2}\{1 + \ln(1/\delta)\}$; *and* $\|\boldsymbol{A}_6\|_{\max} \lesssim (nT)^{-1}J^{-2\kappa}\{\ln(T)\}^{3/r_2}$.

Proof. By Lemma B.1 in Fan et al. [17], with probability at least $1 - \delta$, $\|\widetilde{\mathbf{U}}\mathbf{P}\widetilde{\mathbf{U}}\|_{\max} \lesssim \sqrt{n}(T + p^2n^{2\alpha})\{1 + \ln(1/\delta)\}$. Also, the proof of Lemma D.2 in Wang and Fan [43] implies that $\|\mathbf{U}^\top\boldsymbol{\Phi}\mathbf{B}\|_\infty \lesssim \sqrt{n}T$. Hence, with probability at least $1 - \delta$, $\|\widetilde{\mathbf{U}}^\top\boldsymbol{\Phi}\mathbf{B}\|_\infty \lesssim \sqrt{n}(T + p^2n^{2\alpha})\{1 + \ln(1/\delta)\}$ by Lemma 5. Then, the results follow from that $\|\mathbf{F}\|_{\max} \lesssim \{\ln(T) + \ln(1/\delta)\}^{1/r_2}$ with probability at least $1 - \delta$. □

Proposition 1. *Given* $Jd + 1 < n$ *and* $\kappa \geq 1$,

(i) *With probability at least* $1 - 12\delta$, $T^{-1}\|\widehat{\boldsymbol{F}} - \boldsymbol{FH}\|_{\mathbb{F}}^2 \lesssim (n^{-1} + \{n^{1-2\alpha}T\}^{-1}p^2 + J^{-\kappa})\{1 + \sqrt{\ln(1/\delta)}\}^2\{1 + n^{-1}\ln(1/\delta)\}$.
(ii) *With probability at least* $1 - 8\delta$, $\|\widehat{\boldsymbol{F}} - \boldsymbol{FH}\|_{\max} \lesssim (n^{-1/2} + \{\sqrt{n^{1-2\alpha}T}\}^{-1}p)\{\ln(T)\}^{2/r_2}\{1 + \ln(1/\delta)\}$.

Proof. By Lemma 6, $\|\mathbf{K}^{-1}\|_2 \lesssim 1 + n^{-1}\ln(1/\delta)$ with probability at least $1 - 5\delta$. The result follows from Lemmas 7 and 8. □

Lemma 9. *With probability at least* $1-20\delta$, *(i)* $T^{-1}\|\boldsymbol{A}_1\|_{\mathbb{F}}^2 \lesssim \left\{n^{-2} + n^{-1+2\alpha}T^{-1}p^2 + (nTJ^\kappa)^{-1}(T + p^2n^{2\alpha})\right\}\{1+\sqrt{\ln(1/\delta)}\}^2\{1+n^{-1}\ln(1/\delta)\}$, *(ii)* $T^{-2}\|\boldsymbol{F}^\top\boldsymbol{A}_2\|_{\mathbb{F}}^2 \lesssim n^{-1+2\alpha}T^{-1}p^2\{1 + \sqrt{\ln(1/\delta)}\}^2$, *(iii)* $T^{-2}\|\boldsymbol{F}^\top(\widehat{\boldsymbol{F}} - \boldsymbol{FH})\|_{\mathbb{F}}^2 \lesssim \left\{n^{-2} + n^{-1+2\alpha}T^{-1}p^2 + J^{-\kappa}\right\}\{1 + \sqrt{\ln(1/\delta)}\}^2$, *and (iv)* $T^{-2}\|\widehat{\boldsymbol{F}}^\top(\widehat{\boldsymbol{F}} - \boldsymbol{FH})\|_{\mathbb{F}}^2 \lesssim \{n^{-2} + n^{-1+2\alpha}T^{-1}p^2 + J^{-\kappa}\}\{1 + \sqrt{\ln(1/\delta)}\}^2$.

Proof. (i) First, by Lemmas 4 and 6, with probability at least $1-6\delta$, $\|\mathbf{H}\|_2 \leq (nT)^{-1}\|\mathbf{PG}\|_{\mathbb{F}}^2\|\mathbf{F}\|_{\mathbb{F}}\|\widehat{\mathbf{F}}\|_{\mathbb{F}}\|\mathbf{K}^{-1}\|_2 \lesssim 1+n^{-1}\ln(1/\delta)$. Then, by Lemma 5 and Proposition 1, with probability at least $1 - 20\delta$,

$$\begin{aligned}\|\mathbf{B}^\top\boldsymbol{\Phi}^\top\widetilde{\mathbf{U}}\widehat{\mathbf{F}}\|_{\mathbb{F}}^2 &\leq 2\|\mathbf{B}^\top\boldsymbol{\Phi}^\top\widetilde{\mathbf{U}}(\widehat{\mathbf{F}} - \mathbf{FH})\|_{\mathbb{F}}^2 + 2\|\mathbf{B}^\top\boldsymbol{\Phi}^\top\widetilde{\mathbf{U}}\mathbf{FH}\|_{\mathbb{F}}^2\\ &\lesssim \left\{T^2 + p^2n^{2\alpha}T + p^4n^{4\alpha} + nT(T + p^2n^{2\alpha})/J^\kappa\right\}\{1 + \sqrt{\ln(1/\delta)}\}^2\{1 + n^{-1}\ln(1/\delta)\}.\end{aligned}$$

The result follows that $\|\mathbf{F}\|_{\mathbb{F}} = \|\widehat{\mathbf{F}}\|_{\mathbb{F}} = \sqrt{KT}$ with probability 1.

(ii) By Lemma 5, with probability at least $1 - 4\delta$, $T^{-2}\|\mathbf{F}^\top\mathbf{A}_2\|_{\mathbb{F}}^2 \leq n^{-2}T^{-4}\|\mathbf{F}^\top\widetilde{\mathbf{U}}^\top\boldsymbol{\Phi}\mathbf{B}\|_{\mathbb{F}}^2\|\mathbf{F}\|_{\mathbb{F}}^2\|\widehat{\mathbf{F}}\|_{\mathbb{F}}^2 \lesssim (nT)^{-1}p^2\{1 + \sqrt{\ln(1/\delta)}\}^2$.

(iii) Combining (i) and (ii), the result follows from Lemma 7.

