DISSERTATION

NON-ASYMPTOTIC PROPERTIES OF SPECTRAL DECOMPOSITION OF LARGE GRAM-TYPE MATRICES WITH APPLICATIONS TO HIGH-DIMENSIONAL INFERENCE

Submitted by

Lyuou Zhang

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2020

Doctoral Committee:

Advisor: Wen Zhou
Co-Advisor: Haonan Wang

Jay Breidt
Mary Meyer
Liuqing Yang

ABSTRACT

NON-ASYMPTOTIC PROPERTIES OF SPECTRAL DECOMPOSITION OF LARGE GRAM-TYPE MATRICES WITH APPLICATIONS TO HIGH-DIMENSIONAL INFERENCE

Non-Asymptotic Properties of Spectral Decomposition of Large Gram-Type Matrices with Applications to High-Dimensional Inference

Jointly modeling a large and possibly divergent number of temporally evolving subjects arises ubiquitously in statistics, econometrics, finance, biology, and environmental sciences. To circumvent the challenges due to the high dimesionality as well as the temporal and/or contemporaneous dependence, the factor model and its variants have been widely employed. In general, they model the large scale temporally dependent data using some low dimensional structures that capture variations shared across dimensions. In this dissertation, we investigate the non-asymptotic properties of spectral decomposition of high-dimensional Gram-type matrices based on factor models. Specifically, we derive the exponential tail bound for the first and second moments of the deviation between the empirical and population eigenvectors to the right Gram matrix as well as the Berry-Esseen type bound to characterize the Gaussian approximation of these deviations. We also obtain the non-asymptotic tail bound of the ratio between eigenvalues of the left Gram matrix, namely the sample covariance matrix, and their population counterparts regardless of the size of the data matrix. The documented non-asymptotic properties are further demonstrated in a suite of applications, including the non-asymptotic characterization of the estimated number of latent factors in factor models and related machine learning problems, the estimation and forecasting of high-dimensional time series, the spectral properties of large sample covariance matrix such as perturbation bounds and inference on the spectral projectors, and low-rank matrix denoising from temporally dependent data.

Next, we consider the estimation and inference of a flexible subject-specific heteroskedasticity model for large scale panel data, which employs latent semiparametric factor structure to simultaneously account for the heteroskedasticity across subjects and contemporaneous and/or serial correlations. Specifically, the subject-specific heteroskedasticity is modeled by the product of unobserved factor process and subject-specific covariate effect. Serving as the loading, the covariate effect is further modeled via additive models. We propose a two-step procedure for estimation. Theoretical validity of this procedure is documented. By scrupulously examining the non-asymptotic rates for recovering the latent factor process and its loading, we show the consistency and asymptotic efficiency of our regression coefficient estimator in addition to the asymptotic normality. This leads to a more efficient confidence set for the regression coefficient. Using a comprehensive simulation study, we demonstrate the finite sample performance of our procedure, and numerical results corroborate the theoretical findings.

Finally, we consider the factor model-assisted variable clustering for temporally dependent data. The population level clusters are characterized by the latent factors of the model. We combine the approximate factor model with population level clusters to give an integrative group factor model as a background model for variable clustering. In this model, variables are loaded on latent factors and the factors are the same for variables from a common cluster and are different for variables from different groups. The commonality among clusters is modeled by common factors and the clustering structure is modeled by unique factors of each cluster. We quantify the difficulty of clustering data generated from integrative group factor model in terms of a permutation-invariant clustering error. We develop an algorithm to recover clustering assignments and study its minimax-optimality. The analysis of integrative group factor model and our proposed algorithm partitions a two-dimensional phase space into three regions showing the impact of parameters on the possibility of clustering in integrative group factor model and the statistical guarantee of our proposed algorithm. We also obtain the non-asymptotic characterization of the estimated number of latent factors. The model can be extended to the case of diverging number of clusters with similar results.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisors Prof. Wen Zhou and Prof. Haonan Wang for their continuous support of my Ph.D. study and related research, and their endless patience, encouragement and immense knowledge.

I would also like to thank my committee members Prof. F. Jay Breidt, Prof. Mary Meyer, and Prof. Liuqing Yang for their continuous interest in my research and generously offering their time, support and guidance throughout the preparation and review of this manuscript. Without the help and guidance of my committee members, I would never have been able to finish my dissertation.

I am also grateful to Department of Statistics of Colorado State University for the financial support during my Master and Ph.D. studies. I also appreciate all the faculty, staff and my fellow graduate students for their generous help.

Last but not the least, I would also like to thank my parents and brother who always support me, encourage me and give me valuable suggestions when I encounter any difficulties.

DEDICATION

I would like to dedicate this dissertation to my family.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction and Background

1.1 Overview

Factor model is a popular and widely used tool for dimension reduction, exploratory analysis, high dimensional inference, and modeling large scale data with sophisticated dependences. Classical static factor model

$$y_{it} = a_{i1}f_{t1} + \cdots + a_{iK}f_{tK} + u_{it} \tag{1.1.1}$$

with $t = 1, \ldots, T$ and $i = 1, \ldots, p$ has been widely studied (Anderson, 1962; Anderson and Rubin, 1956; Chamberlain and Rothschild, 1983; Lawley and Maxwell, 1962). Here, u_{it} is an idiosyncratic error process, $(f_{t1}, \ldots, f_{tK})^\top$ is a K-dimensional zero mean latent process, and $(a_{i1}, \ldots, a_{iK})^\top$ is referred to as the factor loadings. Let $\mathbf{A} = (a_{ik})_{i=1,k=1}^{p,K}$, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)^\top$ with $\mathbf{f}_t = (f_{t1}, \ldots, f_{tK})^\top$ or equivalently $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K)$ with $\boldsymbol{f}_k = (f_{1k}, \ldots, f_{Tk})^\top$, and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ with $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{pt})^\top$, (1.1.1) can be equivalently expressed as

$$\mathbf{Y} = \mathbf{A}\mathbf{F}^\top + \mathbf{U}. \tag{1.1.2}$$

Assume that $\mathbf{f}_t$ and $\boldsymbol{u}_t$ are uncorrelated and f_{tk} has unit variance (Chamberlain and Rothschild, 1983), (1.1.2) specifies the covariance structure of $(y_{1t}, \ldots, y_{pt})^\top$ as

$$\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top + \boldsymbol{\Sigma}_u, \tag{1.1.3}$$

where $\boldsymbol{\Sigma} = T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)$ and $\boldsymbol{\Sigma}_u = T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$ are the $p \times p$ population covariance matrix of $\boldsymbol{y}_t$ and $\boldsymbol{u}_t$, respectively. Estimations to $\mathbf{A}$ and the diagonal entries of $\boldsymbol{\Sigma}_u$ are well documented by Anderson and Rubin (1956), Anderson (1962), Chamberlain and Rothschild (1983) and Lawley

and Maxwell (1962) for low dimensional data with fixed p. The method of analysis is based on the asymptotic normality of some estimator to $\mathbf{\Sigma}$, such as the sample covariance matrix, and thus is not applicable when dimension p is large and divergent.

Large dimensional static factor model is first discussed by Stock and Watson (Stock and Watson, 1998) and Forni, Hallin, Lippi, and Reichlin (Forni et al., 2000). Compared to the traditional static factor model, the assumption on dimensions is relaxed and it allows both p and T diverge. Thus, large dimensional factor analysis is applicable to modeling the data of large scales. In addition, the idiosyncratic errors are allowed to be weakly correlated both serially and cross-sectionally, so $\mathbf{\Sigma}_u$ is not necessarily a diagonal matrix, which leads to the approximate factor model (Chamberlain and Rothschild, 1983). Combining "largeness" and "approximate" together, the new factor model is known as high dimensional approximate factor model (Bai, 2003; Bai and Ng, 2002; Stock and Watson, 2002a,b).

An important characteristic of the large dimensional approximate factor model is that the largest K population eigenvalues of $\mathbf{\Sigma}$ diverge with rate p, while the remaining eigenvalues of $\mathbf{\Sigma}$, as well as all eigenvalues of $\mathbf{\Sigma}_u$, are bounded, which gives a spike structure (Johnstone, 2001). As a result, under the pervasiveness assumption that the eigenvalues of $p^{-1}\mathbf{A}^\top\mathbf{A}$ are distinct and bounded away from zero and infinity, the eigenvalues of $\mathbf{A}\mathbf{A}^\top$ will diverge with rate p. This phenomenon arise since the information of the common component accumulates as we sum up the observations across subjects while u_{it} are unit-specific errors and summing the errors across subjects does not lead to the same accumulation of information. This makes the large dimensional approximate factor model different from the classical factor model with fixed dimensionality (Anderson, 1962; Anderson and Rubin, 1956; Lawley and Maxwell, 1962) and innately related to principal component analysis (PCA) (Anderson et al., 1963; Anderson and Rubin, 1956). PCA is widely used as a dimension reduction tool by finding a set of orthogonal linear transformations of the original variables such that the transformed variables maintain the information contained in the original variables as much as possible. That is, the principal component $\mathbf{Z}_1 = \boldsymbol{\omega}_1\mathbf{Y}$ is defined to maximize the variance of $\mathbf{Z}_1$ and the principal component $\mathbf{Z}_k = \boldsymbol{\omega}_k\mathbf{Y}$ is defined to maximize

the variance of $\mathbf{Z}_k$ given that $\mathbf{Z}_k$ is orthogonal to $\mathbf{Z}_1, \ldots, \mathbf{Z}_{k-1}$. Specially, $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_K$ are the eigenvectors corresponding to the largest K eigenvalues of $\boldsymbol{\Sigma}$. In practice, $\boldsymbol{\Sigma}$ is always unknown and estimated by $\widehat{\boldsymbol{\Sigma}} = T^{-1}\mathbf{Y}\mathbf{Y}^\top$. As mentioned before, the largest K population eigenvalues of $\boldsymbol{\Sigma}$, as well as the eigenvalues of $p^{-1}\mathbf{A}^\top\mathbf{A}$, diverge with rate p. Thus, by WeylâĂŹs theorem and Davis-Kahan theorem (Davis and Kahan, 1970), the difference between the eigen-decomposition of $\boldsymbol{\Sigma}$ corresponding to the largest K eigenvalues and that of $\mathbf{A}\mathbf{A}^\top$ converge to zero as p goes to infinity, which shows that large dimensional factor analysis and PCA are approximately the same. Similar as classical factor analysis, this method of analysis (Fan et al., 2013; Lam and Yao, 2012) gives consistent estimator of $\mathbf{A}$ and $\boldsymbol{\Sigma}$.

Large dimensional factor analysis is applied in many fields such as economics, psychology and other disciplines of the social sciences, and accurately estimating the latent factor and loadings are very important in statistical applications. For large dimension p and sample size T, a popular approach is to use PCA to simultaneously estimate the latent factor and loadings (Bai and Ng, 2013). Unlike traditional PCA, the authors suggested using eigen-decomposition of $\mathbf{Y}^\top\mathbf{Y}$. The estimated factor matrix, $\widehat{\mathbf{F}}$, is defined as $\sqrt{T}$ times the eigenvectors corresponding to the K largest eigenvalues of $\mathbf{Y}^\top\mathbf{Y}$ and the loading matrix is estimated by $\widehat{\mathbf{A}} = T^{-1}\mathbf{Y}^\top\widehat{\mathbf{F}}$. Further, Fan et al. (2016) suggested applying PCA to the data matrix projected onto a linear space spanned by relevant covariates to archive faster convergence rate. Asymptotic analysis of the PCA estimator is given by Bai and Ng (2013) and Fan et al. (2016). In this dissertation, we focus on the non-asymptotic analysis of the estimators and its applications.

To carefully study the spectral decomposition of large Gram matrices, we consider data generated from (1.1.1) or (1.1.2) so that not only the data are of high-dimensional but also allow temporarily dependence. For the right Gram matrix $\mathbf{Y}^\top\mathbf{Y}$, the eigenvectors corresponding to the K largest eigenvalues are of the same direction as $\boldsymbol{f}_k$, where $\boldsymbol{f}_k$ is the kth column of $\mathbf{F}$. Therefore, the spectral decomposition of the right Gram matrix can be investigated using the estimates to latent factor process and loading matrix in (1.1.1). That is, given an estimator to $\boldsymbol{f}_k$, denoted by $\widehat{\boldsymbol{f}}_k$, properties of the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{Y}^\top\mathbf{Y}$ can be studied

from $T^{-1/2}\widehat{\boldsymbol{f}}_k$, and vice versa. Although the consistency of estimating $\mathbf{F}$ has been documented in literature (Bai and Ng, 2013; Fan et al., 2016), non-asymptotic properties of the deviation of $\widehat{\mathbf{F}} = (\widehat{\boldsymbol{f}}_1, \ldots, \widehat{\boldsymbol{f}}_K)$, where $\widehat{\boldsymbol{f}}_k$ is the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{Y}^\top\mathbf{Y}$, from $\mathbf{F}$ have not been fully investigated. We study the non-asymptotic properties of $\widehat{\mathbf{F}} - \mathbf{F}$ as well as the approximated distribution of $\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k$ for each k. Particularly, we relax the condition on $\mathbf{F}$ in the traditional factor model. Compared with Condition PC1 in Bai and Ng (2013), we do not restrict $\mathbf{F}$ on a subspace. Therefore, as an important application in modeling high-dimensional time series, the non-asymptotic characterization of $\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k$ shows the accuracy of $\widehat{\boldsymbol{f}}_k$ as an surrogate to $\boldsymbol{f}_k$ for each k so that the parametric model of the K-dimensional latent processes, if specified in advance, can be easily estimated and therefore can be employed to forecast $\boldsymbol{y}_t$. Compared to the traditional likelihood based approach, this approach is computationally easier and requires very little assumptions on innovations of processes. In addition, we obtain non-asymptotic properties of the deviation between eigenvectors corresponding to the largest K eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$, *i.e.*, the sample covariance matrix, to those of $\boldsymbol{\Sigma}$ in (1.1.3). By considering $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ as a perturbation of $\boldsymbol{\Sigma}$, our result is similar to the Davis-Kahan Theorem (Cai et al., 2017; Davis and Kahan, 1970; Fan et al., 2018b; Yu et al., 2014) or the Wedin Theorem (Wedin, 1972). Our conclusion, however, does not depend on the consistent estimation of $\boldsymbol{\Sigma}$. Hence, for the high-dimensional cases, our result remains valid for the spike part of $\boldsymbol{\Sigma}$ even though it cannot be consistently estimated using $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ without regularization. Another important application of our results is to provide the non-asymptotic characterization of the tail probability of correctly estimating the number of latent factors K in the factor models, without which recovering the latent factor processes and their loadings will be meaningless in practice. For fixed or low dimensions, a variety of subjective methods such as scree plot of eigenvalues, distribution-based tests including Bartlett's test, and computational intensive methods including cross-validation have been employed to determine K (Jolliffe, 2002). For high dimensions with p/T converging to some constant, the information criteria such as AIC and BIC has be employed (Bai and Ng, 2002; Bai et al., 2018). If the data also follows a normal distribution, a sequential Kac-Rice test has been introduced to select K (Choi et al., 2017).

For ultra high dimensions with $p \gg T$, from the fact that the largest K eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ grow rapidly in p while others remain bounded or grow much slower, the consecutive-eigenvalue type estimator is widely used to determine K. For example, Lam and Yao (2012) and Ahn and Horenstein (2013) proposed estimators of K based on the ratios of consecutive eigenvalues. A similar approach is to use the difference of consecutive eigenvalues (Onatski, 2012). These early results focus on the consistency of the estimated number of factors when p and T diverge. To better understand how the dimension and sample size affect the probability of correctly estimating the number of latent factors using those consecutive-eigenvalue type estimators, we first refine results regarding eigenvalues of the sample covariance matrix (Bai and Yin, 1993; Johnstone, 2001). Then, we obtain non-asymptotic properties of the ratio of consecutive eigenvalues of the sample covariance matrix, which further provides the desired exponential tail bound of the probability of correctly estimating K for factor models or related machine learning problems.

As an application of large dimensional factor model, we consider a flexible data-driven model, in which the heteroskedasticity across subjects and serial dependence of ε_{it} is assumed to arise from a product of the subject-specific effect and some latent stationary process. Specifically, motivated by Connor and Linton (2007), Connor et al. (2012), and Fan et al. (2016), we model the subject-specific effect in the covariance structure by $\boldsymbol{g}(\boldsymbol{x}_i) = (g_1(\boldsymbol{x}_i), \ldots, g_K(\boldsymbol{x}_i))^\top$ with time invariant covariates $\boldsymbol{x}_i$ and nonparametric functions $g_1, \ldots, g_K$. In practice, $\boldsymbol{x}_i$ could be the genetic information in health study or market capitalization in finance applications. Then, we consider a K-dimensional zero-mean process $\boldsymbol{f}_t$ with finite variance, and introduce the subject-specific heteroskedasticity model with latent semiparametric factor structure as

$$y_{it} = \boldsymbol{z}_{it}^\top \boldsymbol{\beta} + \boldsymbol{g}(\boldsymbol{x}_i)^\top \boldsymbol{f}_t + u_{it}, \tag{1.1.4}$$

where the residual process u_{it} is independent of $\boldsymbol{f}_t$. Analogous to the traditional factor models, $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ serve as the loading and factor, respectively. Particularly, $\boldsymbol{g}(\boldsymbol{x}_i)$ models the desired heteroskedasticity across subjects and, together with $\boldsymbol{f}_t$, retains the cross-sectional dependence while $\boldsymbol{f}_t$ and u_{it} characterize the serial dependence. Like partially linear model or linear mixed

model, though ordinary least squares estimator of $\boldsymbol{\beta}$ is consistent, it is not efficient without taking the unknown dependence into account. That is, a careful estimation on the unobserved loading $\boldsymbol{g}(\boldsymbol{x}_i)$ and accurate recovery of the latent process $\boldsymbol{f}_t$ are in need to guarantee some sort of efficiency in both estimation and inference on $\boldsymbol{\beta}$. In the literature, there exist a variety of approaches to estimate $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$. For instance, Connor and Linton (2007) employed a kernel method to estimate $\boldsymbol{f}_t$ given $\boldsymbol{x}_i$ with finite values, and Connor et al. (2012) extended such estimate for general $\boldsymbol{x}_i$. Additionally, the consistency on estimating the loading and latent factor, along with an important result that such consistency requires no specific relationship between T and n (Fan et al., 2016), also shed lights upon estimating the large covariance structure under assumptions of factor structures (Fan et al., 2013). Motivated from these pioneer works, we propose a two-stage projection-based estimator for $\boldsymbol{\beta}$, $\boldsymbol{g}(\boldsymbol{x}_i)$, and $\boldsymbol{f}_t$ in model (1.1.4). Roughly speaking, adapting a projection-based principle component type estimator (Bai, 2003; Fan et al., 2016), we first estimate $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ from $y_{it} - \boldsymbol{z}_{it}^\top \widehat{\boldsymbol{\beta}}$ for some initial consistent estimator $\widehat{\boldsymbol{\beta}}$. Next, in the second stage, we update the estimate of $\boldsymbol{\beta}$ with a generalized least squares (GLS) type approach using estimation of $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ from the first-stage.

In addition, we consider model-based clustering, in which we define population level clusters relative to a model. We combine the approximate factor model with population level clusters to give an integrative group factor model as a background model for variable clustering. Although factor analysis is widely used to model high dimensional data with dependence, and group factor analysis is applied to model multiple covariance structures, the previous works do not give clustering recovery for dependent data. Our proposal consider a different case to G-block model (Bunea et al., 2020). In particular, the variables in the same group are allowed to have different variances and covariance swith variables in other groups. In addition, each variable can involve temporal dependence. We provide algorithm of recovering clustering assignments for high dimensional dependent data, along with its optimality. The recovery error rate is defined under a loss function free from label switching (Gao et al., 2018; Lu and Zhou, 2016). Compared to the existing literatures, our proof of minimax lower bound of recovery error rate involves a denser covering free

from the number of groups, so it can be extended to the case the number of groups diverges. Also, we apply Le Cam's method to give a tight bound for variable clustering with respect to covariance structures. The upper bound of recovery error rate is derived through scrupulously examining the non-asymptotic rates for estimating the latent factor process and its loading through PCA procedure (Bai and Ng, 2013; Fan et al., 2016). Lastly, we discover a phase transition in the phase space of signals of unique factors compared with those of common factors, which gives the region for possibility and guarantee of successful clustering. Also, our proposed model allows the number of groups not to be finite constant but a small term with respect to dimension. The technical tools we develop here are not limited to our setting alone, but are applicable to integrative group factor models.

1.2 Outline

In this dissertation, we focus on the high-dimensional inference, multivariate time series, and semiparametric modeling, as well as their applications motivated by the massive data related problems. In Chapter 2, we carefully study non-asymptotic properties of the spectral decomposition of large Gram-type matrices based on data that are not necessarily independent. We also obtain the non-asymptotic tail bound of the ratio between eigenvalues of the left Gram matrix and their population counterparts regardless of the size of the data matrix. The documented non-asymptotic properties are further demonstrated in a suite of applications. In Chapter 3, we consider estimation and inference of a flexible subject-specific heteroskedasticity model for large scale panel data. We propose a two-step procedure for estimation. By scrupulously examining the non-asymptotic rates for recovering the latent factor process and its loading, we show the consistency and asymptotic efficiency of our regression coefficient estimator in addition to the asymptotic normality. In Chapter 4, we combine the approximate factor model with population level clusters to give an integrative group factor model as a background model for variable clustering. We quantify the difficulty of clustering data generated from integrative group factor model in terms of a permutation-invariant clustering error., develop an algorithm to recover clustering assignments and study its minimax-

optimality. The analysis of integrative group factor model and our proposed algorithm partitions a two-dimensional phase space into three regions showing the impact of parameters on the possibility of clustering in integrative group factor model and the statistical guarantee of our proposed algorithm. A short summary and discussion of future work are listed in Chapter 5. The details of proofs and some additional results from simulation studies and real data analysis are given in Appendix 1, 2 and 3.

Chapter 2

Spectral Decomposition of Large Gram-Type Matrices

2.1 Introduction

Gram-type matrix or Gram matrix is fundamental in a wide range of fields including statistics (Shawe-Taylor et al., 2005), applied mathematics (Chen et al., 2011; James and Murphy, 1979; Schölkopf et al., 1999; Shawe-Taylor et al., 2002), machine learning (De Almeida et al., 2008a; Drineas and Mahoney, 2005; Ramona et al., 2012), engineering (De Almeida et al., 2008b), and physics (Stark, 2014). Given a $p \times T$ data matrix $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ with p-dimensional observation $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{pt})^\top$, the *left* and the *right* Gram matrices are $\mathbf{Y}\mathbf{Y}^\top$ and $\mathbf{Y}^\top\mathbf{Y}$, respectively (Horst, 1965; Rummel, 1988). Statistically, the left Gram matrix scaled by the sample size $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ coincides with the sample covariance matrix after ignoring the sample mean. As a bilinear function of the data matrix, Gram matrix retains many important information about data. For example, the right Gram matrix and the data matrix share the common null space while the column space of the left Gram matrix agrees with that of the data matrix. Particularly, the spectral decomposition of Gram matrices is a powerful and popular tool to provide a low-rank representation of the original data yet preserves the information as much as possible. For instance, in the linear model, spectral decomposition of the Gram matrix from the design matrix reveals the direction of space spanned by the projection matrix (Mandel, 1982); in the nonparametric regression, spectral decomposition of the Gram matrix from the spline basis functions provides a complete reconstruction of the functional space (Bialecki and Fairweather, 1995); and in the exploratory analysis, spectral decomposition of the Gram matrix from a general data or feature matrix leads to the principal component analysis (PCA) (Hotelling, 1933; Jolliffe, 2002; Pearson, 1901), kernel PCA, or sparse PCA (Zou et al., 2006; Zou and Xue, 2018). In addition, spectral decomposition of the Gram matrix has

been applied to estimate large covariance matrices (Fan et al., 2013, 2018a) and to extract the latent factors that drive the correlation structure in factor models (Bai, 2003; Bai and Ng, 2013; Bartholomew et al., 2011; Fan et al., 2016). By itself, the spectral decomposition has also been applied to other type of matrices to reveal the underlying structure in data, such as the spectral method along with the graph Laplacian or the adjacency matrix in cluster analysis or network study for the detection of clusters or latent communities (Donath and Hoffman, 1973; Ng et al., 2002).

Gram matrix naturally grows along with the size of data, and not only it may incur computational challenges but also lead to theoretical difficulties. For fixed dimensions, the scaled left Gram matrix or the sample covariance matrix converges to its expectation when T diverges (Bai and Yin, 1993; Bai et al., 1986, 1988). However, both the left and the right Gram matrices, as well as their empirical spectral distributions may fail to converge given simultaneously divergent p and T (Bickel and Levina, 2008a,b; Johnstone and Lu, 2009; Wang and Fan, 2017). Based on the asymptotic normality of sample covariance matrix, Anderson et al. (1963) established the joint distribution of empirical eigenvalues in the asymptotic regime where p remains constant and T diverges. For independent and identically distributed (*i.i.d.*) data with divergent dimensions, which scale with the sample size linearly and vice versa, the limiting distribution of spectral structures of the sample covariance matrix has also been widely studied (Adamczak et al., 2010; Bai and Silverstein, 2010; Bai and Yin, 1993; Bai et al., 1986, 1988; Jonsson, 1982; Wachter, 1978). When p/T diverges, a flexible and common approach is the spike structure model (Johnstone, 2001). That is, among the p eigenvalues of the population covariance matrix of $\boldsymbol{y}_t$, there are K dominant eigenvalues compared to the remains so that the signal of low-rank structure outweighs the noise and therefore can be retrieved from the spectral decomposition. Leveraging this spike structure, Wang and Fan (2017) showed that, for divergent p/T, the eigenvalue and corresponding eigenvector of the sample covariance matrix still converge to their population counterparts whenever the K dominant population eigenvalues diverge in p with certain rate. They also showed that the convergence rates of empirical eigenvalue and eigenvector are controlled by the divergent rate of the corresponding population eigenvalue.

The aforementioned assumption that the first K dominant eigenvalues of the population covariance matrix of $\boldsymbol{y}_t$ have order $O(p)$, together with the assumption that noises admit constant variance, is known as the *pervasiveness assumption* or *strong factor assumption* from the factor model and econometrics literature. Under this assumption, the spike structure can be equivalently written as a factor model (Bai, 2003; Chamberlain and Rothschild, 1983; Lam and Yao, 2012; Stock and Watson, 2002a) for which data satisfies

$$y_{it} = a_{i1}f_{t1} + \cdots + a_{iK}f_{tK} + u_{it} \tag{2.1.1}$$

with $t = 1, \ldots, T$ and $i = 1, \ldots, p$. Here, $(f_{t1}, \ldots, f_{tK})^\top$ is a K-dimensional zero mean latent process and u_{it} is an error process. Model (2.1.1) inherently links to a large number of widely used statistical models and methods, such as the panel data model with unobservable interactive effects (Ahn et al., 2001a; Bai, 2009a; Bai and Li, 2014; Moon and Weidner, 2017a) and PCA (Fan et al., 2018a). In matrix form, (2.1.1) is

$$\mathbf{Y} = \mathbf{A}\mathbf{F}^\top + \mathbf{U}, \tag{2.1.2}$$

where $\mathbf{A} = (a_{ik})_{i=1,k=1}^{p,K}$, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)^\top$ with $\mathbf{f}_t = (f_{t1}, \ldots, f_{tK})^\top$ or equivalently $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K)$ with $\boldsymbol{f}_k = (f_{1k}, \ldots, f_{Tk})^\top$, and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ with $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{pt})^\top$. Assume that $\mathbf{f}_t$ and $\boldsymbol{u}_t$ are uncorrelated and $\mathbb{E}(\mathbf{f}_t\mathbf{f}_t^\top) = \mathbf{I}_K$ for each $t = 1, \ldots, T$ (Chamberlain and Rothschild, 1983), the covariance of $\boldsymbol{y}_t$ is then given by

$$\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top + \boldsymbol{\Sigma}_u, \tag{2.1.3}$$

where $\boldsymbol{\Sigma} = T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)$ and $\boldsymbol{\Sigma}_u = T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$. Model (2.1.2) is called the strict factor model if $T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$ is diagonal, *i.e.*, $u_{1t}, \ldots, u_{pt}$ are uncorrelated with each other; otherwise, it is called the approximate factor model if $T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$ is not diagonal (Chamberlain and Rothschild, 1983). Model (2.1.2) provides an effective dimension reduction by approximating a p-dimensional process

$\boldsymbol{y}_t$ with a K-dimensional process $\mathbf{f}_t$ and a loading matrix matrix $\mathbf{A}$. From (2.1.3), it is easy to see that the largest K eigenvalues of $\boldsymbol{\Sigma}$ increase in p while the remaining eigenvalues are bounded (Bai and Ng, 2008), which mimics the spike structure model with divergent spiked eigenvalues.

For the traditional factor model with fixed p and *i.i.d.* normally distributed $\mathbf{f}_t$ and $\boldsymbol{u}_t$, the column space of loading matrix $\mathbf{A}$ and the diagonal entries of $T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$ can be consistently estimated through either the maximum likelihood estimator (MLE) (Lawley and Maxwell, 1962) or PCA (Anderson, 1962; Anderson and Rubin, 1956), both of which rely on the consistent estimation of $\boldsymbol{\Sigma}$. Though factor models and PCA are not identical in general, they are approximately the same for high-dimensional problems under the pervasiveness assumption (Fan et al., 2013, 2018a). Specially, the principal components $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ are defined as $\mathbf{Z}_k = \boldsymbol{w}_k^\top \mathbf{Y}$, where the projection directions $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K \in \mathbb{R}^p$ are the first K eigenvectors of $\boldsymbol{\Sigma}$. This eigen-decomposition formulation of PCA relates PCA to the singular value decomposition (SVD) of $\mathbf{Y}$ as well as the spectral decomposition of the sample covariance matrix, namely the left Gram matrix of $\mathbf{Y}$ scaled by sample size T.

In this paper, to carefully study the spectral decomposition of large Gram matrices, we consider data generated from (2.1.1) or (2.1.2) so that not only the data are of high-dimensional but also allow temporally dependence. For the right Gram matrix $\mathbf{Y}^\top\mathbf{Y}$, the eigenvectors corresponding to the K largest eigenvalues are of the same direction as $\boldsymbol{f}_k$, where $\boldsymbol{f}_k$ is the kth column of $\mathbf{F}$. Therefore, the spectral decomposition of the right Gram matrix can be investigated using the estimates to latent factor process and loading matrix in (2.1.1). That is, given an estimator to $\boldsymbol{f}_k$, denoted by $\widehat{\boldsymbol{f}}_k$, properties of the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{Y}^\top\mathbf{Y}$ can be studied from $T^{-1/2}\widehat{\boldsymbol{f}}_k$, and vice versa. Although the consistency of estimating $\mathbf{F}$ has been documented in literature (Bai and Ng, 2013; Fan et al., 2016), non-asymptotic properties of the deviation of $\widehat{\mathbf{F}} = (\widehat{\boldsymbol{f}}_1, \ldots, \widehat{\boldsymbol{f}}_K)$, where $\widehat{\boldsymbol{f}}_k$ is the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{Y}^\top\mathbf{Y}$, from $\mathbf{F}$ have not been fully investigated. Our *main contribution* in this paper is to study the non-asymptotic properties of $\widehat{\mathbf{F}} - \mathbf{F}$ as well as the approximated distribution of $\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k$ for each k. Particularly, we relax the condition on $\mathbf{F}$ in the traditional factor model. Compared

with Condition PC1 in Bai and Ng (2013), we do not restrict $\mathbf{F}$ on a subspace. Therefore, as an important application in modeling high-dimensional time series, the non-asymptotic characterization of $\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k$ shows the accuracy of $\widehat{\boldsymbol{f}}_k$ as an surrogate to $\boldsymbol{f}_k$ for each k so that the parametric model of the K-dimensional latent processes, if specified in advance, can be easily estimated and therefore can be employed to forecast $\boldsymbol{y}_t$. Compared to the traditional likelihood based approach, this approach is computationally easier and requires very little assumptions on innovations of processes. In addition, we obtain non-asymptotic properties of the deviation between eigenvectors corresponding to the largest K eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$, *i.e.*, the sample covariance matrix, to those of $\boldsymbol{\Sigma}$ in (2.1.3). By considering $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ as a perturbation of $\boldsymbol{\Sigma}$, our result is similar to the Davis-Kahan Theorem (Cai et al., 2017; Davis and Kahan, 1970; Fan et al., 2018b; Yu et al., 2014) or the Wedin Theorem (Wedin, 1972). Our conclusion, however, does not depend on the consistent estimation of $\boldsymbol{\Sigma}$. Hence, for the high-dimensional cases, our result remains valid for the spike part of $\boldsymbol{\Sigma}$ even though it cannot be consistently estimated using $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ without regularization.

Another important application of our results is to provide the non-asymptotic characterization of the tail probability of correctly estimating the number of latent factors K in the factor models, without which recovering the latent factor processes and their loadings will be meaningless in practice. For fixed or low dimensions, a variety of subjective methods such as scree plot of eigenvalues, distribution-based tests including Bartlett's test, and computational intensive methods including cross-validation have been employed to determine K (Jolliffe, 2002). For high dimensions with p/T converging to some constant, the information criteria such as AIC and BIC has be employed (Bai and Ng, 2002; Bai et al., 2018). If the data also follows a normal distribution, a sequential Kac-Rice test has been introduced to select K (Choi et al., 2017). For ultra high dimensions with $p \gg T$, from the fact that the largest K eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ grow rapidly in p while others remain bounded or grow much slower, the consecutive-eigenvalue type estimator is widely used to determine K. For example, Lam and Yao (2012) and Ahn and Horenstein (2013) proposed estimators of K based on the ratios of consecutive eigenvalues. A similar approach is to use the difference of consecutive eigenvalues (Onatski, 2012). These early results focus on the consistency

of the estimated number of factors when p and T diverge. To better understand how the dimension and sample size affect the probability of correctly estimating the number of latent factors using those consecutive-eigenvalue type estimators, we first refine results regarding eigenvalues of the sample covariance matrix (Bai and Yin, 1993; Johnstone, 2001). Then, we obtain non-asymptotic properties of the ratio of consecutive eigenvalues of the sample covariance matrix, which further provides the desired exponential tail bound of the probability of correctly estimating K for factor models or related machine learning problems.

The paper is organized as follows. In Section 2.2, we collect the notation and discuss the preliminary conditions to derive the main results. In Section 2.3, we carry out a non-asymptotic analysis of the spectral decomposition of large Gram matrices and document the main results. In Section 2.4, we discuss a variety of applications of our results to high-dimensional statistics. Section 2.5 presents numerical studies to demonstrate our results in the applications. We conclude the paper in Section 2.6 and relegate all the proofs and technical details to the supplementary file.

2.2 Notation and Preliminary Conditions

We collect notation in Section 2.2.1 that will be used throughout the paper and discuss in details the preliminary assumptions in Section 2.2.2 to establish the main results.

2.2.1 Notation

For p-dimensional vector $\boldsymbol{a} = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$, its ℓ_q-norm is defined by $||\boldsymbol{a}||_q = (\sum_{j=1}^p |a_j|^q)^{1/q}$ with $1 \leqslant q < \infty$. For matrix $\mathbf{M} = (m_{ij})_{1 \leqslant i,j \leqslant p} \in \mathbb{R}^{p \times p}$, $\|\mathbf{M}\|_{\max} = \max_{i,j} |m_{ij}|$ denotes the maximum norm and $||\mathbf{M}||_{\mathbb{F}} = (\sum_{i=1}^p \sum_{j=1}^p m_{ij}^2)^{1/2}$ is the Frobenius 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 $S = \{\boldsymbol{a} \in \mathbb{R}^p : ||\boldsymbol{a}||_2 = 1\}$. Denote the minimum and maximum eigenvalues of $\mathbf{M}$ by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$, respectively. Let $\text{tr}(\mathbf{M}) = \sum_{j=1}^p m_{jj}$ be the trace of $\mathbf{M}$. 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 $a_n \leqslant Cb_n$ for some positive C independent of n; and $a_n \asymp b_n$ if and only if there exist positive constants C and D independent of n such that $Cb_n \leqslant a_n \leqslant Db_n$. Unless specified otherwise, $s > 1$ and $C > 0$ denote generic constants independent of p, T.

2.2.2 Conditions

Suppose one observes data $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{it}, \ldots, y_{pt})$ from model (2.1.1) or (2.1.2) with $t = 1, \ldots, T$. We pose the following conditions throughout the paper.

Condition 2.2.1. *Almost surely, $\mathbf{A}^\top \mathbf{A}$ is a diagonal matrix with distinct entries; for each t, $f_{t1}, \ldots, f_{tK}$ are uncorrelated with each other and have zero mean and unit variance; for each t, $u_{1t}, \ldots, u_{pt}$ have zero mean and finite variances; and $\mathbf{f}_t$ and $\boldsymbol{u}_t$ are independent with each other.*

Condition 2.2.1 is similar to the assumption imposed on the approximate factor model (Chamberlain and Rothschild, 1983), which leads to the decomposition and identification of $\boldsymbol{\Sigma}$ in (2.1.3). The assumption on $\mathbf{A}$ can be viewed as Condition PC1 for the traditional factor models (Bai and Ng, 2013), which is also imposed for the MLE by Lawley and Maxwell (1962).

Condition 2.2.2. *There exist constants $d_1, d_2 > 0$ such that $d_1 \leqslant \lambda_{\min}(p^{-1}\mathbf{A}^\top \mathbf{A}) \leqslant \lambda_{\max}(p^{-1}\mathbf{A}^\top \mathbf{A}) \leqslant d_2$.*

Since the largest K eigenvalues of $\mathbf{A}^\top \mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ are the same, the spiked eigenvalues of $\boldsymbol{\Sigma}$ essentially diverge at rate p under Condition 2.2.2. When the entries of $\mathbf{A}$ remain constants as p diverges, this is always satisfied for a full rank $\mathbf{A}$ under Condition 2.2.1. In general, Condition 2.2.2 implies that, for each $k = 1, \ldots, K$, the mean squared loadings of the kth factor satisfies $p^{-1} \sum_{i=1}^{p} a_{ik}^2 = O(1)$, which can be easily satisfied with high probability if a_{ik} are *i.i.d.* copies from some non-degenerate distribution.

Condition 2.2.3. *Denote $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ the σ-algebra generated by $\{(\mathbf{f}_t, \boldsymbol{u}_t) : t \leqslant 0\}$ and $\{(\mathbf{f}_t, \boldsymbol{u}_t) : t \geqslant T\}$, respectively. Define the mixing coefficient $\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(A \cap B)|$.*

(i) Stationarity: $\{\boldsymbol{u}_t, \mathbf{f}_t\}_{t \leqslant T}$ are weakly stationary.

(ii) Strong mixing across t: There exist $r_1, C_1 > 0$ such that $\alpha(s) < \exp(-C_1 s^{r_1})$ for any $s > 0$.

(iii) Weak dependence in errors: There exist $C_2 > 0$ such that

$$\max_{j \leqslant p} \sum_{i=1}^{p} |\mathbb{E}(u_{it} u_{jt})| < C_2,$$
$$\frac{1}{pT} \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{t=1}^{T} \sum_{s=1}^{T} |\mathbb{E}(u_{it} u_{js})| < C_2,$$
$$\max_{i \leqslant p} \sum_{k=1}^{p} \sum_{m=1}^{p} \sum_{t=1}^{T} \sum_{s=1}^{T} |\operatorname{Cov}(u_{it} u_{kt}, u_{is} u_{ms})| < C_2.$$

(iv) Tail behavior: 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 $i = 1, \ldots, p$, $k = 1, \ldots, K$ and any $s > 0$, $\mathbb{P}(|u_{it}| > s) \leqslant \exp\{-(s/b_1)^{r_2}\}$ and $\mathbb{P}(|f_{tk}| > s) \leqslant \exp\{-(s/b_2)^{r_3}\}$.

Condition 2.2.3 is similar to the standard assumptions for the factor analysis of large scale panel data or high-dimensional time series (Bai, 2003; Fan et al., 2016; Stock and Watson, 2002a). Compared to similar conditions in the literature, we only require $\{\boldsymbol{u}_t, \mathbf{f}_t\}_{t \leqslant T}$ to be weakly stationary rather than strictly stationary in (i) by carefully exploiting Davydov's inequality (Athreya and Lahiri, 2006). In (iii), it suggests that though the common factors explain most dependence within $\boldsymbol{y}_t$, the errors also account for some weak cross-section dependence. It is easy to see $||\boldsymbol{\Sigma}_u||_2 = O(1)$ from (iii), and together with Condition 2.2.2 they are the well-known *pervasiveness assumption*.

It is interesting to notice that the well-known Condition PC1 from Bai and Ng (2013) restricts $\mathbf{F}$ to a subspace $\{\mathbf{F} \in \mathbb{R}^{T \times K} : T^{-1}\mathbf{F}^\top \mathbf{F} = \mathbf{I}_K\}$. However, for an arbitrary K-dimensional process under Condition 2.2.1, $T^{-1}\mathbf{F}^\top \mathbf{F}$ does not necessarily degenerate to its expected value $\mathbb{E}(T^{-1}\mathbf{F}^\top \mathbf{F}) = \operatorname{Var}(\mathbf{f}_1) = \mathbf{I}_K$. To satisfy this subspace restriction, one needs to rescale each realization of $\mathbf{F}$. Since the rescaling operator depends on the realization of $\mathbf{F}$, the rescaled processes no longer follow the original model of $\mathbf{f}_t$ if we assume any. This brings extra challenges to many applications. For example, in Section 2.4.2, this subspace restriction will prevent directly model-

ing $\mathbf{f}_t$ in (2.1.1) with some parametric models to forecast high-dimensional time series. In fact, we notice that this subspace restriction is stringent and can be replaced by the exponential tail bound on the difference between $T^{-1}\mathbf{F}^\top\mathbf{F}$ and its expectation $\mathbf{I}_K$. From the aforementioned well-known conditions, this bound can be easily established with the help of τ-mixing coefficient as defined below.

Definition 2.2.1 (τ-mixing coefficient (Merlevède et al., 2011))**.** *For any real random variable X and σ-algebra $\mathcal{M}$, denote $\mathbb{P}_X$ the distribution of X and $\mathbb{P}_{X|\mathcal{M}}$ the conditional distribution of X on $\mathcal{M}$. The τ-mixing coefficient is defined by*

$$\tau(\mathcal{M}, X) = \sup_{g\in\mathcal{L}_1(\mathbb{R})} \left| \int g(x)\mathbb{P}_{X|\mathcal{M}}(x) - \int g(x)\mathbb{P}_X(x) \right|,$$

where $\mathcal{L}_1(\mathbb{R})$ is the set of 1-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$.

Then, the τ-mixing coefficient of $\{f_{tk}\}$ for each $k = 1, \ldots, K$ is

$$\tau(T) = \sup_{j\geqslant 1} \frac{1}{j} \sup_{s>0, T+s\leqslant t_1<\cdots<t_j} \tau\left(\sigma(f_{tk}, t \leqslant s), (f_{t_1k}, \ldots, f_{t_jk})\right)$$

where $\sigma(f_{tk}, t \leqslant s)$ is the σ-algebra generated from $\{f_{tk}, t \leqslant s\}$. Note that, by Condition 2.2.3 (iv), for each $k = 1, \ldots, K$ and $t = 1, \ldots, T$,

$$Q(x) = \sup_{k,t} \inf\{s > 0 : \mathbb{P}(|f_{tk}^2| > s) \leqslant x\} = b_2^2\{\log(1/x)\}^{2/r_3}.$$

Thus, for $r_4 \in (0, 1)$ and any $x \geqslant 1$,

$$\begin{aligned}\tau(x) &\leqslant 2\int_0^{2\alpha(x)} Q(u)du \\ &\leqslant 4b_2^2 r_4 \left\{\frac{r_3(1-r_4)}{2}\right\}^{2/r_3} \exp\left\{\frac{2}{r_3(1-r_4)}\right\} \{2\alpha(x)\}^{r_4},\end{aligned}$$

which implies that $\mathbf{f}_t$ is τ-mixing by Condition 2.2.3 (ii). Then, following Theorem 1 in Merlevède et al. (2011), with probability at least $1 - T^{-1}$,

$$\|T^{-1}\mathbf{F}^\top\mathbf{F} - \mathbf{I}_k\|_{\mathbb{F}}^2 \lesssim \frac{\log T}{T},$$

which is the desired assumption in place of the subspace restriction on $\mathbf{F}$.

2.3 Main Results

Now we are in position to discuss the main results on non-asymptotic properties of the spectral decomposition of large Gram-type matrices based on (2.1.1) or (2.1.2). Continue to let $\mathbf{Y} = \mathbf{A}\mathbf{F}^\top + \mathbf{U}$, and we denote $T^{-1/2}\widehat{\boldsymbol{f}}_k$ the eigenvector corresponding to the kth largest eigenvalue of the right Gram matrix $\mathbf{Y}^\top\mathbf{Y}$ for $k = 1, \ldots, K$. Then, the loading matrix $\mathbf{A}$ can be estimated by $\widehat{\mathbf{A}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}$, where $\widehat{\mathbf{F}} = (\widehat{\boldsymbol{f}}_1, \ldots, \widehat{\boldsymbol{f}}_K)$. First, we have the following exponential tail bounds on the deviations $\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2$ and $\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\max}$.

Theorem 2.3.1 (Exponential tail bounds on the deviation between $\widehat{\mathbf{F}}$ and $\mathbf{F}$)**.** *Under Conditions 2.2.1-2.2.3, the deviation between* $\widehat{\mathbf{F}}$ *and* $\mathbf{F}$ *satisfies*

(i) with probability at least $1 - e^{-s}$*,*

$$T^{-1}\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2 \lesssim \left(\frac{1}{p} + \frac{1}{T}\right)s^4;$$

(ii) $T^{-1}\mathbb{E}(\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2) \lesssim p^{-1} + T^{-1}$ *and* $T^{-2}\operatorname{Var}(\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2) \lesssim p^{-2} + T^{-2}$*; and*

(iii) with probability at least $1 - e^{-s}$*,*

$$\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\max} \lesssim \left(\frac{1}{\sqrt{p}} + \frac{1}{T}\right)(\log T)^{2/r_3}s.$$

For the approximate factor model, it has been shown that the mean squared error (MSE) $T^{-1}\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2$ converges to zero when p and T diverge, thus $\widehat{\mathbf{F}}$ converges to $\mathbf{F}$ in probability

(Bai and Ng, 2013; Fan et al., 2016). In Theorem 2.3.1, not only have we provided the non-asymptotic characterization on the MSE of $\widehat{\mathbf{F}}$ in the sense that the result holds for finite T and p, but also the convergence of $\widehat{\mathbf{F}}$ to $\mathbf{F}$ is established under a weaker condition on $\mathbf{F}$ compared to Condition PC1 in Bai and Ng (2013) as discussed in Section 2.2.2. Theorem 2.3.1 reveals that the deviation between $\widehat{\mathbf{F}}$ and $\mathbf{F}$ is due to 1) the deviation between $\mathbf{F}$ and its projection onto subspace $\{\mathbf{F} \in \mathbb{R}^{T\times K} : T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K\}$, which is of rate $p^{-1} + T^{-2}$; and 2) the error for estimating this projection, which is of rate $p^{-2} + T^{-1}$. They lead to the non-asymptotic bound on $T^{-1}\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2$ in (i). In addition, $(p+T)^{-1}p\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2$ enjoys a sub-exponential tail with the finite first and second moments from (ii).

Recall that both $\widehat{\mathbf{F}}$ and $\mathbf{F}$ have finite K columns. A by-product of Theorem 2.3.1 is an exponential tail bound on the deviation between the $T^{-1/2}$-scaled kth columns of $\widehat{\mathbf{F}}$, *i.e.*, the kth eigenvector of the right Gram matrix, and its counterpart in $\mathbf{F}$. That is, with probability at least $1 - e^{-s}$, for each $k = 1, \ldots, K$,

$$T^{-1}\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2 \lesssim \left(\frac{1}{p} + \frac{1}{T}\right) s^4.$$

Therefore, $(p+T)^{-1}p\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2$ also admits a sub-exponential tail with the finite first and second moments, which are similar to (ii) in Theorem 2.3.1.

Using the max norm, the error rate remains the same for recovering the projection since it is of finite dimension. On the other hand, the ℓ_∞-deviation between $\mathbf{F}$ and its projection is of rate $(p^{-1/2} + T^{-1})(\log T)^{2/r_3}$, where $\log T$ is due to the maximum inequality to control the maximum among TK entries in $\mathbf{F}$. Result in (iii) provides a non-asymptotic entry-wise bound on the deviation between $\widehat{\mathbf{F}}$ and $\mathbf{F}$. For each $t = 1, \ldots, T$ and $k = 1, \ldots, K$, $|\widehat{f}_{tk} - f_{tk}|(p^{-1/2} + T^{-1})^{-1}\{\log(T)\}^{-2/r_3}$ displays a sub-exponential tail. Thus, following the similar argument in (ii), $(p^{-1/2} + T^{-1})^{-1}\{\log(T)\}^{-2/r_3}|\widehat{f}_{tk} - f_{tk}|$ also has the finite first and second moments for all p and T. By Condition 2.2.3, f_{tk} has the finite first and second moments and so does $\widehat{f}_{tk}$ whenever $(p^{-1/2} + T^{-1})\{\log(T)\}^{2/r_3} = O(1)$ due to the triangle inequality. This nontrivial result makes it

possible to further model the K-dimensional latent process parametrically; see Section 2.4.2 for more details.

Next, to establish the Berry-Esseen type bound for each of the K eigenvectors of the right Gram matrix, we consider the following additional condition.

Condition 2.3.1. *For each t, $u_{1t}, \ldots, u_{pt}$ are independent with each other.*

Condition 2.3.1 is standard for traditional PCA (Jolliffe, 2002) and factor models (Anderson, 1962; Anderson and Rubin, 1956; Lawley and Maxwell, 1962). This stronger condition on $\boldsymbol{u}_t$, compared to (iii) in Condition 2.2.3, enables us to leverage results from random matrix theory to establish the following Berry-Esseen type bound. Theorem 2.3.2 provides the approximation error rate to the distribution of the standardized deviation between $\widehat{\boldsymbol{f}}_k$ and $\boldsymbol{f}_k$ by the standard normal distribution for each k.

Theorem 2.3.2 (Berry-Esseen Type Bound for $||\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k||_2^2$)**.** *Under Conditions 2.2.1, 2.2.2, (i),(ii) and (iv) in 2.2.3, and 2.3.1, for each $k = 1, \ldots, K$, we have*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left\{ \frac{\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2 - \mathbb{E}(\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2)}{\mathrm{Var}^{1/2}(\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2)} \leqslant x \right\} - \Phi(x) \right| \lesssim \frac{\log(T)}{\sqrt{T}} + \frac{1}{\sqrt{p}},$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

From Theorem 2.3.2, with probability at least $1 - e^{-s}$,

$$\left| T^{-1}\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2 - T^{-1}\mathbb{E}(\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2) \right| \lesssim \left(\frac{1}{p} + \frac{1}{T} \right) \sqrt{s}.$$

Compared to Theorem 2.3.1, the above improved sub-Gaussian tail on $\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2$ benefits from the assumption of independent errors in Condition 2.3.1. Theorem 2.3.2 sheds lights on drawing inference on the leading eigenvectors of the covariance matrix for non *i.i.d.* data, which is detailed in Section 2.4.3. For *i.i.d.* data, the traditional rate in the Berry-Esseen bound for Gaussian approximation is $T^{-1/2}$ (Callaert et al., 1978; Chan and Wierman, 1977). In Theorem 2.3.2, $p^{-1/2}$ and $T^{-1/2}$ are due to the uncertainty from $\mathbf{f}_t$ and $\boldsymbol{u}_t$ for computing $\widehat{\boldsymbol{f}}_k$. In addition, the dependence

in $\mathbf{f}_t$ leads to the extra $\log T$ in the bound, which has been observed in literature (Hörmann, 2009; Jirak, 2016).

In the rest of this section, we will study non-asymptotic properties of the eigenvalues of $\mathbf{Y}^\top\mathbf{Y}$. Although the spectral structure of the expected right Gram matrix $\mathbb{E}(\mathbf{Y}^\top\mathbf{Y})$ differs from that of the expected left Gram matrix $\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)$, it is interesting to notice that $\mathbf{Y}^\top\mathbf{Y}$ and $\mathbf{Y}\mathbf{Y}^\top$ share the common non-zero eigenvalues. Hence, we first consider $\mathbf{Y}\mathbf{Y}^\top$, which is conveniently the sample covariance matrix scaled by T. Denote $\{\lambda_i\}_{i=1}^p$ and $\{\boldsymbol{w}_i\}_{i=1}^p$ the eigenvalues (in decreasing order) and corresponding eigenvectors of $\boldsymbol{\Sigma} = T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)$, and let $\{\widehat{\lambda}_i\}_{i=1}^p$ and $\{\widehat{\boldsymbol{w}}_i\}_{i=1}^p$ be the eigenvalues (in decreasing order) and corresponding eigenvectors of $\widehat{\boldsymbol{\Sigma}} = T^{-1}\mathbf{Y}\mathbf{Y}^\top$. We establish the non-asymptotic characterization of $\widehat{\lambda}_i$ relative to λ_i as follows.

Theorem 2.3.3 (Non-asymptotic characterization of $\widehat{\lambda}_i$'s relative to λ_i's)**.** *Under Conditions 2.2.1, 2.2.2, (i),(ii) and (iv) in 2.2.3, and 2.3.1, there exist positive constants C and c that only depend on $\boldsymbol{u}_t$ such that the following results hold.*

(i) If $p < T$, with probability at least $1 - e^{-s}$,

$$
\begin{aligned}
|\widehat{\lambda}_i/\lambda_i - 1| &\leqslant \frac{C}{\sqrt{T}} + \frac{c}{\sqrt{pT}}\sqrt{s}, \qquad i = 1, \ldots, K, \\
|\widehat{\lambda}_i/\lambda_i - 1| &\leqslant \frac{C\sqrt{p}}{\sqrt{T}} + \frac{c}{\sqrt{T}}\sqrt{s}, \qquad i = K+1, \ldots, p.
\end{aligned}
$$

(ii) If $p \geqslant T$, with probability at least $1 - e^{-s}$,

$$
\begin{aligned}
|\widehat{\lambda}_i/\lambda_i - 1| &\leqslant \frac{C}{\sqrt{T}} + \frac{c}{\sqrt{pT}}\sqrt{s}, \qquad i = 1, \ldots, K, \\
\widehat{\lambda}_i/\lambda_i &\geqslant \sqrt{\frac{p}{T}} - C - \frac{c}{\sqrt{T}}\sqrt{s}, \qquad i = K+1, \ldots, T, \\
\widehat{\lambda}_i/\lambda_i &\leqslant \sqrt{\frac{p}{T}} + C + \frac{c}{\sqrt{T}}\sqrt{s}, \qquad i = K+1, \ldots, T.
\end{aligned}
$$

Taking $s = \log T$, Theorem 2.3.3 implies that the first K eigenvalues of the scaled left Gram matrices, *i.e.*, the sample covariance matrix, $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K$ converge to the corresponding eigenvalues

of $\mathbf{\Sigma}$ in probability. When $p < T$, the relative errors of the remaining $p - K$ eigenvalues to their population counterparts are bounded by $T^{-1/2}p^{1/2}$ in probability. By Condition 2.2.1, λ_i is bounded for $i > K$. Thus, the bound on relative error $|\widehat{\lambda}_i/\lambda_i - 1|$ is the same as that of deviation $|\widehat{\lambda}_i - \lambda_i|$ for $i > K$. That is, eigenvalues of $\widehat{\mathbf{\Sigma}}$ converge to those of $\mathbf{\Sigma}$ only if $p/T \to 0$. This agrees with the well known convergence of $\widehat{\mathbf{\Sigma}}$ to $\mathbf{\Sigma}$ in low dimension for *i.i.d.* data (Bien et al., 2016; Bunea and Xiao, 2015).

Different lessons are learned when $p > T$. As $\widehat{\mathbf{\Sigma}}$ is not of full-rank, $\widehat{\lambda}_i$'s with $i > K$ consist of at most $T - K$ non-zeros and at least $p - T$ zeros. For a legitimate covariance $\mathbf{\Sigma}$, at least $p - T$ eigenvalues of $\widehat{\mathbf{\Sigma}}$ are biased for estimating their population counterparts. In addition, the non-zero eigenvalues of $\widehat{\mathbf{\Sigma}}$ could also be biased. For *i.i.d.* data with unit variance and p proportional to T, it is known that non-zero eigenvalues of the sample covariance matrix are spread out and bounded by $(1 - p^{1/2}T^{-1/2})^2$ and $(1 + p^{1/2}T^{-1/2})^2$ (Bai and Yin, 1993; James and Stein, 1992; Johnstone and Paul, 2018; Stein, 1956), which explains the bias in non-zero eigenvalues of the sample covariance matrix compared to their population counterparts (Bai and Yin, 1993; Baik et al., 2005; Johnstone and Paul, 2018). In contrast, the low-rank structure in factor models provides better understanding on eigenvalues of $\widehat{\mathbf{\Sigma}}$. Consider a factor model with $\boldsymbol{u}_t$ assumed to be white noise, Lam and Yao (2012) focused on the cross covariance matrix $\mathbf{M} = \sum_{h=1}^{h_0} \mathbf{\Sigma}(h)\mathbf{\Sigma}(h)^\top$, where $\mathbf{\Sigma}(h)$ is the autocovariance matrix of $\boldsymbol{y}_t$ at lag h. They remarked that asymptotically, spiked eigenvalues of the sample cross covariance matrix converge to the corresponding population eigenvalues, while the non-spiked eigenvalues, although may not converge, are bounded by the ratio of p and T. Theorem 2.3.3 (ii) provides a non-asymptotic characterization of their remarks. First, we confirm that, as expected, $\widehat{\lambda}_i$ fails to converge for $i > K$ if p/T diverges. Also, the non-asymptotic bound in Theorem 2.3.3 shows that the ratio between $\mathbb{E}(\widehat{\lambda}_i)$ and λ_i is bounded above by $2\sqrt{\pi cp/T}\Phi(2^{-1/2}c^{-1}C\sqrt{p})$ for any given p and T. Furthermore, the non-asymptotic bound of $\widehat{\lambda}_i/\lambda_i$ provide a characterization on the closeness between $\mathbb{E}(\widehat{\lambda}_i/\lambda_i)$ and 1 for $i = 1, \ldots, K$. It is easy to see from Theorem 2.3.3 that the deviation between $\mathbb{E}(\widehat{\lambda}_i/\lambda_i)$ and 1 is bounded above by $2\sqrt{\pi c/(pT)}\Phi(2^{-1/2}c^{-1}C\sqrt{p})$. This reflects the asymptotic unbiasedness of $\widehat{\lambda}_i$ for $i = 1, \ldots, K$.

Wang and Fan (2017) considered a noiseless factor model with arbitrary factor strengths, which allows the spiked eigenvalue to be with any rate in p. Compared to their model, (2.1.2) can be viewed as a special case where the spiked eigenvalues are all in the same rate of p if $\boldsymbol{u}_t$ is further modeled by $\mathbf{C}\boldsymbol{f}_t$ with some $\mathbf{C}$ orthogonal to $\mathbf{A}$. The authors showed that eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ are asymptotically unbiased if $pT^{-1}\lambda_i^{-1}$ converge to zero for $i = 1, \ldots, K$. Recall that $\lambda_i = O(p)$ under Condition 2.2.2, so that $pT^{-1}\lambda_i^{-1}$ always converges to zero for (2.1.2). Thus, Theorem 2.3.3 gives a similar result on the asymptotic unbiasedness of $\widehat{\lambda}_i$ as Wang and Fan (2017). Also, the authors established the asymptotic normality of $\widehat{\lambda}_i/\lambda_i - 1$ upon removing the bias. Complement to that, Theorem 2.3.3 (ii) provides a finite sample view on $\widehat{\lambda}_i/\lambda_i$ by showing its non-asymptotic sub-Gaussian tail for $i = 1, \ldots, K$.

2.4 Applications in High-Dimensional Statistics

To demonstrate results in Section 2.3, we consider a number of interesting and widely studied applications in high-dimensional statistics, including the estimation of the number of latent factors in factor models and related machine learning problems, the estimation and forecasting of high-dimensional time series, the spectral properties of large sample covariance matrix such as perturbation bounds and inference on the spectral projectors, and the low-rank matrix denoising from dependent data.

2.4.1 Estimation of the Number of Latent Factors

In high-dimensional factor models or machine learning problems such as PCA, it is necessary to choose the number of latent factors or principal components K before recovering the loading matrix and factors or computing the principal components and scores. Traditional methods to estimate K include, for example, the likelihood ratio test and the screen plot (Jolliffe, 2002). For the high-dimensional data with large covariance matrix, eigenvalues of the sample covariance matrix or their variants have been utilized and the estimation is consistent under certain separation condition of the first K eigenvalues from the remains. A popular approach is based on the ratio of

consecutive eigenvalues (Ahn and Horenstein, 2013; Fan et al., 2016; Lam and Yao, 2012),

$$\widehat{K} = \text{argmax}_{1 \leqslant i < \min(p,T)} \frac{\widehat{\lambda}_i}{\widehat{\lambda}_{i+1}} \tag{2.4.1}$$

where $\widehat{\lambda}_i$ is the ith eigenvalue of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$; while, other methods are based on the eigenvalue differences (Onatski, 2012) or the cumulative magnitude of eigenvalues (Bai and Ng, 2002).

Under the pervasiveness assumption, *i.e.* Condition 2.2.2 and (iii) in Condition 2.2.3, the consistency of $\widehat{K}$ has been established (Fan et al., 2016; Lam and Yao, 2012). However, the rate of the probability of consistent estimation has not been fully explored. Theorem 2.3.3 sheds light on characterizing this rate. In fact, from Theorem 2.3.3, $\widehat{\lambda}_K/\widehat{\lambda}_{K+1}$ is of the order $O_p(p)$ when $p < T$ and $O_p(T)$ when $p > T$. In contrast, $\widehat{\lambda}_i/\widehat{\lambda}_{i+1}$ is $O_p(1)$ for $i \neq K$. As an application, Theorem 2.4.1 establishes the non-asymptotic lower bound of the probability of estimating the correct number of factors.

Theorem 2.4.1. *Under Conditions 2.2.1, 2.2.2, (i),(ii) and (iv) in 2.2.3, and 2.3.1, given* $\mathbf{Y}$ *from* (2.1.1) *or* (2.1.2)*,* $\widehat{K}$ *defined in* (2.4.1) *satisfies*

$$\mathbb{P}(\widehat{K} = K) \geqslant 1 - 2\exp(-\{C_1\sqrt{\max(p,T)} - C_2\sqrt{\min(p,T)}\}^2), \tag{2.4.2}$$

where

$$C_1 = \frac{1}{c}\left[1 - \left\{\frac{\max(p,T)\lambda_{K+1}}{T\lambda_K} \max_{1 \leqslant i < \min(p,T), i \neq K} \frac{\lambda_i}{\lambda_{i+1}}\right\}^{1/4}\right],$$

and $C_2 = c^{-1}C$*, with* C *and* c *defined in Theorem 2.3.3.*

As mentioned in Theorem 2.3.3, C and c are positive constants that only depend on $\boldsymbol{u}_t$ so that $C_2 > 0$ is independent of p and T. Under Conditions 2.2.1 and 2.2.2, $\lambda_i = O(p)$ for $i = 1, \ldots, K$ and $\lambda_i = O(1)$ for $i > K$ so that $C_1 > 0$ for sufficiently large p and T. On the right hand side of (2.4.2), $C_2\sqrt{\min(p,T)}$ is smaller than $C_1\sqrt{\max(p,T)}$ whenever $p \ll T$ or $p \gg T$, so the lower bound is governed by $C_1\sqrt{\max(p,T)}$. It is easy to see that C_1 is large if both λ_{K+1}/λ_K and

$\max_{1\leqslant i<\min(p,T),i\neq K}\lambda_i/\lambda_{i+1}$ are small. That is, it is easy to estimate K if the spiked eigenvalues $\lambda_1,\ldots,\lambda_K$ are close to each other and so do the non-spiked eigenvalues $\lambda_{K+1},\ldots,\lambda_p$. Otherwise, if λ_i/λ_{i+1} is large for some $i\neq K$, C_1 will be small so that the lower bound on the right hand side of (2.4.2) will be away from 1 and implies a more challenging K to be estimated.

When p and T are close, $C_2\sqrt{\min(p,T)}$ is not negligible. Notice that the lower bound in (2.4.2) can be written as $1-2\exp\{-C_1^2(1-C'p^{1/2}T^{-1/2})^2T\}$ for some positive constant C' given $p<T$. When $C_1^2(1-C'p^{1/2}T^{-1/2})^2$ is small, a large T is preferable to drive the lower bound close to 1. If $p\geqslant T$, the lower bound in (2.4.2) can be written as $1-2\exp\{-C_1^2(1-C'T^{1/2}p^{-1/2})^2p\}$ and similarly, a large p is preferable to make the lower bound approaching 1.

An alternative to $\widehat{K}$, proposed by Onatski (2012), is to use the difference of consecutive eigenvalues. That is, for given $\delta>0$ and pre-determined L, one defines

$$\widehat{K}_d=\max\{i\leqslant L:\widehat{\lambda}_i-\widehat{\lambda}_{i+1}\geqslant\delta\},\tag{2.4.3}$$

Similar to Theorem 2.4.1, we have the following result.

Theorem 2.4.2. *Under Conditions 2.2.1, 2.2.2, (i),(ii) and (iv) in 2.2.3, and 2.3.1, given* $\mathbf{Y}$ *from* (2.1.1) *or* (2.1.2)*,* $\widehat{K}_d$ *in* (2.4.3) *satisfies*

$$\mathbb{P}(\widehat{K}_d=K)\geqslant 1-2\sum_{i=1}^{K+1}\exp(-\{C_{1i}\sqrt{T}-C_2\sqrt{p}\}^2),$$

where $C_{1i}=(2c)^{-1}(\lambda_i-\lambda_{i+1}-\delta)$ *for* $i=1,\ldots,K$*,* $C_{1,K+1}=c^{-1}(\delta-\lambda_{K+2}+\lambda_{K+1})$*, and* $C_2=c^{-1}C$*, with* C *and* c *defined in Theorem 2.3.3.*

Under the pervasiveness assumption, Onatski (2012) established the consistency of $\widehat{K}_d$ when p is proportional to T. Theorem 2.4.2 relaxes the restriction on p and T and provides the non-asymptotic characterization of the probability of consistent estimation of K by $\widehat{K}_d$. It suggests that, for carefully selected δ such that $\delta>\lambda_{K+2}-\lambda_{K+1}$, $\widehat{K}_d$ and $\widehat{K}$ have similar rates of the probability of consistent estimation. However, $\widehat{K}_d$ is not tuning free compared to $\widehat{K}$. Onatski (2012)

proposed a data-driven procedure to determine δ. Specifically, an iterative procedure was employed to alternatively update δ and $\widehat{K}_d$ until convergence. Note that $\lambda_{K+2} = \lambda_{K+1}$ if $u_{1t}, \ldots, u_{pt}$ are identical. In this case, an appropriate δ can be easily found. Otherwise, more numerical iterations are required. Sometimes, $\widehat{K}_d$ may perform better than $\widehat{K}$ in practice, which can be explained using the non-asymptotic results from Theorems 2.4.1 and 2.4.2. Consider a special case where $p > T$, $K = 1$, $\lambda_1 = p$, and $\lambda_2 = \cdots = \lambda_p = 1$. The lower bound for $\widehat{K}$ in Theorem 2.4.1 is $1 - 2\exp(-\{c^{-1}(1 - T^{-1/4})\sqrt{p} - c^{-1}C\sqrt{T}\}^2)$ while the lower bound for $\widehat{K}_d$ in Theorem 2.4.2 is $1 - 2\exp(-\{(2c)^{-1}(p - 1 - \delta)\sqrt{T} - c^{-1}C\sqrt{p}\}^2) - 2\exp(-\{(2c)^{-1}\delta\sqrt{T} - c^{-1}C\sqrt{p}\}^2)$. For a divergent p and constant T, $\widehat{K}_d$ outperforms $\widehat{K}$ in terms of a higher rate of the probability of consistent estimation whenever $C > 1 - T^{-1/4}$.

Different from the consecutive eigenvalue based approaches, the information criterion has also been used to estimate K. Some of them can be interpreted as a penalized cumulative magnitude of eigenvalues, such as

$$\mathbb{PC}(k) = \left\{ \frac{1}{pT} \sum_{i \geqslant k} \widehat{\lambda}_i + k\widehat{\sigma}^2 \frac{p+T}{pT} \log\left(\frac{pT}{p+T}\right) \right\}.$$

where $\widehat{\sigma}^2$ is some consistent estimate of $(pT)^{-1} \sum_{i=1,t=1}^{p,T} \mathbb{E}(u_{it}^2)$ (Bai and Ng, 2002). Then, K is estimated by $\widehat{K}_m = \operatorname{argmin}_{k \leqslant L} \mathbb{PC}(k)$ for some pre-determined L. Bai and Ng (2002) further suggested that $\widehat{\sigma}^2$ can be replaced by $(pT)^{-1} \sum_{i>L} \widehat{\lambda}_i$ in practice and the penalty term $(pT)^{-1}(p + T)\log((p+T)^{-1}pT)$ can be replaced by $(pT)^{-1}(p+T)\log(\min(p,T))$ or $\min(p,T)^{-1}\log(\min(p,T))$. They also showed the consistency of $\widehat{K}_m$ when $\widehat{\sigma}^2$ is consistent and the penalty shrinks to zero as p and T diverge. Notice that $\widehat{K}_m$ is entirely based on the empirical distribution of $\widehat{\lambda}_i$ for $i = 1, \ldots, p$. Thus, its non-asymptotic properties such as the rate of the probability of consistent estimation may also be established using Theorem 2.3.3, which we leave to the future work.

2.4.2 Estimation and Forecasting of High-Dimensional Time Series

Making forecast based on high-dimensional time series arises frequently in econometrics, financial analysis, and meteorology. Suppose we observe $\mathbf{Y} \in \mathbb{R}^{p \times T}$, where each entry y_{it} follows

(2.1.1) and the zero mean K-dimensional latent process $\mathbf{f}_t$ is governed by parametric models satisfying Conditions 2.2.1 and 2.2.3. For example, Chen et al. (2018) considered a model similar to (2.1.1) with $\mathbf{f}_t$ following an autoregressive model whose parameters are estimated for predicting $y_{i,s}$ with $s > T$.

As an application of Theorem 2.3.1, we show the consistency on estimating the moments of $\mathbf{f}_t$ using the spectral decomposition of $\mathbf{Y}^\top\mathbf{Y}$, which guarantees the consistency of moment-based estimators to parameters of a large realm of parametric models for $\mathbf{f}_t$. Denote the sample autocovariance function (Brockwell et al., 1991) of $\mathbf{f}_t$ by

$$\widehat{\mathbf{\Gamma}}(h, \mathbf{f}_t) = \frac{1}{T}\sum_{t=1}^{T-|h|}(\mathbf{f}_{t+|h|} - \bar{\mathbf{f}})(\mathbf{f}_t - \bar{\mathbf{f}})^\top,$$

where $\bar{\mathbf{f}} = T^{-1}\sum_{t=1}^{T}\mathbf{f}_t$. Also, let the sample autocovariance function of $\widehat{\mathbf{f}}_t$, the tth row of $\widehat{\mathbf{F}}$, be

$$\widehat{\mathbf{\Gamma}}(h, \widehat{\mathbf{f}}_t) = \frac{1}{T}\sum_{t=1}^{T-|h|}(\widehat{\mathbf{f}}_{t+|h|} - \bar{\widehat{\mathbf{f}}})(\widehat{\mathbf{f}}_t - \bar{\widehat{\mathbf{f}}})^\top,$$

where $\bar{\widehat{\mathbf{f}}} = T^{-1}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t$. In Theorem 2.4.3, we show that the sample autovariance function of $\mathbf{f}_t$ can be consistently recovered by that of $\widehat{\mathbf{f}}_t$.

Theorem 2.4.3. *Under Conditions 2.2.1-2.2.3, given* $\mathbf{Y}$ *from* (2.1.1) *or* (2.1.2)*,* $\widehat{\mathbf{\Gamma}}(h, \mathbf{f}_t)$ *and* $\widehat{\mathbf{\Gamma}}(h, \widehat{\mathbf{f}}_t)$ *defined above satisfy, with probability at least* $1 - e^{-s}$*,*

$$\|\widehat{\mathbf{\Gamma}}(h, \widehat{\mathbf{f}}_t) - \widehat{\mathbf{\Gamma}}(h, \mathbf{f}_t)\|_{\mathbb{F}}^2 \lesssim \frac{1}{T}\left(\frac{1}{p} + \frac{1}{T}\right)s$$

for each $h = -T+1, \ldots, 0, \ldots, T-1$.

Notice that both $\mathbf{F}$ and $\widehat{\mathbf{F}}$ are $K \times K$ matrices. As a direct corollary of Theorem 2.4.3, we can establish the concentration inequality for recovering the temporal dependence structure on each dimension of $\mathbf{f}_t$. For each $k = 1, \ldots, K$, denote the sample autocovariance function of

$\{f_{tk} : t \geqslant 1\}$ as

$$\widehat{\gamma}(h, f_{tk}) = T^{-1} \sum_{t=1}^{T-|h|} (f_{t+|h|,k} - \bar{f}_k)(f_{tk} - \bar{f}_k)^\top,$$

where $\bar{f}_k = T^{-1} \sum_{t=1}^{T} f_{tk}$, and also let the sample autocovariance function of $\{\widehat{f}_{tk} : t \geqslant 1\}$ by

$$\widehat{\gamma}(h, \widehat{f}_{tk}) = T^{-1} \sum_{t=1}^{T-|h|} (\widehat{f}_{t+|h|,k} - \bar{\widehat{f}}_k)(\widehat{f}_{tk} - \bar{\widehat{f}}_k)^\top,$$

where $\bar{\widehat{f}}_k = T^{-1} \sum_{t=1}^{T} \widehat{f}_{tk}$. From Theorem 2.4.3, with probability at least $1 - e^{-s}$, we have

$$|\widehat{\gamma}(h, \widehat{f}_{tk}) - \widehat{\gamma}(h, f_{tk})|^2 \lesssim \frac{1}{T}\left(\frac{1}{p} + \frac{1}{T}\right) s$$

for each $h = -T + 1, \ldots, 0, \ldots, T - 1$. Similarly, denote the sample autocorrelation function (ACF; Brockwell et al., 1991) of $\{f_{tk} : t \geqslant 1\}$ by $\widehat{\rho}(h, f_{tk}) = \{\widehat{\gamma}(0, f_{tk})\}^{-1}\widehat{\gamma}(h, f_{tk})$ and the sample partial autocorrelation function (PACF) by $\widehat{\Psi}(0, f_{tk}) = 1$ and $\widehat{\Psi}(h, f_{tk})$ being the hth entry of $\widehat{\boldsymbol{\Psi}}(f_{tk})$ where $\widehat{\boldsymbol{\Psi}}(f_{tk}) = \widehat{\mathbf{R}}_h^{-1}(f_{tk})\widehat{\boldsymbol{\rho}}_h(f_{tk})$ with $\widehat{\mathbf{R}}_h(f_{tk}) = \{\widehat{\rho}((i-j), f_{tk})\}_{i,j=1}^{h}$ and $\widehat{\boldsymbol{\rho}}_h(f_{tk}) = (\widehat{\rho}(1, f_{tk}), \ldots, \widehat{\rho}(h, f_{tk}))^\top$. Likewise, we denote the sample ACF of $\{\widehat{f}_{tk} : t \geqslant 1\}$ by $\widehat{\rho}(h, \widehat{f}_{tk}) = \{\widehat{\gamma}(0, \widehat{f}_{tk})\}^{-1}\widehat{\gamma}(h, \widehat{f}_{tk})$, let the sample PACF of $\{\widehat{f}_{tk} : t \geqslant 1\}$ be $\widehat{\Psi}(0, \widehat{f}_{tk}) = 1$, and let $\widehat{\Psi}(h, \widehat{f}_{tk})$ be the hth entry of $\widehat{\boldsymbol{\Psi}}(\widehat{f}_{tk})$, where $\widehat{\boldsymbol{\Psi}}(\widehat{f}_{tk}) = \widehat{\mathbf{R}}_h^{-1}(\widehat{f}_{tk})\widehat{\boldsymbol{\rho}}_h(\widehat{f}_{tk})$, $\widehat{\mathbf{R}}_h^{-1}(\widehat{f}_{tk}) = \{\widehat{\rho}((i-j), \widehat{f}_{tk})\}_{i,j=1}^{h}$ and $\widehat{\boldsymbol{\rho}}_h(\widehat{f}_{tk}) = (\widehat{\rho}(1, f_{tk}), \ldots, \widehat{\rho}(h, \widehat{f}_{tk}))^\top$. From Theorem 2.4.3, we have the following results.

Theorem 2.4.4. *Under Conditions 2.2.1-2.2.3, given* $\mathbf{Y}$ *from* (2.1.1) *or* (2.1.2)*, for each* $k = 1, \ldots, K$ *and* $h = -T + 1, \ldots, T - 1$*, with probability at least* $1 - e^{-s}$*,*

$$|\widehat{\rho}(h, \widehat{f}_{tk}) - \widehat{\rho}(h, f_{tk})|^2 \lesssim \frac{1}{T}\left(\frac{1}{p} + \frac{1}{T}\right) s,$$
$$|\widehat{\Psi}(h, \widehat{f}_{tk}) - \widehat{\Psi}(h, f_{tk})|^2 \lesssim \frac{1}{T}\left(\frac{1}{p} + \frac{1}{T}\right) s.$$

Theorem 2.4.4 shows that sample ACF and PACF of $\{f_{tk} : t \geqslant 1\}$ can be consistently recovered by those of $\{\widehat{f}_{tk} : t \geqslant 1\}$. In addition, Theorem 2.4.4 implies that the sample ACF of f_{tk} and $\widehat{f}_{tk}$ have the common asymptotic distribution. Similar conclusions are also true for the sample PACF. These results will have wide applications in modeling and forecasting high-dimensional time series by (2.1.1) along a broad class of parametric models on $\mathbf{f}_t$. For instance, for the autoregression models, the sample PACF's give the Yule-Walker estimator to the autoregressive coefficients; and for the moving average models, the innovation estimator, which is computed from the sample ACF's, can be employed to estimate the moving average coefficients.

2.4.3 Spectral Properties of Large Sample Covariance Matrices

Extending results in Section 2.3 on eigenvectors $\widehat{\boldsymbol{f}}_k$ of the scaled right Gram matrix $T^{-1}\mathbf{Y}^\top\mathbf{Y}$, we study eigenvectors of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{w}}_i$ for $i = 1, \ldots, p$. First, as an application of Theorem 2.3.1, we characterize the deviation between $\widehat{\boldsymbol{w}}_i$ and $\boldsymbol{w}_i$ in Theorem 2.4.5.

Theorem 2.4.5. *Under Conditions 2.2.1-2.2.3, given* $\mathbf{Y}$ *from* (2.1.1) *or* (2.1.2), $\mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2) \lesssim p^{-1} + T^{-1}$ *and* $\mathrm{Var}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2) \lesssim p^{-2} + T^{-2}$ *for each* $i = 1, \ldots, K$.

From Theorem 2.4.5, the first K eigenvectors of the sample covariance matrix converge to those of $\boldsymbol{\Sigma}$ in probability. Together with Theorems 2.3.3, we establish the consistency on estimating the spectral structure corresponding to the first K eigenvalues of $\boldsymbol{\Sigma}$ specified by (2.1.3). Notice that no restrictions on p and T are imposed on this consistency. By Theorem 2.3.3, for $p < T$, $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathbb{F}} \lesssim T^{-1/2}p^{3/2} + T^{-1/2}p\sqrt{s}$ with probability at least $1 - e^{-s}$. Thus, from the Davis-Kahan Theorem (Cai et al., 2017; Davis and Kahan, 1970; Fan et al., 2018b; Yu et al., 2014) and Condition 2.2.1, we have the following corollary.

Corollary 2.4.1. *Let* $\Theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i) = \cos^{-1}(\widehat{\boldsymbol{w}}_i^\top \boldsymbol{w}_i)$ *be the angle between* $\widehat{\boldsymbol{w}}_i$ *and* $\boldsymbol{w}_i$. *Under Conditions 2.2.1, 2.2.2, (i),(ii) and (iv) in 2.2.3, and 2.3.1, given* $\mathbf{Y}$ *from* (2.1.1) *or* (2.1.2), *it satisfies*

$$\mathbb{E}\{\sin\Theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i)\} \lesssim \frac{\mathbb{E}(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathbb{F}})}{\min_{j\neq i}|\lambda_j - \lambda_i|} \lesssim T^{-1/2} + p^{-1/2}$$

for each $i = 1, \ldots, K$.

Since $\sin\Theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i) \leqslant \|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2 \leqslant 2\sin\Theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i)$ with properly chosen direction of $\widehat{\boldsymbol{w}}_i$, Corollary 2.4.1 gives a similar result to the Davis-Kahan Theorem in low dimension. However, when $p > T$, as shown in Theorem 2.3.3, not all eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ necessarily converge to those of $\boldsymbol{\Sigma}$ and neither does $\widehat{\boldsymbol{\Sigma}}$ converge to $\boldsymbol{\Sigma}$. Then, the Davis-Kahan Theorem cannot be directly applied to $\widehat{\boldsymbol{\Sigma}}$. Instead, with the low-rank structure in (2.1.2), we can establish similar results for an alternative estimator to $\boldsymbol{\Sigma}$. We start with eigenvectors corresponding to the first K largest eigenvalues of $\mathbf{Y}^\top\mathbf{Y}$, *i.e.*, the PCA estimator to the latent factor matrix and loading matrix. Under Condition 2.3.1, $\boldsymbol{\Sigma}$ in (2.1.3) can be estimated by $\widehat{\boldsymbol{\Sigma}}_{\mathrm{PCA}} = \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top + \widehat{\boldsymbol{\Sigma}}_u$, where $\widehat{\mathbf{A}}$ is defined in Section 2.3, $\widehat{\boldsymbol{\Sigma}}_u$ is a diagonal matrix with diagonal entries $\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_p^2$, $\widehat{\sigma}_i^2 = T^{-1}\sum_{t=1}^T \widehat{u}_{it}^2$ for $i = 1, \ldots, p$ and $\widehat{u}_{it}$ is the entry in the ith row and tth column of $\widehat{\mathbf{U}} = \mathbf{Y} - \widehat{\mathbf{A}}\widehat{\mathbf{F}}^\top$. Then, similar to Corollary 2.4.1, we have the following result.

Corollary 2.4.2. *Given* $\mathbf{Y}$ *from* (2.1.1) *or* (2.1.2)*, let* $\Theta(\widetilde{\boldsymbol{w}}_{i,\mathrm{PCA}}, \boldsymbol{w}_i) = \cos^{-1}(\widetilde{\boldsymbol{w}}_{i,\mathrm{PCA}}^\top \boldsymbol{w}_i)$ *be the angle between* $\widetilde{\boldsymbol{w}}_{i,\mathrm{PCA}}$ *and* $\boldsymbol{w}_i$*, where* $\widetilde{\boldsymbol{w}}_{i,\mathrm{PCA}}$ *is the eigenvector corresponding to the* i*th largest eigenvalue of* $\widehat{\boldsymbol{\Sigma}}_{\mathrm{PCA}}$*. Then, under Conditions 2.2.1, 2.2.2, (i),(ii) and (iv) in 2.2.3, and 2.3.1, for each* $i = 1, \ldots, K$*,*

$$\mathbb{E}\{\sin\Theta(\widetilde{\boldsymbol{w}}_{i,\mathrm{PCA}}, \boldsymbol{w}_i)\} \lesssim \frac{\mathbb{E}(\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{PCA}} - \boldsymbol{\Sigma})\|_{\mathbb{F}})}{\min_{j\neq i}|\lambda_j - \lambda_i|} \lesssim p^{-1/2}T^{-1/2} + p^{-1}.$$

Next, as an application of Theorem 2.3.2, we will show the approximation error rate to the distribution of the standardized deviation between $\boldsymbol{w}_i$ and $\widehat{\boldsymbol{w}}_i$ by the standard normal distribution, namely the Berry-Esseen type bound. First, we consider $\mathbf{P}_i = \mathbf{I}_p - \boldsymbol{w}_i(\boldsymbol{w}_i'\boldsymbol{w}_i)^{-1}\boldsymbol{w}_i'$ and $\widehat{\mathbf{P}}_i = \mathbf{I}_p - \widehat{\boldsymbol{w}}_i(\widehat{\boldsymbol{w}}_i'\widehat{\boldsymbol{w}}_i)^{-1}\widehat{\boldsymbol{w}}_i'$, which are the projectors onto the orthogonal spaces of $\boldsymbol{w}_i$ and $\widehat{\boldsymbol{w}}_i$. First, we will obtain the Berry-Esseen type bound for $\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2$ in Theorem 2.4.6.

Theorem 2.4.6. *Under Conditions 2.2.1, 2.2.2, (i),(ii) and (iv) in 2.2.3, and 2.3.1, given* $\mathbf{Y}$ *from* (2.1.1) *or* (2.1.2)*, for each* $i = 1, \ldots, p$*,*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left\{ \frac{\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2 - \mathbb{E}(\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2)}{\mathrm{Var}^{1/2}(\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2)} \leqslant x \right\} - \Phi(x) \right| \lesssim \frac{1}{B_i} + \frac{\log(T)}{\sqrt{T}} + \frac{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}}{T^{1/8} B_i}.$$

where $B_i = 2\sqrt{2}\|\mathbf{P}_i \mathbf{\Sigma} \mathbf{P}_i\|_2 \|\mathbf{Q}_i \mathbf{\Sigma} \mathbf{Q}_i\|_2$ *and* $\mathbf{Q}_i = \sum_{j \neq i} (\lambda_i - \lambda_j)^{-1} \mathbf{P}_j$*.*

A similar result has been documented for independent data in literature (Koltchinskii and Lounici, 2016); while, Theorem 2.4.6 is more general by allowing temporal dependence in data. In fact, the third term on the right hand side above quantifies the effect of temporal dependence, and as a result, the convergence rate is slightly compromised compared to the rate under independence. Theorem 2.4.6 leads to the following corollary, which extends the Berry-Esseen bound for random vectors (Bobkov et al., 2018; Bobkov and Chistyakov, 2015; Goldstein et al., 2009).

Corollary 2.4.3. *Under Conditions 2.2.1, 2.2.2, (i),(ii) and (iv) in 2.2.3, and 2.3.1, given* $\mathbf{Y}$ *from* (2.1.1) *or* (2.1.2)*, for any matrix* $\mathbf{C}$ *and* $i = 1, \ldots, p$*,*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left\{ \frac{\|\widehat{\mathbf{P}}_i \mathbf{C} - \mathbf{P}_i \mathbf{C}\|_2^2 - \mathbb{E}(\|\widehat{\mathbf{P}}_i \mathbf{C} - \mathbf{P}_i \mathbf{C}\|_2^2)}{\mathrm{Var}^{1/2}(\|\widehat{\mathbf{P}}_i \mathbf{C} - \mathbf{P}_i \mathbf{C}\|_2^2)} \leqslant x \right\} - \Phi(x) \right| \lesssim \frac{1}{B_i} + \frac{\log(T)}{\sqrt{T}} + \frac{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}}{T^{1/8} B_i}.$$

In addition, for each $i = 1, \ldots, p$*,*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left\{ \frac{\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2 - \mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2)}{\mathrm{Var}^{1/2}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2)} \leqslant x \right\} - \Phi(x) \right| \lesssim \frac{1}{B_i} + \frac{\log(T)}{\sqrt{T}} + \frac{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}}{T^{1/8} B_i}.$$

Note that $B_i = O(\sqrt{p})$ for $i = 1, \ldots, K$. Thus, Corollary 2.4.3 provides a uniform normal approximation to standardized $\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2$ for $i = 1, \ldots, K$. However, $B_i = O(1)$ for $i > K$

so that the upper bounds in both Theorem 2.4.6 and Corollary 2.4.3 do not necessarily shrink to zero. Therefore, as noted by Koltchinskii and Lounici (2016), the normal approximation to $\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2$ for $i > K$ may fail to hold. Together with Theorem 2.3.3, Corollary 2.4.3 shows that, the spectral structures corresponding to the spiked eigenvalues, *i.e.* the first K eigenvalues of the sample covariance matrix, provide good estimates to the corresponding spectral structures of $\boldsymbol{\Sigma}$, even for $p > T$ for which $\widehat{\boldsymbol{\Sigma}}$ is no longer consistent to $\boldsymbol{\Sigma}$.

Remark 2.4.1. *In practice,* $\mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2)$ *and* $\mathrm{Var}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2)$ *are unknown. To use Corollary 2.4.3 for inference, we need estimate them. Koltchinskii and Lounici (2017) offered a data-splitting procedure which splits the sample into three subsamples: the first for estimating the expectation, the second for estimating the variance, and the third for building the confidence set. In addition, since* $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ *is naturally an empirical process, the multiplier bootstrap can been employed to build the confidence set of* $\boldsymbol{w}_i$ *for each* $i = 1, \ldots, K$ *without data splitting for i.i.d data (Naumov et al., 2019). Under Condition 2.2.3,* $\boldsymbol{y}_t$ *from* (2.1.1) *is weakly temporal dependent and can be approximated by some* m*-dependent time series* $\widetilde{\boldsymbol{y}}_t$ *in the following sense,*

$$
\begin{aligned}
|\mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2|\boldsymbol{y}_t) - \mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2|\widetilde{\boldsymbol{y}}_t)| &\lesssim \frac{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}}{T^{9/8}}, \\
|\,\mathrm{Var}^{1/2}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2|\boldsymbol{y}_t) - \mathrm{Var}^{1/2}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2|\widetilde{\boldsymbol{y}}_t)| &\lesssim \frac{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}}{T^{9/8}};
\end{aligned}
$$

and $\|\widehat{w}_i - w_i\|_2^2$ *based on* $\boldsymbol{y}_t$ *and* $\widetilde{\boldsymbol{y}}$ *have the similar normal approximations (Chen et al., 2004, 2007; Zhang and Cheng, 2018). Therefore, we can employ the following blockwise multiplier bootstrap procedure to draw inference on* $\boldsymbol{w}_i$ *(Zhang and Cheng, 2018), whose guarantee is provided by Corollary 2.4.3 and the above approximation using* $\widetilde{\boldsymbol{y}}_t$*.*

Algorithm: Blockwise multiplier bootstrap procedure for the inference of $\boldsymbol{w}_i$.

Input: Observations $\{y_{it}\}_{i=1,t=1}^{p,T}$.

Step 1. Pre-specify integers b_T and l_T such that $T = b_T l_T$ based on the nonparametric plug-in method (Bühlmann and Künsch, 1999), the empirical criteria-based method (Hall et al., 1995) or the algorithm in Zhang and Cheng (2018).

Step 2. Generate e_{js} *i.i.d.* from $\mathcal{N}(1,1)$ for $j = 1, \ldots, B$ and $s = 1, \ldots, l_T$.

Step 3. For each j, calculate $\boldsymbol{\Sigma}_j^{\mathrm{BS}} = T^{-1} \sum_{s=1}^{l_T} e_{js} \sum_{t=(s-1)b_T+1}^{sb_T} \boldsymbol{y}_t \boldsymbol{y}_t'$.

Step 4. For each $i = 1, \ldots, K$, denote $\boldsymbol{w}_{i,j}^{\mathrm{BS}}$ the eigenvector corresponding to the ith largest eigenvalue of $\boldsymbol{\Sigma}_j^{\mathrm{BS}}$ and define $\gamma_\alpha^{\mathrm{BS}}$ as the $1-\alpha$ percentile of $\{\|\boldsymbol{w}_{i,j}^{\mathrm{BS}} - \widehat{\boldsymbol{w}}_i\|_2^2\}_{j=1}^B$.