(iv) The result follows from $T^{-1}\|\widehat{\mathbf{F}}^\top(\widehat{\mathbf{F}} - \mathbf{FH})\|_{\mathbb{F}} \leq T^{-1}\|\widehat{\mathbf{F}} - \mathbf{FH}\|_{\mathbb{F}}^2 + T^{-1}\|\mathbf{H}^\top\mathbf{F}^\top(\widehat{\mathbf{F}} - \mathbf{FH})\|_{\mathbb{F}}$. □

Lemma 10. *With probability at least* $1 - 20\delta$,

$$\|\boldsymbol{H}^\top\boldsymbol{H} - \boldsymbol{I}_K\|_{\mathbb{F}}^2 \lesssim \left(\frac{1}{n^2} + \frac{p^2}{n^{1-2\alpha}T} + \frac{1}{J^\kappa}\right)\{1 + \sqrt{\ln(1/\delta)}\}^2\{1 + n^{-1}\ln(1/\delta)\}.$$

Proof. By Condition 2, $\mathbf{F}^\top\mathbf{F} = T\mathbf{I}_K$ with probability 1 and $\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}} = T\mathbf{I}_K$. So $\mathbf{H}^\top\mathbf{H} = T^{-1}(\mathbf{FH})^\top\mathbf{FH} = T^{-1}(\mathbf{FH} - \widehat{\mathbf{F}})^\top\mathbf{FH} + T^{-1}\widehat{\mathbf{F}}^\top(\mathbf{FH} - \widehat{\mathbf{F}}) + \mathbf{I}_K$ and $\|\mathbf{H}^\top\mathbf{H} - \mathbf{I}_K\|_{\mathbb{F}} \leq T^{-1}\|(\widehat{\mathbf{F}} - \mathbf{FH})^\top\mathbf{F}\|_{\mathbb{F}}\|\mathbf{H}\|_2 + T^{-1}\|\mathbf{F}^\top(\widehat{\mathbf{F}} - \mathbf{FH})\|_{\mathbb{F}}$, which gives the desired result. □

Define $\widehat{\mathbf{B}} = T^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\widetilde{\mathbf{Y}}\widehat{\mathbf{F}}$ so that $\widehat{\mathbf{G}} = T^{-1}\mathbf{P}\widetilde{\mathbf{Y}}\widehat{\mathbf{F}} = \boldsymbol{\Phi}\widehat{\mathbf{B}}$, we have $\widehat{\mathbf{B}} - \mathbf{BH} = \sum_{i=1}^4\mathbf{C}_i$ where $\mathbf{C}_1 = T^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathbf{RF}^\top\widehat{\mathbf{F}}$, $\mathbf{C}_2 = T^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\widetilde{\mathbf{U}}\mathbf{FH}$, $\mathbf{C}_3 = T^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\widetilde{\mathbf{U}}(\widehat{\mathbf{F}} - \mathbf{FH})$, and $\mathbf{C}_4 = T^{-1}\mathbf{BF}^\top(\widehat{\mathbf{F}} - \mathbf{FH})$.

Proposition 2. *With probability at least* $1 - 20\delta$,

(i) $\|\widehat{\boldsymbol{B}} - \boldsymbol{BH}\|_{\mathbb{F}}^2 \lesssim \left\{n^{-2}J + n^{-1+2\alpha}T^{-1}p^2J + n^{-2+4\alpha}T^{-2}p^4J + J^{-\kappa+1}\right\}\{1 + J\ln(J^2/\delta)/n\}^3\{1 + \sqrt{\ln(1/\delta)}\}^4$,
(ii) $n^{-1}\|\widehat{\boldsymbol{G}} - \boldsymbol{GH}\|_{\mathbb{F}}^2 \lesssim \left(n^{-2}J + n^{-1+2\alpha}T^{-1}p^2J + n^{-2+4\alpha}T^{-2}p^4J + J^{-\kappa+1}\right)\{1 + J\ln(J^2/\delta)/n\}^4\{1 + \sqrt{\ln(1/\delta)}\}^4$.

Proof. (i) By Lemmas 1, 5 and 9, with probability at least $1 - 20\delta$,

$$\|\mathbf{C}_1\|_{\mathbb{F}}^2 \lesssim \frac{1}{J^\kappa}\{1 + J\ln(J^2/\delta)/n\}^3,$$

$$\|\mathbf{C}_2\|_{\mathbb{F}}^2 \lesssim \frac{p^2J}{n^{2\alpha}T}\{1 + J\ln(J^2/\delta)/n\}^2\{1 + \sqrt{\ln(1/\delta)}\}^2,$$

$$\|\mathbf{C}_3\|_{\mathbb{F}}^2 \lesssim \left(\frac{J}{n^2}+\frac{p^2J}{n^{2-2\alpha}T}+\frac{p^4J}{n^{2-4\alpha}T^2}+\frac{T+p^2}{nTJ^{\kappa-1}}\right)\{1+J\ln(J^2/\delta)/n\}^2\{1+\sqrt{\ln(1/\delta)}\}^4,$$

$$\|\mathbf{C}_4\|_{\mathbb{F}}^2 \lesssim \left(\frac{J}{n^2}+\frac{p^2J}{n^{1-2\alpha}T}+\frac{1}{J^{\kappa-1}}\right)\{1+J\ln(J^2/\delta)/n\}^3\{1+\sqrt{\ln(1/\delta)}\}^2.$$

So $\|\widehat{\mathbf{B}}-\mathbf{BH}\|_{\mathbb{F}}^2 \lesssim \left\{n^{-2}J+n^{-1+2\alpha}T^{-1}p^2J+n^{-2+4\alpha}T^{-2}p^4J+J^{-\kappa+1}\right\}\{1+J\ln(J^2/\delta)/n\}^3\{1+\sqrt{\ln(1/\delta)}\}^4$.

(ii) The result follows from $n^{-1}\|\widehat{\mathbf{G}}-\mathbf{GH}\|_{\mathbb{F}}^2 \le n^{-2}\|\boldsymbol{\Phi}(\widehat{\mathbf{B}}-\mathbf{BH})\|_{\mathbb{F}}^2+n^{-2}\|\mathbf{RH}\|_{\mathbb{F}}^2$. □

Proposition 3. *With probability at least* $1-20\delta$, *(i)* $\|\widehat{\boldsymbol{B}}-\boldsymbol{BH}\|_{\max} \lesssim n^{-1/2}T^{-1}\{(T+p^2n^{2\alpha})\ln(n)\}^{1/2}\{1+\ln(1/\delta)\}$, *(ii)* $\|\widehat{\boldsymbol{G}}-\boldsymbol{GH}\|_{\max} \lesssim T^{-1}\{(T+p^2n^{2\alpha})\ln(n)\}^{1/2}\{1+\ln(1/\delta)\}$, *and (iii)* $\|\widehat{\boldsymbol{G}}-\boldsymbol{GH}^{-1}\|_{\max} \lesssim T^{-1}\{(T+p^2n^{2\alpha})\ln(n)\}^{1/2}\{1+\ln(1/\delta)\}$.

Proof. (i) By Lemma B.1 in Fan et al. [17], with probability at least $1-\delta$, $\|\widetilde{\mathbf{FU}}\|_{\max} \lesssim \sqrt{(T+p^2)\ln(n)}\{1+\ln(1/\delta)\}$. Then, by Lemmas 1, 5, and 9, with probability at least $1-20\delta$, $\|\mathbf{C}_1\|_{\max} \lesssim \{\sqrt{n}TJ^{\kappa}\}^{-1}\{\ln(T)\}^{2/r_2}\{1+\ln(1/\delta)\}$, $\|\mathbf{C}_2\|_{\max} \lesssim \{\sqrt{n}T\}^{-1}\sqrt{(T+p^2n^{2\alpha})\ln(n)}\{1+\ln(1/\delta)\}$, $\|\mathbf{C}_3\|_{\max} \lesssim \{nT^2\}^{-1}(T+p^2n^{2\alpha})\{\ln(T)\}^{2/r_2}\{1+\ln(1/\delta)\}\delta$, and $\|\mathbf{C}_4\|_{\max} \lesssim \{\sqrt{n}T^2J^{\kappa}\}^{-1}(T+p^2n^{2\alpha})\{\ln(T)\}^{2/r_2}\{1+\ln(1/\delta)\}\delta$. So $\|\widehat{\mathbf{B}}-\mathbf{BH}\|_{\max} \lesssim n^{-1/2}T^{-1}\{(T+p^2n^{2\alpha})\ln(n)\}^{1/2}\{1+\ln(1/\delta)\}$.

(ii) The result follows from $\|\widehat{\mathbf{G}}-\mathbf{GH}\|_{\max} \le n^{-2}\|\boldsymbol{\Phi}(\widehat{\mathbf{B}}-\mathbf{BH})\|_{\max}+n^{-2}\|\mathbf{RH}\|_{\max}.\delta$

(iii) The result follows from $\widehat{\mathbf{G}}-\mathbf{GH}^{-1}=T^{-1}\mathbf{GH}^{-1}(\mathbf{HF}^{\top}-\widehat{\mathbf{F}}^{\top})\widehat{\mathbf{F}}+T^{-1}\widetilde{\mathbf{PU}}(\widehat{\mathbf{F}}-\mathbf{FH})+T^{-1}\widetilde{\mathbf{PU}}\mathbf{FH}$. □

Proposition 4. *With probability at least* $1-20\delta$,

$$\|\boldsymbol{H}-\boldsymbol{I}_K\|_{\mathbb{F}}^2 \lesssim \left(\frac{1}{n^2}+\frac{p^2}{n^{1-2\alpha}T}+\frac{1}{J^{\kappa}}\right)\{1+\sqrt{\ln(1/\delta)}\}^2\{1+n^{-1}\ln(1/\delta)\}.$$

Proof. Note that $\mathbf{HK}=n^{-1}\mathbf{B}^{\top}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\mathbf{B}(T^{-1}\mathbf{F}^{\top}\widehat{\mathbf{F}}-\mathbf{H})+n^{-1}\mathbf{B}^{\top}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\mathbf{BH}$. By Lemma 9, with probability at least $1-20\delta$,

$$\|n^{-1}\mathbf{B}^{\top}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\mathbf{B}(T^{-1}\mathbf{F}^{\top}\widehat{\mathbf{F}}-\mathbf{H})\|_{\mathbb{F}} \lesssim \{n^{-1}+p(n^{1-2\alpha}T)^{-1/2}+J^{-\kappa/2}\}\{1+\sqrt{\ln(1/\delta)}\}\sqrt{1+n^{-1}\ln(1/\delta)}.$$

In addition, by Conditions 3 and 5, $\|\mathbf{G}^{\top}\mathbf{G}-\mathbf{B}^{\top}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\mathbf{B}\|_{\mathbb{F}} \lesssim nJ^{-\kappa/2}$. Therefore, with probability at least $1-20\delta$,

$$\|n^{-1}\mathbf{G}^{\top}\mathbf{GH}-\mathbf{HK}\|_{\mathbb{F}} \lesssim \{n^{-1}+p(n^{1-2\alpha}T)^{-1/2}+J^{-\kappa/2}\}\{1+\sqrt{\ln(1/\delta)}\}\sqrt{1+n^{-1}\ln(1/\delta)}.$$

This implies that with probability at least $1-20\delta$, $\mathbf{H}$ (up to an error term) is a matrix consisting of eigenvectors of $n^{-1}\mathbf{G}^{\top}\mathbf{G}$. By Condition 2, $\mathbf{G}^{\top}\mathbf{G}$ is a diagonal matrix with distinct eigenvalues with probability 1. Thus, each eigenvalue is associated with a unique unitary eigenvector up to a sign change and each eigenvector has a single non-zero entry. Thus, with probability at least $1-20\delta$,

$$\|\mathbf{H}-\mathbf{D}\|_{\mathbb{F}} \lesssim \{n^{-1}+p(n^{1-2\alpha}T)^{-1/2}+J^{-\kappa/2}\}\{1+\sqrt{\ln(1/\delta)}\}\sqrt{1+n^{-1}\ln(1/\delta)}$$

for some diagonal matrix $\mathbf{D}$. By Lemma 10, with probability at least $1-20\delta$, for each $k\in\{1,\dots,K\}$,

$$|\lambda_k(\mathbf{H})-\eta| \lesssim \{n^{-1}+p(n^{1-2\alpha}T)^{-1/2}+J^{-\kappa/2}\}\{1+\sqrt{\ln(1/\delta)}\}\sqrt{1+n^{-1}\ln(1/\delta)},$$

where η is either 1 or -1. Without loss of generality , we can assume that all entries of $\mathbf{H}$ is positive (otherwise we can multiply the corresponding columns of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{G}}$ by -1). Hence, with probability at least $1-20\delta$,

$$\|\mathbf{H}-\mathbf{I}_K\|_{\mathbb{F}}^2=\sum_{i\neq j}h_{ij}^2+\sum_{i=1}^{K}(h_{ii}-1)^2 \lesssim \{n^{-2}+p^2(n^{1-2\alpha}T)^{-1}+J^{-\kappa}\}\{1+\sqrt{\ln(1/\delta)}\}^2\{1+n^{-1}\ln(1/\delta)\}.$$ □

Finally, we present technical results for establishing Theorem 2 in Lemmas 11–14. Recall that $\mathcal{V}(\boldsymbol{f}_t)=T^{-2}\sum_{t=-T+1}^{T-1}(T-|t|)\widehat{\boldsymbol{\Sigma}}_f(t)$ as defined in Section 2.2.2, where $\widehat{\boldsymbol{\Sigma}}_f(s)=(T-s)^{-1}\sum_{t=1}^{T-s}(\boldsymbol{f}_t-\bar{\boldsymbol{f}})(\boldsymbol{f}_{t+s}-\bar{\boldsymbol{f}})^{\top}$ and $\widehat{\boldsymbol{\Sigma}}_f(-s)=(T-s)^{-1}\sum_{t=s}^{T}(\boldsymbol{f}_{t-s}-\bar{\boldsymbol{f}})(\boldsymbol{f}_t-\bar{\boldsymbol{f}})^{\top}$ for $s\ge 0$, respectively.

Lemma 11. *Under Condition 2, with probability at least* $1-\delta$,

$$\|\mathcal{V}(\widehat{\boldsymbol{f}}_t)-\mathcal{V}(\boldsymbol{f}_t)\|_{\mathbb{F}} \lesssim \frac{1}{T}\left(\frac{1}{\sqrt{n}}+\frac{p}{\sqrt{n^{1-2\alpha}T}}+\frac{1}{J^{\kappa/2}}\right)\{1+\sqrt{\ln(20/\delta)}\}.$$

Proof.