Output: Confidence set of $\boldsymbol{w}_i$ as $\{\boldsymbol{w} : \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_i\|_2^2 \leqslant \gamma_\alpha^{\mathrm{BS}}\}$ for $i = 1, \ldots, K$.

2.4.4 Low-rank Matrix Denoising based on Temporally Dependent Data

Low-rank matrix denoising has numerous applications such as robust video restoration (Ji et al., 2011), hyperspectral image restoration (He et al., 2015; Zhang et al., 2013), and underdetermined direction of arrival estimation (Pal and Vaidyanathan, 2014). Lately, the low-rank matrix denoising in the presence of both heteroskedastic errors and dependent samples has attracted great attention in literature (Zhang et al., 2018). Suppose we observe time series

$$y_{it} = x_{it} + u_{it}$$

for $i = 1, \ldots, p$ and $t = 1, \ldots, T$, which can be written as

$$\mathbf{Y} = \mathbf{X} + \mathbf{U}$$

where $\mathbf{Y} = \{y_{it}\}_{i=1,t=1}^{p,T}$, $\mathbf{X} = \{x_{it}\}_{i=1,t=1}^{p,T}$ is a fixed rank-K matrix, and $\mathbf{U} = \{u_{it}\}_{i=1,t=1}^{p,T}$. Assume noise matrix $\mathbf{U}$ satisfies Condition 2.2.3. Let $\mathbf{X} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{V}'$ be the SVD, where $\mathbf{W}$ is a $p \times K$ orthogonal matrix and $\mathbf{V}$ is a $T \times K$ orthogonal matrix. Note that the column space of $\mathbf{W}$ is

essentially that of $\mathbf{A}$ in (2.1.2) under Condition 2.2.1. Then we can use PCA to estimate $\mathbf{W}$ by $\widehat{\mathbf{W}} = (\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}})^{-1/2}\widehat{\mathbf{A}}$ with the following theoretical guarantees.

Corollary 2.4.4. *Suppose that $p \lesssim \lambda_{\min}(\mathbf{\Lambda}) \lesssim \lambda_{\max}(\mathbf{\Lambda}) \lesssim p$. Then $\mathbf{W}$ and $\widehat{\mathbf{W}}$ satisfy*

$$\mathbb{E}\{\|\sin\Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_{\mathbb{F}}\} \lesssim \frac{1}{\sqrt{p}} + \frac{1}{\sqrt{T}},$$

where $\|\sin\Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_{\mathbb{F}} \overset{d}{=} \|\mathbf{W}_\perp^\top \widehat{\mathbf{W}}\|_{\mathbb{F}}$ and $\mathbf{W}_\perp$ is a $p \times (p-K)$ orthogonal matrix such that $(\mathbf{W}, \mathbf{W}_\perp)$ is a $p \times p$ orthogonal matrix.

In Corollary 2.4.4, we consider a spike model with potentially heteroskedastic errors. Like the approximate factor model, the spiked singular values of $\mathbf{X}$ provide stronger signals compared to the model used in traditional matrix denoising (Cai and Zhang, 2016; Zhang et al., 2018). To compare, for the non-spiked signal matrix $\mathbf{X}$ and homoskedastic variance of $\mathbf{U}$, the optimal rate of matrix denoising using the regular SVD is $\mathbb{E}(\|\sin\Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_{\mathbb{F}}) \lesssim \min(p, T)^{-1/2}$ (Theorems 3 and 4, Cai and Zhang, 2016). Thus, Corollary 2.4.4 gives similar results to the regular SVD (Cai and Zhang, 2016) and the diagonal-deletion SVD (Florescu and Perkins, 2016). In addition, Theorem 4 in Zhang et al. (2018) showed that the heteroskedastic PCA can obtain the optimal rate of matrix denoising for non-spiked signal matrix $\mathbf{X}$ with heteroskedastic errors. It is easy to see that if the variance of u_{it} is bounded for each i and t, the optimal rate in Zhang et al. (2018) is also $\mathbb{E}(\|\sin\Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_{\mathbb{F}}) \lesssim \min(p, T)^{-1/2}$. Hence, our result also matches the heteroskedastic PCA (Zhang et al., 2018) in the presence of heteroskedastic errors.

2.5 Numerical studies

In this section, we perform simulation studies to further illustrate results displayed in Sections 2.3, 2.4.1, and 2.4.2.

We first conduct numerical experiments to demonstrate Theorem 2.3.3. Consider model (2.1.1) with $K = 1$, $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.01)$, and three settings for latent process $\mathbf{f}_t$: (1) AR(1) with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0, 1)$ innovation; (2) AR(1) with autoregressive coefficient

$\phi = 0.5$ and t_8 innovation; and (3) ARMA$(1,1)$ with autoregressive coefficient $\phi = 0.5$, moving average coefficient $\theta = 0.5$, and $\mathcal{N}(0,1)$ innovation. Two scenarios on p and T are considered, $p = \lfloor 2T^{1/2} \rfloor$ and $T = \lfloor 2p^{1/2} \rfloor$. Based on 100 replicates, the simulation results are displayed in Figure 2.1. From panels (a1), (b1), and (c1), we can see that $\widehat{\lambda}_i$ converges to λ_i when $p < T$. The relative error $|\widehat{\lambda}_i/\lambda_i - 1|$ for $i = 1$ converges to zero faster than those for $i = 2$ and 10 since λ_1 diverges in p while λ_2 and λ_{10} remain in constants. In addition, from panels (a2), (b2), and (c2), it is noticed that $\widehat{\lambda}_1$ still converges to λ_1 even for $p > T$ while the deviations of other eigenvalues diverge as p and T diverge. These patterns are commonly observed for all three settings on $\mathbf{f}_t$. This matches results in Theorem 2.3.3.

Next, we demonstrate the influence of p, T, and eigenvalues of $\mathbf{\Sigma}$ on the probability of estimating the correct number of factors using the ratio of consecutive eigenvalues in (2.4.1). Consider model (2.1.1) with $K = 3$ factors and $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 25)$. The three components in $\mathbf{f}_t$ are independent and identically follow AR(1) process with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0,1)$ innovation. We further set $\mathbf{A}$ such that $p^{-1}\mathbf{A}^\top\mathbf{A}$ has diagonal entries $\{16, 4, 1\}$ (panels (a1) and (a2) in Figure 2.2), $\{16, 4, 2\}$ (panels (b1) and (b2) in Figure 2.2), and $\{32, 4, 2\}$ (panels (c1) and (c2) in Figure 2.2). For p and T, two settings are reported: (1) T is fixed, $p = 100, 200, \ldots, 1000$; and (2) p is fixed, $T = 100, 200, \ldots, 1000$. Based on 500 replicates, results on $\log(1 - \mathbb{P}\{\widehat{K} = K\})$ are displayed in Figure 2.2. In Figure 2.2, we notice that $\log(1 - \mathbb{P}\{\widehat{K} = K\})$ decreases faster for greater λ_K/λ_{K+1} and smaller $\max_{i \neq K} \lambda_i/\lambda_{i+1}$. In fact, from Theorem 2.4.1, $\log(1 - \mathbb{P}\{\widehat{K} = K\})$ is bounded by a quadratic function of $\sqrt{\max(p,T)}$ with C_1 and C_2 defined in Theorem 2.4.1. Since c and C in Theorem 2.3.3 only depend on the distribution of $\boldsymbol{u}_t$, C_2 is same for different $\mathbf{A}$. On the other hand, as λ_K/λ_{K+1} increases and $\max_{i \neq K} \lambda_i/\lambda_{i+1}$ decreases, C_1 increases so that the quadratic function of $\sqrt{\max(p,T)}$ has a smaller vertex and greater quadratic coefficient. Thus, Figure 2.2 demonstrates the conclusion in Theorem 2.4.1.

Finally, we study the estimation of moments of latent factor process $\mathbf{f}_t$ to demonstrate Theorem 2.4.4. Still consider model (2.1.1) with $K = 1$ factor and $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.01)$. Also, we set three models for $\mathbf{f}_t$: (1) AR(1) with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0,1)$ innovation; (2)

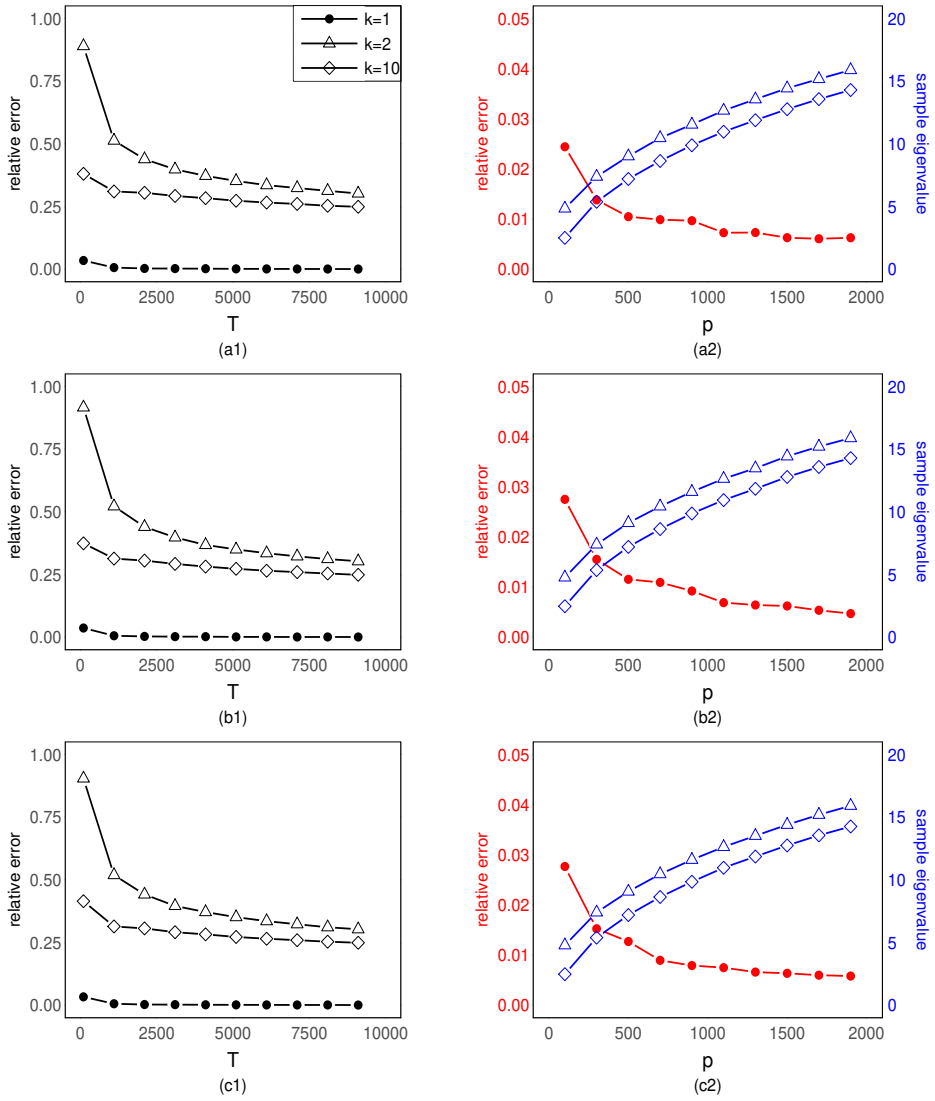

Figure 2.1: In the left column, $p = \lfloor 2T^{1/2} \rfloor$ ($p < T$), and in the right column $T = \lfloor 2p^{1/2} \rfloor$ ($p > T$). In panels (a1) and (a2), latent process $\mathbf{f}_t$ follows setting (1); in panels (b1) and (b2), latent process $\mathbf{f}_t$ follows setting (2); and in panels (c1) and (c2), latent process $\mathbf{f}_t$ follows setting (3). In panels (a1), (b1), and (c1), the relative errors $|\widehat{\lambda}_i/\lambda_i - 1|$ for $i = 1, 2, 10$ are displayed. In panels (a2), (b2), and (c2), the relative errors are displayed for λ_1 and the sample eigenvalues are displayed for λ_2 and λ_{10} to show that they are unbounded in p.

AR(1) with autoregressive coefficient $\phi = 0.5$ and t_8 innovation; and (3) ARMA$(1, 1)$ with autoregressive coefficient $\phi = 0.5$, moving average coefficient $\theta = 0.5$ and $\mathcal{N}(0, 1)$ innovation. Two settings about p and T are considered: $p = 200$ with $T = 100, 200, \ldots, 1000$; and $T = 200$ with $p = 100, 200, \ldots, 1000$. Based on 100 replicates, $|\widehat{\rho}(h, \widehat{f}_{tk}) - \widehat{\rho}(h, f_{tk})|$ and $|\widehat{\Psi}(h, \widehat{f}_{tk}) - \widehat{\Psi}(h, f_{tk})|$ versus T and p are displayed in log-log scale in Figures 2.3 and 2.4. For all settings, the squared

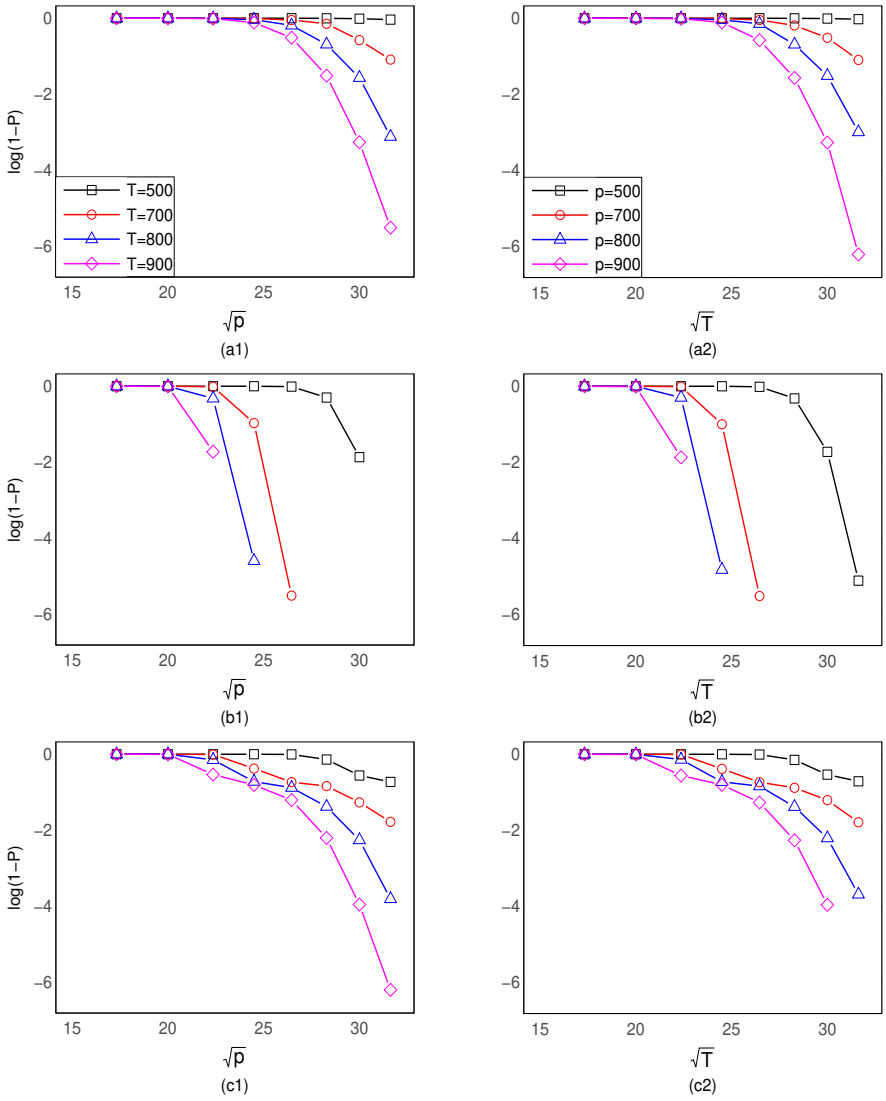

Figure 2.2: Plots about $\log(1-\mathbb{P}\{\widehat{K}=K\})$ for $p=100,200,\ldots,1000$, $T=500,700,800,900$ (left column), and $T=100,200,\ldots,1000$, $p=500,700,800,900$ (right column). The diagonal entries in $p^{-1}\mathbf{A}^\top\mathbf{A}$ are $\{16,4,1\}$ (panels (a1) and (a2)), $\{16,4,2\}$ (panels (b1) and (b2)), and $\{32,4,2\}$ (panels (c1) and (c2)). Points are omitted when $\log(1-\mathbb{P}\{\widehat{K}=K\})=-\infty$, *i.e.* $\mathbb{P}(\widehat{K}=K)=1$.

differences for both ACF and PACF shrink to zero as p and T diverge. Also, in all settings, the slopes of the log difference of ACF or PACF versus $\log T$ or $\log p$ are $-1/2$ (the red lines), which confirms the established rates of convergence in Theorem 2.4.4.

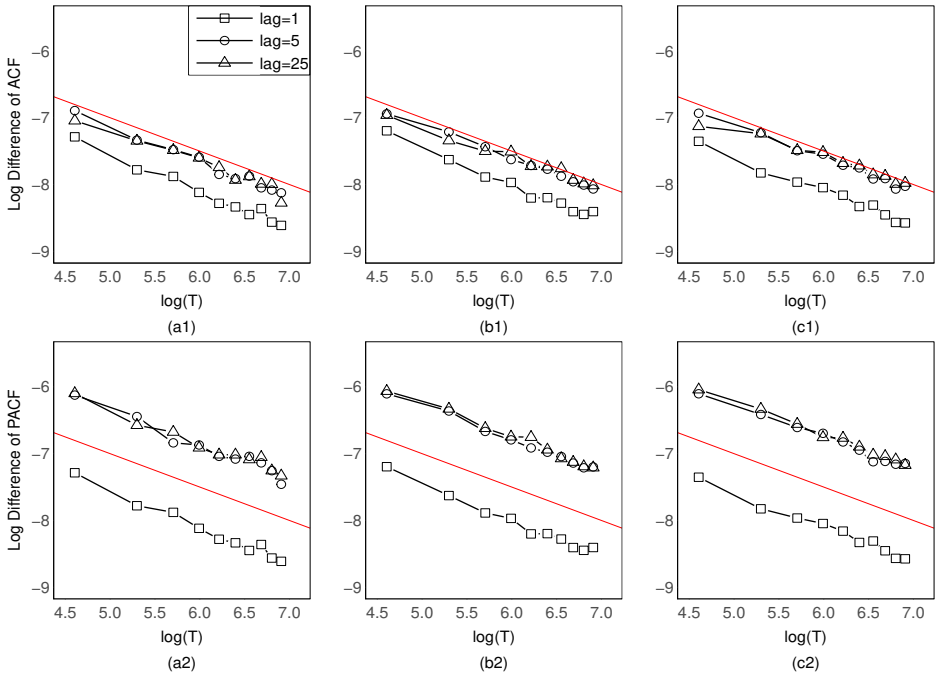

Figure 2.3: Log differences of ACF (first row) and PACF (second row) of $\{f_{t1} : t \geqslant 1\}$ at lag $h = 1$, lag $h = 5$, and lag $h = 25$ for $p = 200$ and $T = 100, 200, \ldots, 1000$. The latent process follows AR(1) process with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0, 1)$ innovation in panels (a1) and (a2); it follows AR(1) process with autoregressive coefficient $\phi = 0.5$ and t_8 innovation in panels (b1) and (b2); and it follows ARMA$(1, 1)$ with autoregressive coefficient $\phi = 0.5$, moving average coefficient $\theta = 0.5$, and $\mathcal{N}(0, 1)$ innovation in panels (c1) and (c2). The red solid line has slope $-1/2$.

2.6 Conclusions

In this paper, we scrupulously study the non-asymptotic properties of the spectral decomposition of large Gram-type matrices under the assumption that data matrix $\mathbf{Y}$ is governed by a factor model. As a result, we establish the exponential tail bound for the first and second moments of the deviation between the empirical and population eigenvectors to the right Gram matrix as well as the Berry-Esseen type bound to characterize the Gaussian approximation of these deviations. Technically, we successfully relax the assumption upon latent factors in the factor model, so that the latent factor processes are no longer restricted to a subspace as stated by Condition PC1 in Bai and Ng (2013). We also obtain the non-asymptotic tail bound of the ratio between eigenvalues of the sample covariance matrix, and their population counterparts regardless of the size of the data matrix. This extends the works of Bai and Yin (1993), Lam and Yao (2012), and Wang and Fan (2017).

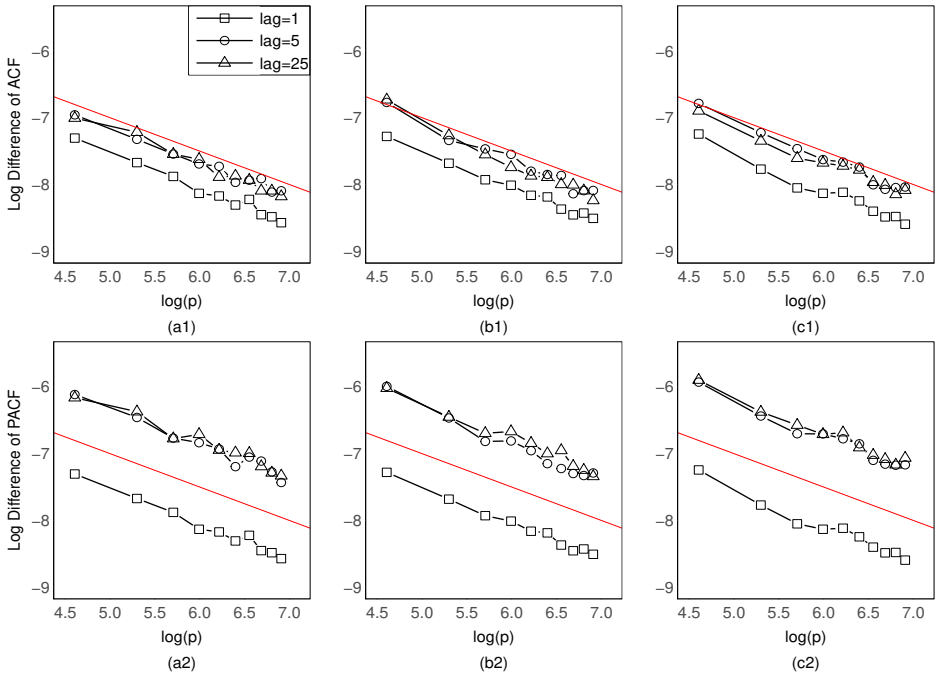

Figure 2.4: Log differences of ACF (first row) and PACF (second row) of $\{f_{t1} : t \geqslant 1\}$ at lag $h = 1$, lag $h = 5$, and lag $h = 25$ for $T = 200$ and $p = 100, 200, \ldots, 1000$. The latent process follows AR(1) process with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0, 1)$ innovation in panels (a1) and (a2); it follows AR(1) process with autoregressive coefficient $\phi = 0.5$ and t_8 innovation in panels (b1) and (b2); and it follows ARMA$(1, 1)$ with autoregressive coefficient $\phi = 0.5$, moving average coefficient $\theta = 0.5$, and $\mathcal{N}(0, 1)$ innovation in panels (c1) and (c2). The red solid line has slope $-1/2$.

With the derived non-asymptotic properties of eigenvalues of the sample covariance matrix, we provide the non-asymptotic characterization of different consecutive-eigenvalues-based methods to estimate the number of latent factors in factor models and relate machine learning problems. The established non-asymptotic lower bound of the probability of estimating the correct number of factors reveal the influence of p, T and eigenvalues of $\mathbf{\Sigma}$ on different methods. In addition, as an application of our main results, we provide statistical guarantees on estimating the parametric models for the latent process in dynamic or approximate factor models, so that one can make forecast based on the factor models and high-dimensional time series. We also obtain non-asymptotic properties of the spectral structure of large sample covariance matrices, including the Davis-Kahan type perturbation result and the approximation error rate to the distribution of the standardized deviation between $\boldsymbol{w}_i$ and $\widehat{\boldsymbol{w}}_i$ by the standard normal distribution, *i.e.* the Berry-Esseen type bound. Based on these results, it is possible to construct confidence sets for the leading eigenvectors of $\mathbf{\Sigma}$

using the multiplier bootstrap. Finally, we apply our results to the low-rank matrix denoising in the presence of heteroskedastic errors and temporal dependence in data.

Chapter 3

Estimation and Inference of Semiparametric Factor Model

3.1 Introduction

Jointly modeling a large and possibly divergent number of temporally evolving subjects arise ubiquitously in statistics, econometrics, finance, biology, and environmental sciences. Statistical analysis has been successfully adopted to explain the interactions and co-movements among the temporally evolving subjects (Hsiao, 2014; Lam and Yao, 2012; Lütkepohl, 2006; Stock and Watson, 2002a). A prototype model with both the modulating or systematic and dependence components is the linear model $y_{it} = \boldsymbol{z}_{it}'\boldsymbol{\beta} + \varepsilon_{it},\ i = 1, \ldots, n,\ t = 1, \ldots, T$, 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 the p-dimensional covariate vector that might evolve in time, and $(\varepsilon_{1t}, \ldots, \varepsilon_{nt})'$ is a vector time series with possible contemporaneous correlation. Here, the number of subjects n might diverge much faster than the number of time points T and p is low-dimension or fixed. To name a few applications in practice, y_{it} can be used to model the gene expression level or protein abundance of the ith marker in a time course experiment (e.g. Desai and Storey, 2012), or the concentration of certain air pollutant in county i at day t (e.g. Lindström et al., 2014), or daily closing prices for asset i on market (e.g. Connor et al., 2012). As n rapidly grows, heteroscedasticity across subjects becomes inevitable and brings substantial challenges to modeling, estimation and inference (Arellano and Bond, 1991; Fan et al., 2014; Hayakawa and Pesaran, 2015). Ignoring the subject-specific heteroscedasticity is known to lead inefficient estimation and inference on the regression components. To battle with such challenges, carefully modeling ε_{it} is needed to characterize the remaining contemporaneous and serial correlations as well as heteroscedasticity across subjects.

In this paper, we introduce a flexible data-driven model, in which 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. This approach is rooted in the idea of approximate factor structure by Chamberlain and Rothschild (1983). That is, by separating the eigenvalues of covariance matrix into divergent and non-divergent groups, the observed process can be approximated by a latent factor process along with its loading (Bai, 2003; Bai and Ng, 2013; Lam and Yao, 2012; Stock and Watson, 2002a; Wang and Wang, 2018). Specifically, motivated by Connor and Linton (2007), Connor et al. (2012), and Fan et al. (2016), we model the subject-specific effect in the covariance by $\boldsymbol{g}(\boldsymbol{x}_i) = (g_1(\boldsymbol{x}_i), \ldots, g_K(\boldsymbol{x}_i))'$ with time invariant covariates $\boldsymbol{x}_i$ and nonparametric functions $g_1, \ldots, g_K$. In practice, $\boldsymbol{x}_i$ could be the genetic information in the health study or the market capitalization in finance applications. Then, we consider a K-dimensional zero-mean process $\boldsymbol{f}_t$, and introduce the subject-specific heteroscedasticity model with latent semiparametric factor structure as

$$y_{it} = \boldsymbol{z}_{it}'\boldsymbol{\beta} + \boldsymbol{g}(\boldsymbol{x}_i)'\boldsymbol{f}_t + u_{it}, \tag{3.1.1}$$

where the residual process u_{it} is independent of $\boldsymbol{f}_t$. Analogous to the traditional factor models, $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ serve as the loading and factor, respectively. Particularly, $\boldsymbol{g}(\boldsymbol{x}_i)$ models the desired 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 (3.1.1) features a large number of widely used statistical models. For example, when $\boldsymbol{f}_t$ is degenerate, (3.1.1) reduces to the partially linear additive models (Bouzebda and Chokri, 2014; Tan et al., 2016); when $\boldsymbol{g}(\boldsymbol{x}_i)$ is known and $\boldsymbol{f}_t$ follows a Gaussian distribution, (3.1.1) is a linear mixed model (Rabe-Hesketh and Skrondal, 2004); when index i is replaced by a one-dimensional spatial location, (3.1.1) is analogous to the spatio-temporal model (Lu et al., 2009); when $\boldsymbol{g}(\cdot)$ degenerate to constant functions, (3.1.1) is equivalent to the traditional factor models (Bai, 2003; Chamberlain and Rothschild, 1983; Lam and Yao, 2012; Stock and Watson, 2002a) or the panel data model with unobservable interactive effects (Ahn et al., 2001b; Bai, 2009b; Bai et al., 2014; Moon and Weidner, 2017b).

Like the partially linear model or the linear mixed model, though ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ is consistent, it is not efficient without taking the unknown dependence into account. That is, a careful estimation on the unobserved loading $\boldsymbol{g}(\boldsymbol{x}_i)$ and accurate recovery of the latent process $\boldsymbol{f}_t$ are in need to guarantee some sort of efficiency in both estimation and inference on $\boldsymbol{\beta}$. In the literature, there exist a variety of approaches to estimate $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$. For instance, Connor and Linton (2007) employed a kernel method to estimate $\boldsymbol{f}_t$ given $\boldsymbol{x}_i$ with finite values, and Connor et al. (2012) extended such estimate for general $\boldsymbol{x}_i$. Additionally, the consistency on estimating the loading and latent factor, along with an important result that such consistency requires no specific relationship between T and n (Fan et al., 2016), also shed lights upon estimating the large covariance matrix under assumptions of factor structures (Fan et al., 2013). Motivated from these pioneering works, we propose a two-stage projection-based estimator for $\boldsymbol{\beta}$, $\boldsymbol{g}(\boldsymbol{x}_i)$, and $\boldsymbol{f}_t$ in model (3.1.1). Roughly speaking, adapting a projection-based principal component type estimator (Bai, 2003; Fan et al., 2016), we first estimate $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ from $y_{it} - \boldsymbol{z}_{it}'\widehat{\boldsymbol{\beta}}^0$ for some initial consistent estimator $\widehat{\boldsymbol{\beta}}^0$. Next, in the second stage, we update the estimate of $\boldsymbol{\beta}$ with a generalized least squares (GLS) type approach using estimates of $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$ from the first-stage.

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, particularly the efficiency, of the second-stage estimator on $\boldsymbol{\beta}$ (Baltagi, 2008; Greene, 2003). To circumvent these challenges, a major theoretical contribution of this paper is to carefully carry out the non-asymptotic analysis on the projection-based estimator on $\boldsymbol{g}(\boldsymbol{x}_i)$ 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. Then, with the derived exponential-type bounds on estimating $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$, we characterize the non-asymptotic deviation of the proposed two-stage estimator on $\boldsymbol{\beta}$ from the oracle GLS with full access to $\boldsymbol{g}(\boldsymbol{x}_i)$ and $\boldsymbol{f}_t$. These nontrivial non-asymptotic results show that our proposed two-stage estimator of $\boldsymbol{\beta}$ is overwhelmingly close to the oracle GLS, and so do their first and second moments. We thereby establish the efficiency on the proposed estimator on $\boldsymbol{\beta}$. Also, we obtain the asymptotic normality of the two-stage estimator on $\boldsymbol{\beta}$ for drawing inference.

The paper is organized as follows. In Section 3.2.1, we detail our model and discuss conditions for its identification. In Section 3.2.2, we introduce the two-stage projection-based estimation procedure on the loading, latent factor processes, and regression coefficients. We carry out the non-asymptotic analysis of our estimator and carefully explore its efficiency on estimating the regression coefficients in Section 3.3. Inference on the regression coefficients is presented in Section 3.4. In Section 3.5, we discuss the determination of the unknown dimension K of the latent process $\boldsymbol{f}_t$ and introduce a novel data-drive approach different from the existing eigenvalue-ratio based procedures. Sections 3.6 and 3.7 present extensive numerical studies and an application on air quality and energy consumption data in the United States to demonstrate the proposed method. The paper concludes with some discussions in Section 3.8. Technical details, proofs of main theorems, and extra simulations including the empirical performance of our procedure on selecting K are retained in the supplementary files.

3.2 Methodology

3.2.1 A heteroscedasticity model with latent semiparametric factor structure

Consider an $n \times 1$ vector of temporally evolving subjects $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{nt})'$ along with p-dimensional covariates $\boldsymbol{z}_{it}$ and d-dimensional time invariant factors $\boldsymbol{x}_i$ 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 accountable via $\boldsymbol{x}_i$. In our baseline formulation, each subject is modeled by a multi-factor linear model $y_{it} = \boldsymbol{z}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$ (Bianchi et al., 2019) for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a p-dimensional vector of regression coefficients common across subjects. As discussed in the introduction, we adopt the semiparametric factor model $\varepsilon_{it} = \boldsymbol{g}(\boldsymbol{x}_i)'\boldsymbol{f}_t + u_{it}$, where the 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. This leads to model (3.1.1).

The unknown smooth functions $g_k(\boldsymbol{x}_i)$'s can be further modeled in the additive fashion on $\mathbb{R}^d$ without losing flexibility yet providing some concision in techniques. That is, $g_k(\boldsymbol{x}_i) = \sum_{\ell=1}^{d} g_{k\ell}(x_{i\ell})$; see, for instance, Hastie and Tibshirani (1986) and Connor et al. (2012). Function $\boldsymbol{g}$ provides structure flexible enough to allow dependence between $\{\boldsymbol{z}_{it}\}_{i\leqslant n, t\leqslant T}$ and $\{\boldsymbol{x}_i\}_{i\leqslant n}$. For instance, consider $\boldsymbol{z}_{it} = \boldsymbol{z}_i^{(0)} + \boldsymbol{z}_{it}^{(1)}$ where $\boldsymbol{z}_i^{(0)}$ and $\boldsymbol{x}_i$ are jointly distributed, and $\boldsymbol{z}_{it}^{(1)}$ is some independent process. By assuming $\boldsymbol{g}$ from a Hölder space with no linear functions dwelling in (Condition 3.3.4 in Section 3.3), (3.1.1) remains identifiable. In addition, assume that f_{kt} has zero mean and finite variance for each k, t, the error process u_{it} has zero mean and finite variance for each i, t and is independent from $\boldsymbol{f}_t$, and $\boldsymbol{f}_t, u_{it}$ are independent from $\boldsymbol{x}_i$ and $\boldsymbol{z}_{it}$, the cross covariance of $\boldsymbol{y}_t$ is $\mathrm{Cov}(y_{it}, y_{js}|\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{g}(\boldsymbol{x}_i)' \mathrm{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_s)\boldsymbol{g}(\boldsymbol{x}_j) + \mathrm{Cov}(u_{it}, u_{js})$ for any i, j, t, s. Enlightened by this discussion, our proposed model reaches beyond the existing literature (Bai et al., 2014; Bianchi et al., 2019) in the way that the intertemporal and intratemporal dependence as well as the subject-specific heteroscedasticity are modeled simultaneously by $\boldsymbol{f}_t$ and $\boldsymbol{g}$. Our model also enriches the toolkit for modeling multivariate time series. Compared to the traditional models (Basu and Reinsel, 1993; Lütkepohl, 2006), our framework remains valid even when the number of time series is much larger than the number of time points.

For each t, let $\mathbf{Z}_t = (\boldsymbol{z}_{1t}, \ldots, \boldsymbol{z}_{nt})'$, $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{nt})'$, and denote the $n \times K$ matrix of $g_k(\boldsymbol{x}_i)$ by $\mathbf{G} = (\boldsymbol{g}(\boldsymbol{x}_1), \ldots, \boldsymbol{g}(\boldsymbol{x}_n))'$, (3.1.1) can be re-written in a more compact form

$$\boldsymbol{y}_t = \mathbf{Z}_t\boldsymbol{\beta} + \mathbf{G}\boldsymbol{f}_t + \boldsymbol{u}_t. \tag{3.2.1}$$

Similar to the traditional factor models, $\mathbf{G}$ and $\boldsymbol{f}_t$ are not separately identifiable. We need the following conditions on the model structures to control the rank and scale of latent loading function and factor process for model identification.

Condition 3.2.1. *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. In addition,* $\mathbf{Z}_t$, $\boldsymbol{f}_t$ *and* $\boldsymbol{u}_t$ *are uncorrelated from each other.*

Condition 3.2.1 corresponds to conditions discussed after (1.1) in Chamberlain and Rothschild (1983) and also Condition (C1) in Lam and Yao (2012) (with diagonal $\mathbf{\Sigma}_\varepsilon$ 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 unnecessarily independent replicates $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ and $\mathbb{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_T)$. Let $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)'$ and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$, (3.2.1) reads

$$\mathbf{Y} = \mathbb{Z}(\mathbf{I}_T \otimes \boldsymbol{\beta}) + \mathbf{G}\mathbf{F}' + \mathbf{U} \tag{3.2.2}$$

where $\otimes$ denotes the Kronecker product. The following condition guarantees the identification of $\mathbf{G}$ in (3.2.2) and therefore that in (3.1.1).

Condition 3.2.2. *Almost surely,* $T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_k$ *and* $\mathbf{G}'\mathbf{G}$ *is diagonal with distinct entries.*

First part of this condition is the PC1 condition of Bai and Ng (2013) and has been commonly adopted in factor analysis (Hallin and Liška, 2008; Moench and Ng, 2011; Wang, 2008). Note that the identification condition on $\mathbf{F}$ is compatible with Condition 3.2.1 as $T^{-1}\mathbf{F}'\mathbf{F}$ is an estimator of $\mathrm{Var}(\boldsymbol{f}_t)$. Under Condition 3.2.2, we can identify $\mathbf{GH}$ and $\mathbf{FH}$ for some $K \times K$ orthogonal matrix $\mathbf{H}$ with $\mathbf{H} = \mathbf{I} + o(\min(n, T)^{-1})$ (Bai and Ng, 2013). The distinction among entries of $\mathbf{G}'\mathbf{G}$ prevents rotational indeterminacy.

In contrast to the approximate factor model that allows cross-sectional dependence among $\boldsymbol{u}_t$, the assumption on u_{it} in Condition 3.2.1 is designated for efficiently estimating $\boldsymbol{\beta}$ without any restrictions on n and T. In fact, in the absence of the modulating component in (3.2.2), mild cross-sectional dependence of u_{it} across i will not affect the estimation on $\mathbf{G}$ and $\mathbf{F}$. On the other hand, without Condition 3.2.1 on u_{it}, a consistent estimate on $\mathrm{Cov}(\boldsymbol{u}_t)$ is required for efficiently estimating $\boldsymbol{\beta}$. This will demand some conditions on n and T, such as $\sqrt{n}\log(n) = o(T)$ (Fan et al., 2013; Wang and Fan, 2017), which is more stringent in comparison to those in Section 3.3.

3.2.2 Two-stage projection-based estimation

First-stage estimation: projection-based estimator of G and F

Given some preliminary estimator $\widehat{\boldsymbol{\beta}}^0$ satisfying $\|\widehat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}\|_2 = O_P(n^{-1/2+\alpha}T^{-1/2})$ for $\alpha \in [0, 1/2)$, let $\widetilde{\mathbf{Y}} = \mathbf{Y} - \mathbb{Z}(\mathbf{I}_T \otimes \widehat{\boldsymbol{\beta}}^0)$ and $\widetilde{\mathbf{U}} = \mathbf{U} + \mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^0)\}$. In general, such $\widehat{\boldsymbol{\beta}}^0$ exists as discussed in Section B.2.5 in the supplementary file. Thus, (3.1.1), or equivalently (3.2.2), can be expressed as

$$\widetilde{\mathbf{Y}} = \mathbf{G}\mathbf{F}' + \widetilde{\mathbf{U}}. \tag{3.2.3}$$

A naive approach is to estimate $\mathbf{G}$ and $\mathbf{F}$ via principal component analysis (PCA). That is, the columns of $\mathbf{F}/\sqrt{T}$ are estimated using eigenvectors corresponding to the first K largest eigenvalues of the $T \times T$ matrix $\widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}$, 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 $\boldsymbol{g}$ in $\mathbf{G}$ or the smooth variation of $\{\widetilde{y}_{it}\}_{i=1}^n$ from (3.2.3) against $\boldsymbol{x}_i$ at each t. Fan et al. (2016) 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 error-less.

To begin with, let $\mathbb{H}$ be a linear space spanned by a sequence of orthonormal basis functions $\{\phi_0(x) \equiv 1, \phi_1(x), \phi_2(x), \dots \phi_J(x)\}$. For each $k = 1, \dots, K$, $i = 1, \dots, n$, and $\ell = 1, \dots, d$, 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 $\{b_{j,k\ell}\}_{j \leqslant J}$ are the coefficients and $R_{k\ell}$ is the approximation or projection error. Assume $Jd + 1 < n$ so that the coefficients are estimable. Denote, for each $k = 1, \dots, K$ and $i = 1, \dots, n$, $\boldsymbol{b}_k = (b_{0,k}, b_{1,k1}, \dots, b_{J,k1}, \dots, b_{1,kd}, \dots, b_{J,kd})'$, where $b_{0,k} = \sqrt{J}\sum_{\ell=1}^d b_{0,k\ell}$, and
$\boldsymbol{\varphi}_i = (1/\sqrt{J}, \phi_1(x_{i1}), \dots, \phi_J(x_{i1}), \dots, \phi_1(x_{id}), \dots, \phi_J(x_{id}))'$. Then, it admits $g_k(\boldsymbol{x}_i) = \boldsymbol{\varphi}_i'\boldsymbol{b}_k +$

$\sum_{\ell=1}^{d} R_{k\ell}(x_{i\ell})$ and (3.2.3) can be rewritten as

$$\widetilde{\mathbf{Y}} = (\mathbf{\Phi B} + \mathbf{R})\mathbf{F}' + \widetilde{\mathbf{U}}, \tag{3.2.4}$$

where $\mathbf{\Phi} = (\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n)'$, $\mathbf{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K)$, and $\mathbf{R} = \{\sum_{\ell=1}^{d} R_{k\ell}(x_{i\ell})\}_{i=1,k=1}^{n,K}$. Then, we let $\mathbf{P} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'$ and apply the PCA procedure to projected data matrix $\mathbf{P}\widetilde{\mathbf{Y}}$. That is, we estimate $\widehat{\mathbf{F}}$ by letting the columns of $\widehat{\mathbf{F}}/\sqrt{T}$ be the eigenvectors corresponding to the first K largest eigenvalues of $\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}$ and estimate $\mathbf{G}$ by $\widehat{\mathbf{G}} = T^{-1}\mathbf{P}\widetilde{\mathbf{Y}}\widehat{\mathbf{F}}$. Moreover, $\mathbf{B}$ is estimated by $\widehat{\mathbf{B}} = (\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\widetilde{\mathbf{Y}}\widehat{\mathbf{F}}$.

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

First, consider an averaged version of (3.2.2) over t, $\bar{\boldsymbol{y}} = \mathbb{Z}_0'\boldsymbol{\beta} + \mathbf{G}T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t + T^{-1}\sum_{t=1}^{T}\boldsymbol{u}_t$, 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 3.2.1 implies that the variance of $n \times 1$ vector $\bar{\boldsymbol{y}}$ is

$$\mathbf{V} = \mathbf{G}\,\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t\right)\mathbf{G}' + \boldsymbol{\mathcal{D}}, \tag{3.2.5}$$

where the $n \times n$ diagonal matrix $\boldsymbol{\mathcal{D}}$ has diagonal entries $\mathrm{Var}(T^{-1}\sum_{t=1}^{T}u_{1t}), \ldots, \mathrm{Var}(T^{-1}\sum_{t=1}^{T}u_{nt})$. Then, (3.2.5) naturally leads to the oracle GLS-type estimate of $\boldsymbol{\beta}$,

$$\widetilde{\boldsymbol{\beta}} = \left(\mathbb{Z}_0'\mathbf{V}^{-1}\mathbb{Z}_0\right)^{-1}\mathbb{Z}_0'\mathbf{V}^{-1}\bar{\boldsymbol{y}}. \tag{3.2.6}$$

With the full knowledge on $\mathbf{G}$ and $\mathbf{F}$ in (3.2.2), $\mathbf{V}$ in (3.2.6) can be estimated as following. Let $\bar{\boldsymbol{f}} = T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t$, it is known that $\mathrm{Var}(T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t) = T^{-2}\sum_{t=-T+1}^{T-1}(T-|t|)\mathbf{\Sigma}_f(t)$, where $\mathbf{\Sigma}_f(s) = \mathrm{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_{t+s})$ and $\mathbf{\Sigma}_f(-s) = \mathrm{Cov}(\boldsymbol{f}_{t-s}, \boldsymbol{f}_t)$ can be estimated by $\widehat{\mathbf{\Sigma}}_f(s) = (T-s)^{-1}\sum_{t=1}^{T-s}(\boldsymbol{f}_t - \bar{\boldsymbol{f}})(\boldsymbol{f}_{t+s} - \bar{\boldsymbol{f}})'$ and $\widehat{\mathbf{\Sigma}}_f(-s) = (T-s)^{-1}\sum_{t=s}^{T}(\boldsymbol{f}_{t-s} - \bar{\boldsymbol{f}})(\boldsymbol{f}_t - \bar{\boldsymbol{f}})'$ for $s \geqslant 0$, respectively. Hence, operator

$$\mathcal{V}(\boldsymbol{f}_t) = \frac{1}{T^2}\sum_{t=-T+1}^{T-1}(T-|t|)\widehat{\mathbf{\Sigma}}_f(t) \tag{3.2.7}$$

defines an estimator of $\mathrm{Var}(T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t)$ in (3.2.5). Similarly, we define $\mathcal{V}(u_{it})$ as the estimator of $\mathrm{Var}(T^{-1}\sum_{t=1}^{T}u_{it})$ for each $i = 1,\ldots,n$ and the $n \times n$ diagonal matrix with diagonals $\mathcal{V}(u_{1t}),\ldots,\mathcal{V}(u_{nt})$ provides estimate of $\boldsymbol{\mathcal{D}}$ in (3.2.5).

The oracle GLS estimator $\widetilde{\boldsymbol{\beta}}$ is not accessible as it depends on the full knowledge on $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$. Motivated by $\widetilde{\boldsymbol{\beta}}$, an improved estimate for $\boldsymbol{\beta}$ can be obtained by replacing $\mathbf{G}$ and $\mathbf{F}$ with $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{F}}$ in (3.2.6), respectively. Specifically, with $\widehat{\mathbf{F}}$ from the first-stage, we can further approximate $\mathcal{V}(\boldsymbol{f}_t)$ and $\mathcal{V}(u_{it})$ in (3.2.5) by $\mathcal{V}(\widehat{\boldsymbol{f}}_t)$ and $\mathcal{V}(\widehat{u}_{it})$ respectively, where $\mathcal{V}(\cdot)$ is defined in (3.2.7), $\widehat{\boldsymbol{f}}_t$ is the tth row of $\widehat{\mathbf{F}}$, and $\widehat{\boldsymbol{u}}_t$ is the tth column of corresponding $\widehat{\mathbf{U}} = \widetilde{\mathbf{Y}} - \widehat{\mathbf{G}}\widehat{\mathbf{F}}'$. Then, we define the estimator of $\mathbf{V}$ by

$$\widehat{\mathbf{V}} = \widehat{\mathbf{G}}\mathcal{V}(\widehat{\boldsymbol{f}}_t)\widehat{\mathbf{G}}' + \widehat{\boldsymbol{\mathcal{D}}}, \tag{3.2.8}$$

where $\widehat{\boldsymbol{\mathcal{D}}}$ is the $n \times n$ diagonal matrix with diagonals $\mathcal{V}(\widehat{u}_{1t}),\ldots,\mathcal{V}(\widehat{u}_{nt})$ and arrive at the **T**w**O**-stage **P**rojection-based **E**stimator (**TOPE**) of $\boldsymbol{\beta}$

$$\bar{\boldsymbol{\beta}} = \left(\mathbb{Z}_0'\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0\right)^{-1}\mathbb{Z}_0'\widehat{\mathbf{V}}^{-1}\bar{\boldsymbol{y}}. \tag{3.2.9}$$

The detailed computation is summarized in Algorithm 1.

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

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

Procedure:

1: For a given preliminary estimator $\widehat{\boldsymbol{\beta}}^0$, 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 $\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}$ and estimate $\mathbf{G}$ by $\widehat{\mathbf{G}} = \mathbf{P}\widetilde{\mathbf{Y}}\widehat{\mathbf{F}}/T$.

3: Second-stage: compute $\widehat{\mathbf{V}} = \widehat{\mathbf{G}}\mathcal{V}(\widehat{\boldsymbol{f}}_t)\widehat{\mathbf{G}}' + \widehat{\boldsymbol{\mathcal{D}}}$ as in (3.2.8), 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** in (3.2.9).

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

Alternative to TOPE, one can first project $\mathbf{Y}$ using $(\mathbf{I}_n - \mathbf{P})$, where $\mathbf{P} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}$. Then, (3.1.1) or (3.2.2) leads to $(\mathbf{I}_n - \mathbf{P})\mathbf{Y} = (\mathbf{I}_n - \mathbf{P})\mathbb{Z}(\mathbf{I}_T \otimes \boldsymbol{\beta}) + \mathbf{R}\mathbf{F}' + (\mathbf{I}_n - \mathbf{P})\mathbf{U}$, and $\boldsymbol{\beta}$ can be directly estimated via OLS. This is similar to the procedure of profile likelihood (Fan et al.,

2005) or restricted maximum likelihood (Jiang et al., 1996). However, the validity of this approach relies on the assumption that $\mathbb{Z}$ and $\mathbf{\Phi}$ are linearly independent, which is more restricted than that of TOPE. Another seemingly straightforward approach is to project $\mathbf{Y}$ using $(\mathbf{I}_n - \widetilde{\mathbf{P}}_{\mathbf{Z}})$ where $\widetilde{\mathbf{P}}_{\mathbf{Z}} = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'$ and perform PCA on $(\mathbf{I}_n - \widetilde{\mathbf{P}}_{\mathbf{Z}})\mathbf{Y} \approx (\mathbf{I}_n - \widetilde{\mathbf{P}}_{\mathbf{Z}})\mathbf{G}\mathbf{F}'$ to estimate the loading and latent process. Though such an estimate of $\mathbf{F}$ remains consistent, as also noted by Wang et al. (2017), this approach only identify the part of the latent structure that is orthogonal to $\mathbb{Z}$. That is, one can only obtain a consistent estimate of $(\mathbf{I}_n - \widetilde{\mathbf{P}}_{\mathbf{Z}})\mathbf{G}$, and in particular $\widehat{\mathbf{G}} + \widetilde{\mathbf{P}}_{\mathbf{Z}}\mathbf{A}$ is also a valid estimator for any $n \times K$ matrix $\mathbf{A}$.

3.3 Theoretical properties

We first collect some notation throughout the remaining sections. For vector $\boldsymbol{a} = (a_1, \ldots, a_p)' \in \mathbb{R}^p$, its ℓ_q-norm is defined by $||\boldsymbol{a}||_q = (\sum_{j=1}^p |a_j|^q)^{1/q}$ with $1 \leqslant q < \infty$. For a matrix $\mathbf{M} = (m_{ij})_{1 \leqslant i,j \leqslant p} \in \mathbb{R}^{p \times p}$, write $\|\mathbf{M}\|_{\max} = \max_{i,j} |m_{ij}|$ to be the maximum norm and $||\mathbf{M}||_{\mathbb{F}} = (\sum_{i=1}^p \sum_{j=1}^p m_{ij}^2)^{1/2}$ to be the Frobenius norm. The spectral norm of matrix $\mathbf{M}$ corresponds to its largest singular value, defined as $||\mathbf{M}||_2 = \sup_{\boldsymbol{a} \in S} ||\mathbf{M}\boldsymbol{a}||_2$, where $S = \{\boldsymbol{a} \in \mathbb{R}^p : ||\boldsymbol{a}||_2 = 1\}$. We write $\mathbf{I}$ for an identify 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 $a_n \leqslant C b_n$ for some C independent of n; and $a_n \asymp b_n$ if and only if there exists C, D independent on n such that $C|b_n| \leqslant |a_n| \leqslant D|b_n|$. Denote $\xrightarrow{p}$ and $\xrightarrow{d}$ as the convergence in probability and in distribution, respectively. Unless specified otherwise, $\delta > 0$ and $C > 0$ denote generic constants independent of n, T, p.

3.3.1 Preliminaries

We impose the following conditions on our model, in addition to Condition 3.2.1.

Condition 3.3.1. *With probability at least* $1-\delta$, $1 - n^{-1}\log(1/\delta) \lesssim \lambda_{\min}(n^{-1}\mathbf{G}'\mathbf{G}) \leqslant \lambda_{\max}(n^{-1}\mathbf{G}'\mathbf{G}) \lesssim 1 + n^{-1}\log(1/\delta)$.

Condition 3.3.2. *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 3.3.3. *(Accuracy of the sieve approximation)*

(i) For each $\ell = 1, \ldots, d$, $k = 1, \ldots, K$, the loading function $g_{k\ell}(\cdot)$ belongs to a Hölder class $\mathcal{G} = \{g : |g^{(r)}(s) - g^{(r)}(t)| \leqslant L|s-t|^{\alpha}\}$ for some $L > 0$.

(ii) For $\kappa = 2(r+\alpha) \geqslant 4$, $\sup_{x \in \mathcal{X}} \left| g_{k\ell}(x) - \sum_{j=1}^{J} b_{k,j\ell}\phi_j(x) \right|^2 \lesssim J^{-\kappa}$.

(iii) It admits $\max_{k,j,\ell} b_{k,j\ell}^2 < \infty$.

Condition 3.3.1 is similar to the pervasive condition on loading matrix in the traditional factor model (Stock and Watson, 2002a). Since $\mathbf{G}\mathbf{G}'$ and $\mathbf{G}'\mathbf{G}$ have their first K largest eigenvalues in common, the K largest eigenvalues of $\mathbf{G}'\mathbf{G}$ also diverges in n. This condition ensures that $\boldsymbol{x}_i$ has non-vanishing explaining power on loading so that $\mathbf{G}'\mathbf{G}$ has spiked eigenvalues. Condition 3.3.2 is standard in the literature of nonparametric and semiparametric statistics (Huang et al., 2004; Liang et al., 2009; Stone, 1985). Here, our model allows $\boldsymbol{x}_i$ to be dependent across subjects and non-stationary in i. The accuracy of sieve approximation in Condition 3.3.3 can be obtained by common basis like polynomial or B-splines (Chen, 2007; Fan et al., 2016; Lorentz, 1966).

Condition 3.3.4. *For $\mathbb{Z}_0 = T^{-1}\sum_{t=1}^{T} \mathbf{Z}_t$, almost surely, we have*

(i) for each n and T, eigenvalues of $n^{-1}\mathbb{Z}_0'\mathbb{Z}_0$ are bounded away from 0 and infinity;

(ii) $\|\mathbf{P}_Z\mathbf{G}\|_{\mathbb{F}} = O(n^{\alpha})$ for each n and T and some $\alpha \in [0, 1/2)$, where $\mathbf{P}_Z$ is the projection matrix on $\mathbb{Z}_0$.

Condition 3.3.4 (i) is similar to the standard condition on the design matrix in linear model that $\mathbb{Z}_0'\mathbb{Z}_0/n$ converges in n. Similar to conditions for semiparametric models in Robinson (1988), (ii) guarantees identifications between the parametric and nonparametric parts in our model. Particularly, it allows dependence between $\boldsymbol{z}_{it}$ and $\boldsymbol{x}_i$. If $\boldsymbol{x}_i$ and $\mathbb{Z}$ are dependent, as mentioned in Section 3.2.1, function class $\mathcal{G}$ in Condition 3.3.3 must exclude linear functions, where (ii) is the empirical

characterization of such an exclusion condition. Overall, Condition 3.3.4 guarantees the existence of the consistent preliminary estimator $\widehat{\boldsymbol{\beta}}^0$ in the first stage of TOPE.

At last, we impose some widely-used conditions (Bai, 2003; Stock and Watson, 2002a) 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 $\{(\boldsymbol{f}_t, \boldsymbol{u}_t) : t \leqslant 0\}$ and $\{(\boldsymbol{f}_t, \boldsymbol{u}_t) : t \geqslant T\}$, and recall the α-mixing coefficient as $\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(A \cap B)|$.

Condition 3.3.5. *(Serial dependence, stationarity, and tail behavior)*

(i) $\{\boldsymbol{u}_t, \boldsymbol{f}_t\}_{t \leqslant T}$ *are strictly stationary with zero mean and finite long run variances.*

(ii) There exist $r_1, C_1 > 0$ *such that for all* $T > 0$, $\alpha(T) < \exp(-C_1 T^{r_1})$.

(iii) Exponential tail: 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* $s > 0$, $\mathbb{P}(|u_{it}| > s) \leqslant \exp\{-(s/b_1)^{r_2}\}$ *and* $\mathbb{P}(|f_{tk}| > s) \leqslant \exp\{-(s/b_2)^{r_3}\}$.

3.3.2 Statistical guarantees

To establish the statistical guarantees of TOPE for (3.2.2) as well as (3.1.1), we carry out a non-asymptotic analysis of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{G}}$ first, then derive a non-asymptotic bound of the error for estimating $\boldsymbol{\beta}$ using $\bar{\boldsymbol{\beta}}$. Following this, we obtain a non-asymptotic result of $\text{Var}(\bar{\boldsymbol{\beta}})$ in comparison to that of GLS estimator $\widetilde{\boldsymbol{\beta}}$ to study the efficiency of TOPE.

Theorem 3.3.1. *Suppose that Conditions 3.2.1, 3.2.2, and 3.3.1-3.3.5 hold, and assume* $J = o(n^{1-2\alpha})$. *With probability at least* $1 - \delta$, *we have*

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

In contrast to the known asymptotic properties of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{G}}$ for traditional and semiparametric factor models with divergent n an T (Bai and Ng, 2013; Fan et al., 2016), Theorem 3.3.1 provides finite sample performance of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{G}}$. Given a finite p, the rates obtained in Theorem 3.3.1 agree with the asymptotic results in Fan et al. (2016) as expected. 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, the non-asymptotic results in Theorem 3.3.1 make it possible to establish the following finite sample results on both $\bar{\boldsymbol{\beta}}$ and its variance-covariance matrix with respect to GLS estimator $\widetilde{\boldsymbol{\beta}}$ as defined in (3.2.6).

Theorem 3.3.2. *Under conditions in Theorem 3.3.1, with probability at least* $1-\delta$*,*

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

where $\widetilde{\boldsymbol{\beta}}$ *is the GLS estimator of* $\boldsymbol{\beta}$ *with full knowledge of* $\mathbf{G}$ *and* $\mathbf{F}$ *as in Section 3.2.2. 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 3.3.2 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 $\|\mathbf{A}\|_{\mathbf{S}} := n^{-1/2}\|\mathbf{S}^{-1/2}\mathbf{A}\mathbf{S}^{-1/2}\|_2$ and define a class of estimator to $\boldsymbol{\beta}$ with respect to working covariance $\mathbf{V}_\zeta$ by $\Theta_\zeta = \{\check{\boldsymbol{\beta}}_\zeta = (\mathbb{Z}_0'\mathbf{V}_\zeta^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\mathbf{V}_\zeta^{-1}\bar{\boldsymbol{y}} : \|\mathbf{V}_\zeta - \mathbf{V}\|_{\mathbf{V}} \lesssim \zeta\}$, where GLS estimator $\widetilde{\boldsymbol{\beta}} \in \Theta_0$, the TOPE $\bar{\boldsymbol{\beta}} \in \Theta_{\vartheta_{n,T,J}}$ by Theorem 3.3.1, and OLS estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} \in \Theta_{n^\alpha}$ by Proposition B.2.5 in the supplementary file. From the proof of Theorem 3.3.2, $\|\check{\boldsymbol{\beta}}_\zeta - \widetilde{\boldsymbol{\beta}}\|_2 = O_p(n^{-1/2}T^{-1/2}\zeta)$ 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 $\|\check{\boldsymbol{\beta}}_\zeta - \widetilde{\boldsymbol{\beta}}\|_2 = o_p(n^{-1/2}T^{-1/2})$

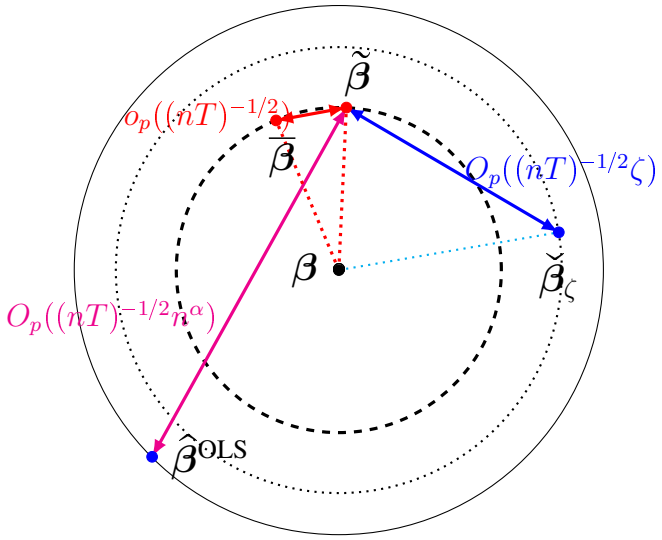

Figure 3.1: A schematic about different estimators to $\boldsymbol{\beta}$ in (3.1.1), where $\overline{\boldsymbol{\beta}}$ is the TOPE estimator and $\widetilde{\boldsymbol{\beta}}$ is the oracle GLS estimator with full knowledge on $\mathbf{G}$ and $\mathbf{F}$.

if $\zeta = o(1)$. With heteroscedasticity across subjects and/or autocorrelation, GLS is known to be efficient in general (Baltagi, 2008; Greene, 2003). Particularly, for (3.2.2), 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 $t = 1-T, \ldots, T-1$. Therefore, Theorem 3.3.2 implies that the 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 Figure 3.1.

As a final remark, the following theorem establishes results analogous to Theorem 3.3.1 in the max norm and shares common observations with Wang and Fan (2017) and Barigozzi et al. (2018).

Theorem 3.3.3. *For model* (3.2.2)*, under the same conditions of Theorem 3.3.1, with probability at least* $1-\delta$*,*

$$\begin{aligned}
\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\max} &\lesssim \left(\frac{1}{\sqrt{n}} + \frac{p^2}{\sqrt{n^{1-\alpha}T}} + \frac{1}{J^{\kappa/2}}\right)\{\log(T)\}^{2/r_3}\log(1/\delta), \\
\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\max} &\lesssim \left\{\sqrt{\frac{(T+p^2n^\alpha)\log(n)}{T^2}} + \frac{1}{\sqrt{n}} + \frac{p^2}{\sqrt{n^{1-\alpha}T}} + \frac{1}{J^{\kappa/2}}\right\}\log(1/\delta), \\
\|\widehat{\mathbf{B}} - \mathbf{B}\|_{\max} &\lesssim \left\{\sqrt{\frac{(T+p^2n^\alpha)\log(n)}{nT^2}} + \frac{1}{\sqrt{n}} + \frac{p^2}{\sqrt{n^{1-\alpha}T}} + \frac{1}{J^{\kappa/2}}\right\}\log(1/\delta).
\end{aligned}$$

3.4 TOPE-based inference

The following theorem paves a way for drawing inference on $\boldsymbol{\beta}$ based on the TOPE. In general, the expectation with respect to $\{\mathbf{Z}_t, \boldsymbol{x}_i\}$ is unknown as their joint distribution is not accessible. To draw inference about $\boldsymbol{\beta}$ in practice, (ii) below establishes the asymptotic distribution of $\bar{\boldsymbol{\beta}}$ conditional on $\{\mathbf{Z}_t, \boldsymbol{x}_i\}$.

Theorem 3.4.1. *Under Conditions 3.2.1 and 3.3.1-3.3.5 and $J = o(\sqrt{n})$, we have*

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

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

Replacing $\mathbf{V}$ in Theorem 3.4.1 (ii) by $\widehat{\mathbf{V}} = \widehat{\mathbf{G}}\mathcal{V}(\widehat{\boldsymbol{f}}_t)\widehat{\mathbf{G}}' + \widehat{\boldsymbol{\mathcal{D}}}$ from (3.2.8), for any estimable $\mathbf{C}\boldsymbol{\beta}$ with $q \times p$ matrix $\mathbf{C}$ and $q < p$, a $100(1-\alpha)\%$ confidence set is given by

$$\mathrm{CS}_{\mathbf{C}} = \left\{\mathbf{C}\boldsymbol{\beta} : (\mathbf{C}\boldsymbol{\beta} - \mathbf{C}\bar{\boldsymbol{\beta}})'\{\mathbf{C}(\mathbb{Z}_0'\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}\mathbf{C}'\}^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{C}\bar{\boldsymbol{\beta}}) < \chi^2_{q,1-\alpha}\right\} \tag{3.4.1}$$

where $\chi^2_{q,1-\alpha}$ is the $100(1-\alpha)\%$ quantile of χ^2_q distribution. Furthermore, for each $\ell = 1, \ldots, p$, denote $\widehat{\sigma}^2_\ell$ the ℓth diagonal entry of $(\mathbb{Z}_0'\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}$, a $100(1-\alpha)\%$ confidence interval for the ℓth entry of $\boldsymbol{\beta}$, β_ℓ, is given by

$$\mathrm{CI}_\ell = [\widehat{\beta}_\ell - \widehat{\sigma}_\ell\Phi^{-1}(1-\alpha/2), \widehat{\beta}_\ell + \widehat{\sigma}_\ell\Phi^{-1}(1-\alpha/2)]. \tag{3.4.2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal.

When rows of $\mathbf{C}$ are the natural basis of $\mathbb{R}^p$, (3.4.1) provides a confidence set of a subset of $\boldsymbol{\beta}$. To draw inference on individual entries of $\boldsymbol{\beta}$, (3.4.2) provides a natural alternative to (3.4.1). For a subset of $\boldsymbol{\beta}$ with multiple components, correction such as the Bonferroni procedure can be applied to (3.4.2) to control family-wise error rate. Moreover, Theorem 3.4.1 implies that $\mathbb{P}\left(\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > \varepsilon\right) < p\exp\left(-\varepsilon^2 p^{-1}\sigma^{-2}\right)$, where σ^2 can be estimated by the minimum diagonal of $(\mathbb{Z}_0'\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}$. A uniform confidence set for $\boldsymbol{\beta}$ at level $100(1-\alpha)\%$ is given by $\mathrm{CI}' = \{\boldsymbol{\beta} : |\beta_\ell - \widehat{\beta}_\ell| \leqslant \widehat{\sigma}\sqrt{p\log(p/\alpha)},\ l = 1, \ldots, p\}$.

To draw inference on the explaining power of covariates $\boldsymbol{x}_i$ on the dependence structure of data, Fan et al. (2016) proposed a semiparametric specification testing statistic $\mathrm{S}_G = \mathrm{tr}\{(\widetilde{\mathbf{F}}'\widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}\widetilde{\mathbf{F}})^{-1}\widetilde{\mathbf{F}}'\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}\widetilde{\mathbf{F}}\}$, where columns of $\widetilde{\mathbf{F}}/\sqrt{T}$ are the eigenvectors corresponding to the K largest eigenvalues of $\widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}$. In addition to Conditions 3.3.1-3.3.5, assuming $T^{2/3} = o(n)$, $n\{\log(n)\}^4 = o(T^2)$, $J = o(\min\{\sqrt{n}, \sqrt{T}\})$, 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. Thus, S_G provides a diagnostic tool for the proposed model (3.1.1) or (3.2.2).

3.5 Determining the number of factors K

In our model, the dimension of latent process $\boldsymbol{f}_t$ or the number of loading functions $g_1(\boldsymbol{x}_i), \ldots, g_K(\boldsymbol{x}_i)$, K is unknown in practice and needs to be estimated. Once a consistent estimator of K is obtained, all results achieved can be naturally carried over using a standard conditioning argument.

When the number of subjects n is much less than the number of time points T, subjective methods such as scree plot of eigenvalues, distribution-based test such as Bartlett's test, and computational intensive method such as cross-validation can be employed to determine K (Jolliffe, 2002). When $n \gg T$, relying on the fact that the K largest eigenvalues of the sample covariance matrix grow fast as n increases and others remain slowly growing or bounded, the eigenvalue ratio estimator/procedure has been widely used to provide consistent estimation of K. Specifically, Lam and Yao (2012) and Ahn and Horenstein (2013) proposed to select K corresponding to the largest ratio of the adjacent eigenvalues of $\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}'$. For model (3.1.1), it is naturally to work with $\mathbf{P}\widetilde{\mathbf{Y}}$ (Fan et al., 2016). In fact, as the non-vanishing eigenvalues of $\mathbf{P}\widetilde{\mathbf{Y}}(\mathbf{P}\widetilde{\mathbf{Y}})'$ and $(\mathbf{P}\widetilde{\mathbf{Y}})'\mathbf{P}\widetilde{\mathbf{Y}}$ are same, it suffices to focus on the eigenvalues of matrix $\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}$. That is, denote $\lambda_k(\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}})$ the k-th largest eigenvalue of $\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}$. Assuming $K < Jd/2$, which can be achieved by increasing J, the eigenvalue ratio procedure selects K as

$$\widehat{K} = \mathrm{argmax}_{0<k<Jd/2} \frac{\lambda_k(\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}})}{\lambda_{k+1}(\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}})}.$$

Under Conditions 3.2.1, 3.3.1-3.3.5, and Assumption 6.1 in Fan et al. (2016), as n and T go to infinity, it is known that $\mathbb{P}(\widehat{K} = K) \to 1$ if $J = o(\min\{\sqrt{n}, \sqrt{T}\})$ and $K < Jd/2$. Hence, the eigenvalue ratio procedure provides a consistent estimator on K.

3.5.1 A high-dimensional white noise (HDWN) testing-based procedure

In this section, motivated from the recent development on testing high-dimensional white noise (Chang et al., 2017; Li et al., 2019), we propose a different procedure to determine K compared to the eigenvalue ratio method. To motivate our method, we first consider a detrended model that $\boldsymbol{y}_t = \mathbf{G}\boldsymbol{f}_t + \boldsymbol{u}_t$ with independent process $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$, where $\boldsymbol{u}_t$ is white noise but not $\boldsymbol{f}_t$. Then, $(\mathbf{I}_n - \mathbf{P})\boldsymbol{y}_t$ is a white noise if and only if $\mathbf{PG} = \mathbf{G}$. That is, any projection $\mathbf{P}$ making $(\mathbf{I}_n - \mathbf{P})\boldsymbol{y}_t$ white noise must admit rank greater than or equal to $n - K$. Therefore, the determination of K can be achieved via sequentially testing multivariate, potentially high-dimensional, white noise with respect to $\mathbf{P}$'s.

Motivated from above, for the proposed model (3.1.1) or (3.2.2), denote $\widetilde{\mathbf{V}}$ and $\widehat{\boldsymbol{\beta}}$ some consistent estimators of $\mathbf{V}$ and $\boldsymbol{\beta}$, respectively, such as thresholding estimator (Bickel and Levina, 2008b) or POET-estimator (Fan et al., 2013) and OLS $\widehat{\boldsymbol{\beta}}^{\text{OLS}}$. Given each $K_0 \geqslant 1$, denote $\widehat{\boldsymbol{\gamma}}_i$ the eigenvector associate to the i-th largest eigenvalue of $\widetilde{\mathbf{V}}$ and consider projection

$$\mathbf{P}_{\widetilde{\mathbf{V}}} = \mathbf{I}_n - (\widehat{\boldsymbol{\gamma}}_{K_0+1}, \ldots, \widehat{\boldsymbol{\gamma}}_n)'.$$

Then, let $\widetilde{\boldsymbol{w}}_t = (\mathbf{I}_n - \mathbf{P}_{\widetilde{\mathbf{V}}})\widetilde{\boldsymbol{y}}_t$ for $t = 1, \ldots, T$, where $\widetilde{\boldsymbol{y}}_t = \boldsymbol{y}_t - \mathbf{Z}_t\widehat{\boldsymbol{\beta}}$. For the to-be-determined dimension K of factor process $\boldsymbol{f}_t$ in (3.1.1) or (3.2.2) and a prescribed K_0, testing

$$H_0(K_0) : K \leqslant K_0 \text{ v.s. } H_1(K_0) : K > K_0$$

is therefore equivalent to test

$$\widetilde{H}_0(K_0) : \widetilde{\boldsymbol{w}}_t \text{ is white noise v.s. } \widetilde{H}_1(K_0) : \text{not } \widetilde{H}_0(K_0). \tag{3.5.1}$$

Therefore, we will test (3.5.1) for each $K_0 \geqslant 1$ and let

$$\widehat{K} = \arg\min\{0 < K_0 < \min(n, T) : \text{ such that } \widetilde{H}_0(K_0) \text{ fails to be rejected}\}.$$

To that end, we employed the test by Chang et al. (2017). For each $s = 1, \ldots, S$ with prescribed integer $S > 1$, denote $\widehat{\boldsymbol{\Sigma}}(s) = T^{-1} \sum_{t=1}^{T-s} \widetilde{\boldsymbol{w}}_{t+s} \widetilde{\boldsymbol{w}}_t'$ and $\widehat{\rho}_{ij}(s)$ the entry of $\mathrm{diag}\{\widehat{\boldsymbol{\Sigma}}(0)\}^{-1/2} \widehat{\boldsymbol{\Sigma}}(s)$ $\mathrm{diag}\{\widehat{\boldsymbol{\Sigma}}(0)\}^{-1/2}$. Consider testing statistic

$$\zeta_T = \sqrt{T} \max_{1 \leqslant s \leqslant S} \max_{1 \leqslant i,j \leqslant n-K_0} |\widehat{\rho}_{ij}(s)|.$$

To specify the critical value cv_α that $\mathbb{P}(\zeta_T > \mathrm{cv}_\alpha) = \alpha$, similar to Chang et al. (2017), we first show the following result on normal approximation.

Theorem 3.5.1. *Under Conditions 3.2.1 and 3.3.5, assume that* $\log(n) \lesssim T^{\iota_1}$ *for some* $\iota_1 > 0$. *Then there exists* $\boldsymbol{\mathcal{G}} \sim N(\mathbf{0}, \boldsymbol{\Xi}_T)$ *with* $\boldsymbol{\Xi}_T = (\mathbf{I}_S \otimes \boldsymbol{\Gamma}) \mathbb{E}(\boldsymbol{\xi}_T \boldsymbol{\xi}_T')(\mathbf{I}_S \otimes \boldsymbol{\Gamma})$, *where* $\boldsymbol{\Gamma} = \mathrm{diag}\{\boldsymbol{\Sigma}(0)\}^{-1/2} \otimes$ $\mathrm{diag}\{\boldsymbol{\Sigma}(0)\}^{-1/2}$ *and* $\boldsymbol{\xi}_T = \sqrt{T}[vec\{\widehat{\boldsymbol{\Sigma}}(1)\}', \ldots, vec\{\widehat{\boldsymbol{\Sigma}}(S)\}']'$, *such that under* $\widetilde{H}_0(K_0)$,

$$\sup_{K_0 \leqslant K} \sup_{s \geqslant 0} |\mathbb{P}(\zeta_T > s) - \mathbb{P}(|\boldsymbol{\mathcal{G}}|_\infty > s)| \to 0 \text{ as } T \to \infty.$$

By Theorem 3.5.1 and Lemma 3.1 in Chernozhukov et al. (2013), cv_α can be empirically determined by Monte Carlo samples drawn from $N(\mathbf{0}, \widehat{\boldsymbol{\Xi}}_T)$, where $\widehat{\boldsymbol{\Xi}}_T$ is some estimate of $\boldsymbol{\Xi}_T$ (Andrews, 1991; Chang et al., 2017). Then, we will reject $\widetilde{H}_0(K_0)$ whenever ζ_T exceeds such a $\widehat{\mathrm{cv}}_\alpha$. The following theorem from Chang et al. (2017) guarantees the validity of this test for (3.5.1) with each prescribed K_0.

Algorithm 1: HDWN Testing-Based Procedure for Selecting K

Input: Observations $\{(y_{it}, \mathbf{Z}_{it})\}_{i=1,t=1}^{n,T}$, $\widetilde{\mathbf{V}}$ and $\widehat{\boldsymbol{\beta}}$ as the estimators to $\mathbf{V}$ and $\boldsymbol{\beta}$ discussed in Section 3.5.1, lag S, and $\alpha_n = Cn^{-\iota}$ for constants $C > 0$ and $\iota > 0$.

WN testing statistic

1: Detrend data by $\widetilde{\boldsymbol{y}}_t := \boldsymbol{y}_t - \mathbf{Z}_t\widehat{\boldsymbol{\beta}}$ for $t = 1, \ldots, T$.

2: For each $K_0 \geqslant 1, 1 \leqslant s \leqslant S$, compute $\{\widehat{\rho}_{ij}(s)\}_{i,j=1}^{n-K_0}$.

3: Calculate the WN testing statistic $\zeta_T = \sqrt{T}\max_{1\leqslant s\leqslant S}\max_{1\leqslant i,j\leqslant n-K_0}|\widehat{\rho}_{ij}(s)|$.

Critical value

1: Let $\mathbf{W}_t = (\text{vec}(\widetilde{\boldsymbol{w}}_{t+1}\widetilde{\boldsymbol{w}}_t')', \ldots, \text{vec}(\widetilde{\boldsymbol{w}}_{t+S}\widetilde{\boldsymbol{w}}_t')')'$ for $t = 1, \ldots, T-S$.

2: For $\mathcal{K}(x) = 25(12\pi^2x^2)^{-1}\{(6\pi x/5)^{-1}\sin(6\pi x/5) - \cos(6\pi x/5)\}$ with $\mathcal{K}(0) = 1$ and data-driven bandwidth b_T (Andrews, 1991), calculate

$$\widehat{\mathbf{S}}(s) = \sum_{t=|s|+1}^{T-S}(T-S)^{-1}\mathbf{W}_t\mathbf{W}_{t-s}', \ \widehat{\mathbf{J}}_T = \sum_{t=-T+S+1}^{T-S-1}\mathcal{K}(t/b_T)\widehat{\mathbf{S}}(t).$$

3: Compute $\widehat{\boldsymbol{\Xi}}_T = (\mathbf{I}_S \otimes \widehat{\boldsymbol{\Gamma}})\widehat{\mathbf{J}}_T(\mathbf{I}_S \otimes \widehat{\boldsymbol{\Gamma}})$ where $\widehat{\boldsymbol{\Gamma}} = \text{diag}\{\widehat{\boldsymbol{\Sigma}}(0)\}^{-1/2} \otimes \text{diag}\{\widehat{\boldsymbol{\Sigma}}(0)\}^{-1/2}$ and $\widehat{\boldsymbol{\Sigma}}(s) = \sum_{t=1}^{T-s}\widetilde{\boldsymbol{w}}_{t+s}\widetilde{\boldsymbol{w}}_t'/T$ for each s.

4: Generate $\boldsymbol{\mathcal{G}}_1, \ldots, \boldsymbol{\mathcal{G}}_B$ from $N(\mathbf{0}, \widehat{\boldsymbol{\Xi}}_T)$ and compute $\widehat{\text{cv}}_{\alpha_n}$ as the $[B\alpha_n]$th largest value among $|\boldsymbol{\mathcal{G}}_1|_\infty, \ldots, |\boldsymbol{\mathcal{G}}_B|_\infty$.

Selection of K

1: For each $K_0 \geqslant 1$, perform the WN test, that is rejecting $H_0(K_0)$ if $\zeta_T > \widehat{\text{cv}}_{\alpha_n}$.

2: Stop at the first K_0 that $H_0(K_0)$ fails to be rejected and let $\widehat{K} = K_0$.

Output: Estimated $\widehat{K}$.

Theorem 3.5.2. *Consider the estimator of $\boldsymbol{\Xi}_T$ in the following algorithm with $|\mathcal{K}(x)| \asymp |x|^{-\tau}$ as $|x| \to \infty$ for $\tau > 1$ and $b_T \asymp T^\rho$ for $0 < \rho < \min\{(\tau-1)/(3\tau), r_1/(2r_1+1)\}$. Under Conditions*

3.2.1 and 3.3.5, assume that $\log(n) \lesssim T^{\iota_2}$ *with* $\iota_2 > 0$*. Let* α_n *be a sequence approaching zero (e.g.* $\alpha_n = O(n^{-\iota_3})$ *for* $\iota_3 > 0$*), we have*

(1) $\mathbb{P}(\zeta_T > \widehat{\mathrm{cv}}_{\alpha_n}) \to \alpha_n$ *under* $\widetilde{H}_0(K_0)$ *as* $T \to \infty$*, and*

(2) assume that $\max_{1\leqslant s\leqslant S} \max_{1\leqslant i,j\leqslant n} |\rho_{ij}(s)| > \eta^{1/2}(1+\epsilon_T)T^{-1/2}\lambda(n,\alpha_n)$*, where* η *is the maximum diagonal of* $\mathbf{\Xi}_T$*,* $\lambda(n,\alpha_n) = \{2\log(n^2 S)\}^{1/2} + \{2\log(1/\alpha_n)\}^{1/2}$*, and* ϵ_T *satisfies that* $\epsilon_T \to 0$ *and* $\epsilon_T^2 \log n \to \infty$*, then* $\mathbb{P}(\zeta_T > \widehat{cv}_{\alpha_n}) \to 1$ *under* $\widetilde{H}_1(K_0)$ *as* $T \to \infty$*.*

Then, the following theorem establishes the validity of the proposed HDWN testing-based procedure for selecting K.

Theorem 3.5.3. *Under conditions in Theorems 3.5.1 and 3.5.2,* $\inf_n \mathbb{P}(\widehat{K} = K) \to 1$ *as* $T \to \infty$*.*

To conclude, we provide some remarks here. The K largest eigenvalues of $\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}$ diverge in rate n while the rest remains in constant. However, compared to the discrepancy among divergent eigenvalues, the gaps between the divergent and non-divergent eigenvalues are not necessarily large for finite sample when n is only moderately large. This mimics the scenario for testing high-dimensional hypotheses with strong and sparse signals. As discussed in Chang et al. (2017), statistic ζ_T is particular powerful against such an alternative, which is also supported by empirical numerical studies in Section B.3.4. Finally, the procedure is summarized in Algorithm 1.

3.6 Numerical studies

3.6.1 Simulation settings

For model (3.1.1) or (3.2.2), we demonstrate the finite sample performance of TOPE for both estimation and inference in comparison to three competing methods: OLS estimator, which is computed by ignoring heteroscedasticity across subjects and dependence; GLS estimator, which is the traditional GLS estimator naively utilizing the first K components of $T^{-1}\sum_{t=1}^{T}(\boldsymbol{y}_t - \bar{\boldsymbol{y}})(\boldsymbol{y}_t - \bar{\boldsymbol{y}})'$ as $\widehat{\mathbf{V}}$; and last, oracle estimator, which is the TOPE with known $\mathbf{G}$ without using functional approximation. To implement the TOPE, we employ the OLS as the preliminary estimator $\widehat{\boldsymbol{\beta}}^0$.

The mean squared error (MSE) and the empirical coverage probability (ECP) of the confidence region for $\boldsymbol{\beta}$ are used to compare different procedures. 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 the TOPE. The empirical maximum marginal length of the confidence set (MML) is used to show the efficiency; that is, the confidence set with ECP agreeing to the nominal level and small MML is more preferable. For clear presentation, we display MML of different methods normalized by the largest one (the MML of OLS, in general).

We consider $n = 50, 100, 200, 500, 1000, 2000$ and $T = 20, 50, 100, 200, 500$; also, we set $p = 4$ with $\boldsymbol{\beta} = (1, 1, 1, 1)'$ and generate *i.i.d.* $z_{i\ell,t} \sim N(3\exp(t/30), 1)$ for each $i = 1, \ldots, n$, $\ell = 1, \ldots, p$, and $t = 1, \ldots, T$. A similar setting was used in Huang et al. (2004). 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 Fan et al. (2016), with the initial realization $\mathbf{G}_0$ for g_1, g_2 and g_3, we further compute $\mathbf{H}_{\mathbf{G}} = \mathbf{G}_0'\mathbf{G}_0$ and set $\mathbf{G} = \mathbf{G}_0\mathbf{H}_{\mathbf{G}}$ in simulations so that Condition 3.2.2 is satisfied.

The latent process $\boldsymbol{f}_t$ consists of $K = 3$ independent univariate time series generated from the same model. Specifically, we consider three dependence structures: independent in t, AR(1) with autoregressive coefficient $\rho = 0.5$, and ARMA$(1, 1)$ with autoregressive coefficient $\rho = 0.5$ and moving average coefficient $\theta = 0.5$. In addition, three innovations are considered, including the standard normal, centered χ_5^2, and t_8. 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 $(\chi_5^2 - 5)/10$. For each setting, 500 simulations are conducted.

3.6.2 Results of TOPE

Figure 3.2 displays the MSE with respect to the $\log(nT)$ on the logarithm scale when $T = 20$ and $\boldsymbol{f}_t$ are independent in t or follow ARMA$(1, 1)$ model with $\mathcal{N}(0, 1)$ or t_8 innovations. Additional simulation results are included in the supplementary file. In Figure 3.2, the MSEs of all

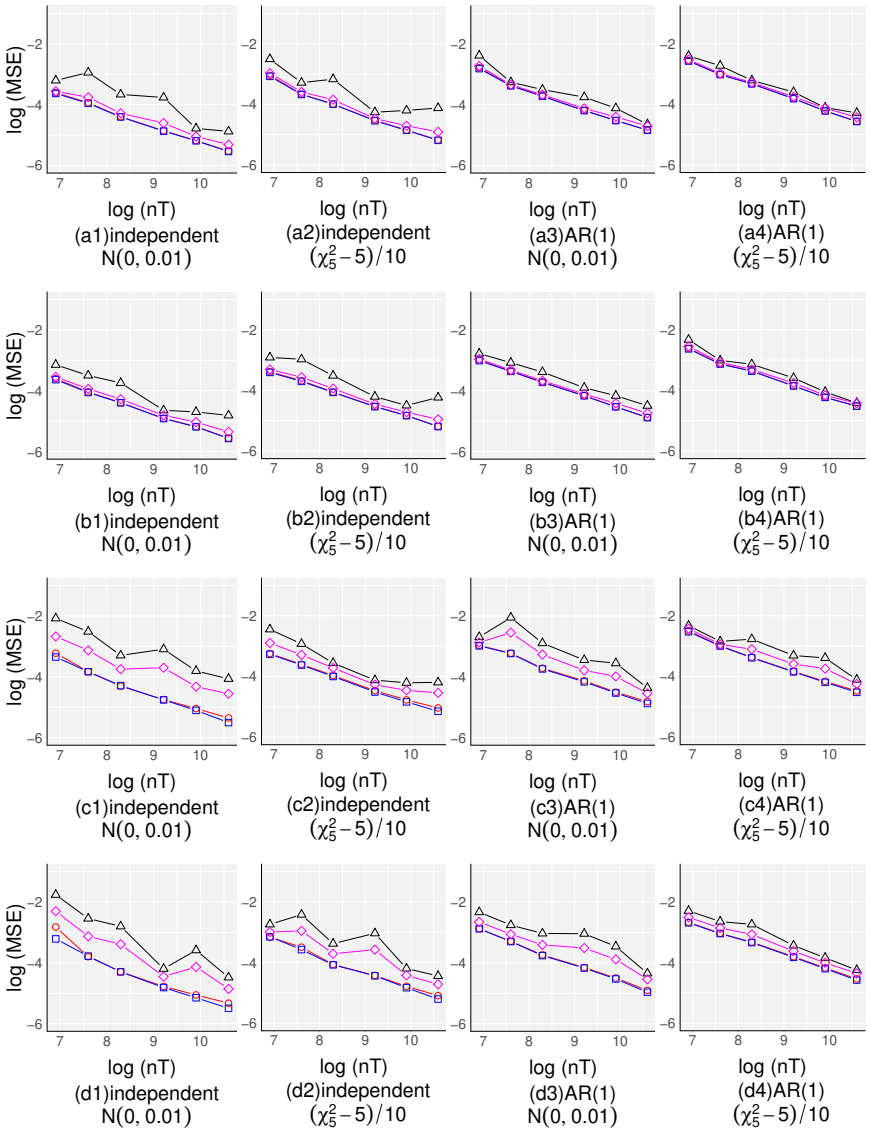

Figure 3.2: Comparisons of the logarithm of MSE for estimating $\boldsymbol{\beta}$ by TOPE ("–○–") along those of the oracle estimator (" –▫–"), the GLS estimator (" –◇–"), and the OLS ("–△–"). Results are about $T = 20$. In plots (a1)-(a4), $f_{kt} \sim N(0,1)$ are independent in k, t. In plots (b1)-(b4), $f_{kt} \sim t_8$ are independent in k, t. In plots (c1)-(c4), $\boldsymbol{f}_k$ follows ARMA$(1,1)$ with $N(0,1)$ innovation for each k. In plots (d1)-(d4), $\boldsymbol{f}_k$ follows ARMA$(1,1)$ with t_8 innovation for each k. Distributions and serial correlations of $\boldsymbol{u}_i$ are displayed in the plots.

estimators reduce as n increasing. Both the TOPE and GLS perform similarly as the oracle estimator when $\boldsymbol{f}_t$ is independent in t (plots (a1)-(a4), (b1)-(b4)), and all outperform the OLS; on the other hand, sophisticated dependence on $\boldsymbol{u}_t$ slightly increases the MSE but does not alter the convergence rate (plots (c1)-(c4), (d1)-(d4)). In addition, in the presence of dependence of $\boldsymbol{f}_t$ in t,

the GLS is outperformed as well while the TOPE remains its performance compared to the oracle estimator (plots (c1)-(c4), (d1)-(d4)). In the supplementary file, additional numerical results for settings similar to those in Figure 3.2 but with $T = 100, 500$ are reported in Figures B.4-B.7, and results for latent factor processes $\boldsymbol{f}_t$ following AR(1) settings are reported in Figures B.8-B.10 for $T = 20, 100, 500$. Similar observations are obtained for $T = 20$ with different settings for $\boldsymbol{f}_t$, and overall, the differences among distinct estimators decrease as T increasing.

Figure 3.3 displays the estimation error of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{G}}$ in terms of $\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}/\sqrt{T}$ and $\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathbb{F}}/\sqrt{n}$, which both decrease to zero when n increases. Error of $\widehat{\mathbf{G}}$ decreases as T increasing while error of $\widehat{\mathbf{F}}$ admits similar patterns when n is large yet it slightly inflates for small n and large T. This observation reflects the need of a relatively large number of subjects n to recover the latent factor processes satisfactorily when T is large.

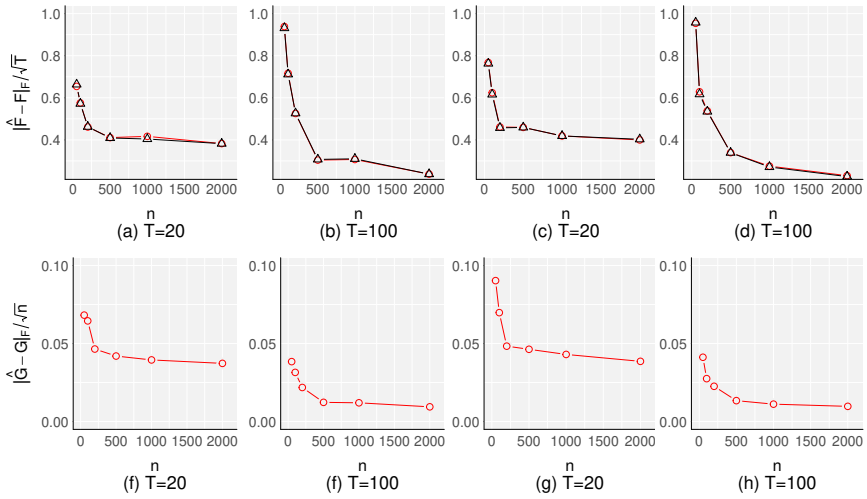

Figure 3.3: $\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathrm{F}}/\sqrt{T}$ by TOPE ("–○–") and oracle case ("–△–") and $\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathbb{F}}/\sqrt{n}$ by TOPE. In (a), (b), (e), and (f), $f_t \sim N(0, 1)$ and are independent in t. In (c), (d), (g), and (h), $f_t \sim (\chi^2_5 - 5)$ and are independent in t; $u_{it} \sim N(0, 0.1)$ are independent in t.