Note that $\mathcal{V}(\boldsymbol{f}_t)=T^{-2}\sum_{t,s=1}^{T}(\boldsymbol{f}_t-\bar{\boldsymbol{f}})(\boldsymbol{f}_s-\bar{\boldsymbol{f}})^{\top}=T^{-2}\mathbf{F}^{\top}\mathbf{P}_1\mathbf{F}$, where $\mathbf{P}_1$ is the projection matrix onto $(1,\dots,1)^{\top}\in\mathbb{R}^T$. Thus, by Theorem 1

$$\|\mathcal{V}(\widehat{\boldsymbol{f}}_t)-\mathcal{V}(\boldsymbol{f}_t)\|_{\mathbb{F}}^2=\frac{1}{T^4}\|\widehat{\mathbf{F}}\mathbf{P}_1\widehat{\mathbf{F}}^{\top}-\mathbf{F}\mathbf{P}_1\mathbf{F}^{\top}\|_{\mathbb{F}}^2 \lesssim \frac{1}{T^2}\left(\frac{1}{n}+\frac{p^2}{n^{1-2\alpha}T}+\frac{1}{J^{\kappa}}\right)\{1+\sqrt{\ln(20/\delta)}\}^2.$$

The conclusion follows. □

Lemma 12. *Under Conditions 1, 2, and 7,* $\|\mathcal{V}(\boldsymbol{f}_t) - \mathrm{Var}(T^{-1}\sum_{t=1}^T \boldsymbol{f}_t)\|_{\mathbb{F}} \lesssim T^{-2}$.

Proof. Recall that $\mathrm{Var}(T^{-1}\sum_{t=1}^T \boldsymbol{f}_t) = T^{-2}\sum_{t,s}\mathrm{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_s)$ and $\mathcal{V}(\boldsymbol{f}_t) = T^{-2}\sum_{t,s}(\boldsymbol{f}_t - \bar{\boldsymbol{f}})(\boldsymbol{f}_s - \bar{\boldsymbol{f}})^\top$. By Davydov's inequality [1], for each $k \in \{1,\dots,K\}$ and $t,s \in \{1,\dots,T\}$, $|\mathrm{E}(f_{tk}f_{sk})^2| \lesssim \{\alpha(|t-s|)\}^{1/r_1}\{\mathrm{E}(|f_{tk}|^{2q_1})\}^{1/q_1}\{\mathrm{E}(|f_{sk}|^{2q_2})\}^{1/q_2}$, for some $q_1, q_2 > 0$ such that $1/r_1 + 1/q_1 + 1/q_2 = 1$, where $\alpha(\cdot)$ is the α-mixing coefficient. By Condition 7, $\mathrm{E}(|f_{tk}|^{q_1})$ and $\mathrm{E}(|f_{sk'}|^{q_2})$ exist for each $t \in \{1,\dots,T\}$ and $\alpha(|t-s|) < \exp(-C_1|t-s|^{r_1})$, so $|\mathrm{E}(f_{tk}f_{sk})^2| \lesssim \exp(-|t-s|)$. Thus, $\|\mathrm{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_s)\|_{\mathbb{F}} \lesssim \exp(-|t-s|)$ and $\|\mathcal{V}(\boldsymbol{f}_t) - \mathrm{Var}(T^{-1}\sum_{t=1}^T \boldsymbol{f}_t)\|_{\mathbb{F}} = \|T^{-2}\sum_{t,s}\mathrm{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_s)\|_{\mathbb{F}} \lesssim T^{-2}\sum_{t=1}^T \exp(-t) \lesssim T^{-2}$. □

Lemma 13. *For each $i \in \{1,\dots,n\}$, with probability at least $1-\delta$, $|\mathcal{V}(\widehat{u}_{it}) - \mathrm{Var}(T^{-1}\sum_{t=1}^T u_{it})| \lesssim T^{-1}[\{\sqrt{n}T\}^{-1} + n^{-3/2+\alpha} + \{T^{1/2}n^{3/2}\}^{-1}p + \{\sqrt{n^2T}\}^{-1}\{(T+p^2)\ln(n)\}^{1/4} + n^{-1}J^{-\kappa/2}]\{1 + \sqrt{\ln(21/\delta)}\}$, where $\mathcal{V}(\widehat{u}_{1t})$ is defined in Section 2.2.2.*

Proof. Denote $\widehat{\mathbf{U}} = \{\widehat{u}_{it}\}_{i=1,t=1}^{n,T}$. Note that $\mathbf{U} - \widehat{\mathbf{U}} = (\widehat{\mathbf{G}} - \mathbf{G}\mathbf{H}^{-1})(\widehat{\mathbf{F}}^\top - \mathbf{H}\mathbf{F}^\top) + \mathbf{G}\mathbf{H}^{-1}(\widehat{\mathbf{F}}^\top - \mathbf{H}\mathbf{F}^\top) + (\widehat{\mathbf{G}} - \mathbf{G}\mathbf{H}^{-1})\mathbf{H}\mathbf{F}^\top$. By Propositions 1 and 3, with probability at least $1 - 20\delta$, $T^{-1}\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\mathbb{F}}^2 \lesssim \{n^{-1} + \{n^{1-2\alpha}T\}^{-1}p^2 + T^{-1}\sqrt{(T+p^2n^{2\alpha})\ln(n)} + J^{-\kappa}\}\{1 + \ln(1/\delta)\}^2$. Thus, similarly to the proof of Lemmas 11 and 12, with probability at least $1-\delta$, $|\mathcal{V}(\widehat{u}_{it}) - \mathcal{V}(u_{it})| \lesssim (nT)^{-1}[n^{-1/2} + (n^{1-2\alpha}T)^{-1/2}p + T^{-1/2}\{(T+p^2n^{2\alpha})\ln(n)\}^{1/4} + J^{-\kappa/2}]\{1 + \sqrt{\ln(20/\delta)}\}$ and $|\mathcal{V}(u_{it}) - \mathrm{Var}(T^{-1}\sum_{t=1}^T u_{it})| \lesssim \{\sqrt{n}T^2\}^{-1}\{1 + \sqrt{\ln(1/\delta)}\}$. The conclusion follows. □

Lemma 14. *With probability at least $1-\delta$,*

$$\|\boldsymbol{V}^{-1}(\widehat{\boldsymbol{V}} - \boldsymbol{V})\|_2 \lesssim \left\{\frac{\sqrt{J}}{n} + \frac{1}{\sqrt{n}} + \frac{1}{T} + \frac{p\sqrt{J}}{\sqrt{n^{1-2\alpha}T}} + \frac{1}{J^{(\kappa-1)/2}}\right\}\{1 + \ln(21/\delta)\}.$$

Proof. Recall that $\mathbf{V} = \mathbf{G}\,\mathrm{Var}(T^{-1}\sum_{t=1}^T \boldsymbol{f}_t)\mathbf{G}^\top + \boldsymbol{\mathcal{D}}$, so $\lambda_{\min}(\mathbf{V}) \geq \lambda_{\min}\{\mathbf{G}\,\mathrm{Var}(T^{-1}\sum_{t=1}^T \boldsymbol{f}_t)\mathbf{G}^\top\} + \lambda_{\min}(\mathbf{D}) \gtrsim T^{-1}$. Note that $\widehat{\mathbf{V}} - \mathbf{V} = \mathbf{G}\{\mathcal{V}(\widehat{\boldsymbol{f}}_t) - \mathrm{Var}(T^{-1}\sum_{t=1}^T \boldsymbol{f}_t)\}\mathbf{G}^\top + (\widehat{\mathbf{G}} - \mathbf{G})\mathcal{V}(\widehat{\boldsymbol{f}}_t)\widehat{\mathbf{G}}^\top + \mathbf{G}\mathcal{V}(\widehat{\boldsymbol{f}}_t)(\widehat{\mathbf{G}} - \mathbf{G})^\top + (\widehat{\boldsymbol{\mathcal{D}}} - \boldsymbol{\mathcal{D}})$. In addition, by the proof of Theorem 2 in Fan et al. [15], $\|\mathbf{G}\mathbf{V}^{-1}\mathbf{G}\|_2 = O(T)$. Thus,

$$\|\mathbf{V}^{-1}(\widehat{\mathbf{V}} - \mathbf{V})\|_2 \leq \|\mathcal{V}(\widehat{\boldsymbol{f}}_t) - \mathrm{Var}(T^{-1}\sum_{t=1}^T \boldsymbol{f}_t)\|_{\mathbb{F}} + 2\|\mathcal{V}(\widehat{\boldsymbol{f}}_t)\|_{\mathbb{F}}\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathbb{F}} + \|\widehat{\boldsymbol{\mathcal{D}}} - \boldsymbol{\mathcal{D}}\|_{\mathbb{F}}.$$

From Lemmas 11 and 12, with probability at least $1-\delta$,

$$\left\|\mathcal{V}(\widehat{\boldsymbol{f}}_t) - \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}_t\right)\right\|_{\mathbb{F}} \lesssim \frac{1}{T}\left(\frac{1}{\sqrt{n}} + \frac{1}{T} + \frac{p}{\sqrt{n^{1-2\alpha}T}} + \frac{1}{J^{\kappa/2}}\right)\{1 + \sqrt{\ln(21/\delta)}\},$$

and

$$\|\widehat{\boldsymbol{\mathcal{D}}} - \boldsymbol{\mathcal{D}}\|_{\mathbb{F}} \lesssim \frac{1}{T}\left[\frac{1}{T} + \frac{1}{n} + \frac{p}{n^{1-2\alpha}T^{1/2}} + \frac{\{(T+p^2n^{2\alpha})\ln(n)\}^{1/4}}{\sqrt{nT}} + \frac{1}{\sqrt{n}J^{\kappa/2}}\right]\{1 + \sqrt{\ln(21/\delta)}\}$$

which leads to the desired assertion by Lemma 13 and Theorem 1. □

As a straightforward corollary to Lemma 14, with probability at least $1-\delta$, $\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\mathbf{V},\mathbb{F}} \lesssim \{n^{-1}\sqrt{J} + n^{-1/2} + T^{-1} + (\sqrt{n^{1-2\alpha}T})^{-1}p\sqrt{J} + J^{-(\kappa-1)/2}\}\sqrt{\ln(1/\delta)}$, where $\|\mathbf{A}\|_{\mathbf{S},\mathbb{F}} := n^{-1/2}\|\mathbf{S}^{-1/2}\mathbf{A}\mathbf{S}^{-1/2}\|_{\mathbb{F}}$. If $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$ are independent across t, then $\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\mathbf{V},\mathbb{F}} \lesssim \{n^{-1}\sqrt{J} + p\sqrt{J}n^{-1/2+\alpha}T^{-1/2} + J^{-(\kappa-1)/2}\}\sqrt{\ln(1/\delta)}$, which mimics the optimal rate from Fan et al. [18] and Wang and Fan [43].

8.3. Discussions on legitimate preliminary $\widehat{\boldsymbol{\beta}}^0$

In this section, we will discuss some preliminary estimators $\widehat{\boldsymbol{\beta}}^0$ that satisfy the condition of TOPE, i.e., $\|\widehat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}\|_2 = O_P(n^{-1/2+\alpha}T^{-1/2})$ for $\alpha \in [0, 1/2)$ in Section 2.2.2. In fact, Conditions 1, 2, and 6 guarantee the existence of such a preliminary $\widehat{\boldsymbol{\beta}}^0$. We start with an OLS estimator based on an average version of (3) over time,

$$\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (\mathbb{Z}_0^\top\mathbb{Z}_0)^{-1}\mathbb{Z}_0^\top\tilde{\boldsymbol{y}}. \tag{14}$$

Before showing that $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ is a legitimate preliminary in Proposition 5, we first collect some technical results in Lemmas 15–17.

Lemma 15. *Under Condition 7, $v_i(T) = T^{-1/2}\sum_{t=1}^T u_{it}$ is sub-exponential for each $i \in \{1,\dots,n\}$.*

Proof. Note that $\mathrm{E}(|u_{it}|^{4+\delta_1}) < \infty$ for any $t \in \{1, \ldots, T\}$, $i \in \{1, \ldots, n\}$ and $\delta_1 > 0$ and $\sum_{T=0}^{\infty} \alpha(T)^{1/3} < \sum_{t=0}^{\infty} \exp(-C T^{r_1}/3) < \infty$. By Theorem 4 in Tikhomirov [42], $|\Pr\{v_i(T) < s\} - \Pr(W_i < s)| \leq C_1 T^{-1/2}(1 + |s|)^{-4}\{\ln(T)\}^3$ for each $i \in \{1, \ldots, n\}$ and any s, where $W \sim N(0, \sigma_i^2)$ and $\sigma_i^2 = \mathrm{E}(u_{i1}^2) + 2\sum_{t=2}^{\infty} \mathrm{E}(u_{i1}u_{it})$. Thus, we have

$$\Pr\{|v_i(T)| > s\} = \Pr(|T^{-1/2}\sum_{t=1}^{T} u_{it}| > s) \leq 2\exp\{-s^2/(2\sigma_i^2)\} + C_1 T^{-1/2}(1+s)^{-4}\{\ln(T)\}^3$$

for any T and constants $C_1 > 0$. Furthermore, for any $k \in \{1, 2, \ldots\}$,

$$\begin{aligned}\mathrm{E}\{|v_i(T)|^k\} &= \int_0^1 \Pr\{|v_i(T)| > s^{1/k}\}ds + \int_1^{\infty} \Pr\{|v_i(T)| > s^{1/k}\}ds \\ &\leq 1 + 2(2\sigma_i^2)^{k/2}k\Gamma(k/2) + C_1\pi T^{-1/2}\{\ln(T)\}^3 k!,\end{aligned}$$

so that $\mathrm{E}[\exp\{s v_i(T)\}] \leq 1 + \sum_{k=2}^{\infty} |s|^k \mathrm{E}\{|v_i(T)|^k\}/k! \lesssim \exp\{2\sigma_i^2 s^2 + C_1\pi T^{-1/2}\{\ln(T)\}^3\}$ for $|s| < \min\{1/\sigma_i, 1\}$. The assertion follows from the definition of sub-exponential distributions. □