Figures 3.4 and 3.5 display the ECP and MML with respect to different T and n for different estimators. The nominal level is 0.95. In Figure 3.4, 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 deviation is substantial when n increases.

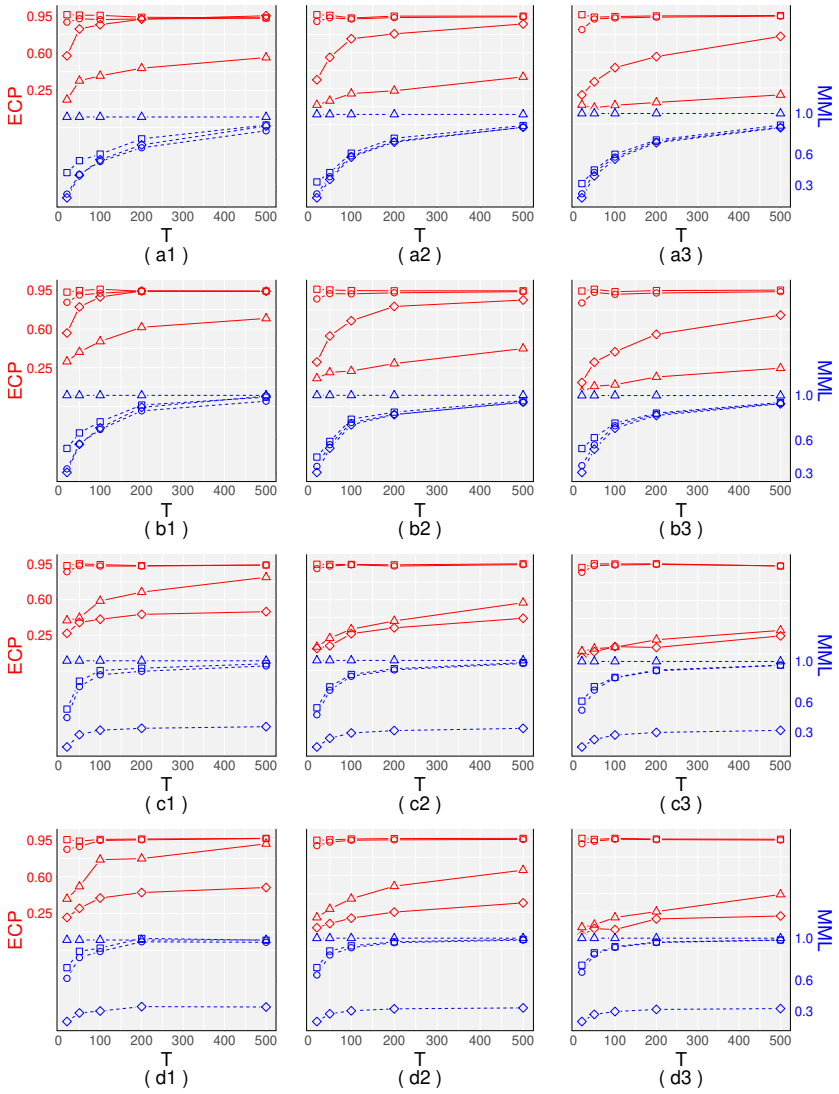

Figure 3.4: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $f_{kt} \sim N(0,1)$ are independent in k, t; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In plots (a1)-(a4) $u_{it} \sim N(0, 0.01)$ are independent in i, t. In plots (b1)-(b4), $u_{it} \sim (\chi^2_5 - 5)/10$ are independent in i, t. In plots (c1)-(c4) $\boldsymbol{u}_i$ follows the AR(1) model with $N(0, 0.01)$ innovation while same model is used for $\boldsymbol{u}_i$ in plots (d1)-(d4) with $(\chi^2_5 - 5)/10$ innovation.

Also, when dependence of $\boldsymbol{u}_t$ in t is introduced, the 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 $\mathbf{F}$ of TOPE prefers large n (see plots (c1) and (c2), (d1)-(d2) in Figure 3.4 for example). In the presence of the dependence of $\boldsymbol{f}_t$ in t, the TOPE performs remarkably well in terms of maintaining small MML and its ECP quickly converges to the nominal level in

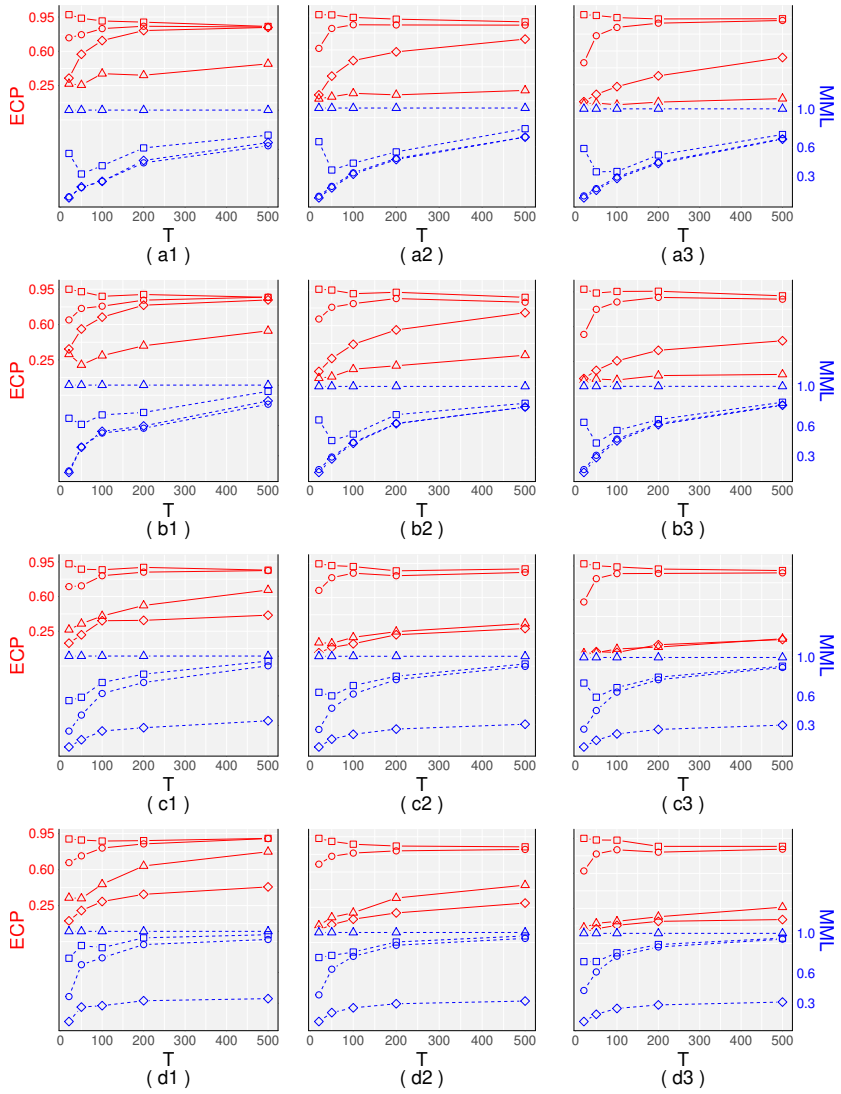

Figure 3.5: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows AR(1) with t_8 innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In plots (a1)-(a4), $u_{it} \sim N(0, 0.01)$ are independent in i, t. In plots (b1)-(b4), $u_{it} \sim (\chi^2_5 - 5)/10$ are independent in i, t. In plots (c1)-(c4), $\boldsymbol{u}_i$ follows the AR(1) model with $N(0, 0.01)$ innovation while same model is used for $\boldsymbol{u}_i$ in plots (d1)-(d4) with $(\chi^2_5 - 5)/10$ innovation.

T (Figure 3.5). Meanwhile, given the heteroscedasticity across subjects and serial/cross-sectional dependence, both GLS and OLS hardly maintain their ECP to the nominal level. More simulation results are retained in the supplementary file and provide similar observations. Specifically, Figures B.11-B.14 displays results for independent f_{kt} in k, t with either independent u_{it} in i, t or $\boldsymbol{u}_i$

following AR(1) model with different innovations. Results with $\boldsymbol{f}_t$ following the AR(1) model or the ARMA$(1,1)$ model with different innovations are included in Figures B.15-B.24.

3.7 Study on air quality and energy consumption data using the TOPE

In this section, we implement our proposed method to analyze an 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, which is extracted from 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 is known as a potential factor to affect concentration of air pollutants. For this illustrative study, as covariates, we include the annual state-level energy consumption proportions of three major sources out of all possible resources, namely coal, natural gas, and petroleum, in 2015 (https://www.eia.gov/electricity/data/browser/). For analysis, we take logarithm transformation on the air pollutant data and remove potential seasonality. Also, we transform the latitude and longitude to keep their values within $[0,1]$.

From Figures 3.6 and B.25 in the supplementary file, it is observed that both geographical variables (latitude and longitude) and energy consumption proportions can help explaining the observed heteroscedasticity across monitoring sites so that we will consider them as $\boldsymbol{x}_i$ in (3.1.1). 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 (3.1.1).

To determine the dimension K of latent factor process, we apply both eigenvalue-ratio procedure and the proposed HDWN testing-based procedure (detailed in Section 3.5 in the supplementary file). Ratios of the first ten adjacent eigenvalues of $\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}$ 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 (3.5.1) 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 (3.5.1). 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 3.4, 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 (3.1.1) for performing analysis on this data is

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

where $\ln(\text{PM2.5}_{it})$ is the log concentration of PM2.5 from the monitoring site i at time t; $\ln(\text{NO}_{2,it})$, $\ln(\text{SO}_{2,it})$, and $\ln(\text{Oz}_{it})$ 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.

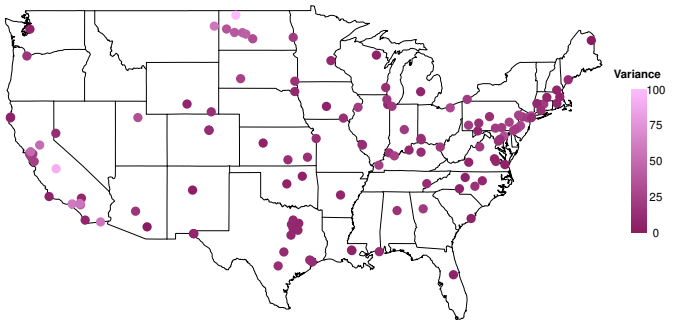

Figure 3.6: Variance of the mean PM2.5 concentration (over all time points) at 129 monitoring sites across the United States.

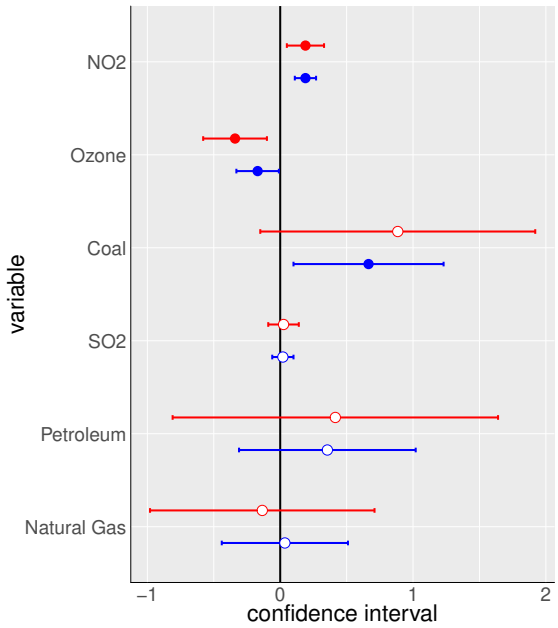

Figure 3.7: The 95% confidence intervals (the OLS estimator in red and the TOPE for (3.1.1) in blue) of the effects of energy consumption proportions of coal, natural gas, and petroleum and daily max 1-hour concentration of NO_2, SO_2, and ozone on the PM2.5 concentration.

For $g_{k\ell}$ in the above model ($\ell = 1, \ldots, 5$), we use cubic spline with 11 knots to construct $\mathbf{\Phi}$ for projection. We fit the above model using the TOPE and draw inference as proposed in Section 3.4 to inspect the effects of covariates on the PM2.5 concentration. As an expected advantage, no further restrictions need to be imposed to model (3.1.1) and the TOPE. In Figure 3.7, the 95% confidence intervals for estimated coefficients using the TOPE and the OLS estimator (by ignoring the variance components) are displayed for comparison. It reflects the efficiency of the TOPE in the presence of heteroscedasticity across monitoring sites and serial/contemporaneous correlations discussed in Section 3.3. Specifically, the confidence intervals constructed by the TOPE are shorter than those by the OLS estimator uniformly. Both resutls suggest significant positive correlation between daily max 1-hour concentration of NO_2 and PM2.5 concentration, $[0.05, 0.33]$ for the OLS estimator and $[0.11, 0.27]$ for the TOPE, and significant negative correlation between ozone concentration to PM2.5 concentration, $[-0.58, -0.10]$ for the OLS estimator and $[-0.33, -0.01]$ for the TOPE. However, the TOPE displays a significant positive correlation between coal consumption and PM2.5 concentration while the OLS estimator does not. This agrees with Liang et al.

(2015) that coal consumption positively contribute to PM2.5 concentration. Also, the recovered $\boldsymbol{g}_k$ for $k = 1, 2, 3$ display clear non-linearity and are depicted in Figure B.26 in the supplementary file.

3.8 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 develop a two-stage projection-based estimator for both the modulating and dependence components of the model, and establish an inference procedure for regression coefficients. Theoretically, 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 OLS and naive GLS estimators.

The widely-used Condition 3.2.2 essentially restricts $\mathbf{F}$ to subspace $\{\mathbf{F} \in \mathbb{R}^{T\times K} : T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K\}$, which might be stringent for some applications. In fact, we notice that it can be greatly relaxed by a concentration assumption of $T^{-1}\mathbf{F}'\mathbf{F}$ to $\mathbf{I}_K$, which can be derived from Condition 3.3.5 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 3.3.1. Furthermore, as noted after Condition 3.2.2, 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, $\max_{j\leqslant n}\sum_{i=1}^{n}|\mathbb{E}(u_{it}u_{jt})| < C_2$, $\max_{i\leqslant n}\sum_{k=1}^{n}\sum_{m=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T}|\operatorname{cov}(u_{it}u_{kt}, u_{is}u_{ms})| < C_2$, and $(nT)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T}$ $|\mathbb{E}(u_{it}u_{js})| < C_2$ for some $C_2 > 0$. However, as a result, the $n \times n$ covariance matrix $\operatorname{Cov}(\boldsymbol{u}_t)$ must be used in place of $\boldsymbol{\mathcal{D}}$ in (3.2.5) to retain the efficiency of the TOPE. For that purpose, both the weighted PCA (Jolliffe, 2002) and the estimator using thresholding principal orthogonal com-

plements (Fan et al., 2013) can be employed in conjunction with the 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 3.3. Finally, from its construction, the TOPE also paves a potentially effective way, which is free from performing sophisticated constraint likelihood estimation, to make predictions on $\boldsymbol{y}_t$ using our proposed model in conjunction with some parametric assumptions on $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$. We will extend our work to these questions in future efforts.

Chapter 4

Integrative Group Factor Model for Variable Clustering on Temporally Dependent Data

4.1 Introduction and Notation

We encounter large scale data with potential temporal dependence from multiple resources or multiple modalities. Such a new data structure often bears similarity as well as uniqueness among variables, which results in multiple or diverging numbers of covariance structures. In numerous empirical studies, this data structure has been emerging in various big data applications in a wide range of scientific fields such as bioinformatics and biology (Ernst et al., 2005; Möller-Levet et al., 2003; Pyatnitskiy et al., 2014), genetics (Fujita et al., 2012; Subhani et al., 2010) and multimedia (Niennattrakul and Ratanamahatana, 2006, 2007; Ratanamahatana and Keogh, 2005). Previous works have shown that, by combining diverse but usually complementary information from different covariance structures, an integrative analysis of large scale data is often beneficial for understanding the underlying structions; See Klami et al. (2014), Li and Li (2019), Wang et al. (2019) and Bunea et al. (2020) for examples.

To learn the covariance structure from the large scale data, a number of statistical methods have recently been developed, such as high-dimensional covariance estimation (Cai et al., 2016; Fan et al., 2011; Wang and Fan, 2017). Among these models, an important class of such approaches is the factor model, which model large scale data by the components that capture joint variation shared across variables (Anderson, 1962; Anderson and Rubin, 1956; Chamberlain and Rothschild, 1983; Lawley and Maxwell, 1962). Factor model is useful in dimension reduction. When it comes to high dimension case, large dimensional static factor model (Forni et al., 2000; Stock and Watson, 1998) is applied. However, these methods of factor analysis are based on the assumption that all variables share the same covariance structure or in particular, the same factors. To allow for

different factors, canonical correlation analysis (CCA) (Thompson, 1984, 2005) and inter-battery factor analysis (Browne, 1979; Tucker, 1958) are applied to two groups of variables, and multi-battery factor analysis (Browne, 1980) and group factor analysis (Klami et al., 2014) are applied to multiple groups. These methods work for more than one group of variables, but are based on known group assignments.

A more ubiquitous interest with more challenge is to recover unknown clustering assignments of a large number of variables with potential temporal dependence from multiple resources or multiple modalities. There are some unique characteristics that make the problem challenging. Most importantly, large scale data are often correlated, both cross-sectionally and temporally. This phenomenon has been constantly observed, and is actually the base upon which those matrix or tensor factorization solutions are built (Bai, 2003; Bai and Ng, 2013; Fan et al., 2016; Lock et al., 2013). The cross-sectional correlation brings challenge to many standard high-dimensional models such as LASSO (Tibshirani, 1996), as they usually require the predictors not to be highly correlated in order to achieve the desired statistical properties. To deal with the cross-sectional correlation, a popular approach is to treat recovering unknown clustering assignments as a variable clustering question, as given by Klami et al. (2014), Wang et al. (2019) and Bunea et al. (2020). Variable clustering is also of great challenge for large scale data. First, large scale data often refers to high-dimension case, where even a single group contains more variables than the sample size. Second, multiple resources arise new questions; for instance, how to define and recover clustering assignments based on covariance structure. In addition, multiple resources usually results in similarities as well as unique characteristics among variables. Since the clustering structures are given by unique characteristics and masked by errors and similarities, careful modeling is required to separate the unique characteristics from both errors and similarities. Finally, we need to deal with temporal correlation, which always gives more difficulty in modeling and clustering.

To deal with the temporal correlation, we consider using time series, which is naturally high dimensional, large in data size and dependent (Keogh and Kasetty, 2003; Lin et al., 2004; Rani and Sikka, 2012). Time series clustering is a widely used approach to recover clustering assign-

ments for large dependent data. For time series with difference on means, MacQueen et al. (1967) applied *k-means* algorithm with representing clusters by the mean value of the objects in the cluster. Kaufman and Rousseeuw (2009) applied *k-medoids* algorithm with representing clusters by the most centrally located object in a cluster. For time series with difference on covariances, Karypis et al. (1999), Guha et al. (1998) and Zhang et al. (1996) applied a hierarchical clustering method (Johnson, 1967) by grouping dataobjects into a tree of clusters, and Jebara et al. (2007), Yin and Yang (2005) and Alzate et al. (2009) applied a spectral clustering method (Ng et al., 2002; Von Luxburg, 2007). However, these procedures are all data-based and their statistical properties are not clear. Compared with data-based variablie clustering, model-based clustering enjoys advantage of clearly defined population-level clusters, which enables us to interpret the clusters and check the quality of a particular clustering algorithm. Model-based variable clustering is studied by Bunea et al. (2020). The author proposed an algorithm to recover clustering assignments and showed its minimax-optimality, which sheds a light on model-based variable clustering for high-dimensional data. Similar approaches of variable clustering through estimation of covariance matrix can also be seen in Wagaman and Levina (2009), Ieva et al. (2016) and Hallac et al. (2017). Despite these efforts, however, there is little work for temporally dependent data.

In this paper, we aim to bridge the gap. We consider integrative group factor model built upon the latent factors extracted from the large scale data. In particular, the factors are characterized to common factors for all groups and unique factors for each single group. We show that the model works for high dimension and high correlation, both cross-sectional and temporal. Based on this model, we first show the minimax lower bound of clustering recovery rate. The minimax lower bound is achieved by a similar but denser covering to that in Lu and Zhou (2016) and Gao et al. (2018), and given by Le Cam's method; see Theorem 4.3.1 for details. Second, we give estimation based on principal component analysis (PCA) procedure (Bai and Ng, 2013; Fan et al., 2016) to the latent factors and loadings, and show its non-asymptotic statistical guarantee. If the clustering assignments are known, the common factors and unique factors are estimated group-wise by PCA or partial common PCA (Wang et al., 2019). In practice, the clustering assignments are often

unknown. Thus, all factors are estimated simultaneously. See Section 4.4.1 for details. Then, we propose an algorithm to apply the estimation of latent factors and loadings to recover clustering assignments and gives its upper bound of clustering recovery rate. The clustering recovery works for high dimensional data with cross-sectional and temporal correlation; see Theorem 4.4.4 for details. The minimax lower bound and upper bound of clustering recovery rate combine together to give the optimality of our proposed algorithm. More importantly, from the minimax lower bound of clustering recovery rate, we find the precise demarcation for the *Region of Possibility* and *Region of Impossibility* in the two-dimensional phase space of the signals of unique factors in the largest cluster and the signals of unique factors in the smallest cluster compared with those of common factors. In the former, signals of unique factors are strong enough to allow successful clustering. In the latter, signals of unique factors are too weak for successful clustering. Also, in the two-dimensional phase space, we find the precise demarcation for the *Region of Guarantees* and *Region of Unknown Guarantees* from the upper bound of clustering recovery rate. In the former, our algorithm is guaranteed to give successful clustering. In the latter, the algorithm for clustering is not clear. From the four regions above, we discover a phase transition for variable clustering. The phase space partitions into three disjoint regions. In the first one, it is impossible to do clustering. In the second one, it is possible to do clustering and our algorithm is guaranteed. In the third, although it is possible to do clustering, the clustering algorithm is not clear. See Figure 4.1 for details. Also, we propose eigenvalue-ratio test (Ahn and Horenstein, 2013; Fan et al., 2016; Lam and Yao, 2012) to determine number of latent factors, both common and unique, in the integrative group factor model. We show the non-asymptotic properties of the estimator, which guarantees the convergence; see Theorem 4.4.5 for details. Finally, we generalize our proposed method to the case where the number of groups is not finite but diverges with respect to dimension. Similar results of latent factors and loadings estimation, clustering recovery and its optimality are given.

Our proposal contributes on several fronts. Although factor analysis is widely used to model high dimensional data with dependence, and group factor analysis is applied to model multiple

covariance structures, the previous works do not give clustering recovery for dependent data. Our proposal consider a different case to G-block model (Bunea et al., 2020). In particular, the variables in the same group are allowed to have different variances and covariance swith variables in other groups. In addition, each variable can involve temporal dependence. We provide algorithm of recovering clustering assignments for high dimensional dependent data, along with its optimality. The recovery error rate is defined under a loss function free from label switching (Gao et al., 2018; Lu and Zhou, 2016). Compared to the existing literatures, our proof of minimax lower bound of recovery error rate involves a denser covering free from the number of groups, so it can be extended to the case the number of groups diverges. Also, we apply Le Cam's method to give a tight bound for variable clustering with respect to covariance structures. The upper bound of recovery error rate is derived through scrupulously examining the non-asymptotic rates for estimating the latent factor process and its loading through PCA procedure (Bai and Ng, 2013; Fan et al., 2016). Lastly, we discover a phase transition in the phase space of signals of unique factors compared with those of common factors, which gives the region for possibility and guarantee of successful clustering. Also, our proposed model allows the number of groups not to be finite constant but a small term with respect to dimension. The technical tools we develop here are not limited to our setting alone, but are applicable to integrative group factor models.

We employ the following notation throughout this article. For a real number x, let $[x]$ be the largest integer no larger than x. Denote $\boldsymbol{a}_p = (a, \ldots, a)^\top \in \mathbb{R}^p$. For a p-dimensional vector $\boldsymbol{a} = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$, its ℓ_q-norm is defined by $||\boldsymbol{a}||_q = (\sum_{j=1}^p |a_j|^q)^{1/q}$ with $1 \leqslant q < \infty$. For a matrix $\mathbf{M} = (m_{ij})_{1 \leqslant i,j \leqslant p} \in \mathbb{R}^{p \times p}$, write $\|\mathbf{M}\|_{\max} = \max_{i,j} |m_{ij}|$ to be the maximum norm and $||\mathbf{M}||_{\mathbb{F}} = (\sum_{i=1}^p \sum_{j=1}^p m_{ij}^2)^{1/2}$ to be the Frobenius norm. The spectral norm of matrix $\mathbf{M}$ corresponds to its largest singular value, defined as $||\mathbf{M}||_2 = \sup_{\boldsymbol{a} \in S} ||\mathbf{M}\boldsymbol{a}||_2$, where $S = \{\boldsymbol{a} \in \mathbb{R}^p : ||\boldsymbol{a}||_2 = 1\}$. Denote the minimum and maximum eigenvalues of matrix $\mathbf{M}$ by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$, respectively. Let $\mathrm{tr}(\mathbf{M}) = \sum_{j=1}^p m_{jj}$ be the trace of $\mathbf{M}$. 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 $a_n \leqslant Cb_n$ for some C independent of n; and $a_n \asymp b_n$ if and only if there exists positive C and D independent on n such that $Cb_n \leqslant a_n \leqslant Db_n$. Unless specified otherwise, $s > 1$ and $C > 0$ denote generic constants independent of p, T.

The paper is organized as follows. In Section 4.2, we detail the integrative group factor model and discuss the preliminary conditions to derive the main results. In Section 4.3, we show the minimax lower bound of recovery error rate. In Section 4.4, we propose an estimation procedure on the latent factor processes and their loadings, an algorithm to recover clustering assignments based on the estimation, and determination of number of latent factors. We carry out a non-asymptotic analysis of our proposed estimator in Section 4.4.1, explore the upper bound of recovery error rate in Section 4.4.2 and derive convergence rate of determining of number of latent factors in Section 4.4.5. The results in Sections 4.3 and 4.4.2 combine to give the optimality and phase transition of our proposed algorithm. In Section 4.5, we discuss an alternative case where the number of groups diverges. Similar results of the upper bound and the minimax lower bound of recovery error rate are given. We conclude the paper with some discussions in Section 4.6.

4.2 Model

In this section, we introduce the model lucubrated in this paper and make some reasonable and necessary assumptions about the models and groups of the integrative group factor model. First, we denote a p-dimensional multivariate time series with T observations as y_{it} for $i = 1, \ldots, p$ and $t = 1, \ldots, T$. Unlike traditional clustering analysis that is based on different means for different groups (Gao et al., 2018; Lu and Zhou, 2016; Zhang et al., 2018, 2016), we assume that y_{it} is a stationary time series with zero mean. Instead, variables in different groups are defined by their variance and covariance with others. In particular, two curves of time series y_{it} and y_{jt} in two different groups are "weakly" correlated in the sense that there exists a process $\boldsymbol{f}_t$ satisfying that $\mathrm{cov}(y_{it}, y_{jt}|\boldsymbol{f}_t) = 0$ and $\mathrm{Var}(y_{it}|\boldsymbol{f}_t), \mathrm{Var}(y_{jt}|\boldsymbol{f}_t) \neq 0$, while such a process does not exist for two curves of time series in the same cluster. In other words, we define the groups of time series based on their covariance structure. First, we set up condition of the cluster assignments. Suppose that

the p dimensions are split into m disjoint groups by clustering assignment $\boldsymbol{z} = (z_1, \ldots, z_p) \in \{1, \ldots, m\}^p$ as follows:

$$\mathcal{V} = \mathcal{V}^{(1)} \cup \cdots \cup \mathcal{V}^{(m)},$$

where $\mathcal{V} = \{1, \ldots, p\}$, $\mathcal{V}^{(j)} = \{i : z_i = j\}$ and m is a constant independent of p and T. Let $p_j = |\mathcal{V}^{(j)}|$ be the number of curves in the jth cluster for $j = 1, \ldots, m$. Throughout the paper, we focus on the relatively "balanced" groups with the following condition.

Condition 4.2.1. *For each $j = 1, \ldots, m$, $p_j \asymp p$.*

Condition 4.2.1 illustrates what we mean by the balanced groups aforementioned, which basically assumes that each cluster has a size proportional to p. This assumption ensures the number of curves in each cluster is not too small, and thus it guarantees the accuracy of clustering. Note that if we only have one group, we are dealing with a high dimensional covariacne structure. Thus, further condition on the covariance structure is needed to identify the covaraiance and cluster structure. As suggested by Johnstone (2001), to simplify the model, we consider spike covariance structure model, which is important and useful for high dimensional data. According to Chamberlain and Rothschild (1983), Stock and Watson (2002a), Bai (2003) and Lam and Yao (2012), the spike covariance structure can be modeled by an approximate factor model. Thus, we start from an approximate factor model.

4.2.1 Approximate Factor Model

An approximate factor model (Bai, 2003; Chamberlain and Rothschild, 1983; Lam and Yao, 2012; Stock and Watson, 2002a) is

$$y_{it} = a_{i1} f_{t1} + \cdots + a_{ir} f_{tr} + u_{it},$$

for $t = 1, \ldots, T$ and $i = 1 \ldots, p$, where y_{it} is an observation from the ith variable at time t, $(f_{t1}, \ldots, f_{tr})^\top$ is a r-dimensional process and u_{it} is an error process. The model can be written as

in matrix format

$$\mathbf{Y} = \mathbf{A}\mathbf{F}^\top + \mathbf{U},$$

where $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ with $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{pt})^\top$, $\mathbf{A} = \{a_{ik}\}_{i=1,k=1}^{p,r}$, $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)^\top$ with $\boldsymbol{f}_t = (f_{t1}, \ldots, f_{tr})^\top$ and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ with $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{pt})^\top$. In the approximate factor model, we assume that:

Condition 4.2.2. *For each $t = 1, \ldots, T$, $f_{t1}, \ldots, f_{tr}$ are uncorrelated with each other and have zero mean and unit variance; for each t, $u_{1t}, \ldots, u_{pt}$ have zero mean and finite variances; and $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$ are independent with each other.*

Condition 4.2.3. $\mathbf{A}^\top\mathbf{A}$ *is a diagonal matrix with non-zero distinct entries and there exist constants $d_1, d_2 > 0$ such that $d_1 \lesssim \lambda_{\min}(p^{-1}\mathbf{A}^\top\mathbf{A}) \leqslant \lambda_{\max}(p^{-1}\mathbf{A}^\top\mathbf{A}) \lesssim d_2$.*

Condition 4.2.4. *Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ denote the σ-algebras generated by $\{(\boldsymbol{f}_t, \boldsymbol{u}_t) : t \leqslant 0\}$ and $\{(\boldsymbol{f}_t, \boldsymbol{u}_t) : t \geqslant T\}$, respectively. Define the mixing coefficient $\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(A \cap B)|$. The data are generated as follows.*

(i) Stationarity. $\{\boldsymbol{u}_t, \boldsymbol{f}_t\}_{t \leqslant T}$ is weak stationary;

(ii) Strong mixing. There exist $q_1, C_1 > 0$ such that for any $s > 0$, $\alpha(s) < \exp(-C_1 s^{q_1})$;

(iii) Exponential tail. There exist $q_2, q_3 > 1$ with $q_1^{-1} + q_2^{-1} + q_3^{-1} > 1$ and $b_1, b_2 > 0$ such that for each $i = 1, \ldots, p$, and any $s > 0$, $\mathbb{P}(|u_{it}| > s) \leqslant \exp\{-(s/b_1)^{q_2}\}$ and $\mathbb{P}(|f_{tk}| > s) \leqslant \exp\{-(s/b_2)^{q_3}\}$.

Condition 4.2.2 are conditions on the latent factor process and error process in the approximate factor model. In the model, the factors $f_{t1}, \ldots, f_{tK}$ are assumed to be uncorrelated. Thus, the covariance structure of model is given by

$$\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top + \boldsymbol{\Sigma}_u,$$

where $\mathbf{\Sigma} = T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)$ and $\mathbf{\Sigma}_u = T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$ are the population covariance matrixs of the data matrix $\mathbf{Y}$ and error matrix $\mathbf{U}$. We do not assume that $u_{1t}, \ldots, u_{pt}$ are independent with each other, so $\mathbf{\Sigma}_u$ is allowed not to be a diagonal matrix. Condition 4.2.3 is similar to the PC1 condition in Bai and Ng (2013). Unlike PC1 condition, we relax the restriction on $\mathbf{F}$ and do not require that almost surely, $T^{-1}\mathbf{F}^\top\mathbf{F} = \mathbf{I}_r$. Instead, we only require that with probability at least $1 - e^{-s}$, $\|T^{-1}\mathbf{F}^\top\mathbf{F} - \mathbf{I}_k\|_{\mathbb{F}}^2 \lesssim T^{-1}s$, which is satisfied for $\mathbf{F}$ under Condition 4.2.2 and 4.2.4 (Zhang et al., 2020). Since $\mathbf{A}\mathbf{A}^\top$ and $\mathbf{A}^\top\mathbf{A}$ shares the same non-zero eigenvalues, condition 4.2.3 gives that the first K eigenvalues of $\mathbf{A}\mathbf{A}^\top$ diverge with p. This ensures the identifiability of the approximate factor model (Chamberlain and Rothschild, 1983). In addition, the serial covariance between two variables y_{it} and y_{jt},

$$\operatorname{cov}(y_{it}, y_{jt}) = \sum_{k=1}^{r} a_{ik}a_{jk} + \operatorname{cov}(u_{it}, u_{jt})$$

which is mainly contributed by the loading matrix $\mathbf{A}$. Condition 4.2.4 is commonly imposed in high-dimensional factor analysis (e.g. Bai, 2003; Stock and Watson, 2002a) that requires weak serial dependency of the latent factor process and error process. The exponential tail is satisfied for some heavy tailed distribution such as t-distribution or gamma distribution. Thus, we relax the normality assumption in traditional factor models (Bai, 2003; Bai and Ng, 2013).

4.2.2 Integrative Group Factor Model

An integrative group factor model is an extension of the approximate factor model that the variables in each cluster follows the approximate factor model, and can be useful in many fields such as psychology (Ramírez et al., 2018). It is easy to see that, if we ignore the error process, two variables will be uncorrelated if they are loaded on different factors. Thus, we allow the existence of common factors and unique factors across variables. For simplicity, we propose the following model

$$y_{it} = a_{i1}f_{t1}^{(0)} + \cdots + a_{ir_{z_i}}f_{tr_{z_i}}^{(0)} + b_{i1}f_{i1}^{(z_i)} + \cdots + b_{ir_{z_i}}f_{ir_{z_i}}^{(z_i)} + u_{it} \tag{4.2.1}$$

for $t = 1, \ldots, T$ and $i = 1 \ldots, p$, where $z_1, \ldots, z_p \in \{1, \ldots, m\}$ are underlying labels. Here, y_{it} is an observation from the ith variable at time t, $(f_{t1}^{(0)}, \ldots, f_{tr_0}^{(0)})^\top$ is a r_0-dimensional process, $(f_{t1}^{(j)}, \ldots, f_{tr_j}^{(j)})^\top$ is a r_j-dimensional process for $j = 1, \ldots, m$ and u_{it} is an error process. Denote $\{i_1, \ldots, i_{p_j}\} := \{i : z_i = j\}$ for each $j = 1, \ldots, m$. Then the model for group j can be written as

$$\boldsymbol{y}_t^{(j)} = \mathbf{A}_j \boldsymbol{f}_t^{(0)} + \mathbf{B}_j \boldsymbol{f}_t^{(j)} + \boldsymbol{u}_t^{(j)}, \tag{4.2.2}$$

where $\mathbf{A}_j = \{a_{i_\ell k}\}_{\ell=1,k=1}^{p_j, r_0}$, $\mathbf{B}_j = \{b_{i_\ell k}\}_{\ell=1,k=1}^{p_j, r_j}$, $\boldsymbol{y}_t^{(j)} = (y_{i_1 t}^{(j)}, \ldots, y_{i_{p_j} t}^{(j)})^\top$, $\boldsymbol{f}_t^{(j)} = (f_{t1}^{(j)}, \ldots, f_{tr_j}^{(j)})^\top$ and $\boldsymbol{u}_t^{(j)} = (u_{i_1 t}, \ldots, u_{i_{p_j} t})^\top$ for $j = 0, 1, \ldots, m$ and $t = 1, \ldots, T$. In the integrative group factor model, we apply the following condition:

Condition 4.2.5. *Condition 4.2.2 and 4.2.4 hold for* $\boldsymbol{f}_t = (f_{1t}^{(0)}, \ldots, f_{r_0 t}^{(0)}, \ldots, f_{1t}^{(m)}, \ldots, f_{r_m t}^{(m)})^\top$ *and* $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{pt})^\top$.

Similar as Condition 4.2.2, Condition 4.2.5 ensures the identifiability of the integrative group factor model. We assume that all factors are uncorrelated with each other, so that serial covariance between two variables y_{it} and y_{jt},

$$\operatorname{cov}(y_{it}, y_{jt}) = \sum_{k=1}^{r_0} a_{ik} a_{jk} + \sum_{k=1}^{r_{z_i}} b_{ik} b_{jk} + \operatorname{cov}(u_{it}, u_{jt}) \tag{4.2.3}$$

if $z_i = z_j$ and

$$\operatorname{cov}(y_{it}, y_{jt}) = \sum_{k=1}^{r_0} a_{ik} a_{jk} + \operatorname{cov}(u_{it}, u_{jt}) \tag{4.2.4}$$

if $z_i \neq z_j$. Thus, we simultaneously models the variance structure within groups and covariance between groups. In addition, if $z_i \neq z_j$, the covariance between y_{it} and y_{jt} conditional on $\boldsymbol{f}_t^{(0)}$ is

$$\operatorname{cov}(y_{it}, y_{jt} | \boldsymbol{f}_t^{(0)}) = \operatorname{cov}(u_{it}, u_{jt}),$$

while their variance conditional on $\boldsymbol{f}_t^{(0)}$ is

$$\mathrm{Var}(y_{it}|\boldsymbol{f}_t^{(0)}) = \sum_{k=1}^{r_{z_i}} b_{ik}^2 + \mathrm{Var}(u_{it}).$$

This matches our definition about groups in Section 4.2. Further, we denote $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)^\top$. For the sake of identification and consistent estimate, we propose the following conditions.

Condition 4.2.6. *There exist positive constant R such that $r_j \leqslant R$ for each $j = 0, 1, \ldots, m$. For each $j = 1, \ldots, m$ and $i = 1, \ldots, p_j$, $\sum_{\ell=1}^{r_0} \mathcal{I}(a_{i\ell} \neq 0) \geqslant 1$ and $\sum_{\ell=1}^{r_j} \mathcal{I}(b_{i\ell}^{(j)} \neq 0) \geqslant 1$. For each $j = 1, \ldots, m$, there exist $i \in \{1, \ldots, p_j\}$ such that $\sum_{\ell=1}^{r_j} \mathcal{I}(b_{i\ell}^{(j)} \neq 0) = r_j$.*

Condition 4.2.7. *For each $j = 1, \ldots, m$, $\mathbf{A}_j^\top \mathbf{A}_j$ and $\mathbf{B}_j^\top \mathbf{B}_j$ are diagonal matrices with non-zero distinct entries and $\mathbf{A}_j^\top \mathbf{B}_j = \mathbf{0}$. There exist constants $d_1, d_2 > 0$ such that $d_2/d_1 < m$, $d_1 \leqslant \lambda_{\min}(p^{-1}\mathbf{A}_j^\top \mathbf{A}_j) \leqslant \lambda_{\max}(p^{-1}\mathbf{A}_j^\top \mathbf{A}_j) \leqslant d_2$ and $d_1 \leqslant \lambda_{\min}(p^{-1}\mathbf{B}_j^\top \mathbf{B}_j) \leqslant \lambda_{\max}(p^{-1}\mathbf{B}_j^\top \mathbf{B}_j) \leqslant d_2$ for each $j = 1, \ldots, m$.*

Condition 4.2.6 gives assumption on the loadings for the integrative group factor model. The upper bound of factors ensures the total number of factors $K = \sum_{j=0}^{m} r_j$ is bounded, which is essential in statistical guarantee in estimating K. The lower bound of non-zero elements in loading matrix separates the loaded and unloaded factors for each variable. This condition is essential in estimating the latent factors and loadings as well as recovering the latent cluster assignments. Condition 4.2.7 is similar to Condition 4.2.3. Besides the condition on eigenvalues of $\mathbf{A}_j^\top \mathbf{A}_j$ and $\mathbf{B}_j^\top \mathbf{B}_j$, we put an additional condition on the bound of eigenvalues that $d_2/d_1 < m$. This condition requires the lower bound of eigenvalues not to be too small and the upper bound not to be too large. Unlike approximate factor model, in integrative group factor model, we have common factors as well as unique factors. Note that by Condition 4.2.7, $d_1 m \leqslant \lambda_{\min}(p^{-1}\sum_{j=1}^{m} \mathbf{A}_j^\top \mathbf{A}_j) \leqslant \lambda_{\max}(p^{-1}\sum_{j=1}^{m} \mathbf{A}_j^\top \mathbf{A}_j) \leqslant d_2 m$, which implies that the common factors and unique factors are of the same strength. Thus, the common factors and unique factors are not distinguishable without additional condition on the strength, which is crucial in estimating number of common and unique factors and the factor matrices. Hence, we give an additional

condition that $d_2/d_1 < m$ in Condition 4.2.7. See Section 4.4.5 for more details. We assume the columns of loading matrix to be orthogonal and columns of factor matrix to be orthonormal, which ensures the unbiasedness of PCA procedure (Bai and Ng, 2013). Let

$$\widetilde{\mathbf{C}} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1 & & \\ \vdots & & \ddots & \\ \mathbf{A}_m & & & \mathbf{B}_m \end{bmatrix}. \tag{4.2.5}$$

It is easy to write equations (4.2.3) and (4.2.4) in matrix form as

$$\mathrm{Var}(\boldsymbol{y}_t) = \widetilde{\mathbf{C}}\widetilde{\mathbf{C}}^\top + \mathrm{Var}(\boldsymbol{u}_t), \tag{4.2.6}$$

where $\boldsymbol{y}_t = (\boldsymbol{y}_t^{(1)\top}, \ldots, \boldsymbol{y}_t^{(m)\top})$ and $\boldsymbol{u}_t = (\boldsymbol{u}_t^{(1)\top}, \ldots, \boldsymbol{u}_t^{(m)\top})$. Note that by Condition 4.2.7, $d_1 m \leqslant \lambda_{\min}(p^{-1}\sum_{j=1}^m \mathbf{A}_j^\top \mathbf{A}_j) \leqslant \lambda_{\max}(p^{-1}\sum_{j=1}^m \mathbf{A}_j^\top \mathbf{A}_j) \leqslant d_2 m$ and $d_1 \leqslant \lambda_{\min}(p^{-1}\mathbf{B}_j^\top \mathbf{B}_j) \leqslant \lambda_{\max}(p^{-1}\mathbf{B}_j^\top \mathbf{B}_j) \leqslant d_2$ for each $j = 1, \ldots, m$, which implies that $d_1 \leqslant \lambda_{\min}(p^{-1}\widetilde{\mathbf{C}}^\top\widetilde{\mathbf{C}}) \leqslant \lambda_{\max}(p^{-1}\widetilde{\mathbf{C}}^\top\widetilde{\mathbf{C}}) \leqslant d_2 m$. Then, by the identifiability of approximate factor model (Theorem 4, Chamberlain and Rothschild, 1983), K is uniquely determined by number of diverging eigenvalues of $\mathrm{Var}(y_t)$. In addition, the decomposition of $\mathrm{Var}(y_t)$ in (4.2.6) is unique in the sense that suppose there exist $p \times K$ matrix $\mathbf{V}$ and $p \times p$ positive definite matrix $\mathbf{W}$ with uniformly bounded eigenvalues such that $\mathrm{Var}(y_t) = \mathbf{V}\mathbf{V}^\top + \mathbf{W}$, then $\mathbf{V}\mathbf{V}^\top = \widetilde{\mathbf{C}}\widetilde{\mathbf{C}}^\top$ and $\mathbf{W} = \mathrm{Var}(\boldsymbol{u}_t)$. That is, the column space of $\widetilde{\mathbf{C}}$ is uniquely determined by $\mathrm{Var}(y_t)$. Also, Condition 4.2.7 shows that $\widetilde{\mathbf{C}}^\top\widetilde{\mathbf{C}}$ is a diagonal matrix, which further gives K^2 equations upon $\widetilde{\mathbf{C}}$ ($K(K-1)/2$ equations from that $\widetilde{\mathbf{C}}^\top\widetilde{\mathbf{C}}$ is a diagonal matrix and $K(K+1)/2$ equations from that $T^{-1}\mathbf{F}^\top\mathbf{F}$ converges to $\mathbf{I}_K$ in probability) and uniquely determines $\widetilde{\mathbf{C}}$ from its column space. Thus, it is clear that any $p \times K$ matrix $\widetilde{\mathbf{C}}$ satisfying Condition 4.2.7 and (4.2.6) must have the same block structure as (4.2.5) gives. In addition, by Condition 4.2.7, the columns of $\widetilde{\mathbf{C}}$ corresponding to the largest r_0 eigenvalues are the columns of loadings of common factors. This shows the identifiability of the integrative group factor model.