Lemma 16. *Under Conditions 1 and 7, for any $s > 0$, $p \times n$ matrix $\boldsymbol{A}$ and $\sigma^2 = \max_i \sigma_i^2$ with σ_i^2 defined in Lemma 15, $\Pr\{\|\boldsymbol{A}\sum_{t=1}^{T} \boldsymbol{u}_t/T\|_2 > s\|\boldsymbol{A}\|_{\mathbb{F}}/\sqrt{T}\} < 2p\exp\{-s^2/(2\sigma^2)\}$.*

Proof. Write $\mathbf{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p)^{\top}$, where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p$ are n-dimensional vectors. For each $m \in \{1, \ldots, p\}$ and $w \geq 0$, by Conditions 1, 2, and 7, Lemma 15, and Corollary 4 in Samson [36], $\Pr(|\boldsymbol{a}_m^{\top}\sum_{t=1}^{T} \boldsymbol{u}_t/T| \geq s) = \Pr(|\sum_{i=1}^{n} a_{mi}v_i(T)| \geq s\sqrt{T}) \leq 2\exp\{-s^2T/(2\sigma^2\|\boldsymbol{a}_m\|_2^2)\}$. Hence, $\Pr\{|\boldsymbol{a}_m^{\top}\sum_{t=1}^{T} \boldsymbol{u}_t/T| > \|\boldsymbol{a}_m\|_2 s/\sqrt{T}\} \leq 2\exp\{-s^2/(2\sigma^2)\}$ for any $m \in \{1, \ldots, p\}$ and $\Pr\{\|\mathbf{A}\sum_{t=1}^{T} \boldsymbol{u}_t/T\|_2 > \|\mathbf{A}\|_{\mathbb{F}} s/\sqrt{T}\} \leq 2p\exp\{-s^2/(2\sigma^2)\}$. □

Conclusion in Lemma 16 remains valid for correlated $\{u_{it}\}_{t=1}^{T}$ over i. In fact, if one assumes cross-sectional dependence of $\{u_{it}\}$ over i by letting $\max_{j\leq n}\sum_{i=1}^{n}|\mathrm{E}(u_{it}u_{jt})| < C_2$, $\max_{i\leq n}\sum_{k=1}^{n}\sum_{m=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T}|\mathrm{cov}(u_{it}u_{kt}, u_{is}u_{ms})| < C_2$, and $(nT)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T}|\mathrm{E}(u_{it}u_{js})| < C_2$ for some $C_2 > 0$, Corollary 4 in Samson [36] still applies.

Lemma 17. *For $p \times K$ matrix $\boldsymbol{A}$, under Conditions 1 and 7, $\Pr\{\|\boldsymbol{A}\sum_{t=1}^{T} \boldsymbol{f}_t/T\|_2 > s\|\boldsymbol{A}\|_{\mathbb{F}}/\sqrt{T}\} \leq 2pC_3\exp(-C_4 s^2/2)$ for constants $C_3, C_4 > 0$.*

Proof. The proof is similar to that of Lemma 16 and omitted here. □

Proposition 5. *Under Conditions 1, 2, and 6, with probability at least $1 - \delta$,*

$$\|\widehat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta}\|_2^2 \lesssim \frac{p^2}{n^{1-2\alpha}T}\ln(1/\delta).$$

Proof. Combining (14) and $\bar{\boldsymbol{y}} = \mathbb{Z}_0^{\top}\boldsymbol{\beta} + \mathbf{G}T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t + T^{-1}\sum_{t=1}^{T}\boldsymbol{u}_t$, we have $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \boldsymbol{\beta} + (\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}\mathbf{G}(T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t) + (\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}(T^{-1}\sum_{t=1}^{T}\boldsymbol{u}_t) \equiv \boldsymbol{\beta} + (\mathrm{I}) + (\mathrm{II})$, where $(\mathrm{I}) = (\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}\mathbf{G}(T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t)$ and $(\mathrm{II}) = (\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}(T^{-1}\sum_{t=1}^{T}\boldsymbol{u}_t)$. By Condition 6, with probability 1, $\|\mathbf{P}_Z\mathbf{G}\|_{\mathbb{F}}^2 \lesssim n^{2\alpha}$. In addition, eigenvalues of $n^{-1}\mathbb{Z}_0^{\top}\mathbb{Z}_0$ is bounded away from 0 and infinity almost surely by Condition 6(i). Thus, eigenvalues of $(n^{-1}\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}$ are bounded away from 0 and infinity almost surely. That is, $\|(\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}\|_{\mathbb{F}}^2 = \mathrm{tr}\{(\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\} \lesssim n^{-1}p$, and thus, $\|(\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}\mathbf{G}\|_{\mathbb{F}}^2 \leq \|(\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}\|_{\mathbb{F}}^2 \|\mathbf{P}_Z\mathbf{G}\|_{\mathbb{F}}^2 \lesssim n^{-1+2\alpha}p^2$ by Cauchy–Schwarz inequality. In light of Lemma 17, we have $\Pr\{\|(\mathrm{I})\|_2 > sT^{-1/2}\|(\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}\mathbf{G}\|_{\mathbb{F}}\} < C_1\exp(-C_2 s^2)$. By Lemma 16, it holds $\Pr\{\|(\mathrm{II})\|_2 > sT^{-1/2}\|(\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}\|_{\mathbb{F}}\} < C_1\exp(-C_2 s^2)$. Thus, we have $\Pr[\|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}\|_2 > sT^{-1/2}\{\|(\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}\mathbf{G}\|_{\mathbb{F}} + \|(\mathbb{Z}_0^{\top}\mathbb{Z}_0)^{-1}\mathbb{Z}_0^{\top}\|_{\mathbb{F}}\}] < 2C_1\exp(-C_2 s^2)$. □

Proposition 5 implies that $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ is a legitimate preliminary for TOPE. Alternatively, one may consider the following choice on $\widehat{\boldsymbol{\beta}}^0$. Rewrite $\boldsymbol{g}(\boldsymbol{x}_i)$ as $\boldsymbol{g}(\boldsymbol{x}_i) = \mathbf{A}\boldsymbol{z}_{i\cdot} + \boldsymbol{g}_0(\boldsymbol{x}_i)$, where $\mathbf{A}$ is a $K \times p$ matrix and $\boldsymbol{z}_{i\cdot} = T^{-1}\sum_{t=1}^{T}\boldsymbol{z}_{it}$ is the average of $\boldsymbol{z}_{it}$ over time. Then, model (1) can be rewritten as $y_{it} = \boldsymbol{z}_{it}^{\top}\boldsymbol{\beta} + \boldsymbol{z}_{i\cdot}^{\top}\boldsymbol{\eta}_t + \boldsymbol{g}_0(\boldsymbol{x}_i)^{\top}\boldsymbol{f}_t + u_{it}$, where $\boldsymbol{\eta}_t = \mathbf{A}^{\top}\boldsymbol{f}_t$. Under Condition 1, $\boldsymbol{g}_0(\boldsymbol{x}_i)^{\top}\boldsymbol{f}_t + u_{it}$ is uncorrelated with the regressors $\boldsymbol{z}_{it}$. Hence, we can use the following random-effects GLS [37] to estimate

$(\boldsymbol{\beta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)$ by $(\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\eta}}_1^\top, \ldots, \widehat{\boldsymbol{\eta}}_T^\top)^\top = (\mathbf{W}^\top \widehat{\boldsymbol{\Sigma}}_R^{-1} \mathbf{W})^{-1} \mathbf{W}^\top \widehat{\boldsymbol{\Sigma}}_R^{-1} \mathbf{y}$, where $\mathbf{y} = (y_{11}, \ldots, y_{n1}, \ldots, y_{1T}, \ldots, y_{nT})^\top$,

$$\mathbf{W} = \begin{bmatrix} \boldsymbol{z}_{11}^\top & \boldsymbol{z}_{1\cdot}^\top & & \\ \vdots & \vdots & & \\ \boldsymbol{z}_{n1}^\top & \boldsymbol{z}_{n\cdot}^\top & & \\ \vdots & & \ddots & \\ \boldsymbol{z}_{1T}^\top & & & \boldsymbol{z}_{1\cdot}^\top \\ \vdots & & & \vdots \\ \boldsymbol{z}_{nT}^\top & & & \boldsymbol{z}_{n\cdot}^\top \end{bmatrix},$$

and $\widehat{\boldsymbol{\Sigma}}_R$ is an estimator of $\boldsymbol{\Sigma}_R$, the covariance matrix of $\mathbf{v} = (\boldsymbol{g}_0(\boldsymbol{x}_1)^\top \boldsymbol{f}_1 + u_{11}, \ldots, \boldsymbol{g}_0(\boldsymbol{x}_n)^\top \boldsymbol{f}_1 + u_{n1}, \ldots, \boldsymbol{g}_0(\boldsymbol{x}_1)^\top \boldsymbol{f}_T + u_{1T}, \ldots, \boldsymbol{g}_0(\boldsymbol{x}_n)^\top \boldsymbol{f}_T + u_{nT})^\top$. Under Condition 1, $\boldsymbol{\Sigma}_R$ is a block diagonal matrix $\mathrm{diag}(\boldsymbol{\Sigma}_{R,1}, \ldots, \boldsymbol{\Sigma}_{R,T})$ with $\boldsymbol{\Sigma}_{R,t} = \mathrm{E}\{(\boldsymbol{g}_0(\boldsymbol{x}_1), \ldots, \boldsymbol{g}_0(\boldsymbol{x}_n))^\top (\boldsymbol{g}_0(\boldsymbol{x}_1), \ldots, \boldsymbol{g}_0(\boldsymbol{x}_n))\} + \sigma_u^2 \mathbf{I}_n$ for each $t \in \{1, \ldots, T\}$, where $\mathrm{var}(u_{it}) = \sigma_u^2$. There are a variety of estimators of $\widehat{\boldsymbol{\Sigma}}_{R,1}$. For instance, Bai [3] and Schmidheiny and Basel [37] estimated $\widehat{\boldsymbol{\Sigma}}_{R,1}$ by first estimating $\mathbf{v}$, which is achieved via the OLS estimator. This is the so-called feasible GLS estimator [3,26,27] and can be extended to the iterative feasible GLS estimator [3,34]. That is, we can update $\widehat{\boldsymbol{\Sigma}}_{R,1}^{\mathrm{new}}$ using $(\widehat{\boldsymbol{\beta}}^{\mathrm{old}}, \widehat{\boldsymbol{\eta}}_1^{\mathrm{old}}, \ldots, \widehat{\boldsymbol{\eta}}_T^{\mathrm{old}})$ from the previous step and iteratively update $(\widehat{\boldsymbol{\beta}}^{\mathrm{new}}, \widehat{\boldsymbol{\eta}}_1^{\mathrm{new}}, \ldots, \widehat{\boldsymbol{\eta}}_T^{\mathrm{new}})$ using the update $\widehat{\boldsymbol{\Sigma}}_{R,1}^{\mathrm{new}}$. The update $(\widehat{\boldsymbol{\beta}}^{\mathrm{new}}, \widehat{\boldsymbol{\eta}}_1^{\mathrm{new}}, \ldots, \widehat{\boldsymbol{\eta}}_T^{\mathrm{new}})$ admits the following shrinkage of errors.

Proposition 6 (*Lemma 1 in Phillips [34]*). *Under Conditions C1 to C3 in Phillips [34], if $T \geq p + 1$ and $\boldsymbol{A}_0 = \mathrm{E}(\boldsymbol{W}^\top \boldsymbol{\Sigma}_R^{-1} \boldsymbol{W})$ is nonsingular, $\sqrt{n}\{((\widehat{\boldsymbol{\beta}}^{new})^\top, (\widehat{\boldsymbol{\eta}}_1^{new})^\top, \ldots, (\widehat{\boldsymbol{\eta}}_T^{new})^\top)^\top - (\boldsymbol{\beta}^\top, \boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_T^\top)^\top\} = 2(T-1)^{-1}\sqrt{n}T\{((\widehat{\boldsymbol{\beta}}^{old})^\top, (\widehat{\boldsymbol{\eta}}_1^{old})^\top, \ldots, (\widehat{\boldsymbol{\eta}}_T^{old})^\top)^\top - (\boldsymbol{\beta}^\top, \boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_T^\top)^\top\}\boldsymbol{\psi}\boldsymbol{A}_0^{-1}\boldsymbol{\psi} + o_P(1)$, where $\boldsymbol{\psi}$ is given in Phillips [34].*

Together along with $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leq \|(\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\eta}}_1^\top, \ldots, \widehat{\boldsymbol{\eta}}_T^\top)^\top - (\boldsymbol{\beta}^\top, \boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_T^\top)^\top\|_2$, Proposition 6 implies that the iterative feasible GLS estimator improves as the iteration grows. Thus, upon some iterations, the iterative feasible GLS estimator also provide a legitimate preliminary estimator for TOPE.

CRediT authorship contribution statement

Lyuou Zhang: Methodology, Theoretical analysis, Numerical study, Data analysis, Writing – original draft, Writing – review & editing. **Wen Zhou:** Conceptualization, Methodology, Theoretical analysis, Writing – original draft, Writing – review & editing, Supervision. **Haonan Wang:** Conceptualization, Methodology, Theoretical analysis, Writing – original draft, Writing – review & editing, Supervision.

Acknowledgments

The authors thank the Editor-in-Chief, an Associate Editor, and two reviewers for many helpful and constructive comments. The work of Wen Zhou was partially supported by Department of Energy, USA grant DE-SC0018344 and National Science Foundation, USA grants IIS-1545994 and IOS-1922701. The research of Haonan Wang was partially supported by National Science Foundation, USA grants DMS-1737795, DMS-1923142 and CNS-1932413.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2021.104786.