Combining model (4.2.2) with the cluster assignments $\boldsymbol{z}$, we have a model with parameter

$$(\boldsymbol{z}, \mathbf{C}) \in (\mathcal{Z}, \mathcal{C}) = \Big\{ \boldsymbol{z} \in \{1, \ldots, m\}^p, \mathbf{C} = \{c_{ik}\}_{i,k=1}^{p,K} \in \mathbb{R}^{p \times K} : \\ p_j(\boldsymbol{z}) \asymp p, c_{ik} = 0 \text{ for } k \notin \{1 + \sum_{\ell=0}^{j-1} r_j, \ldots, \sum_{\ell=0}^{j} r_j\} \text{ if } z_i = j \Big\},$$

where $\mathcal{Z}$ is the set of all labels $\boldsymbol{z} = (z_1, \ldots, z_p) \in \{1, \ldots, m\}^p$ which satisfy Condition 4.2.1 and $\mathcal{C}$ is the set of matrices which have the form in (4.2.5) and satisfy Condition 4.2.7. Assuming that the cluster assignments satisfy that $z_1 \leqslant \cdots \leqslant z_p$, we can write model (4.2.1) in a matrix form as

$$\mathbf{Y} = \widetilde{\mathbf{C}}\mathbf{F}^\top + \mathbf{U}, \tag{4.2.7}$$

where $\widetilde{\mathbf{C}}$ is defined in (4.2.5). In practice, the group assignments are always unknown. To account for the latent group assignments, we introduce a $p \times p$ permutation matrix $\mathbf{\Pi}$. Then, model (4.2.1) can be written as

$$\mathbf{Y} = \mathbf{\Pi}\widetilde{\mathbf{C}}\mathbf{F}^\top + \mathbf{U} \tag{4.2.8}$$

where $\mathbf{Y} = \{y_{it}\}_{i=1,t=1}^{p,T}$, $\mathbf{U} = \{u_{it}\}_{i=1,t=1}^{p,T}$ or

$$\mathbf{Y} = \mathbf{C}\mathbf{F}^\top + \mathbf{U}, \tag{4.2.9}$$

where $\mathbf{C} = \mathbf{\Pi}\widetilde{\mathbf{C}}$. The permutation matrix $\mathbf{\Pi}$ is not unique for model (4.2.2). However, the cluster assignments given by the permutation matrix $\mathbf{\Pi}$ is unique up to label switching.

The integrative group factor model given above enjoys some commonalities with previous works in existing literatures. For instantce, Wang et al. (2019) propose partial common PCA in modeling multiple covariance matrices. Based on equal group size and known clustering assign-

ments, the covariance matrix for the jth cluster is decomposed as

$$\boldsymbol{\Sigma}_j = \boldsymbol{\Gamma}\boldsymbol{\Lambda}_j\boldsymbol{\Gamma}^\top + \boldsymbol{\Psi}_j,$$

where $\boldsymbol{\Gamma}$ is a $p \times K$ orthogonal matrix and $\boldsymbol{\Lambda}_j$ is a diagonal matrix. Under the same conditions of equal group size and known clustering assignments, the covariance matrix for the jth cluster of integrative group factor model with no common factor is given by

$$\boldsymbol{\Sigma}_j = \mathbf{B}_j\mathbf{B}_j^\top + \boldsymbol{\Psi}_j.$$

Thus, under additional condition that $\mathbf{B}_1, \ldots, \mathbf{B}_m$ share the same column space, partial common PCA and integrative group factor model are equivalent. Thus, partial common PCA is actually a special case of integrative group factor model. Also, Bunea et al. (2020) proposed an approximate G-block model for model-based variable clustering. In particular, the variables are clustered by partition $\boldsymbol{z}$ and the covariance matrix $\boldsymbol{\Sigma}$ is decomposed as

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\mathbf{V}\boldsymbol{\Gamma}^\top + \boldsymbol{\Psi}, \tag{4.2.10}$$

where $\boldsymbol{\Gamma}$ is a 0-1 membership matrix relative to $\boldsymbol{z}$, $\mathbf{V}$ is a symmetric $m \times m$ matrix and $\boldsymbol{\Psi}$ has small off-diagonal entries. In the approximate G-block model, the clustering assignments are given by the membership matrix $\mathbf{A}$ and the variance and covariance of variables are given by $\mathbf{V}$. Thus, all variables in a cluster share the same variance and covariance, which is different from the variance and covariance of integrative group factor model given by (4.2.3) and (4.2.4). Thus, approximate G-block model and integrative group factor model are different in most cases, and equivalent in some special cases. For example, we consider a special case, where there is no common factor and only one unique factor with the same loading among all variables in the each cluster in integrative group factor model, and $\mathbf{V}$ is a diagonal matrix in approximate G-block model. In this case, the loading matrix $\mathbf{B}_j$ of unique factors in the jth cluster is $b_j\mathbf{1}_{p_j}$. Then, the covariance matrix for

integrative group factor model is given by

$$\begin{aligned}\boldsymbol{\Sigma} =&\{\mathrm{diag}(b_1\mathbf{1}_{p_1},\ldots,b_m\mathbf{1}_{p_m})\}\{\mathrm{diag}(b_1\mathbf{1}_{p_1},\ldots,b_m\mathbf{1}_{p_m})\}^\top+\boldsymbol{\Sigma}_u\\ =&\{\mathrm{diag}(\mathbf{1}_{p_1},\ldots,\mathbf{1}_{p_m})\}\{\mathrm{diag}(b_1^2,\ldots,b_m^2)\}\{\mathrm{diag}(\mathbf{1}_{p_1},\ldots,\mathbf{1}_{p_m})\}^\top+\boldsymbol{\Sigma}_u.\end{aligned}$$

Note that $\mathrm{diag}(\mathbf{1}_{p_1},\ldots,\mathbf{1}_{p_m})$ is a 0-1 membership matrix and $\mathrm{diag}(b_1^2,\ldots,b_m^2)$ is a $m\times m$ diagonal matrix. Thus, approximate G-block model and integrative group factor model are equivalent in this case.

4.3 Minimax Lower Bound of Clustering Recovery

First, we consider detecting the latent clustering assignments of variables in model (4.2.1). Since an ordering of the variables is not available in many applications such as genetics, social, finical and economic data, methods invariant to variable permutations are appropriate for such applications (Wagaman and Levina, 2009). Thus, as suggested by Jin et al. (2015), the clustering recovery errors should not depend on how we label each of the m groups. To do variable clustering that is permutation invariant, we introduce a loss function that is also permutation invariant as follows. Denote $\mathcal{S}_m$ as the set of all permutations of $\{1,\ldots,m\}$. Then, the $0-1$ loss function for estimated labels $\widehat{\boldsymbol{z}}$ is defined as

$$L(\widehat{\boldsymbol{z}},\boldsymbol{z})=\inf_{\boldsymbol{\Pi}\in\mathcal{S}_m}\left[\frac{1}{p}\sum_{i=1}^{p}\mathcal{I}\{\boldsymbol{\Pi}(\widehat{z}_i)\neq z_i\}\right]. \tag{4.3.1}$$

Note that in model (4.2.2), the cluster structure is invariant to the permutations of label symbols, so we do not distinguish for cluster label switching. In practice, we only care about which curves are in the same cluster, instead of the exact cluster labels, so the actual labels used in defining the cluster assignments should be inconsequential. Thus, we introduce the permutation $\boldsymbol{\Pi}$ and the minimization over all permutations to avoid the error from label switching. The loss function in (4.3.1) has been previously used in the investigation of clustering in mixture models (Lu and Zhou, 2016) and stochastic block models (Gao et al., 2017, 2018; Zhang et al., 2016). Alternatively,

Bunea et al. (2020) considers recovering exact group assignments, which treats label switching as errors, which results in different group recovery errors. They may give slight different minimax lower bound, which is essentially caused by the difference of loss functions instead of difficulty of modeling.

4.3.1 Minimax Lower Bound for Integrative Group Factor Model

As discussed in Section 4.2.2, model (4.2.2) models the variance structure within groups and covariance between groups simultaneously. Thus, in this paper, we consider separating variables through multiple groups through the covariance matrix of variables. In particular, we put variables with relatively strong correlation into the same cluster. Unlike normal clustering based on mean of each cluster, we apply clustering based on variance and covariance matrix of groups. Covariance-type clustering has been studied before by Ieva et al. (2016) and Hallac et al. (2017). However, the error rate of covariance-type clustering is not given, which is not trivial. We will first show the lower bound of the probability of failing to estimate correct cluster assignments in the following theorem. To get minimax lower bound of the expected value of the $0-1$ loss function, we first choose $p/4$ elements in $\{1,\ldots,m\}^p$, which is the space of m cluster assignments for p variables, as separations. Similar approach can be seen in Lu and Zhou (2016) and Gao et al. (2018). Compared to those authors, we choose a denser covering that is independent on m, so the covering works for finite m. Then, the distances between each pair of the $p/4$ are quantified by the K-L divergence between the corresponding distributions. To get a tighter lower bound, we apply Le Cam's method and give the following minimax lower bound.

Theorem 4.3.1. *Let $\mathcal{Z}$ be the set of all labels $(z_1,\ldots,z_p) \in \{1,\ldots,m\}$ which satisfy Condition 4.2.1 and $\mathcal{C}$ be the set of matrices which have the form in* (4.2.5) *and satisfy Condition 4.2.7. Further, let $D_A^2 = \max_j \lambda_{\max}(p^{-1}\mathbf{A}_j^\top \mathbf{A}_j)$ and $d_B^2 = \min_j \lambda_{\min}(p^{-1}\mathbf{B}_j^\top \mathbf{B}_j)$. Then, the signal-noise ratio for model* (4.2.8) *is defined as $\theta = (D_A^2 r_0 + d_B^2 \min_j r_j)^{-1} d_B^2 r_0 \max_j r_j$. Under Condi-*

tions 4.2.1, 4.2.5, 4.2.6 and 4.2.7, for some $\varepsilon \in (\log(2)/\log(p/4), 1)$*, we have*

$$\inf_{\widehat{\boldsymbol{z}}} \sup_{\boldsymbol{z}\in\mathcal{Z},\mathbf{C}\in\mathcal{C}} \mathbb{E}\{L(\widehat{\boldsymbol{z}}, \boldsymbol{z})\} \geqslant \frac{\{\varepsilon \log(p/4) - \log(2)\}(1-\varepsilon)m}{16\theta pT} \wedge \frac{1}{64}, \tag{4.3.2}$$

where the infimum is taken over all label estimators $\widehat{\boldsymbol{z}}$*.*

For $p > 8$, the maximum of (4.3.2) about ε is achieved by letting $\varepsilon = \{2\log(p/4)\}^{-1}\log(p/2)$, we have

$$\inf_{\widehat{\boldsymbol{z}}} \sup_{\boldsymbol{z}\in\mathcal{Z},\mathbf{C}\in\mathcal{C}} \mathbb{E}\{L(\widehat{\boldsymbol{z}}, \boldsymbol{z})\} \geqslant \frac{\{\log^2(p/4) + \log^2(2)\}m}{32\theta p\log(p/4)T} \wedge \frac{1}{64}, \tag{4.3.3}$$

where the infimum is taken over all label estimators $\widehat{\boldsymbol{z}}$. To obtain a non-trivial lower bound in (4.3.3), we require sample size T to be no larger than $\theta^{-1}m^{-1}p\log(p/8)$. If sample size is larger than $\theta^{-1}m^{-1}p\log(p/8)$, Theorem 4.3.1 provides a negative, and thus trivial lower bound. Note that this constrain does not give an upper bound the dimension p. Thus, the lower bound still holds for high dimensional setting. The signal-noise ratio $\theta = (D_A^2 r_0 + d_B^2 \min_j r_j)^{-1} d_B^2 r_0 \max_j r_j$ shows the strength of unique factors compared with common factors. Recall that the covariance of variables are decomposed into two parts, one from the common factors and the other from the unique factors. In addition, we separate variables through multiple groups through the covariance matrix of variables. If the signal-noise ratio θ is big, that is, the loadings of unique factors are big compared with the loadings of common factors, the covariance will mainly be attributed by the unique factors. Since unique factors for different groups are uncorrelated, this gives a strong correlation between variables in the same cluster and a weak correlation between variables in different groups. In this case, it is easy to estimate correct cluster assignments and a small lower bound does not require huge sample size T or dimension p. Conversely, if the signal-noise ratio θ is small, that is, the covariance of variables is mainly attributed by the common factors, a correct estimation of cluster assignments requires a large sample size T and dimension p.

4.3.2 Difficulty of Clustering Recovery for Approximate G-block Model and Integrative Group Factor Model

Similar results of minimax lower bound of clustering recovery rate for approximate G-block model (4.2.10) are given by Bunea et al. (2020). However, the author derived the lower bound based on different loss function, which considers label switching as errors. Thus, the minimax lower bound cannot be compared with our result directly. To compared approximate G-block model in Bunea et al. (2020) and integrative group factor model, we consider the minimax lower bound of clustering recovery rate based on the loss function in (4.3.1) for approximate G-block model. We apply similar proof as Theorem 4.3.1 with the same covering to give the following corollary.

Corollary 4.3.1. *Let $\mathcal{Z}$ be the set of all labels $(z_1, \ldots, z_p) \in \{1, \ldots, m\}$ which satisfy Condition 4.2.1 and $\mathcal{V}$ be the set of $m \times m$ symmetric matrices with positive diagonal entries. Further, we let a be the smallest diagonal element and b be the largest off-diagonal element of matrices in $\mathcal{V}$. For model* (4.2.10)*, under Conditions 4.2.1, 4.2.5, 4.2.6 and 4.2.7, for some $\varepsilon \in (\log(2)/\log(p/4), 1)$, we have*

$$\inf_{\widehat{\boldsymbol{z}}} \sup_{\boldsymbol{z}\in\mathcal{Z}, \mathbf{V}\in\mathcal{V}} \mathbb{E}\{L(\widehat{\boldsymbol{z}}, \boldsymbol{z})\} \geqslant \frac{(b-a)\{\varepsilon \log(p/4) - \log(2)\}(1-\varepsilon)m}{16pT} \wedge \frac{1}{64}, \tag{4.3.4}$$

where the infimum is taken over all label estimators $\widehat{\boldsymbol{z}}$.

Similar as Theorem 4.3.1, for $p > 8$, the maximum of (4.3.4) about ε is achieved by letting $\varepsilon = \{2\log(p/4)\}^{-1}\log(p/2)$, we have

$$\inf_{\widehat{\boldsymbol{z}}} \sup_{\boldsymbol{z}\in\mathcal{Z}, \mathbf{V}\in\mathcal{V}} \mathbb{E}\{L(\widehat{\boldsymbol{z}}, \boldsymbol{z})\} \geqslant \frac{\{\log^2(p/4) + \log^2(2)\}m}{32(b-a)p\log(p/4)T} \wedge \frac{1}{64}, \tag{4.3.5}$$

where the infimum is taken over all label estimators $\widehat{\boldsymbol{z}}$. It can be seen that the signal-noise ratio is different for approximate G-block model and integrative group factor model. Since signal-noise ratio of integrative group factor model is more complicate than that of approximate G-block model,

the parameter of integrative group factor model is more complicated. However, the convergence rates of their minimax lower bound of clustering recovery rate are the same, so the two models have the same difficulty of clustering recovery.

4.4 Methodology for Clustering

Next, we consider detecting the latent cluster assignments. By (4.2.3) and (4.2.4), it is easy to see that the covariance structure of the observation process y_{it} is given by the cluster assignments. Thus, it is straightforward to estimate the cluster assignments by the covariance structure. However, for high dimensional data or temporally dependent data, the covariance structure cannot be consistently estimated without further assumption. On the other hand, recall that in Section 4.2, we define two curves of time series in different groups if they are uncorrelated conditional on some factors, which are the common factors in model (4.2.1). That is, if two curves of time series are loaded on the same factors, they are in the same cluster. Conversely, if there exist some factors that either of the two curves is not loaded on, they are in different groups. Thus, the cluster assignment is given by the structure of the loading matrix $\mathbf{C}$ as shown in (4.2.5). Unlike the covariance structure, the loading matrix is consistently estimated for high dimensional data with weak temporal dependence (Bai and Ng, 2013; Fan et al., 2016). Thus, we start from estimating latent factors and loadings.

4.4.1 Estimating Latent Factors and Loadings

Note that for the $T \times T$ matrix $\mathbf{Y}^\top \mathbf{Y}$, the eigenvectors corresponding to the K largest eigenvalues are approximately the same direction as $\boldsymbol{f}_k$, the column vectors of $\mathbf{F}$. Therefore, the spectral decomposition of $\mathbf{Y}^\top \mathbf{Y}$ can be investigated using the estimates to latent factor process and loading matrix in (4.2.9) and we employ principal component analysis (PCA) approach (Bai and Ng, 2013; Fan et al., 2016) to estimate the factors and loadings. Here, the latent factors and loadings can be easily estimated via matrix eigen-decomposition. Particularly, we let $T^{-1/2}\widehat{\boldsymbol{v}}_k$ be the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{Y}^\top \mathbf{Y}$ for $k = 1, \ldots, K$. Then,

the common factors $\mathbf{F}_0$ is estimated by $\widehat{\mathbf{F}}_0 = (\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_{r_0})$. Consequently, the common loading matrix $\mathbf{A} = (\mathbf{A}_1^\top, \ldots, \mathbf{A}_m^\top)^\top$ is estimated by right projecting $T^{-1}\mathbf{Y}$ onto the estimated $\mathbf{F}_0$, i.e. $\widehat{\mathbf{A}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}_0$. To estimate the unique factors and loadings, we consider two cases, where the cluster assignments are known or unknown. With the information of cluster assignments, we can then estimate the unique factors and their loadings for each group. For each $j = 1, \ldots, m$, let $\mathbf{Y}_j = (\boldsymbol{y}_1^{(j)}, \ldots, \boldsymbol{y}_T^{(j)})$ be the data matrix of group j. Then, we estimate $\mathbf{F}_j$ and $\mathbf{B}_j$ by PCA approach again. We let $T^{-1/2}\widehat{\boldsymbol{w}}_k^{(j)}$ be the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{Y}_j^\top \mathbf{Y}_j$ for $k = r_0 + 1, \ldots, r_0 + r_j$. Then, $\mathbf{F}_j$, the unique factors for group j, is estimated by $\widehat{\mathbf{F}}_j = (\widehat{\boldsymbol{w}}_1^{(j)}, \ldots \widehat{\boldsymbol{w}}_{r_j}^{(j)})$. Consequently, the corresponding loading matrix $\mathbf{B}_j$ is estimated by $\widehat{\mathbf{B}}_j = T^{-1}\mathbf{Y}_j\widehat{\mathbf{F}}_j$. If the cluster assignments are unknown, we first estimate all factors together and then eliminate the common factors. All factors $\mathbf{F}$ is estimated by $\widehat{\mathbf{F}} = (\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_K)$ and their loading matrix $\mathbf{C}$ is estimated by $\widehat{\mathbf{C}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}$. Similar as Bai and Ng (2013) and Fan et al. (2016), statistical guarantee of estimating latent factors and loading are given in the following theorems in terms of mean squared errors of the estimating procedure.

Theorem 4.4.1 (First moment of $\widehat{\mathbf{F}}_0$ and $\widehat{\mathbf{A}}$)**.** *Under Conditions 4.2.1, 4.2.5, 4.2.6 and 4.2.7, denoting* $d_A^2 = \min_j \lambda_{\min}(p^{-1}\mathbf{A}_j^\top \mathbf{A}_j)$*, with probability at least* $1 - e^{-s}$*,*

$$\begin{aligned}\frac{1}{T}\|\widehat{\mathbf{F}}_0 - \mathbf{F}_0\|_{\mathbb{F}}^2 \lesssim & \frac{D_A^2}{d_A^2}\left(\frac{1}{p} + \frac{1}{T}\right)s^3,\\ \frac{1}{p}\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathbb{F}}^2 \lesssim & \frac{D_A^2}{d_A^2}\left(\frac{1}{p} + \frac{1}{T}\right)s^4.\end{aligned}$$

For the unique factors and loadings, similar results of convergence as Theorem 4.4.1 are given in the following theorem.

Theorem 4.4.2 (First moment of $\widehat{\mathbf{F}}_j$ and $\widehat{\mathbf{B}}_j$)**.** *Under Conditions 4.2.1, 4.2.5, 4.2.6 and 4.2.7, denoting* $d_{B_j}^2 = \lambda_{\min}(p^{-1}\mathbf{B}_j^\top \mathbf{B}_j)$ *for each* $j = 1, \ldots, m$*, with probability at least* $1 - e^{-s}$*, for each* $j = 1, \ldots, m$*,*

$$\frac{1}{T}\|\widehat{\mathbf{F}}_j - \mathbf{F}_j\|_{\mathbb{F}}^2 \lesssim \log(m)\frac{D_A^2 r_0}{d_{B_j}^2 r_j}\left(\frac{m}{p} + \frac{1}{T}\right)s^3,$$

$$\frac{1}{p}\|\widehat{\mathbf{B}}_j - \mathbf{B}_j\|_{\mathbb{F}}^2 \lesssim \log(m)\frac{D_A^2 r_0}{d_{B_j}^2 r_j}\left(\frac{m}{p}+\frac{1}{T}\right)s^4.$$

For all factors and loadings, similar as Theorem 4.4.1, we have the following guarantee.

Theorem 4.4.3 (First moment of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{C}}$)**.** *Under Conditions 4.2.1, 4.2.5, 4.2.6 and 4.2.7, with probability at least* $1-e^{-s}$*,*

$$\frac{1}{T}\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2 r_0}{d_B^2 \min_j r_j}\left(\frac{m}{p}+\frac{1}{T}\right)s^3,$$
$$\frac{1}{p}\|\widehat{\mathbf{C}} - \mathbf{C}\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2 r_0}{d_B^2 \min_j r_j}\left(\frac{m}{p}+\frac{1}{T}\right)s^4.$$

In contrast to the known asymptotic properties of estimating latent factor matrix and loading matrix for traditional and semiparametric factor models with divergent number of variables p and time points T (Bai and Ng, 2013; Fan et al., 2016),Theorems 4.4.1 to 4.4.3 provides similar results as estimating latent factors and loadings for approximate factor model, both asymptotically as in Bai and Ng (2013) and Fan et al. (2016) and non-asymptotically as in Zhang et al. (2019) and Zhang et al. (2020). Since D_A^2, d_A^2, d_B^2, r_0, $\min_j r_j$ and $\max_j r_j$ are constant, they are always omitted in previous results. As shown in Theorem 4.3.1, these constants play important roles in the possibility of estimating the cluster assignments. Thus, it is direct to think that the constants also play roles in the statistical guarantee of estimating the latent factors and loadings as well as the cluster assignments. Therefore, unlike previous results, we keep the constants to see the relationship among the constants, dimension p and sample size T. In Theorem 4.4.1, the result is established under a weaker condition on $\mathbf{F}_0$ compared to Condition PC1 in Bai and Ng (2013) as discussed in Section 4.2.1. Besides the deviation between $\mathbf{F}_0$ and its projection onto subspace $\{\mathbf{F} \in \mathbb{R}^{T\times K} : T^{-1}\mathbf{F}_0'\mathbf{F}_0 = \mathbf{I}_{r_0}\}$, which is of rate p^{-1} as given in Bai and Ng (2013),Fan et al. (2016) and Zhang et al. (2019), we also account for the error for estimating this projection, which is of rate $p^{-2}+T^{-1}$. This leads to the slower convergence rate on $T^{-1}\|\widehat{\mathbf{F}}_0 - \mathbf{F}_0\|_{\mathbb{F}}^2$. Similar results can be seen in Zhang et al. (2020).

4.4.2 Data Driven Recovery of Clustering Assignments

Next, we will provide an estimator of the cluster assignments based on the estimator of loading matrix in Section 4.4.1, and show the statistical guarantee based on Theorem 4.4.3. First, we consider a block diagonal matrix $\mathbf{B} = \mathrm{diag}(\mathbf{B}_1, \ldots, \mathbf{B}_m)$, which is the (r_0+1)th to the Kth columns of $\mathbf{C} = \{c_{ik}\}_{i=1,k=1}^{p,K}$, and its estimate $\widehat{\mathbf{B}}$, which is the (r_0+1)th to the Kth columns of $\widehat{\mathbf{C}} = \{\widehat{c}_{ik}\}_{i=1,k=1}^{p,K}$. Denote $\mathbf{B} = \{b_{ik}\}_{i=1,k=1}^{p,K-r_0}$, $\widehat{\mathbf{B}} = \{\widehat{b}_{ik}\}_{i=1,k=1}^{p,K-r_0}$ and $\mathbf{E} = \{e_{ik}\}_{i=1,k=1}^{p,K-r_0}$. Note that we are dealing with the possible range for each element of $\widehat{\mathbf{B}}$ simultaneously. Thus, unlike Theorems 4.4.1 to 4.4.3, to start with, we use maximum norm to quantify the deviation of $\widehat{\mathbf{C}}$ from $\mathbf{C}$ in the following corollary.

Corollary 4.4.1. *Conditions 4.2.1, 4.2.5, 4.2.6 and 4.2.7, with probability at least* $1 - 10e^{-s}$*,*

(i) $\|\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2\|_{\max} \lesssim D_A\sqrt{r_0}d_B^{-1}(\min_j r_j)^{-1/2}p^{-1/2}\{\log(T)\}^{2/r_2}s$;

(ii) $\|\widehat{\mathbf{C}} - \mathbf{C}\mathbf{H}_2^{-1}\|_{\max} \lesssim D_A\sqrt{r_0}d_B^{-1}(\min_j r_j)^{-1/2}(T^{-1/2}\sqrt{\log(p)}s$;

(iii) $\|\widehat{\mathbf{C}} - \mathbf{C}\|_{\max} \lesssim D_A\sqrt{r_0}d_B^{-1}(\min_j r_j)^{-1/2}(T^{-1/2}\sqrt{\log(p)} + p^{-1/2})s$.

Different from Theorem 4.4.3, Corollary 4.4.1 provides the union maximum bound of each element of $\widehat{\mathbf{C}} - \mathbf{C}$. Similar results can be seen in Wang and Fan (2017), Barigozzi et al. (2018) and Zhang et al. (2019). However, since D_A, d_B, r_0 and $\min_j r_j$ are constant, they are also omitted in previous results. Similar as Theorems 4.4.1 to 4.4.3, we keep the constants to see the relationship among the constants, dimension p and sample size T. Then, we apply an estimator of cluster assignments based on $\widehat{\mathbf{B}}$. In following algorithm, we first apply thresholding on each element of $\widehat{\mathbf{B}}$ based on their large deviations given in Corollary 4.4.1, and get a matrix $\widehat{\mathbf{I}}$ consisting of only 0 and 1. Then, $\widehat{\boldsymbol{z}}$, the estimator of cluster assignments, is given by row-wise screening of $\widehat{\mathbf{I}}$.

Algorithm 4 Cluster Assignments Estimation

Input: $p \times (K - r_0)$ matrix $\widehat{\mathbf{B}} =: \{\widehat{b}_{ik}\}_{i,k=1}^{p,K-r_o}$ and pre-determined δ.

1: Let $\tau = \delta \log\{m^{-1}\log^{-1}(p)pT\}(T^{-1/2}\sqrt{\log(p)} + p^{-1/2})$. For $k = 1, \ldots, K - r_0$, let $\widehat{\boldsymbol{i}}_k = (\mathcal{I}(|\widehat{b}_{1k}| > \tau), \ldots, \mathcal{I}(|\widehat{b}_{pk}| > \tau)^\top$ and $\widehat{\mathbf{I}} = (\widehat{\boldsymbol{i}}_1, \ldots, \widehat{\boldsymbol{i}}_{K-r_0}) =: \{\widehat{i}_{ik}\}_{i=1,k=1}^{p,K-r_0}$. Denote the rows of $\widehat{\mathbf{I}}$ as $\widehat{\mathbf{i}}_1, \ldots, \widehat{\mathbf{i}}_p$ and the kth element of $\widehat{\mathbf{i}}_1$ as $\widehat{\mathbf{i}}_1(k)$.

2: If there exists k such that $\widehat{\mathbf{i}}_1(k) = \widehat{\mathbf{i}}_2(k) = 1$, let $\widehat{z}_1 = \widehat{z}_2 = 1$ and $\widehat{\mathbf{i}}_{(1)} = \widehat{\mathbf{i}}_1 + \widehat{\mathbf{i}}_2$. Else, let $\widehat{z}_1 = 1$, $\widehat{z}_2 = 2$, $\widehat{\mathbf{i}}_{(1)} = \widehat{\mathbf{i}}_1$ and $\widehat{\mathbf{i}}_{(2)} = \widehat{\mathbf{i}}_2$.

3: If there exists j and k such that $\widehat{\mathbf{i}}_3(k) \neq 0$ and $\widehat{\mathbf{i}}_{(j)}(k) \neq 0$, let $\widehat{z}_3 = j$ and $\widehat{\mathbf{i}}_{(j)} = \sum_{i:z_i=j} \widehat{\mathbf{i}}_i$. Else, let $\widehat{z}_3 = \max_{i<3} \widehat{z}_i + 1$ and $\widehat{\mathbf{i}}_{(\widehat{z}_3)} = \widehat{\mathbf{i}}_3$.

4: Repeat Step 3 for $i = 4, \ldots, p$.

Output: Cluster assignments estimator $\widehat{\boldsymbol{z}} = (\widehat{z}_1, \ldots, \widehat{z}_p)^\top$.

In Algorithm 4, the constant δ reflects the balance of the difference between two $0-1$ matrices, $\widehat{\mathbf{I}} = (\widehat{\boldsymbol{i}}_1, \ldots, \widehat{\boldsymbol{i}}_{K-r_0})$ and $\mathbf{I} = (\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{K-r_0})$, where $\boldsymbol{i}_k = (\mathcal{I}(b_{1k} \neq 0), \ldots, \mathcal{I}(b_{pk} \neq 0))^\top$ and $\mathbf{B} = \{b_{ik}\}_{i=1,k=1}^{p,K-r_0}$ is the $(r_0 + 1)$th to the Kth columns of $\mathbf{C}$. A big δ will decrease $\mathbb{P}(|\widehat{b}_{ik}| > \tau | b_{ik} \neq 0)$ and increase $\mathbb{P}(|\widehat{b}_{ik}| < \tau | b_{ik} = 0)$. If we want to prevent two curves of time series in the same cluster from being mistakenly assigned in different groups, we will choose a big δ. Conversely, a small δ prevents two curves of time series in different groups from being mistakenly assigned in the same cluster. Since both probability converge to zero as p and T go to infinity, the choice of δ will not affect the convergence of Algorithm 4. Thus, in practice, we can choose $\delta = 1$. Algorithm 4 requires the pairwise comparisons of p vectors with dimension $K - r_0$ so the computation complexity is $O(pK)$. We show the statistical guarantee of $\widehat{\boldsymbol{z}}$ by the following theorem.

Theorem 4.4.4. *For model* (4.2.8)*, under Conditions 4.2.1, 4.2.5, 4.2.6 and 4.2.7, the* 0-1 *loss function in* (4.3.1) *for* $\widehat{\boldsymbol{z}}$ *defined in Algorithm 4 satisfy that,*

$$\mathbb{E}\{L(\widehat{\boldsymbol{z}}, \boldsymbol{z})\} \leqslant \frac{10 \exp\{-(D_A\sqrt{r_0})^{-1} d_B \sqrt{\min_j r_j} \min_{b_{ik}\neq 0} |b_{ik}|\} m \log(p)}{pT}.$$

It is easy to see that if we apply the estimator in Algorithm 4 upon the oracle loading matrix, that is, the (r_0+1)th to the Kth rows of $\mathbf{C}$, under Condition 4.2.6, the estimator will be exactly the true cluster assignments up to label switch. Consequentially, in Theorem 4.4.4, the error rate of estimating cluster assignments is given by the difference between two $0-1$ matrices, $\widehat{\mathbf{I}} = (\widehat{\boldsymbol{i}}_1, \ldots, \widehat{\boldsymbol{i}}_{K-r_0})$ and $\mathbf{I} = (\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{K-r_0})$, where $\boldsymbol{i}_k = (\mathcal{I}(b_{1k} \neq 0), \ldots, \mathcal{I}(b_{pk} \neq 0))^\top$ and $\mathbf{B} = \{b_{ik}\}_{i=1,k=1}^{p,K-r_0}$ is the (r_0+1)th to the Kth columns of $\mathbf{C}$. Thus, the error rate of estimating cluster assignments is given by the error rate of estimating loading matrix. In the right hand side of the inequality in Theorem 4.4.4, the rate $\log(p)mp^{-1}T^{-1}$, reflects the deviation of $\widehat{\mathbf{B}}$ being the same as $\mathbf{B}$, and the constant $10\exp\{-(D_A\sqrt{r_0})^{-1} d_B \sqrt{\min_j r_j} \min_{b_{ik}\neq 0} |b_{ik}|$ in the error rate reflects the tolerance of $\widehat{\mathbf{B}}$ deviating from $\mathbf{B}$.

4.4.3 Upper Bound of Group Recovery and Optimality

In addition, Theorem 4.4.4 gives an upper bound of miss-clustering error with the same rate as the minimax lower bound in Theorem 4.3.1 up to a constant. Thus, the cluster assignment estimator $\widehat{\boldsymbol{z}}$ in Algorithm 4 is the optimal cluster assignment estimator.

The signal-noise ratio θ in Theorem 4.3.1 can be written as

$$\theta = \frac{D_A^{-2} d_B^2 \max_j r_j}{1 + D_A^{-2} r_0^{-1} d_B^2 \min_j r_j} := \frac{\zeta_1}{1+\zeta_2},$$

where $\zeta_1 = D_A^{-2} d_B^2 \max_j r_j$ and $\zeta_2 = D_A^{-2} r_0^{-1} d_B^2 \min_j r_j$. Here, ζ_1 quantifies the smallest signal of unique factors in the largest cluster compared with the largest signal of common factors and ζ_2 quantifies the smallest signal of unique factors in the smallest cluster compared with the largest signal of common factors. It is easy to see that θ is monotone increasing with respect to ζ_1 and

monotone decreasing with respect to ζ_2. That is, the signal-noise ratio is greater if the unique factors in the largest cluster is separable form the common factors. Also, in Theorem 4.3.1 , if $\theta = O(\log(p)mp^{-1}T^{-1})$, the the lower bound in (4.3.3) is lower bounded by a constant. Thus, for ζ_1 and ζ_2 satisfying that $\theta = O(\log(p)mp^{-1}T^{-1})$, it is not strongly consistent to estimate cluster assignments. Theorem 4.3.1 indicates the necessity of separation conditions in estimating cluster assignments. If ζ_1 is small or ζ_2 is large, the signal from common factors is so strong compared with the signal from unique factors that it is impossible to identify unique factors or separate groups based on unique factors. Combining Theorem 4.4.4 and 4.3.1, we have the region of $D_A^{-2}d_B^2 \max_j r_j$ and $D_A^{-2}(\max_j r_j)^{-1}d_B^2 r_0$ for the possibility and guarantee of estimating cluster assignments as shown in Figure 4.1. In Figure 4.1, the bottom right region of impossibility is the region of ζ_1 and ζ_2 that $\theta = O(\log(p)mp^{-1}T^{-1})$ and thus the the lower bound in Theorem 4.3.1 is lower bounded by a constant. The top left region of possibility and guarantee is the region that both the upper bound in Theorem 4.4.4 lower bound in Theorem 4.3.1 converge to zero as p and T go to infinity. The bottom left region of possibility and unknown guarantee indicates the case that it is possible to estimate cluster assignments but our method in Algorithm 4 may not work. In general, if d_B^2 and $\max_j r_j$ are large, and D_A^2 and r_0 are small, the signal of unique factors for each cluster is strong compared with the signal of common factors, so it is possible and guaranteed to estimate cluster assignments.

4.4.4 Upper Bound of Clustering Recovery for COD

For model-based covariance-type variable clustering, Bunea et al. (2020) proposed covariance difference (COD) algorithm based on scaled covariance difference (sCOD)

$$\mathrm{sCOD}(i,j) = \max_{\ell \neq i,j} \left| \frac{\mathrm{cov}(y_i - y_j, y_\ell)}{\sqrt{\mathrm{Var}(y_i - y_j)\,\mathrm{Var}(y_\ell)}} \right|$$

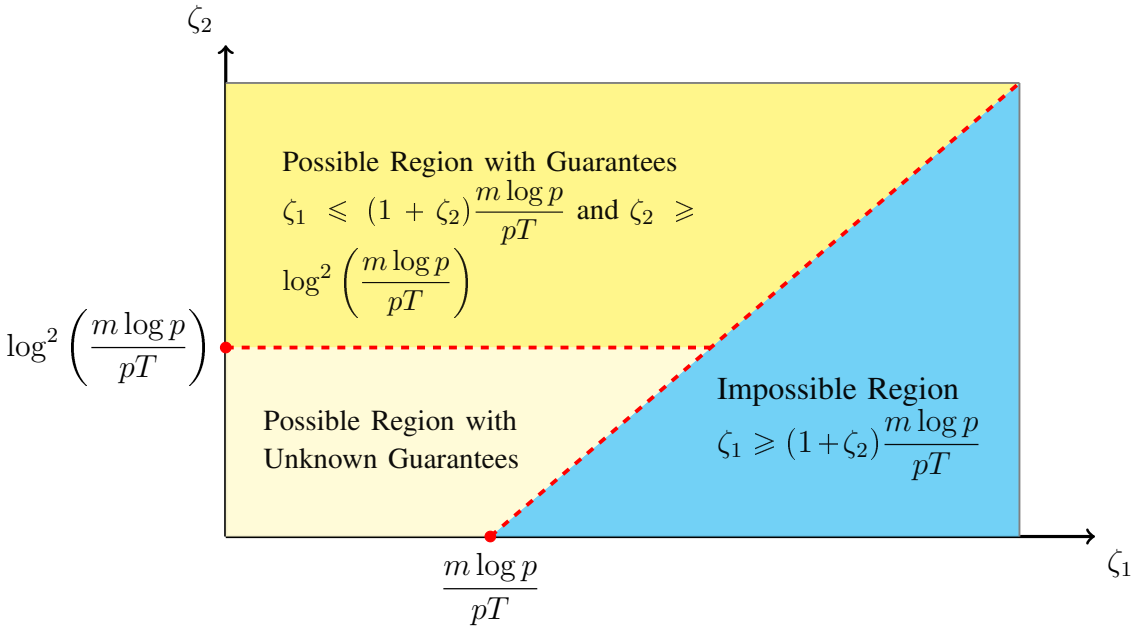

Figure 4.1: Region for possibility and guarantee of estimating cluster assignments. It is possible to estimate cluster assignments if $\zeta_1/(1+\zeta_2) \leqslant p^{-1}T^{-1}m\log(p)$ and guaranteed if $\exp(\sqrt{\zeta_2}) \geqslant p^{-1}T^{-1}m\log(p)$ where $\zeta_1 = D_A^{-2}d_B^2 \max_j r_j$ and $\zeta_2 = D_A^{-2}(\min_j r_j)^{-1}d_B^2 r_0$.

and its estimator

$$\widehat{\mathrm{sCOD}}(i,j) = \max_{\ell \neq i,j} \left| \frac{\hat{\mathbf{\Sigma}}_{i\ell} - \hat{\mathbf{\Sigma}}_{j\ell}}{\sqrt{(\hat{\mathbf{\Sigma}}_{ii} + \hat{\mathbf{\Sigma}}_{jj} - 2\hat{\mathbf{\Sigma}}_{ij})\hat{\mathbf{\Sigma}}_{\ell\ell}}} \right|,$$

where $\hat{\mathbf{\Sigma}}$ is an estimator of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}_{ij}$ is the element of $\hat{\mathbf{\Sigma}}$ in the ith row and jth column. Then, the clustering assignments are estimated by the following algorithm.

Algorithm 5 COD algorithm

Input: $p \times p$ matrix $\widehat{\mathbf{\Sigma}}$ and pre-determined δ.

1: Initialize $\mathcal{S} = \{1, \ldots, p\}$ and $k = 1$.

2: If $|\mathcal{S}| = 1$, let $\widehat{G}_k = \mathcal{S}$. If $|\mathcal{S}| > 1$, let $(i_k, j_k) = \text{argmin}_{i,j \in \mathcal{S}, i \neq j} \widehat{\text{sCOD}}(i, j)$. If $\widehat{\text{sCOD}}(i_k, j_k) > \delta$, let $\widehat{G}_k = i_k$. If $\widehat{\text{sCOD}}(i_k, j_k) \leqslant \delta$, let $\widehat{G}_k = \{\ell \in \mathcal{S} : \widehat{\text{sCOD}}(i_k, \ell) \wedge \widehat{\text{sCOD}}(j_k, \ell) \leqslant \delta\}$. Replace $\mathcal{S}$ by $\mathcal{S} \backslash \widehat{G}_k$.

3: Replace k by $k + 1$ and repeat Step 2.

4: For each $i = 1, \ldots, p$, let $\widehat{z}_i = k$ if $i \in \widehat{G}_k$.

Output: Cluster assignments estimator $\widehat{\boldsymbol{z}} = (\widehat{z}_1, \ldots, \widehat{z}_p)^\top$.

Since COD algorithm is highly dependent on covariance matrix estimation, which may fail to converge for high-dimensional dependent data, we consider estimating $\mathbf{\Sigma}$ through estimated loading matrix $\widehat{\mathbf{C}}$, that is

$$\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top + \widehat{\sigma}_u^2 \mathbf{I}.$$

To show the properties of COD algorithm for integrative group factor model, first, we consider a special case, where $r_0 = r_1 = \cdots = r_j = 1$, $\mathbf{A}_j = a_j \mathbf{1}_{p_j}$ and $\mathbf{B}_j = b_j \mathbf{1}_{p_j}$ for some a_j and b_j and each $j = 1, \ldots, m$. Then, we slightly change the special case to a second case where $r_0 = r_1 = \cdots = r_j = 1$, but $\mathbf{A}_j = \mathbf{0}$ for each $j = 1, \ldots, m$, $\mathbf{B}_j = b_j \mathbf{1}_{p_j}$ for some b_j and each $j \geqslant 2$ and $\mathbf{B}_1 = (1, \ldots, p_1)^\top$. In both cases, we provide properties of COD algorithm in the following corollary.

Corollary 4.4.2. *(i) For model* (4.2.10)*, if* $r_0 = r_1 = \cdots = r_j = 1$*,* $\mathbf{A}_j = a_j \mathbf{1}_{p_j}$ *and* $\mathbf{B}_j = b_j \mathbf{1}_{p_j}$ *for some* a_j *and* b_j *and each* $j = 1, \ldots, m$*, the* 0-1 *loss function in* (4.3.1) *for* $\widehat{\boldsymbol{z}}$ *defined in*

Algorithm 5 using satisfy that,

$$\mathbb{E}\{L(\widehat{\boldsymbol{z}}, \boldsymbol{z})\} \leqslant \frac{10\exp\{-(D_A\sqrt{r_0})^{-1}d_B\sqrt{\min_j r_j}\max_{i,k}|c_{ik}|\}m\log(p)}{pT}.$$

(ii) For model (4.2.10)*, if* $r_0 = r_1 = \cdots = r_j = 1$*,* $\mathbf{A}_j = \mathbf{0}$ *for each* $j = 1, \ldots, m$*,* $\mathbf{B}_j = b_j\mathbf{1}_{p_j}$ *for some* b_j *and each* $j \geqslant 2$ *and* $\mathbf{B}_1 = (1, \ldots, p_1)^\top$*, the* 0-1 *loss function in* (4.3.1) *for* $\widehat{\boldsymbol{z}}$ *defined in Algorithm 5 using oracle* $\boldsymbol{\Sigma}$ *satisfy that, for some positive constant* C*,*

$$\mathbb{E}\{L(\widehat{\boldsymbol{z}}, \boldsymbol{z})\} \geqslant \frac{C}{m}.$$

Corollary 4.4.2 (i) shows the convergence of COD algorithm similar to Algorithm 4. Although both algorithms are minimax optimal, Algorithm 4 is free from covariance estimation, and thus more flexible to high-dimensional dependent data. However, note that m is a fixed constant here. Thus, the lower bound of clustering recovery rate in Corollary 4.4.2(ii) does not converge to 0 even using oracle covariance matrix $\boldsymbol{\Sigma}$. Compared with the conditions of Corollary 4.4.2(i), we can see that COD algorithm does not work even when we only change loadings for factors in a single cluster, which shows that, the convergence of COD algorithm is restricted for integrative group factor model. When the variances of variables in the same cluster are allowed to be different, COD algorithm may not work.

4.4.5 Determining Number of Factors

In practice, the number of factors is always unknown and it is necessary to choose the number of latent factors, both common and unique, before estimating the latent factors and their loadings and apply the estimator of loading matrix upon Algorithm 4 to estimate cluster assignments. Traditional methods to estimate K include, for example, the likelihood ratio test and the screen plot (Jolliffe, 2002). For the high-dimensional data with large covariance matrix, eigenvalues of the sample covariance matrix or their variants have been utilized and the estimation is consistent under certain separation condition of the first K eigenvalues from the remains. A popular approach is

based on the ratio of consecutive eigenvalues (Ahn and Horenstein, 2013; Fan et al., 2016; Lam and Yao, 2012). Let $\widehat{\lambda}_k$ be the kth largest eigenvalue of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$. Similar as approximate factor model, by Condition 4.2.7, the first K largest eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ will diverge with rate p, the rest stays constant if $p = O(T)$ and diverge with rate p/T if $T = o(p)$. Thus, the ratio of the Kth and $(K+1)$th eigenvalues $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ will diverge with rate $\min(p, T)$ as p and T go to infinity, while the other eigenvalue-ratios stay constant. Thus, the total number of factors K is determined by the largest gap between eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ as suggested by Lam and Yao (2012), Ahn and Horenstein (2013) and Fan et al. (2016). Unlike usual factor models, we still need to determine the number of common factors r_0 to consistently estimate the common factors and their loadings. As discussed above, the ratio of the r_0th and (r_0+1)th eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ will not diverge as p and T go to infinity. However, by Condition 4.2.7, the ratio of the r_0th and (r_0+1)th eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ is greater than other ratios among the first K eigenvalues, so the largest gap among the largest K eigenvalues can be used to estimate r_0. Since in practice, K is always unknown, we replace it by $\widehat{K}$ given by the discussion before. That is, instead of choosing the largest gap between eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ as suggested by Lam and Yao (2012), Ahn and Horenstein (2013) and Fan et al. (2016), we choose two largest gaps, and define

$$\begin{aligned}\widehat{K} &= \operatorname{argmax}_{1\leqslant k<\min(p,T)} \frac{\widehat{\lambda}_k}{\widehat{\lambda}_{k+1}},\\ \widehat{r}_0 &= \operatorname{argmax}_{1\leqslant k<\widehat{K}} \frac{\widehat{\lambda}_k}{\widehat{\lambda}_{k+1}}.\end{aligned} \tag{4.4.1}$$

Then, we provide the following statistical guarantee of $\widehat{K}$ and $\widehat{r}_0$.