References

[1] K.B. Athreya, S.N. Lahiri, Measure Theory and Probability Theory, Springer Science & Business Media, New York, 2006.
[2] J. Bai, Inferential theory for factor models of large dimensions, Econometrica 71 (2003) 135–171.
[3] J. Bai, Panel data models with interactive fixed effects, Econometrica 77 (2009) 1229–1279.
[4] J. Bai, K. Li, Theory and methods of panel data models with interactive effects, Ann. Statist. 42 (2014) 142–170.
[5] J. Bai, S. Ng, Principal components estimation and identification of static factors, J. Econometrics 176 (2013) 18–29.
[6] B.H. Baltagi, Econometrics, fourth ed., Springer-Verlag, New York, 2008.
[7] M. Barigozzi, H. Cho, P. Fryzlewicz, Simultaneous multiple change-point and factor analysis of high-dimensional time series, J. Econometrics 206 (2018) 87–225.
[8] D. Bianchi, M. Billio, R. Casarin, M. Guidolin, Modeling systemic risk with Markov switching graphical SUR models, J. Econometrics 210 (2019) 58–74.
[9] P. Billingsley, Probability and Measure, John Wiley & Sons, New York, 2012.
[10] M. Cao, W. Zhou, F.J. Breidt, G. Peers, Large scale maximum average power multiple inference on time-course count data with application to RNA-seq analysis, Biometrics 76 (2020) 9–22.

[11] G. Chamberlain, M. Rothschild, Arbitrage, factor structure, and mean–variance analysis on large asset markets, Econometrica 51 (1983) 1281–1304.
[12] G. Connor, M. Hagmann, O. Linton, Efficient semiparametric estimation of the fama-french model and extensions, Econometrica 80 (2012) 713–754.
[13] G. Connor, O. Linton, Semiparametric estimation of a characteristic-based factor model of common stock returns, J. Empir. Financ. 14 (2007) 694–717.
[14] Z.J. Daye, J. Chen, H. Li, High-dimensional heteroscedastic regression with an application to eQTL data analysis, Biometrics 68 (2012) 316–326.
[15] J. Fan, Y. Fan, J. Lv, High dimensional covariance matrix estimation using a factor model, J. Econometrics 147 (2008) 186–197.
[16] J. Fan, T. Huang, Profile likelihood inference on semiparametric varying-coefficient partially linear models, Bernoulli 11 (2005) 1031–1057.
[17] J. Fan, Y. Liao, M. Mincheva, High dimensional covariance matrix estimation in approximate factor models, Ann. Statist. 39 (2011) 3320–3356.
[18] J. Fan, Y. Liao, M. Mincheva, Large covariance estimation by thresholding principal orthogonal complements, J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (2013) 603–680.
[19] J. Fan, Y. Liao, W. Wang, Projected principal component analysis in factor models, Ann. Statist. 44 (2016) 219–254.
[20] M. Hallin, R. Liška, Dynamic factors in the presence of block structure, J. Econometrics 163 (2011) 29–41.
[21] T. Hastie, R. Tibshirani, Generalized additive models, Statist. Sci. 1 (1986) 297–318.
[22] R. Hauser, H. Liang, J.D. Meeker, H. Su, S.W. Thurston, Empirical likelihood based inference for additive partial linear measurement error models, Stat. Interface 2 (2009) 83–90.
[23] J.Z. Huang, C.O. Wu, L. Zhou, Polynomial spline estimation and inference for varying coefficient models with longitudinal data, Statist. Sinica 14 (2004) 763–788.
[24] J. Jiang, REML estimation: asymptotic behavior and related topics, Ann. Statist. 24 (1996) 255–286.
[25] I.T. Jolliffe, Principal Component Analysis, second ed., Springer-Verlag, New York, 2002.
[26] C. Lam, Q. Yao, Factor modeling for high-dimensional time series: inference for the number of factors, Ann. Statist. 40 (2012) 694–726.
[27] J.T. Leek, J.D. Storey, Capturing heterogeneity in gene expression studies by surrogate variable analysis, PLoS Genet. 3 (2007) e161,.
[28] X. Liang, T. Zou, B. Guo, S. Li, H. Zhang, S. Zhang, H. Huang, S.X. Chen, Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 471 (2015) 2015.0257.
[29] J. Lindström, A.A. Szpiro, P.D. Sampson, A.P. Oron, M. Richards, T.V. Larson, L. Sheppard, A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates, Environ. Ecol. Stat. 21 (2014) 411–433.
[30] G.G. Lorentz, Approximation of Functions, second ed., American Mathematical Society, 1986.
[31] Z. Lu, D.J. Steinskog, D. Tjøstheim, Q. Yao, Adaptively varying-coefficient spatiotemporal models, J. R. Stat. Soc. Ser. B 71 (2009) 859–880.
[32] F. Merlevède, M. Peligrad, E. Rio, A Bernstein type inequality and moderate deviations for weakly dependent sequences, Probab. Theory Related Fields 151 (2011) 435–474.
[33] G. Motta, H. Ombao, Evolutionary factor analysis of replicated time series, Biometrics 68 (2012) 825–836.
[34] R.F. Phillips, Iterated feasible generalized least-squares estimation of augmented dynamic panel data models, J. Bus. Econom. Statist. 28 (2010) 410–422.
[35] P.M. Robinson, Root-N-consistent semiparametric regression, Econometrica 56 (1988) 931–954.
[36] P.-M. Samson, Concentration of measure inequalities for Markov chains and Φ-mixing processes, Ann. Probab. 28 (2000) 416–461.
[37] K. Schmidheiny, U. Basel, Panel data: fixed and random effects, Short Guides To Microeconometrics 7 (2011) 2–7.
[38] J.H. Stock, M.W. Watson, Forecasting using principal components from a large number of predictors, J. Amer. Statist. Assoc. 97 (2002) 1167–1179.
[39] C.J. Stone, Additive regression and other nonparametric models, Ann. Statist. 10 (1985) 689–705.
[40] J.D. Storey, W. Xiao, J.T. Leek, R.G. Tompkins, R.W. Davis, Significance analysis of time course microarray experiments, Proc. Natl. Acad. Sci. USA 102 (2005) 12837–12842.
[41] Z. Tan, G. Qin, H. Zhou, Estimation of a partially linear additive model for data from an outcome-dependent sampling design with a continuous outcome, Biostatistics 17 (2016) 663–676.
[42] A.N. Tikhomirov, On the convergence rate in the central limit theorem for weakly dependent random variables, Theory Probab. Appl. 25 (1981) 790–809.
[43] W. Wang, J. Fan, Asymptotics of empirical eigenstructure for high dimensional spiked covariance, Ann. Statist. 45 (2017) 1342–1374.
[44] F. Wang, H. Wang, Modelling non-stationary multivariate time series of counts via common factors, J. R. Stat. Soc. Ser. B Stat. Methodol. 80 (2018) 769–791.
[45] J. Wang, Q. Zhao, T. Hastie, A.B. Owen, Confounder adjustment in multiple hypothesis testing, Ann. Statist. 45 (2017) 1863–1894.
[46] L. Wang, X.-H. Zhou, Assessing the adequacy of variance function in heteroscedastic regression models, Biometrics 63 (2007) 1218–1225.