Theorem 4.4.5. *Under Conditions 4.2.1, 4.2.5, 4.2.6 and 4.2.7, we have*

$$\begin{aligned}\mathbb{P}(\widehat{K} = K) &\geqslant 1 - 2\exp\left[-C_3\left\{\sqrt{\max(p,T)} - C_4\sqrt{\min(p,T)}\right\}^2\right],\\ \mathbb{P}(\widehat{r}_0 = r_0) &\geqslant 1 - 2\exp\left[-C_1\left\{\sqrt{\max(p,T)} - C_2\sqrt{\min(p,T)}\right\}^2\right],\end{aligned}$$

where C_1, C_2, C_3 and C_4 are positive constants.

Theorem 4.4.5 gives the non-asymptotic properties in estimating r_0 and K. As mentioned before, the first K largest eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ will diverge with rate p. Thus, in Theorem 4.4.5, the error rate of $\widehat{K}$ and $\widehat{r}_0$ corresponds to p. By Theorem 4.4.5, the error rates of estimating K and $\widehat{r}_0$ is large when p are much larger or smaller than T. Thus, eigenvalue-ratio estimator 4.5.3 works well in low dimension case ($p \ll T$) and ultra high dimension case ($p \gg T$). In the case where p is close to T, the error rate of estimating K and $\widehat{r}_0$ may not converge, which has been pointed out before by Bai and Yin (1993). In addition, with Theorem 4.4.5, we will assume that r_0 and K are known when estimating the latent factors, loadings and cluster assignments. Otherwise, we will use $\widehat{r}_0$ and $\widehat{K}$ given above and all results are conditional on the event $\{\widehat{r}_0 = r_0, \widehat{K} = K\}$.

Although the number of factors can be properly estimated, the estimation to the number of groups m is still not clear. By the definition that $K = \sum_{j=0}^{m} r_j$ and consistent estimator $\widehat{K}$ and $\widehat{r}_0$, an estimated upper bound of m is given by $\widehat{K} - \widehat{r}_0$. However, since the number of unique factors in each cluster is not determined, this upper bound cannot give a consistent estimate of m. In practice, we will assume a known number of groups, or choose a large number to avoid mistakenly assigning two curves of time series in different groups in the same cluster.

4.5 The Recovery of Divergent Number of Groups

In this section, we introduce the integrative group factor model with diverging number of groups, that is, m is not a constant but diverge with respect to p. This is a different approach from traditional clustering analysis, where the number of groups m are always assumed to be finite. However, this constriction is not always satisfied. By allowing the number of groups to diverge, we enable the analysis to work for a large number of groups and a rather small number of curves, such as 10 groups of 100 curves. In this case, the number of curves within each cluster is a small term of p. Recall that, in Section 4.2, we let the curves of time series be stationary with different covariance structure, which is modeled through an approximate factor model. In the case m diverges, we still model each cluster by an approximate factor model with finite number of factors, but the total number of factors will diverge with the number of groups. Thus, by allow-

ing for diverging number of groups as well as factors, we have a larger parameter space and less information about it, which results in difficulty in estimating cluster assignments and estimating latent factors and loadings. We will show that most approaches for finite the number of groups still works for the case where the number of groups diverges, but with a smaller convergence rate. Similar as the integrative group factor model with finite number of groups, we denote a p-dimensional multivariate time series with T observations as y_{it} for $i = 1, \ldots, p$ and $t = 1, \ldots, T$ and split the p dimensions into m disjoint groups as follows:

$$\mathcal{V} = \mathcal{V}^{(1)} \cup \cdots \cup \mathcal{V}^{(m)}$$

where $\mathcal{V} = \{1, \ldots, p\}$ and m diverges with p. Let $p_j = |\mathcal{V}^{(j)}|$ be the number of curves in the kth cluster for $j = 1, \ldots, m$. Similar as before, we focus on the relatively "balanced" groups with the following condition.

Condition 4.5.1. *There exist constant $\gamma \in (0, 1)$ such that $m \asymp p^{\gamma}$ and $p_j \asymp p^{1-\gamma}$ for each $j = 1, \ldots, m$.*

Condition 4.5.1 illustrates what we mean by the balanced groups aforementioned. In Condition 4.5.1, m is proportional to p, which is not essential for the case m diverges. In fact, m can be any small term of p. However, the condition of p_j is necessary by assuming that each cluster has a size proportional to $p^{1-\gamma}$ for some constant γ between 0 and 1, that is, the size of each cluster is of the same scale. This assumption ensures the number of curves in each cluster is not too small or too large. It is easy to see that, if we assume there is no sparsity in the loading matrices $\mathbf{A}_1, \ldots, \mathbf{A}_m$ and $\mathbf{B}_1, \ldots, \mathbf{B}_m$, the number of common and unique factors are finite, and the elements of the loading matrices are constants, the strength of a factor is proportional to the number of curves loaded on the factor. Thus, by assuming each cluster has a size proportional to $p^{1-\gamma}$, we let the unique factors have the same strength, which is different from the strength of common factors. Thus, it guarantees the separation of common and unique factors and the accuracy of clustering. Similar as the integrative group factor model with finite number of groups, we propose Conditions 4.2.5

and 4.2.6 for the possibility and guarantee of recovering cluster assignments and estimating latent factors and loadings. These conditions are also essential in the case m diverges. In addition, we propose the following condition.

Condition 4.5.2. *For each $j = 1, \ldots, m$, $\mathbf{A}_j^\top \mathbf{A}_j$ and $\mathbf{B}_j^\top \mathbf{B}_j$ are diagonal matrices with non-zero distinct entries and $\mathbf{A}_j^\top \mathbf{B}_j = \mathbf{0}$. There exist constants $d_1, d_2 > 0$ such that $d_2/d_1 < m$, $d_1 \leqslant \lambda_{\min}(p^{-1+\gamma}\mathbf{A}_j^\top \mathbf{A}_j) \leqslant \lambda_{\max}(p^{-1+\gamma}\mathbf{A}_j^\top \mathbf{A}_j) \leqslant d_2$ and $d_1 \leqslant \lambda_{\min}(p^{-1+\gamma}\mathbf{B}_j^\top \mathbf{B}_j) \leqslant \lambda_{\max}(p^{-1+\gamma}\mathbf{B}_j^\top \mathbf{B}_j) \leqslant d_2$ for each $j = 1, \ldots, m$.*

Different from Condition 4.2.7, in Condition 4.5.2, we divide $\mathbf{A}_j^\top \mathbf{A}_j$ and $\mathbf{B}_j^\top \mathbf{B}_j$ by $p^{1-\gamma}$ instead of p, since by Condition 4.5.1, p_j, the number of curves in each cluster, is proportional to $p^{1-\gamma}$. Note that by Condition 4.5.2, $d_1 \leqslant \lambda_{\min}(p^{-1}\sum_{j=1}^m \mathbf{A}_j^\top \mathbf{A}_j) \leqslant \lambda_{\max}(p^{-1}\sum_{j=1}^m \mathbf{A}_j^\top \mathbf{A}_j) \leqslant d_2$, which implies that the common factors and unique factors are of different strength (p versus $p^{1-\gamma}$. Thus, unlike the case of finite m, the common factors and unique factors are distinguishable without any further conditions. Recall that each cluster only have finite number of factors, so given the cluster assignments, we can estimate the latent factors and loadings in each cluster by PCA procedure as given in Section 4.4.1. Motivated by this, similar as before, we proposed PCA procedure to estimate the latent factors and loadings, and estimate cluster assignments based on the estimation.

Similar as the case m is finite, we let $T^{-1/2}\widehat{\boldsymbol{v}}_k$ be the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{Y}^\top \mathbf{Y}$ for $k = 1, \ldots, K$. Then, the common factors $\mathbf{F}_0$ is estimated by $\widehat{\mathbf{F}}_0 = (\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_{r_0})$ and the common loading matrix $\mathbf{A} = (\mathbf{A}_1^\top, \ldots, \mathbf{A}_m^\top)^\top$ is estimated by $\widehat{\mathbf{A}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}_0$. With the information of cluster assignments, for each $j = 1, \ldots, m$, we let $\mathbf{Y}_j = (\boldsymbol{y}_1^{(j)}, \ldots, \boldsymbol{y}_T^{(j)})$ be the data matrix of group j. We let $T^{-1/2}\widehat{\boldsymbol{w}}_k^{(j)}$ be the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{Y}_j^\top \mathbf{Y}_j$ for $k = r_0 + 1, \ldots, r_0 + r_j$. Then, $\mathbf{F}_j$, the unique factors for group j, is estimated by $\widehat{\mathbf{F}}_j = (\widehat{\boldsymbol{w}}_1^{(j)}, \ldots \widehat{\boldsymbol{w}}_{r_j}^{(j)})$ and the corresponding loading matrix $\mathbf{B}_j$ is estimated by $\widehat{\mathbf{B}}_j = T^{-1}\mathbf{Y}_j\widehat{\mathbf{F}}_j$. If the cluster assignments are unknown, we first estimate all factors together and then eliminate the common factors. All factors $\mathbf{F}$ is estimated by $\widehat{\mathbf{G}} = (\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_K)$ and their loading matrix $\mathbf{C}$ is estimated by $\widehat{\mathbf{C}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}$. Similar results of statistical guarantee of

estimating latent factors and loading are given in the following theorems in terms of mean squared errors of the estimating procedure.

Theorem 4.5.1. *Under Conditions 4.5.1, 4.2.5, 4.2.6 and 4.5.2, denoting* $d_A^2 = \min_j \lambda_{\min}(p^{-1}\mathbf{A}_j^\top \mathbf{A}_j)$ *and* $d_{B_j}^2 = \lambda_{\min}(p^{-1}\mathbf{B}_j^\top \mathbf{B}_j)$ *for each* $j = 1, \ldots, m$*, with probability at least* $1 - e^{-s}$*,*

$$
\begin{aligned}
\frac{1}{T}\|\widehat{\mathbf{F}}_0 - \mathbf{F}_0\|_{\mathbb{F}}^2 \lesssim & \frac{D_A^2}{d_A^2}\left(\frac{1}{p} + \frac{1}{T}\right)s^3, \\
\frac{1}{p}\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathbb{F}}^2 \lesssim & \frac{D_A^2}{d_A^2}\left(\frac{1}{p} + \frac{1}{T}\right)s^4, \\
\frac{1}{T}\|\widehat{\mathbf{F}}_j - \mathbf{F}_j\|_{\mathbb{F}}^2 \lesssim & \frac{D_A^2 r_0 \log(m)}{d_{B_j}^2 r_j}\left(\frac{1}{p^{1-\gamma}} + \frac{1}{T}\right)s^3, \\
\frac{1}{p}\|\widehat{\mathbf{B}}_j - \mathbf{B}_j\|_{\mathbb{F}}^2 \lesssim & \frac{D_A^2 r_0 \log(m)}{d_{B_j}^2 r_j}\left(\frac{1}{p^{1-\gamma}} + \frac{1}{T}\right)s^4, \\
\frac{1}{T}\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2 \lesssim & \frac{D_A^2 r_0}{d_B^2 \min_j r_j}\left(\frac{1}{p^{1-\gamma}} + \frac{1}{T}\right)s^3, \\
\frac{1}{p}\|\widehat{\mathbf{C}} - \mathbf{C}\|_{\mathbb{F}}^2 \lesssim & \frac{D_A^2 r_0}{d_B^2 \min_j r_j}\left(\frac{1}{p^{1-\gamma}} + \frac{1}{T}\right)s^4.
\end{aligned}
$$

Note that the number of curves in each cluster is $p^{1-\gamma}$ rather than p. Thus, the convergence rate of estimating the unique factors are smaller than that for the case where m is fixed. Since the identifiability condition 4.2.5 works for both common factors and unique factors, the slower convergence rate holds for estimating unique factors and all factors simultaneous, as shown in Theorem 4.4.2 and 4.4.3. In addition, since sample size does not affect the strength of signal, the convergence rate with respect to sample size T is the same for estimating loadings of common factors and unique factors. Note that when we are estimating unique factors separately for each cluster, the dimension is smaller compared with that when estimating common factors ($p^{1-\gamma}$ versus p), the convergence rate of unique factors and loadings is smaller than that of common factors and loadings. Similarly, estimating $\mathbf{C}$ and $\mathbf{F}$ requires a PCA procedure with diverging number of factors, so the convergence rate or all factors and loadings is smaller compared with the result of common factors and loadings. Also, it is shown in Theorem 4.5.1 that the estimation error of latent

factors and loadings is large if the strongest signal is strong (D_A^2 and r_0 are large) and the weakest signal is weak (d_B^2 and $\min_j r_j$ is small). Thus, the estimation error is small if all signals are approximately of the same strength. Also, we apply Algorithm 4 to estimate cluster assignments with the following statistical guarantee.

Theorem 4.5.2. *For model* (4.2.8)*, under Conditions 4.5.1, 4.2.5, 4.2.6 and 4.5.2, the* 0-1 *loss function in* (4.3.1) *for* $\widehat{\boldsymbol{z}}$ *defined in Algorithm 4 satisfy that,*

$$\mathbb{E}\{L(\widehat{\boldsymbol{z}}(\widehat{\mathbf{I}}), \boldsymbol{z})\} \leqslant \frac{10\exp\{-(CD_A\sqrt{r_0})^{-1}d_B\sqrt{\min_j r_j}\min_{b_{ik}\neq 0}|b_{ik}|\}\log(p)}{p^{1-\gamma}T}.$$

Recall that in the case m diverges, estimating $\mathbf{C}$ and $\mathbf{F}$ requires a PCA procedure with diverging number of factors, so the convergence rate or all factors and loadings is smaller with respect to p ($p^{1-\gamma}$ versus p) compared with the result of common factors and loadings. Thus, the upper bound of estimating cluster assignment is smaller when m diverges than that when m is finite.

4.5.1 Minimax Lower Bound of Group Recovery and Optimality when m Diverges

To show the optimality of the upper bound in Theorem 4.5.2, the minimax lower bound of group recovery is given by the following theorem for the case m diverges. Similar as the techniques of Theorem 4.3.1, we first choose $p/4$ elements in $\{1,\ldots,m\}^p$ as covering and quantify the distances between each pair by the K-L divergence between the corresponding distributions. Since the covering we find in the proof of Theorem 4.3.1 does not depend on m, it works for both the case m is finite and the case m diverges. Also, we apply Le Cam's method again.

Theorem 4.5.3. *Let* $\mathcal{Z}$ *be the set of all labels* $(z_1,\ldots,z_p) \in \{1,\ldots,m\}$ *which satisfy Condition 4.5.1 and* $\mathcal{C}$ *be the set of matrices which have the form in* (4.2.5) *and satisfy Condition 4.2.7. Further, let* $D_A^2 = \max_j \lambda_{\max}(p^{-1+\gamma}\mathbf{A}_j^\top\mathbf{A}_j)$ *and* $d_B^2 = \min_j \lambda_{\min}(p^{-1+\gamma}\mathbf{B}_j^\top\mathbf{B}_j)$*. Then, the signal-noise ratio for model* (4.2.8) *is defined as* $\theta = (D_A^2 r_0 + d_B^2\min_j r_j)^{-1}d_B^2 r_0\max_j r_j$*. Under Condi-*

tions 4.5.1, 4.2.5, 4.2.6 and 4.5.2, for some $\varepsilon \in (\log(2)/\log(p/4), 1)$*, we have*

$$\inf_{\widehat{\boldsymbol{z}}} \sup_{\boldsymbol{z}\in\mathcal{Z},\mathbf{C}\in\mathcal{C}} \mathbb{E}\{L(\widehat{\boldsymbol{z}}, \boldsymbol{z})\} \geqslant \frac{\{\varepsilon \log(p/4) - \log(2)\}(1-\varepsilon)}{16\theta p^{1-\gamma}T} \wedge \frac{1}{64}, \tag{4.5.1}$$

where the infimum is taken over all label estimators $\widehat{\boldsymbol{z}}$*.*

By letting $\varepsilon = \{2\log(p/4)\}^{-1}\log(p/2)$, we have

$$\inf_{\widehat{\boldsymbol{z}}} \sup_{\boldsymbol{z}\in\mathcal{Z},\mathbf{C}\in\mathcal{C}} \mathbb{E}\{L(\widehat{\boldsymbol{z}}, \boldsymbol{z})\} \geqslant \frac{\{\log^2(p/4) + \log^2(2)\}}{32\theta\log(p/4)p^{1-\gamma}T} \wedge \frac{1}{64}, \tag{4.5.2}$$

where the infimum is taken over all label estimators $\widehat{\boldsymbol{z}}$. The possible and guaranteed region of estimating cluster assignments are given in the following figure. From Theorems 4.5.2 and 4.5.3, we can see that the lower bound and upper bound of clustering is smaller when m diverges. The change of possible and guaranteed region of clustering is shown in Figure 4.2. In Figure 4.2, the red dash line shows the regions when m is finite. As shown in Figure 4.2, when m diverges, the impossible region gets greater and the possible and guaranteed region gets smaller, which shows that it is harder to estimate cluster assignments when m diverges.

4.5.2 Determining Number of Factors when m Diverges

However, when m diverges with p, the number of all factors $K = \sum_{j=0}^{m} r_j$ also diverge, even if r_j is finite for each $j = 0, 1, \ldots, m$. Thus, although traditional methods in estimating K such as eigenvalue-ratio test (Ahn and Horenstein, 2013; Fan et al., 2016; Lam and Yao, 2012), eigenvalue-difference test (Onatski, 2012) and likelihood based test (Jolliffe, 2002) may still work for the case m diverges, their statistical guarantee is not clear. In particular, by Condition 4.2.7, the first r_0 largest eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ will diverge with rate p, the $r_0 + 1$ to K largest eigenvalues will diverge with rate $p^{1-\gamma}$, the rest stays constant if $p = O(T)$ and diverge with rate p/T if $T = o(p)$. Thus, we let $\widehat{\lambda}_k$ be the kth largest eigenvalue of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$, and define

$$\widehat{K}_1 = \text{argmax}_{1\leqslant k<\min(p,T)} \frac{\widehat{\lambda}_k}{\widehat{\lambda}_{k+1}},$$

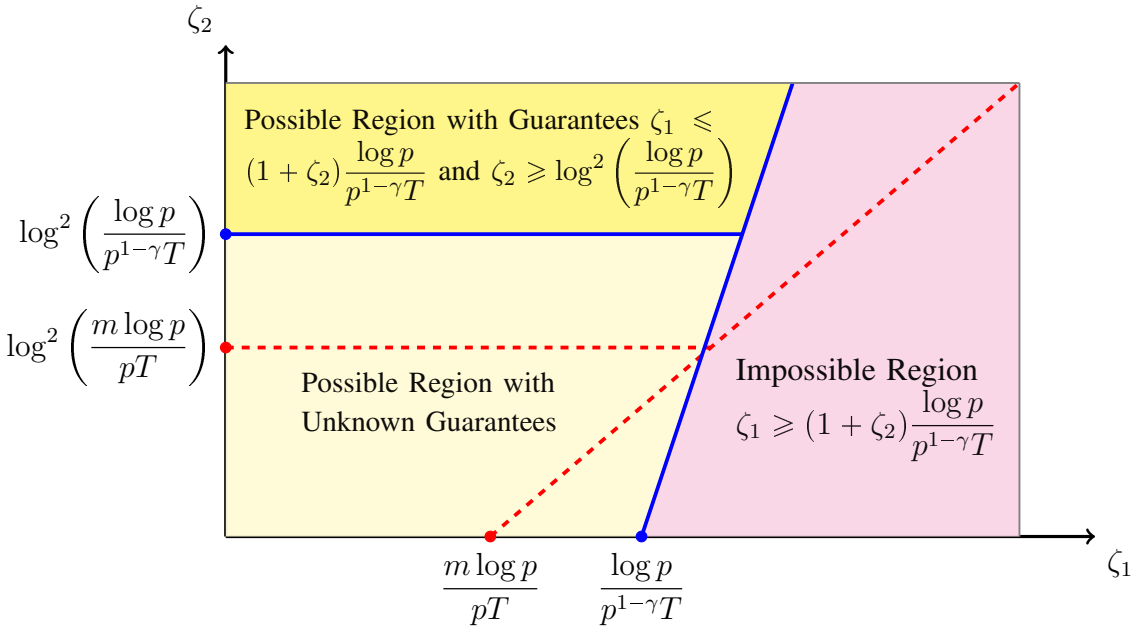

Figure 4.2: Region for possibility and guarantee of estimating cluster assignments. In the case m diverges, it is possible to estimate cluster assignments if $\zeta_1/(1+\zeta_2) \leqslant p^{-1+\gamma}T^{-1}\log(p)$ and guaranteed if $\exp(\sqrt{\zeta_2}) \geqslant p^{-1+\gamma}T^{-1}\log(p)$ where $\zeta_1 = D_A^{-2}d_B^2 \max_j r_j$ and $\zeta_2 = D_A^{-2}(\min_j r_j)^{-1}d_B^2 r_0$.

$$\widehat{K}_2 = \text{argmax}_{1\leqslant k<\min(p,T),k\neq\widehat{K}_1} \frac{\widehat{\lambda}_k}{\widehat{\lambda}_{k+1}}. \tag{4.5.3}$$

Then, r_0 and K are estimated by $\widehat{r}_0 = \min(\widehat{K}_1, \widehat{K}_2)$ and $\widehat{K} = \max(\widehat{K}_1, \widehat{K}_2)$. Alternatively, given additional condition that $\boldsymbol{f}_t$ is temporally correlated and $\boldsymbol{u}_t$ is a white noise process, we can conduct sequential test

$$H_0(K_0) : K \leqslant K_0 \quad v.s. \quad H_1(K_0) : K > K_0,$$

with K_0 some pre-specified positive integer and estimate K by

$$\widehat{K} = \text{argmin}_{K_0}\{H_0(K_0) \text{ fail to be rejected}\}.$$

The test can be given by likelihood ratio test (Jolliffe, 2002) or white noise test (Chang et al., 2017; Li et al., 2019; Zhang et al., 2019). Since we are expected to conduct K tests, a false discovery rate

(FDR) controlling procedure is applied. However, as discussed above, K diverges with respect to p, so the convergence of $\widehat{K}$ to K is not guaranteed by existing theories. Thus, in the case where m diverges, we will estimate the latent factors and loadings and estimate cluster assignments based on known number of all factors K.

4.6 Conclusions and Discussions

In this paper, we consider an integrative group factor model to do covariance-type variable clustering of a $p \times T$ data matrix, viewed as p variables with T replicates, which makes it different from stochastic block models (Gao et al., 2018; Zhang et al., 2018, 2016) where a $p \times p$ binary adjacency matrix is analyzed. Although we could have applied available clustering procedures tailored for stochastic block models to the empirical covariance matrix $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ or covariance matrix estimated by PCA procedure $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top + \widehat{\sigma}_u^2\mathbf{I}$, by treating it as some sort of weighted adjacency matrix, it turns out that applying verbatim the spectral clustering procedure would lead to poor results (Bunea et al., 2020). Also, different from mean-type model for variable clustering (Lu and Zhou, 2016), we assume the data to be zero mean and define the clustering structure based on variances and covariances of variable. Compared with existing covariance-type model for variable clustering such as approximate G-block model (Bunea et al., 2020), we relax the condition of data matrix by allowing the T replicates to be temporally dependent. More importantly, we allow the variables in the same cluster to have different variances and covariances, which gives more flexibility to the model and increases the difficulty of identifiability and estimation as well. In addition, we consider a permutation invariant loss function of clustering assignments to give a permutation invariant clustering error rate. The invariant clustering error rate is appropriate for applications in which an ordering of the variables is not available (Jin et al., 2015; Wagaman and Levina, 2009), such as genetics, social, finical and economic data

In the integrative group factor model, the commonality among clusters are modeled through some common factors. The factor structure and test for loadings of factors (Fan et al., 2016) shed a light on testing for commonality among clusters. We continue to formalize the notion of common

factors between different clusters of variables and propose to use it as a general approach to study the structure of factors. By testing on the loadings of common factors, we propose inference on the commonality among clusters, which leads to better understanding of the clustering structure. Since the latent factors integrative group factor model consist of common factors as well as unique factors, it is crucial to separate common factors from unique factors, especially when the signals of common factors are not very strong. Thus, we will first propose the test for commonality conditional on perfect separation of common factors from unique factors and then extend it to unconditional case. We will extend our work to the question in future efforts.

In addition, to demonstrate our method of clustering recovery, we will propose numerical and real-data studies to compare our propose Algorithm 4 with other variable clustering algorithms, such as *k-means* algorithm (MacQueen et al., 1967), *k-medoids* algorithm (Kaufman and Rousseeuw, 2009), hierarchical clustering method (Guha et al., 1998; Karypis et al., 1999; Zhang et al., 1996), spectral clustering method (Alzate et al., 2009; Jebara et al., 2007; Yin and Yang, 2005), group factor analysis (Klami et al., 2014), partial common PCA (Wang et al., 2019), COD and PECOK algorithm (Bunea et al., 2020). We will consider some simple cases such as the special case of both integrative group factor model and approximate G-block model given in Section 4.2.2, and some complicated cases of integrative group factor model where there are multiple factors in each cluster. Also, we will generate data from different dependence structures to show the properties of our method under temporal dependence. A real data analysis will be conducted using multinational macroeconomics indices and Fama-French series.

Chapter 5

Conclusion and Future Work

In this dissertation, we studied non-asymptotic properties of estimation and inference of large dimensional factor model and their applications to high-dimensional inference, multivariate time series, and semiparametric modeling. In Chapter 2, we carefully study the non-asymptotic properties of spectral decomposition of large Gram-type matrices based on data generated from large dimensional factor model. Specifically, we obtain the tail bound and rate of convergence of the first and second moments for deviations between the empirical and population eigenvectors to the right Gram matrix as well as the Berry-Esseen type bound to characterize the asymptotic distribution of these deviations. We also derive the non-asymptotic tail bound of the ratio of eigenvalues for the left Gram matrix, namely the sample covariance matrix, to their population counterparts regardless of the dimension and size of data matrix. The documented non-asymptotic properties are further applied to non-asymptotically characterize the property of the recovered number of latent factors in PCA, which gives the bound of error rate for finite dimension and sample size. In Chapter 3, we consider a flexible subject-specific heteroskedasticity model for large scale panel data, which employs latent semiparametric factor structure to simultaneously account for the heteroskedasticity across subjects and contemporaneous and/or serial correlations. We propose a two-step procedure for estimation and show the consistency and asymptotic efficiency of our regression coefficient estimator in addition to the asymptotic normality. This leads to a more efficient confidence set of the regression coefficient. In Chapter 4, we combine the approximate factor model with population level clusters to give an integrative group factor model as a background model for variable clustering. We quantify the difficulty of clustering data generated from integrative group factor model in terms of a permutation-invariant clustering error., develop an algorithm to recover clustering assignments and study its minimax-optimality. The analysis of integrative group factor model and our proposed algorithm partitions a two-dimensional phase space into three regions showing the impact of parameters on the possibility of clustering in integrative group factor model and the

statistical guarantee of our proposed algorithm. In the future, We plan to continue pursuing current and develop new directions to further broaden my research portfolio. We will adopt factor model to model the multivariate or high dimensional time serie with multiple changing points in the covariance structures and therefore pave a path to detect the changing covariance/structure of high dimensional time series. Also, We will propose to extend large dimensional factor model for modeling multivariate time series of count data with potentially growing number of subjects. In addition, we will combine PCA and Canonical Correlation Analysis (CCA) to simultaneously study the sample covariance matrix and the cross-covariance. Lastly, we will study to an alternative estimation to the number of factors for traditional principal component analysis using Kac-Rice statistic.

Bibliography

Bibliography

Adamczak, R., Litvak, A., Pajor, A., and Tomczak-Jaegermann, N. (2010). Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561.

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2001a). GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics*, 101(2):219–255.

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2001b). Gmm estimation of linear panel data models with time-varying individual effects. *Journal of econometrics*, 101(2):219–255.

Alzate, C., Espinoza, M., De Moor, B., and Suykens, J. A. (2009). Identifying customer profiles in power load time series using spectral clustering. In *International Conference on Artificial Neural Networks*, pages 315–324. Springer.

Anderson, T. W. (1962). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Anderson, T. W. et al. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34(1):122–148.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, pages 111–150. University of California Press.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, pages 817–858.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297.

Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. Springer, New York.

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

Bai, J. (2009a). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. (2009b). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. and Li, K. (2014). Theory and methods of panel data models with interactive effects. *The Annals of Statistics*, 42(1):142–170.

Bai, J., Li, K., et al. (2014). Theory and methods of panel data models with interactive effects. *The Annals of Statistics*, 42(1):142–170.

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

Bai, J. and Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics*, 3(2):89–163.

Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29.

Bai, Z., Choi, K. P., and Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *The Annals of Statistics*, 46(3):1050–1076.

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer, New York.

Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294.

Bai, Z., Yin, Y., and Krishnaiah, P. R. (1986). On limiting spectral distribution of product of two random matrices when the underlying distribution is isotropic. *Journal of Multivariate Analysis*, 19(1):189–200.

Bai, Z., Yin, Y., and Krishnaiah, P. R. (1988). On the limiting empirical distribution function of the eigenvalues of a multivariate F matrix. *Theory of Probability and Its Applications*, 32(3):490–500.

Baik, J., Arous, G. B., Péché, S., et al. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697.

Baltagi, B. (2008). *Econometrics. Fourth Edition*. Springer-Verlag, New York.

Barigozzi, M., Cho, H., and Fryzlewicz, P. (2018). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics*, 206(1):187–225.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley, New York.

Basu, S. and Reinsel, G. C. (1993). Properties of the spatial unilateral first-order arma model. *Advances in applied Probability*, 25(3):631–648.

Bialecki, B. and Fairweather, G. (1995). Matrix decomposition algorithms in orthogonal spline collocation for separable elliptic boundary value problems. *SIAM Journal on Scientific Computing*, 16(2):330–347.

Bianchi, D., Billio, M., Casarin, R., and Guidolin, M. (2019). Modeling systemic risk with markov switching graphical sur models. *Journal of econometrics*, 210(1):58–74.

Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Bien, J., Bunea, F., and Xiao, L. (2016). Convex banding of the covariance matrix. *Journal of the American Statistical Association*, 111(514):834–845.

Bobkov, S. G., Chistyakov, G., Götze, F., et al. (2018). Berry-Esseen bounds for typical weighted sums. *Electronic Journal of Probability*, 23.

Bobkov, S. G. and Chistyakov, G. P. (2015). On concentration functions of random variables. *Journal of Theoretical Probability*, 28(3):976–988.

Bouzebda, S. and Chokri, K. (2014). Statistical tests in the partially linear additive regression models. *Statistical Methodology*, 19:4–24.

Brockwell, P. J., Davis, R. A., and Fienberg, S. E. (1991). *Time Series: Theory and Methods: Theory and Methods*. Springer, New York.

Browne, M. W. (1979). The maximum-likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology*, 32(1):75–86.

Browne, M. W. (1980). Factor analysis of multiple batteries by maximum likelihood. *British Journal of Mathematical and Statistical Psychology*, 33(2):184–199.

Bühlmann, P. and Künsch, H. R. (1999). Block length selection in the bootstrap for time series. *Computational Statistics and Data Analysis*, 31(3):295–310.

Bunea, F., Giraud, C., Luo, X., Royer, M., Verzelen, N., et al. (2020). Model assisted variable clustering: minimax-optimal recovery and algorithms. *The Annals of Statistics*, 48(1):111–137.

Bunea, F. and Xiao, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli*, 21(2):1200–1230.

Cai, T., Han, X., and Pan, G. (2017). Limiting laws for divergent spiked eigenvalues and largest non-spiked eigenvalue of sample covariance matrices. *arXiv preprint arXiv:1711.00217*.

Cai, T. T., Ren, Z., Zhou, H. H., et al. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59.

Cai, T. T. and Zhang, A. (2016). Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of multivariate analysis*, 150:55–74.

Callaert, H., Janssen, P., et al. (1978). The Berry-Esseen theorem for u-statistics. *The Annals of Statistics*, 6(2):417–421.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Chan, Y.-K. and Wierman, J. (1977). On the Berry-Esseen theorem for u-statistics. *The Annals of Probability*, pages 136–139.

Chang, J., Guo, B., and Yao, Q. (2018). Principal component analysis for second-order stationary vector time series. *The Annals of Statistics*, 46(5):2094–2124.

Chang, J., Yao, Q., and Zhou, W. (2017). Testing for high-dimensional white noise using maximum cross-correlations. *Biometrika*, 104(1):111–127.

Chen, L., Wang, W., and Wu, W. (2018). Dynamic semiparametric factor model with structural breaks. *Available at SSRN: https://ssrn.com/abstract=3131684 or http://dx.doi.org/10.2139/ssrn.3131684*.

Chen, L. H., Shao, Q.-M., et al. (2004). Normal approximation under local dependence. *The Annals of Probability*, 32(3):1985–2028.

Chen, L. H., Shao, Q.-M., et al. (2007). Normal approximation for nonlinear statistics using a concentration inequality approach. *Bernoulli*, 13(2):581–599.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632.

Chen, X., Womersley, R. S., and Ye, J. J. (2011). Minimizing the condition number of a Gram matrix. *SIAM Journal on Optimization*, 21(1):127–148.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Choi, Y., Taylor, J., and Tibshirani, R. (2017). Selecting the number of principal components: estimation of the true rank of a noisy matrix. *The Annals of Statistics*, 45(6):2590–2617.

Connor, G., Hagmann, M., and Linton, O. (2012). Efficient semiparametric estimation of the Fama–French model and extensions. *Econometrica*, 80(2):713–754.

Connor, G. and Linton, O. (2007). Semiparametric estimation of a characteristic-based factor model of common stock returns. *Journal of Empirical Finance*, 14(5):694–717.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

De Almeida, M. C., Asada, E. N., and Garcia, A. V. (2008a). On the use of gram matrix in observability analysis. *IEEE Transactions on Power Systems*, 23(1):249–251.

De Almeida, M. C., Asada, E. N., and Garcia, A. V. (2008b). Power system observability analysis based on gram matrix and minimum norm solution. *IEEE Transactions on Power Systems*, 23(4):1611–1618.

Desai, K. H. and Storey, J. D. (2012). Cross-dimensional inference of dependent high-dimensional data. *Journal of the American Statistical Association*, 107(497):135–151.

Donath, W. E. and Hoffman, A. J. (1973). Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425.

Drineas, P. and Mahoney, M. W. (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(Dec):2153–2175.

Ernst, J., Nau, G. J., and Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, 21(suppl_1):i159–i168.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2):293–314.

Fan, J., Huang, T., et al. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11(6):1031–1057.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320–3356.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B*, 75(4):603–680.

Fan, J., Liao, Y., and Wang, W. (2016). Projected principal component analysis in factor models. *The Annals of Statistics*, 44(1):219âĂŞ–254.

Fan, J., Sun, Q., Zhou, W., and Zhu, Z. (2018a). Principal component analysis for big data. *Wiley StatsRef: Statistics Reference Online*, pages 1–13.

Fan, J., Wang, W., and Zhong, Y. (2018b). An ℓ^∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42.

Florescu, L. and Perkins, W. (2016). Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, 82(4):540–554.

Fujita, A., Severino, P., Kojima, K., Sato, J. R., Patriota, A. G., and Miyano, S. (2012). Functional clustering of time series gene expression data by granger causality. *BMC systems biology*, 6(1):137.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(1):1980–2024.

Gao, C., Ma, Z., Zhang, A. Y., Zhou, H. H., et al. (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185.

Goldstein, L., Shao, Q.-M., et al. (2009). Berry-Esseen bounds for projections of coordinate symmetric random vectors. *Electronic Communications in Probability*, 14:474–485.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Guha, S., Rastogi, R., and Shim, K. (1998). Cure: an efficient clustering algorithm for large databases. *ACM Sigmod record*, 27(2):73–84.

Hall, P., Horowitz, J. L., and Jing, B. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574.

Hallac, D., Vare, S., Boyd, S., and Leskovec, J. (2017). Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 215–223.

Hallin, M. and Liška, R. (2008). Dynamic factors in the presence of block structure. *European University Institute WP 2008/22*.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–318.

Hayakawa, K. and Pesaran, M. H. (2015). Robust standard errors in transformed likelihood estimation of dynamic panel data models with cross-sectional heteroskedasticity. *Journal of econometrics*, 188(1):111–134.

He, W., Zhang, H., Zhang, L., and Shen, H. (2015). Hyperspectral image denoising via noise-adjusted iterative low-rank matrix approximation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):3050–3061.

Hörmann, S. (2009). Berry-Esseen bounds for econometric time series. *ALEA Lat. Am. J. Probab. Math. Stat*, 6:377–397.

Horst, P. (1965). *Factor Analysis of Data Matrices*. Holt, Rinehart and Winston, New York.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417.

Hsiao, C. (2014). *Analysis of panel data*. Number 54. Cambridge university press.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14:763–788.

Ieva, F., Paganoni, A. M., and Tarabelloni, N. (2016). Covariance-based clustering in multivariate and functional data analysis. *The Journal of Machine Learning Research*, 17(1):4985–5005.

James, G. and Murphy, G. (1979). The determinant of the gram matrix for a specht module. *Journal of Algebra*, 59(1):222–235.

James, W. and Stein, C. (1992). Estimation with quadratic loss. In *Breakthroughs in statistics*, pages 443–460. Springer.

Jebara, T., Song, Y., and Thadani, K. (2007). Spectral clustering and embedding with hidden markov models. In *European Conference on Machine Learning*, pages 164–175. Springer.

Ji, H., Huang, S., Shen, Z., and Xu, Y. (2011). Robust video restoration by joint sparse and low rank matrix approximation. *SIAM Journal on Imaging Sciences*, 4(4):1122–1142.

Jiang, J. et al. (1996). Reml estimation: asymptotic behavior and related topics. *The Annals of Statistics*, 24(1):255–286.

Jin, J. et al. (2015). Fast community detection by score. *The Annals of Statistics*, 43(1):57–89.

Jirak, M. (2016). Berry–Esseen theorems under weak dependence. *The Annals of Probability*, 44(3):2024–2063.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Johnstone, I. M. and Paul, D. (2018). PCA in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292.

Jolliffe, I. (2002). *Principal Component Analysis, 2nd ed*. Springer, New York.

Jonsson, D. (1982). Some limit theorems for the eigenvalues of a sample covariance matrix. *Journal of Multivariate Analysis*, 12(1):1–38.

Karypis, G., Han, E.-H., and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Keogh, E. and Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, 7(4):349–371.

Klami, A., Virtanen, S., Leppäaho, E., and Kaski, S. (2014). Group factor analysis. *IEEE transactions on neural networks and learning systems*, 26(9):2136–2147.

Koltchinskii, V. and Lounici, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(4):1976–2013.

Koltchinskii, V. and Lounici, K. (2017). Normal approximation and concentration of spectral projectors of sample covariance. *The Annals of Statistics*, 45(1):121–157.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

Lawley, D. N. and Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society: Series D*, 12(3):209–229.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9).

Li, Q. and Li, L. (2019). Integrative factor regression and its inference for multimodal data analysis. *arXiv preprint arXiv:1911.04056*.

Li, Z., Lam, C., Yao, J., Yao, Q., et al. (2019). On testing for high-dimensional white noise. *The Annals of Statistics*, 47(6):3382–3412.

Liang, H., Su, H., Thurston, S. W., Meeker, J. D., and Hauser, R. (2009). Empirical likelihood based inference for additive partial linear measurement error models. *Statistics and Its Interface*, 2(1):83–90.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257.

Lin, J., Vlachos, M., Keogh, E., and Gunopulos, D. (2004). Iterative incremental clustering of time series. In *International Conference on Extending Database Technology*, pages 106–122. Springer.

Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., and Sheppard, L. (2014). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and Ecological Statistics*, 21(3):411–433.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523.

Lorentz, G. (1966). *Approximation of Functions, Athena Series*. Holt, Rinehart and Winston, New York.

Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*.

Lu, Z., Steinskog, D. J., Tjøstheim, D., and Yao, Q. (2009). Adaptively varying-coefficient spatiotemporal models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 71(4):859–880.

Lütkepohl, H. (2006). Structural vector autoregressive analysis for cointegrated variables. *Allgemeines Statistisches Arch.*, 90(1):75–88.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Mandel, J. (1982). Use of the singular value decomposition in regression analysis. *The American Statistician*, 36(1):15–24.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3):435–474.

Moench, E. and Ng, S. (2011). A factor analysis of housing market dynamics in the U.S. and the regions. *The Econometrics Journal*, 14:1–24.

Möller-Levet, C. S., Klawonn, F., Cho, K.-H., and Wolkenhauer, O. (2003). Fuzzy clustering of short time-series and unevenly distributed sampling points. In *International symposium on intelligent data analysis*, pages 330–340. Springer.

Moon, H. R. and Weidner, M. (2017a). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory*, 33(1):158–195.

Moon, H. R. and Weidner, M. (2017b). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory*, 33(1):158–195.

Naumov, A., Spokoiny, V., and Ulyanov, V. (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probability Theory and Related Fields*, 174(3-4):1091–1132.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856.

Niennattrakul, V. and Ratanamahatana, C. A. (2006). Clustering multimedia data using time series. In *2006 International Conference on Hybrid Information Technology*, volume 1, pages 372–379. IEEE.

Niennattrakul, V. and Ratanamahatana, C. A. (2007). On clustering multimedia time series data using k-means and dynamic time warping. In *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, pages 733–738. IEEE.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Pal, P. and Vaidyanathan, P. P. (2014). A grid-less approach to underdetermined direction of arrival estimation via low rank matrix denoising. *IEEE Signal Processing Letters*, 21(6):737–741.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Phillips, R. F. (2010). Iterated feasible generalized least-squares estimation of augmented dynamic panel data models. *Journal of Business & Economic Statistics*, 28(3):410–422.

Pyatnitskiy, M., Mazo, I., Shkrob, M., Schwartz, E., and Kotelnikova, E. (2014). Clustering gene expression regulators: new approach to disease subtyping. *PLoS One*, 9(1).

Rabe-Hesketh, S. and Skrondal, A. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman and Hall, London.

Ramírez, D., Santamaria, I., Van Vaerenbergh, S., and Scharf, L. L. (2018). An alternating optimization algorithm for two-channel factor analysis with common and uncommon factors. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 1743–1747. IEEE.

Ramona, M., Richard, G., and David, B. (2012). Multiclass feature selection with kernel Gram-matrix-based criteria. *IEEE Transactions on Neural Networks and Learning Systems*, 23(10):1611–1623.

Rani, S. and Sikka, G. (2012). Recent techniques of clustering of time series data: a survey. *International Journal of Computer Applications*, 52(15).

Ratanamahatana, C. A. and Keogh, E. (2005). Multimedia retrieval using time series representation and relevance feedback. In *International Conference on Asian Digital Libraries*, pages 400–405. Springer.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, pages 931–954.

Rummel, R. J. (1988). *Applied Factor Analysis*. Northwestern University Press, Evanston.

Samson, P.-M. et al. (2000). Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461.

Schmidheiny, K. and Basel, U. (2011). Panel data: fixed and random effects. *Short Guides to Microeconometrics*, 7(1):2–7.

Schölkopf, B., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (1999). Generalization bounds via eigenvalues of the Gram matrix. *Technical Report 99-035, NeuroCOLT*.

Shawe-Taylor, J., Williams, C., Cristianini, N., and Kandola, J. (2002). On the eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In *International Conference on Algorithmic Learning Theory*, pages 23–40. Springer.

Shawe-Taylor, J., Williams, C. K., Cristianini, N., and Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, 51(7):2510–2522.

Stark, C. (2014). Self-consistent tomography of the state-measurement Gram matrix. *Physical Review A*, 89(5):052109.

Stein, C. (1956). Some problems in multivariate analysis, part i. Technical report, STANFORD UNIV CA.

Stock, J. H. and Watson, M. W. (1998). Diffusion indexes. Technical report, National bureau of economic research.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 10:689–705.

Subhani, N., Rueda, L., Ngom, A., and Burden, C. J. (2010). Multiple gene expression profile alignment for microarray time-series data clustering. *Bioinformatics*, 26(18):2281–2288.

Tan, Z., Qin, G., and Zhou, H. (2016). Estimation of a partially linear additive model for data from an outcome-dependent sampling design with a continuous outcome. *Biostatistics*, 17(4):663–676.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Number 47. Sage.

Thompson, B. (2005). Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tikhomirov, A. N. (1981). On the convergence rate in the central limit theorem for weakly dependent random variables. *Theory of Probability and Its Applications*, 25(4):790–809.

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika*, 23(2):111–136.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Wachter, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *The Annals of Probability*, 6(1):1–18.

Wagaman, A. and Levina, E. (2009). Discovering sparse covariance structures with the isomap. *Journal of Computational and Graphical Statistics*, 18(3):551–572.

Wang, B., Luo, X., Zhao, Y., and Caffo, B. (2019). Semiparametric partial common principal component analysis for covariance matrices. *bioRxiv*, page 808527.

Wang, F. and Wang, H. (2018). Modelling non-stationary multivariate time series of counts via common factors. *Journal of the Royal Statistical Society: Series B*, 80(4):769–791.

Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Annals of statistics*, 45(5):1863.

Wang, P. (2008). Large dimensional factor models with a multi-level factor structure: identification, estimation and inference. *Unpublished manuscript, New York University*.

Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45(3):1342–1374.

Wedin, P. Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111.

Yin, J. and Yang, Q. (2005). Integrating hidden markov models and spectral analysis for sensory time series clustering. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE.

Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer.

Yu, Y., Wang, T., and Samworth, R. J. (2014). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

Zhang, A., Cai, T. T., and Wu, Y. (2018). Heteroskedastic pca: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*.

Zhang, A. Y., Zhou, H. H., et al. (2016). Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280.

Zhang, H., He, W., Zhang, L., Shen, H., and Yuan, Q. (2013). Hyperspectral image restoration using low-rank matrix recovery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8):4729–4743.

Zhang, L., Zhou, W., and Wang, H. (2019). Estimation and inference of a heteroskedasticity model with latent semiparametric factors for panel data analysis. *under review*.

Zhang, L., Zhou, W., and Wang, H. (2020). Non-asymptotic properties of spectral decomposition of large Gram-type matrices and applications. *under review*.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: an efficient data clustering method for very large databases. *ACM Sigmod Record*, 25(2):103–114.

Zhang, X. and Cheng, G. (2018). Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli*, 24(4A):2640–2675.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

Zou, H. and Xue, L. (2018). A selective overview of sparse principal component analysis. *Proceedings of the IEEE*, 106(8):1311–1320.

Appendix A

Supplemental materials for Chapter 2

This online supplementary material contains technical results used for the main paper. In Section A.1, we prove the main results in Theorems 2.3.1-2.3.3 in the paper. In Section A.2, we show Theorems 2.4.1-2.4.6 of the main paper. Lastly, in Section A.3, we include technical lemmas and auxiliary results.

Here and after, we constantly explore the tail probability of random variable X in the following sense: with probability at least $1-e^{-s}$, $X \lesssim s$ for $s > 1$. Such an inequality is often proved with probability $1-Ce^{-s}$ instead, where $C > 0$ is an absolute constant. In such cases, it is easy to show that the inequality still holds with the original probability. By replacing s with $s + \log(C)$, we claim that with probability at least $1-e^{-s}$, $X \lesssim s+\log(C) \lesssim \{1+\log(C)\}s$. Thus, it will be said without further explanation that probability bound $1-Ce^{-s}$ can be replaced by $1-e^{-s}$ via adjusting the constant. See Koltchinskii and Lounici (2017); Zhang et al. (2019) for similar discussions. Finally, Conditions 2.2.1-2.3.1 and models (2.1.1)-(2.1.2) are referred to corresponding conditions or models in the main paper.

A.1 Proof of Results in Section 2.3

A.1.1 Proof of Theorem 2.3.1

Proof. The conclusion in (i) follows from Lemmas A.3.3 and A.3.6. By (i), for each $a > 1$, we have

$$
\begin{aligned}
&T^{-1}(p^{-1}+T^{-1})^{-1}\mathbb{E}(\|\widehat{\mathbf{F}}-\mathbf{F}\|_{\mathbb{F}}^2)\\
=&\int_0^\infty \mathbb{P}\{T^{-1}(p^{-1}+T^{-1})^{-1}\|\widehat{\mathbf{F}}-\mathbf{F}\|_{\mathbb{F}}^2 > s\}ds\\
=&\int_0^a \mathbb{P}\{T^{-1}(p^{-1}+T^{-1})^{-1}\|\widehat{\mathbf{F}}-\mathbf{F}\|_{\mathbb{F}}^2 > s\}ds + \int_a^\infty \mathbb{P}\{T^{-1}(p^{-1}+T^{-1})^{-1}\|\widehat{\mathbf{F}}-\mathbf{F}\|_{\mathbb{F}}^2 > s\}ds
\end{aligned}
$$

$$\leqslant a + \int_a^\infty \exp(-Cs^{1/4})ds$$
$$\leqslant a + 4C(a^{3/4} + 3a^{1/2} + 6a^{1/4} + 6)\exp(-Ca^{1/4}),$$

where C is a positive constant. The right hand side of the above inequality is minimized at a positive constant C_1 such that

$$T^{-1}\mathbb{E}(\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2) \leqslant C_1(p^{-1} + T^{-1}).$$

Similarly, we have

$$T^{-2}\operatorname{Var}(\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2) \leqslant \mathbb{E}(\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^4) \leqslant (p^{-2} + T^{-2})\{2 + \log(6)\}C_2,$$

where C_2 is a positive constant. Finally, (iii) follows from Lemmas A.3.6 and A.3.7. □

A.1.2 Proof of Theorem 2.3.2

Proof. For each $k = 1, \ldots, K$, let

$$\widetilde{\mathbf{P}}_k = \mathbf{I}_T - \boldsymbol{f}_k(\boldsymbol{f}_k^\top \boldsymbol{f}_k)^{-1}\boldsymbol{f}_k^\top,\ \bar{\mathbf{P}}_k = \mathbf{I}_T - \widehat{\boldsymbol{f}}_k(\widehat{\boldsymbol{f}}_k^\top \widehat{\boldsymbol{f}}_k)^{-1}\widehat{\boldsymbol{f}}_k^\top,$$

$$\widetilde{\mathbf{Q}}_k = \sum_{j\neq k}\{\lambda_k(\mathbf{A}^\top\mathbf{A}) - \lambda_j(\mathbf{A}^\top\mathbf{A})\}^{-1}\widetilde{\mathbf{P}}_j,$$

$$\widetilde{B}_k = 2\sqrt{2}\|\widetilde{\mathbf{P}}_k p^{-1}\mathbf{F}\mathbf{A}^\top\mathbf{A}\mathbf{F}^\top\widetilde{\mathbf{P}}_k\|_2\|\widetilde{\mathbf{Q}}_k p^{-1}\mathbf{F}\mathbf{A}^\top\mathbf{A}\mathbf{F}^\top\widetilde{\mathbf{Q}}_k\|_2,$$

and

$$\widetilde{C}_k = 2\operatorname{tr}\{\widetilde{\mathbf{P}}_t p^{-1}\mathbf{F}\mathbf{A}^\top\mathbf{A}\mathbf{F}^\top\widetilde{\mathbf{P}}_k\}\operatorname{tr}\{\widetilde{\mathbf{Q}}_k p^{-1}\mathbf{F}\mathbf{A}^\top\mathbf{A}\mathbf{F}^\top\widetilde{\mathbf{Q}}_k\}.$$

Similar to the proof of Theorem 6 in Koltchinskii and Lounici (2017) and Lemma A.3.13,

$$T\widetilde{B}_k^{-1}\left\{\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2 - \mathbb{E}(\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2)\right\} \stackrel{d}{=} \tau + \zeta$$

where $\sup_x |\mathbb{P}(\tau \leqslant x) - \mathbf{\Phi}(x)| \lesssim \widetilde{B}_k^{-1}$ and with probability at least $1 - e^{-s}$ and $|\zeta| \lesssim T^{-1/2}s + \widetilde{B}_k^{-1}T^{-1/2}s^{3/2}$. Thus,

$$\sup_x \left| \mathbb{P}\left[T\widetilde{B}_k^{-1} \left\{ \|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2 - \mathbb{E}(\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2) \right\} \leqslant x \right] - \mathbf{\Phi}(x) \right| \lesssim \frac{1}{\widetilde{B}_k} + \frac{z}{\sqrt{T}} + \frac{z^{3/2}}{\widetilde{B}_k\sqrt{T}} + \exp(-z)$$

for any $z > 0$. By letting $z = \min\{\log(\widetilde{B}_k), \log(T)\}$, we have

$$\sup_x \left| \mathbb{P}\left[T\widetilde{B}_k^{-1} \left\{ \|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2 - \mathbb{E}(\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2) \right\} \leqslant x \right] - \mathbf{\Phi}(x) \right| \lesssim \frac{1}{\widetilde{B}_k} + \frac{\log(T)}{\sqrt{T}}.$$

As $\widetilde{B}_k = O(p)$, the theorem is proved following the similar arguements for Theorem 6 in Koltchinskii and Lounici (2017). □

A.1.3 Proof of Theorem 2.3.3

Proof. Recall that

$$\frac{1}{T}\mathbf{Y}\mathbf{Y}^\top = \mathbf{A}\mathbf{A}^\top + \frac{1}{T}\mathbf{A}\mathbf{F}^\top\mathbf{U}^\top + \frac{1}{T}\mathbf{U}\mathbf{F}\mathbf{A}^\top + \frac{1}{T}\mathbf{U}\mathbf{U}^\top$$

and

$$\frac{1}{T}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top) = \mathbf{A}\mathbf{A}^\top + \frac{1}{T}\mathbb{E}(\mathbf{U}\mathbf{U}^\top).$$

Also, note that $\mathbf{A}\mathbf{A}^\top$, $T^{-1}\mathbf{A}\mathbf{F}^\top\mathbf{U}^\top$, and $T^{-1}\mathbf{U}\mathbf{F}\mathbf{A}^\top$ all have rank K, so that for each $i = K + 1, \ldots, \min(p, T)$, $\widehat{\lambda}_i \asymp \lambda_i(T^{-1}\mathbf{U}\mathbf{U}^\top)$ and $\lambda_i \asymp \lambda_i(\mathbb{E}(T^{-1}\mathbf{U}\mathbf{U}^\top))$.

(i) If $p < T$, by Condition 2.3.1 and Theorem 5.58 in Vershynin (2010), with probability at least $1 - e^{-2s}$,

$$|\lambda_i(T^{-1}\mathbf{U}\mathbf{U}^\top) - \lambda_i(\mathbb{E}(T^{-1}\mathbf{U}\mathbf{U}^\top))| \leqslant \max(\zeta, \zeta^2),$$

for each $i = 1, \ldots, p$, where $\zeta = \sqrt{C}T^{-1/2}\sqrt{p} + \sqrt{c}T^{-1/2}\sqrt{s}$ and C and c are positive constants only depending on $\boldsymbol{u}_t$. Thus, with probability at least $1 - e^{-2s}$,

$$|\lambda_i(T^{-1}\mathbf{U}\mathbf{U}^\top) - \lambda_i(\mathbb{E}(T^{-1}\mathbf{U}\mathbf{U}^\top))| \leqslant C\sqrt{\frac{p}{T}} + \frac{c}{\sqrt{T}}\sqrt{s},$$

(ii) If $p > T$, note that the first T largest eigenvalues of $T^{-1}\mathbf{U}\mathbf{U}^\top$ are the same as those of $T^{-1}\mathbf{U}^\top\mathbf{U}$. By Condition 2.3.1 and Theorem 5.39 in Vershynin (2010), for each $i = 1, \ldots, T$, with probability at least $1 - e^{-2s}$,

$$\sqrt{\frac{p}{T}} - C - \frac{c}{\sqrt{T}}\sqrt{s} \lesssim \lambda_i(T^{-1}\mathbf{U}^\top\mathbf{U}) \lesssim \sqrt{\frac{p}{T}} + C + \frac{c}{\sqrt{T}}\sqrt{s}.$$

By Condition 2.2.2, for each $i = 1, \ldots, K$, $\lambda_i(\mathbf{A}\mathbf{A}^\top) = O(p)$. Also, by Condition 2.2.3 and the discussion above, $p^{-1}\lambda_i(T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top))$ and $p^{-1}\lambda_i(T^{-1}\mathbf{U}\mathbf{U}^\top)$ are bounded. Thus, with probability at least $1 - e^{-2s}$,

$$\begin{aligned}\frac{\widehat{\lambda}_i}{\lambda_i} &\leqslant \frac{\lambda_i(\mathbf{A}\mathbf{A}^\top) + \lambda_{\max}(T^{-1}\mathbf{A}\mathbf{F}^\top\mathbf{U}^\top) + \lambda_{\max}(T^{-1}\mathbf{U}\mathbf{F}\mathbf{A}^\top) + \lambda_{\max}(T^{-1}\mathbf{U}\mathbf{U}^\top)}{\lambda_i(\mathbf{A}\mathbf{A}^\top) + \lambda_{\min}(T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top))} \\ &\lesssim 1 + \frac{C}{\sqrt{T}} + \frac{c}{\sqrt{pT}}\sqrt{s}\end{aligned}$$

and

$$\begin{aligned}\frac{\widehat{\lambda}_i}{\lambda_i} &\geqslant \frac{\lambda_i(\mathbf{A}\mathbf{A}^\top) + \lambda_{\min}(T^{-1}\mathbf{A}\mathbf{F}^\top\mathbf{U}^\top) + \lambda_{\min}(T^{-1}\mathbf{U}\mathbf{F}\mathbf{A}^\top) + \lambda_{\min}(T^{-1}\mathbf{U}\mathbf{U}^\top)}{\lambda_i(\mathbf{A}\mathbf{A}^\top) + \lambda_{\max}(\mathbb{E}(\mathbf{U}\mathbf{U}^\top))} \\ &\gtrsim 1 - \frac{C}{\sqrt{T}} - \frac{c}{\sqrt{pT}}\sqrt{s}.\end{aligned}$$

□

A.2 Proofs of Results in Section 2.4

A.2.1 Proof of Results in Section 2.4.1

Proof of Theorem 2.4.1. By Theorem 2.3.3, if $p < T$, for $s < c^{-2}(\sqrt{T} - C\sqrt{p})^2$, with probability at least $1 - e^{-s}$,

$$\begin{aligned}
\frac{\widehat{\lambda}_i}{\widehat{\lambda}_{i+1}} &\lesssim \frac{\lambda_i}{\lambda_{i+1}} \left\{1 + \frac{C}{\sqrt{T}} + \frac{c}{\sqrt{pT}}\sqrt{s}\right\}^2, \qquad i = 1, \ldots, K-1, \\
\frac{\widehat{\lambda}_K}{\widehat{\lambda}_{K+1}} &\gtrsim \frac{\lambda_K}{\lambda_{K+1}} \left\{1 - \frac{C\sqrt{p}}{\sqrt{T}} - \frac{c}{\sqrt{T}}\sqrt{s}\right\} \left\{1 - \frac{C}{\sqrt{T}} - \frac{c}{\sqrt{pT}}\sqrt{s}\right\}, \\
\frac{\widehat{\lambda}_i}{\widehat{\lambda}_{i+1}} &\lesssim \frac{\lambda_i}{\lambda_{i+1}} \left\{1 + \frac{C\sqrt{p}}{\sqrt{T}} + \frac{c}{\sqrt{T}}\sqrt{s}\right\}^2, \qquad i = K+1, \ldots, L,
\end{aligned}$$

Thus, for $s < c^{-1}T + c^{-1}Cp$, with probability at least $1 - e^{-s}$,

$$\frac{\widehat{\lambda}_K/\widehat{\lambda}_{K+1}}{\max_{i \neq K} \widehat{\lambda}_i/\widehat{\lambda}_{i+1}} \gtrsim \frac{\lambda_K/\lambda_{K+1}}{C_3} \left\{1 - \frac{C\sqrt{p}}{\sqrt{T}} - \frac{c}{\sqrt{T}}\sqrt{s}\right\}^4$$

where $C_3 = \max_{1 \leqslant i \leqslant L, i \neq K} \lambda_i/\lambda_{i+1}$. Thus,

$$\mathbb{P}(\widehat{K} = K) \geqslant 1 - 2\exp\{-(C_1\sqrt{T} - C_2\sqrt{p})^2\},$$

where

$$C_1 = \frac{1}{c}\left\{1 - \left(\frac{\lambda_{K+1}}{\lambda_K} \max_{1 \leqslant i \leqslant L, i \neq K} \lambda_i/\lambda_{i+1}\right)^{1/4}\right\}$$

and $C_2 = c^{-1}C$.

On the other hand, if $p > T$, for $s < c^{-2}(\sqrt{p} - C\sqrt{T})^2$, with probability at least $1 - e^{-s}$,

$$\begin{aligned}
\frac{\widehat{\lambda}_i}{\widehat{\lambda}_{i+1}} &\lesssim \frac{\lambda_i}{\lambda_{i+1}} \left\{1 + \frac{C}{\sqrt{T}} + \frac{c}{\sqrt{pT}}\sqrt{s}\right\}^2, \qquad i = 1, \ldots, K-1, \\
\frac{\widehat{\lambda}_K}{\widehat{\lambda}_{K+1}} &\gtrsim \frac{T\lambda_K}{p\lambda_{K+1}} \left\{1 - \frac{C\sqrt{T}}{\sqrt{p}} - \frac{c}{\sqrt{p}}\sqrt{s}\right\} \left\{1 - \frac{C}{\sqrt{T}} - \frac{c}{\sqrt{pT}}\sqrt{s}\right\},
\end{aligned}$$

$$\frac{\widehat{\lambda}_i}{\widehat{\lambda}_{i+1}} \lesssim \frac{\lambda_i}{\lambda_{i+1}} \left\{1 + \frac{C\sqrt{T}}{\sqrt{p}} + \frac{c}{\sqrt{p}}\sqrt{s}\right\}^2, \qquad i = K+1, \ldots, L.$$

Thus, for $s < c^{-1}p + c^{-1}CT$, with probability at least $1 - e^{-s}$,

$$\frac{\widehat{\lambda}_K/\widehat{\lambda}_{K+1}}{\max_{i\neq K} \widehat{\lambda}_i/\widehat{\lambda}_{i+1}} \gtrsim \frac{T\lambda_K/(p\lambda_{K+1)}}{C_3} \left\{1 - \frac{C\sqrt{T}}{\sqrt{p}} - \frac{c}{\sqrt{p}}\sqrt{s}\right\}^4.$$

Thus,

$$\mathbb{P}(\widehat{K} = K) \geqslant 1 - 2\exp\{-(C_4\sqrt{p} - C_2\sqrt{T})^2\},$$

where

$$C_4 = \frac{1}{c}\left\{1 - \left(\frac{p\lambda_{K+1}}{T\lambda_K} \max_{1\leqslant i\leqslant L, i\neq K} \lambda_i/\lambda_{i+1}\right)^{1/4}\right\}$$

The conclusion follows. □

Proof of Theorem 2.4.2. By Theorem 2.3.3, for each $i = 1, \ldots, L$,

$$|\widehat{\lambda}_i - \lambda_i| \leqslant \frac{C\sqrt{p}}{\sqrt{T}} + \frac{c}{\sqrt{T}}\sqrt{s}.$$

Thus, with probability at least $1 - e^{-s}$,

$$\widehat{\lambda}_i - \widehat{\lambda}_{i+1} \geqslant \lambda_i - \lambda_{i+1} - \frac{2C\sqrt{p}}{\sqrt{T}} - \frac{2c}{\sqrt{T}}\sqrt{s}$$

for $i = 1, \ldots, K$ and

$$\widehat{\lambda}_{K+1} - \widehat{\lambda}_{K+2} \leqslant \lambda_{K+1} - \lambda_{K+2} + \frac{2C\sqrt{p}}{\sqrt{T}} + \frac{2c}{\sqrt{T}}\sqrt{s}.$$

Thus,

$$\begin{aligned}\mathbb{P}(\widehat{K}_d = K) =& \bigcap_{i=1}^{K} \mathbb{P}(\widehat{\lambda}_i - \widehat{\lambda}_{i+1} \geqslant \delta, \widehat{\lambda}_{K+1} - \widehat{\lambda}_{K+2} \leqslant \delta) \\ \geqslant& 1 - 2\sum_{i=1}^{K+1} \exp\{-(C_{1i}\sqrt{T} - C_2\sqrt{p})^2\},\end{aligned}$$

where $C_{1i} = (2c)^{-1}(\lambda_i - \lambda_{i+1} - \delta)$ for $i = 1, \ldots, K$, $C_{1,K+1} = c^{-1}(\delta - \lambda_{K+2} + \lambda_{K+1})$, and $C_2 = c^{-1}C$. The theorem is proved. □

A.2.2 Proof of Results in Section 2.4.2

Proof of Theorem 2.4.3. Note that

$$\begin{aligned}\widehat{\Gamma}(h, \widehat{\mathbf{f}}_t) - \widehat{\Gamma}(h, \mathbf{f}_t) =& \frac{1}{T}\widehat{\mathbf{F}}\mathbf{P}_1 \begin{bmatrix} \mathbf{0}_{h\times(T-h)} & \mathbf{0}_{(T-h)\times(T-h)} \\ \mathbf{I}_{T-h} & \mathbf{0}_{(T-h)\times h} \end{bmatrix} \mathbf{P}_1\widehat{\mathbf{F}} \\ & - \frac{1}{T}\mathbf{F}\mathbf{P}_1 \begin{bmatrix} \mathbf{0}_{h\times(T-h)} & \mathbf{0}_{(T-h)\times(T-h)} \\ \mathbf{I}_{T-h} & \mathbf{0}_{(T-h)\times h} \end{bmatrix} \mathbf{P}_1\mathbf{F}.\end{aligned}$$

The conclusion follows from Lemmas A.3.3 and A.3.6. □

Proof of Theorem 2.4.4. By Theorem 2.4.3, for each $k = 1, \ldots, K$, with probability at least $1 - e^{-s}$,

$$|\widehat{\gamma}(h, \widehat{f}_{tk}) - \widehat{\gamma}(h, f_{tk})|^2 \lesssim \frac{1}{T}\left(\frac{1}{p} + \frac{1}{T}\right)s.$$

Thus, with probability at least $1 - e^{-s}$,

$$|\widehat{\rho}(h, \widehat{f}_{tk}) - \widehat{\rho}(h, f_{tk})|^2 = \left|\frac{\widehat{\gamma}(h, \widehat{f}_{tk})}{\widehat{\gamma}(0, \widehat{f}_{tk})} - \frac{\widehat{\gamma}(h, f_{tk})}{\widehat{\gamma}(0, f_{tk})}\right|^2 \lesssim \frac{1}{T}\left(\frac{1}{p} + \frac{1}{T}\right)s.$$

□

A.2.3 Proof of Results in Section 2.4.3

Proof of Theorem 2.4.5. It is easy to see that $\widehat{\boldsymbol{w}}_i = \|\mathbf{Y}\widehat{\boldsymbol{f}}_i\|_2^{-1}\mathbf{Y}\widehat{\boldsymbol{f}}_i$ for $i \leqslant K$. Also, we have

$$\mathbf{Y}\boldsymbol{f}_i = \mathbf{A}\mathbf{F}^\top\boldsymbol{f}_i + \mathbf{U}\boldsymbol{f}_i = \mathbf{A}(\boldsymbol{f}_1^\top\boldsymbol{f}_i, \ldots, \boldsymbol{f}_K^\top\boldsymbol{f}_i)^\top + (\boldsymbol{u}_1^\top\boldsymbol{f}_i, \ldots, \boldsymbol{u}_n^\top\boldsymbol{f}_i)^\top.$$

By Condition 2.2.1,

$$\mathbb{E}(\mathbf{Y}\boldsymbol{f}_i) = T\mathbf{A}\boldsymbol{e}_i,$$

where $\boldsymbol{e}_i$ is the ith natural basis in $\mathbb{R}^T$, and

$$T^{-1}\|\mathbf{Y}\boldsymbol{f}_i - \mathbb{E}(\mathbf{Y}\boldsymbol{f}_i)\|_2^2 \lesssim T^{-1}\|\mathbf{A}\|_2^2 s,$$

with probability at least $1 - e^{-s}$. Thus, by the proof of Theorem 2.3.1, with probability at least $1 - 4e^{-s}$,

$$T^{-1}\|\mathbf{Y}\boldsymbol{f}_i - \mathbb{E}(\mathbf{Y}\boldsymbol{f}_i)\|_2^2 \lesssim (p^{-1} + T^{-1})\|\mathbf{A}\|_2^2 s.$$

Finally, by Condition 2.2.1, $\mathbf{A}^\top\mathbf{A}$ is a diagonal matrix almost surely. Thus, $\boldsymbol{w}_i = \|\mathbf{A}\boldsymbol{e}_i\|_2^{-1}\mathbf{A}\boldsymbol{e}_i$ almost surely. Similar to the proof of Theorem 2.3.1, the conclusion follows. □

For each $i = 1, \ldots, p$, let

$$B_i = 2\sqrt{2}\|\mathbf{P}_i T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)\mathbf{P}_i\|_2\|\mathbf{Q}_i T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)\mathbf{Q}_i\|_2$$

and

$$C_i = 2\,\mathrm{tr}\{\mathbf{P}_i T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)\mathbf{P}_i\}\,\mathrm{tr}\{\mathbf{Q}_i T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)\mathbf{Q}_i\},$$

where, as defined in Section 2.4.3, $\mathbf{P}_i = \mathbf{I}_p - \boldsymbol{w}_i(\boldsymbol{w}_i^\top\boldsymbol{w}_i)^{-1}\boldsymbol{w}_i^\top$, $\widehat{\mathbf{P}}_i = \mathbf{I}_p - \widehat{\boldsymbol{w}}_i(\widehat{\boldsymbol{w}}_i^\top\widehat{\boldsymbol{w}}_i)^{-1}\widehat{\boldsymbol{w}}_i^\top$, and $\mathbf{Q}_i = \sum_{j\neq i}(\lambda_i - \lambda_j)^{-1}\mathbf{P}_j$.

Proof of Theorem 2.4.6. Similar to the proof of Theorem 2.3.2,by Lemma A.3.12,

$$TB_i^{-1}\left\{\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2 - \mathbb{E}(\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2)\right\} \overset{d}{=} \tau + \zeta$$

where

$$\sup_x |\mathbb{P}(\tau \leqslant x) - \mathbf{\Phi}(x)| \lesssim \frac{1}{B_i}$$

and with probability at least $1 - e^{-s}$,

$$|\zeta| \lesssim \frac{s}{\sqrt{T}} + \frac{s^{3/2}}{B_i\sqrt{T}} + \frac{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}}{T^{1/8}B_i e^{-s/4}}.$$

Thus,

$$\begin{aligned}
&\sup_x \left|\mathbb{P}\left[TB_i^{-1}\left\{\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2 - \mathbb{E}(\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2)\right\} \leqslant x\right] - \mathbf{\Phi}(x)\right| \\
\lesssim &\frac{1}{B_i} + \frac{s}{\sqrt{T}} + \frac{s^{3/2}}{B_i\sqrt{T}} + \frac{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}\exp(-s/4)}{T^{1/8}B_i} + \exp(-s)
\end{aligned}$$

for any $s > 0$. By letting

$$s = \min\left\{\log(B_i), \log(T), \log\left(\frac{T^{1/8}B_i}{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}}\right)\right\},$$

we have

$$\begin{aligned}
&\sup_x \left|\mathbb{P}\left[TB_i^{-1}\left\{\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2 - \mathbb{E}(\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2)\right\} \leqslant x\right] - \mathbf{\Phi}(x)\right| \\
\lesssim &\frac{1}{B_i} + \frac{\log(T)}{\sqrt{T}} + \frac{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}}{T^{1/8}B_i}.
\end{aligned}$$

Similar to the proof of Theorem 6 in Koltchinskii and Lounici (2017), the conclusion follows. □

A.3 Auxiliary Lemmas

Lemma A.3.1. *Under Conditions 2.2.1-2.2.3, given* $\mathbf{Y}$ *in model* (2.1.1) *or* (2.1.2)*,*

(i) $\mathbb{E}(\|\mathbf{F}^\top\mathbf{U}^\top\|_{\mathbb{F}}^2) = O(pT)$*,* $\mathbb{E}(\|\mathbf{U}\|_2^2) = O(p)$*,* $\mathbb{E}(\|\mathbf{A}^\top\mathbf{U}\|_{\mathbb{F}}^2) = O(pT)$ *and* $\mathbb{E}(\|\mathbf{A}^\top\mathbf{U}\mathbf{F}\|_{\mathbb{F}}^2) = O(pT)$*.*

(ii) With probability at least $1-4e^{-s}$*,* $\|\mathbf{F}^\top\mathbf{U}^\top\|_{\mathbb{F}} \lesssim (pT)^{1/2}\sqrt{s}$*,* $\|\mathbf{U}\|_2 \lesssim p^{1/2}\sqrt{s}$*,* $\|\mathbf{A}^\top\mathbf{U}\|_{\mathbb{F}} \lesssim (pT)^{1/2}\sqrt{s}$ *and* $\|\mathbf{A}^\top\mathbf{U}\mathbf{F}\|_{\mathbb{F}} \lesssim (pT)^{1/2}\sqrt{s}$*.*

Proof. (i) The conclusion follows from Lemma D.2 in Wang and Fan (2017).

(ii) For any $x > 0$, it holds

$$\begin{aligned}
\mathbb{P}(\|\mathbf{F}^\top\mathbf{U}^\top\|_{\mathbb{F}}/\sqrt{C_0pT} > M) &\leqslant \exp(-xM)\mathbb{E}[\exp\{x\|\mathbf{F}^\top\mathbf{U}^\top\|_{\mathbb{F}}/\sqrt{C_0pT}\}] \\
&\leqslant \exp(-xM)\mathbb{E}\Big[1 + x\|\mathbf{F}^\top\mathbf{U}^\top\|_{\mathbb{F}}/\sqrt{C_0pT} \\
&\quad + x^2\|\mathbf{F}^\top\mathbf{U}^\top\|_{\mathbb{F}}^2/\{2C_0pT\} + o(x^2\|\mathbf{F}^\top\mathbf{U}^\top\|_{\mathbb{F}}^2/\{2C_0pT\})\Big] \\
&\leqslant \exp\{-xM + x + x^2/2 + o(x^2)\}
\end{aligned}$$

since $\mathbb{E}(\|\mathbf{F}^\top\mathbf{U}^\top\|_{\mathbb{F}}^2) \leqslant C_0pT$ for some $C_0 > 0$. The minimum of right hand side is $\exp\{-(M-1)^2/2\}$. Letting $s = 2^{-1}(M-1)^2$, we have with probability at least $1-e^{-s}$, $\|\mathbf{F}^\top\mathbf{U}^\top\|_{\mathbb{F}} \lesssim \sqrt{pTs}$. The remaining bounds can be derived similarly.

□

Denote $\mathbf{K}$ a $K\times K$ diagonal matrix with diagonals equal to the first K eigenvalues of $(pT)^{-1}\mathbf{Y}^\top\mathbf{Y}$. Then $(pT)^{-1}\mathbf{Y}^\top\mathbf{Y}\widehat{\mathbf{F}} = \widehat{\mathbf{F}}\mathbf{K}$. Let

$$\mathbf{H} = (pT)^{-1}\mathbf{A}^\top\mathbf{A}\mathbf{F}^\top\widehat{\mathbf{F}}\mathbf{K}^{-1}.$$

By model (2.1.2), we have $\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H} = \left(\sum_{i=1}^3 \mathbf{M}_i\right)\mathbf{K}^{-1}$ where

$$\mathbf{M}_1 = \frac{1}{pT}\mathbf{F}\mathbf{A}^\top\mathbf{U}\widehat{\mathbf{F}}, \quad \mathbf{M}_2 = \frac{1}{pT}\mathbf{U}^\top\mathbf{A}\mathbf{F}^\top\widehat{\mathbf{F}}, \quad \mathbf{M}_3 = \frac{1}{pT}\mathbf{U}^\top\mathbf{U}\widehat{\mathbf{F}}.$$

Then, we will provide a bound on $\|\mathbf{H}-\mathbf{I}\|_{\mathbb{F}}$ using Lemmas A.3.2 to A.3.6.

Lemma A.3.2. *Under Conditions 2.2.1-2.2.3, with probability at least* $1-5e^{-s}$, $\|\mathbf{K}^{-1}\|_2 \lesssim 1+T^{-1/2}\sqrt{s}$.

Proof. The K largest eigenvalues of $(pT)^{-1}\mathbf{Y}^\top\mathbf{Y}$ are the same as those of $\mathbf{W}=(pT)^{-1}\mathbf{Y}\mathbf{Y}^\top$. As $\mathbf{Y}=\mathbf{A}\mathbf{F}^\top+\mathbf{U}$, we have $\mathbf{W}=\sum_{i=1}^5\mathbf{W}_i$ where

$$\mathbf{W}_1=\frac{1}{p}\mathbf{A}\mathbf{A}^\top,\quad \mathbf{W}_2=\frac{1}{pT}\mathbf{A}\mathbf{F}^\top\mathbf{U}^\top,\quad \mathbf{W}_3=\mathbf{W}_2^\top,$$
$$\mathbf{W}_4=\frac{1}{pT}\mathbf{U}\mathbf{U}^\top,\quad \mathbf{W}_5=\frac{1}{p}\mathbf{A}\left(\frac{1}{T}\mathbf{F}^\top\mathbf{F}-\mathbf{I}\right)\mathbf{A}^\top.$$

By Lemma A.3.1, with probability at least $1-4e^{-s}$,

$$\|\mathbf{W}_2\|_2 \leqslant (pT)^{-1}\|\mathbf{A}\|_2\left\|\mathbf{F}^\top\mathbf{U}\right\|_{\mathbb{F}} \lesssim T^{-1/2}\sqrt{s},$$

and

$$\|\mathbf{W}_4\|_2 \leqslant (pT)^{-1}\|\mathbf{U}\|_2^2 \lesssim T^{-1}s.$$

By Condition 2.2.1, with probability at least $1-e^{-s}$, $\|\mathbf{W}_5\|_2 \lesssim T^{-1/2}\sqrt{s}$. For $k=1,\ldots,K$, $|\lambda_k(\mathbf{W})-\lambda_k(\mathbf{W}_1)| \leqslant \|\mathbf{W}-\mathbf{W}_1\|_2$. This implies, with probability at least $1-5e^{-s}$, $|\lambda_k(\mathbf{W})-\lambda_k(\mathbf{W}_1)| \lesssim T^{-1/2}\sqrt{s}$ for each $k=1,\ldots,K$. Note that the K largest eigenvalues of $\mathbf{W}_1$ is also the K largest eigenvalues of $p^{-1}\mathbf{A}^\top\mathbf{A}$. Thus, with probability at least $1-5e^{-s}$, $\|\mathbf{K}^{-1}\|_2 \lesssim 1+T^{-1/2}\sqrt{s}$. □

Lemma A.3.3. *Under Conditions 2.2.1-2.2.3, with probability at least* $1-5e^{-s}$,

$$\frac{1}{T}\|\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H}\|_{\mathbb{F}}^2 \lesssim \left(\frac{1}{p}+\frac{1}{T^2}\right)\left(1+\frac{s}{T}\right)s^2.$$

Proof. Note that $\|\widehat{\mathbf{F}}\|_{\mathbb{F}}=\sqrt{KT}$ with probability 1 and by Condition 2.2.1,

$$\|\mathbf{F}\|_{\mathbb{F}} \lesssim \sqrt{T}\{1+T^{-1/2}\sqrt{s}\}$$

with probability at least $1 - e^{-s}$. Then, by Lemma A.3.1, with probability at least $1 - 5e^{-s}$, $\|\mathbf{M}_1\|_{\mathbb{F}}, \|\mathbf{M}_2\|_{\mathbb{F}} \lesssim \sqrt{p^{-1}Ts}$ and $\|\mathbf{M}_3\|_{\mathbb{F}} \lesssim T^{-1/2}s$. Then, the results follows Lemma A.3.2. □

Lemma A.3.4. *Under Conditions 2.2.1-2.2.3, with probability at least* $1 - 5e^{-s}$,

(i) $T^{-1}\|\mathbf{M}_1\|_{\mathbb{F}}^2 \lesssim (p^{-2} + p^{-1}T^{-1})(1 + T^{-1}s)s^3$,

(ii) $T^{-2}\|\mathbf{F}^\top \mathbf{M}_2\|_{\mathbb{F}}^2 \lesssim (pT)^{-1}(1 + T^{-1}s)s$,

(iii) $T^{-2}\|\mathbf{F}^\top(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\mathbb{F}}^2 \lesssim (p^{-2} + p^{-1}T^{-1})(1 + T^{-1}s)s^3$,

(iv) $T^{-2}\|\widehat{\mathbf{F}}^\top(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\mathbb{F}}^2 \lesssim (p^{-2} + p^{-1}T^{-1})(1 + T^{-1}s)s^3$.

Proof. (i) With probability at least $1 - 5e^{-s}$,

$$\|\mathbf{H}\|_2 \leqslant (pT)^{-1}\|\mathbf{A}\|_{\mathbb{F}}^2\|\mathbf{F}\|_{\mathbb{F}}\|\widehat{\mathbf{F}}\|_{\mathbb{F}}\|\mathbf{K}^{-1}\|_2 \lesssim 1 + sT^{-1}$$

by Lemma A.3.2. Then by Lemmas A.3.1 and A.3.3, with probability at least $1 - 5e^{-s}$,

$$\begin{aligned}\|\mathbf{A}^\top\mathbf{U}\widehat{\mathbf{F}}\|_{\mathbb{F}}^2 &\leqslant 2\|\mathbf{A}^\top\mathbf{U}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\mathbb{F}}^2 + 2\|\mathbf{A}^\top\mathbf{U}\mathbf{F}\mathbf{H}\|_{\mathbb{F}}^2 \\ &\lesssim pT\left(\frac{T}{p} + 1\right)s^3 + pTs\left(1 + \frac{s^2}{T^2}\right) \\ &\lesssim (T^2 + pT)s^3.\end{aligned}$$

The result follows that $\|\mathbf{F}\|_{\mathbb{F}} \lesssim \sqrt{T}\{1 + T^{-1/2}\sqrt{s}\}$ with probability at least $1 - e^{-s}$.

(ii) Similar to (i), with probability at least $1 - 5e^{-s}$,

$$\frac{1}{T^2}\|\mathbf{F}^\top\mathbf{M}_2\|_{\mathbb{F}}^2 \leqslant \frac{1}{p^2T^4}\|\mathbf{F}^\top\mathbf{U}^\top\mathbf{A}\|_{\mathbb{F}}^2\|\mathbf{F}\|_{\mathbb{F}}^2\|\widehat{\mathbf{F}}\|_{\mathbb{F}}^2 \lesssim \frac{s}{pT}\left(1 + \frac{s}{T}\right).$$

(iii) Combining (i) and (ii), the result follows from the proof of Lemma A.3.3.

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

□

Lemma A.3.5. *Under Conditions 2.2.1-2.2.3, with probability at least* $1 - 5e^{-s}$*,*

$$\|\mathbf{H}^\top \mathbf{H} - \mathbf{I}_K\|_{\mathbb{F}}^2 \lesssim \left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right)^3.$$

Proof. By Condition 2.2.1, $\|T^{-1}\mathbf{F}^\top\mathbf{F} - \mathbf{I}_K\|_{\mathbb{F}} \lesssim T^{-1/2}\sqrt{s}$ with probability at least $1 - e^{-s}$. Also, $\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}} = T\mathbf{I}_K$. Thus,

$$\mathbf{H}^\top\mathbf{H} - \mathbf{I}_K = \mathbf{H}^\top\left(\mathbf{I}_K - \frac{1}{T}\mathbf{F}^\top\mathbf{F}\right)\mathbf{H} + \frac{1}{T}(\mathbf{F}\mathbf{H} - \widehat{\mathbf{F}})^\top\mathbf{F}\mathbf{H} + \frac{1}{T}\widehat{\mathbf{F}}^\top(\mathbf{F}\mathbf{H} - \widehat{\mathbf{F}}).$$

The result follows from Lemma A.3.4. □

Lemma A.3.6. *Under Conditions 2.2.1-2.2.3, with probability at least* $1 - 5e^{-s}$*,*

$$\|\mathbf{H} - \mathbf{I}_K\|_{\mathbb{F}}^2 \lesssim \left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right).$$

Proof. Note that $p\mathbf{H}\mathbf{K} = \mathbf{A}^\top\mathbf{A}\left(T^{-1}\mathbf{F}^\top\widehat{\mathbf{F}} - \mathbf{H}\right) + \mathbf{A}^\top\mathbf{A}\mathbf{H}$. By Lemma A.3.4, with probability at least $1 - 5e^{-s}$,

$$\begin{aligned}
&\left\|\frac{1}{p}\mathbf{A}^\top\mathbf{A}\left(\frac{1}{T}\mathbf{F}^\top\widehat{\mathbf{F}} - \mathbf{H}\right)\right\|_{\mathbb{F}}^2 \\
\leqslant&\frac{1}{p^2}\|\mathbf{A}^\top\mathbf{A}\|_{\mathbb{F}}^2\frac{1}{T^2}\|\mathbf{F}^\top(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\mathbb{F}}^2 + \frac{1}{p^2}\|\mathbf{A}^\top\mathbf{A}\|_{\mathbb{F}}^2\left\|\mathbf{I}_K - \frac{1}{T}\mathbf{F}^\top\mathbf{F}\right\|_{\mathbb{F}}^2\|\mathbf{H}\|_2^2 \\
\lesssim&\left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right).
\end{aligned}$$

Therefore, with probability at least $1 - 5e^{-s}$,

$$\left\|\frac{1}{p}\mathbf{A}^\top\mathbf{A}\mathbf{H} - \mathbf{H}\mathbf{K}\right\|_{\mathbb{F}}^2 \lesssim \left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right).$$

This implies that with probability at least $1 - 5e^{-s}$, $\mathbf{H}$ (up to an error term) is a matrix consisting of eigenvectors of $p^{-1}\mathbf{A}^\top\mathbf{A}$. By Condition 2.2.1, $\mathbf{A}^\top\mathbf{A}$ 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 $1 - 5e^{-s}$,

$$\|\mathbf{H} - \mathbf{J}\|_{\mathbb{F}}^2 \lesssim \left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right)$$

for some diagonal matrix $\mathbf{J}$. By Lemma A.3.5, with probability at least $1 - 5e^{-s}$, for each $k = 1, \ldots, K$,

$$|\lambda_k(\mathbf{H}) - \eta|^2 \lesssim \left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right)$$

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 $\widehat{\mathbf{A}}$ by -1). Hence, with probability at least $1 - 5e^{-s}$,

$$\|\mathbf{H} - \mathbf{I}_K\|_{\mathbb{F}}^2 = \sum_{i \neq j} h_{ij}^2 + \sum_{i=1}^{K}(h_{ii} - 1)^2 \lesssim \left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right).$$

□

Lemma A.3.7. *Under Conditions 2.2.1-2.2.3, with probability at least* $1 - 6e^{-s}$,

$$\|\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}\|_{\max} \lesssim \left(\frac{1}{\sqrt{p}} + \frac{1}{T}\right)\{\log(T)\}^{2/r_3} s. \tag{A.3.1}$$

Proof. By Lemma D.2 in Fan et al. (2013), with probability at least $1 - 2e^{-s}$,

$$\|\mathbf{U}^\top \mathbf{U}\|_{\max} \lesssim (\sqrt{p}T + p)s$$

and

$$\|\mathbf{A}^\top \mathbf{U}\|_{\max} \lesssim \sqrt{p}Ts.$$

Also, with probability at least $1 - e^{-s}$, $\|\mathbf{F}\|_{\max} \lesssim \{\log(T) + s\}^{1/r_3}$. Thus, with probability at least $1 - 3e^{-s}$, $\|\mathbf{M}_1\|_{\max}, \|\mathbf{M}_2\|_{\max} \lesssim p^{-1/2}$
$\{\log(T)\}^{2/r_3} s$ and $\|\mathbf{M}_3\|_{\max} \lesssim (p^{-1/2} + T^{-1})\{\log(T)\}^{1/r_3} s$. The result follows from Lemma A.3.2 that $\|\mathbf{K}^{-1}\|_2 \lesssim 1 + T^{-1/2}\sqrt{s}$ with probability at least $1 - 5e^{-s}$. □

Recall that $\widehat{\mathbf{A}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}$. We have $\widehat{\mathbf{A}} - \mathbf{AH} = \sum_{i=1}^{3} \mathbf{C}_i$ where

$$\mathbf{C}_1 = \frac{1}{T}\mathbf{A}\mathbf{F}^\top(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}), \quad \mathbf{C}_2 = \frac{1}{T}\mathbf{U}\mathbf{F}\mathbf{H}, \quad \mathbf{C}_3 = \frac{1}{T}\mathbf{U}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}).$$

Lemma A.3.8. *Under Conditions 2.2.1-2.2.3, with probability at least* $1 - 7e^{-s}$*,*

(i) $p^{-1}\|\widehat{\mathbf{A}} - \mathbf{AH}\|_{\mathbb{F}}^2 \lesssim (T^{-1}s + T^{-2}s^2 + p^{-1}T^{-1}s^3 + p^{-2}s^3)(1 + T^{-1}s)$,

(ii) $\|\widehat{\mathbf{A}} - \mathbf{AH}\|_{\max} \lesssim T^{-1/2}\{\log(p)\}^{1/2}s$,

(iii) $\|\widehat{\mathbf{A}} - \mathbf{AH}^{-1}\|_{\max} \lesssim T^{-1/2}\{\log(p)\}^{1/2}s$.

Proof. (i) By Lemmas A.3.1 and A.3.3, with probability at least $1 - 7e^{-s}$, $\|\mathbf{C}_1\|_{\mathbb{F}}^2 \lesssim (p^{-1} + T^{-1})(1 + T^{-1}s)s^3$, $\|\mathbf{C}_2\|_{\mathbb{F}}^2 \lesssim T^{-1}p(1 + T^{-1}s)^2 s$ and $\|\mathbf{C}_3\|_{\mathbb{F}}^2 \lesssim (T^{-1} + pT^{-3})(1 + T^{-1}s)s^3$. So

$$p^{-1}\|\widehat{\mathbf{A}} - \mathbf{AH}\|_{\mathbb{F}}^2 \lesssim (pT^{-1}s + pT^{-2}s^2 + T^{-1}s^3 + p^{-1}s^3)(1 + T^{-1}s).$$

(ii) By Lemma B.1 in Fan et al. (2011), with probability at least $1 - e^{-s}$, $\|\mathbf{F}^\top\mathbf{U}^\top\|_{\max} \lesssim \sqrt{T\log(p)}s$. Then, with probability at least $1 - 7e^{-s}$, $\|\mathbf{C}_1\|_{\max} \lesssim p^{-1}s$, $\|\mathbf{C}_2\|_{\max} \lesssim T^{-1/2}\sqrt{\log(p)}s$, and $\|\mathbf{C}_3\|_{\max} \lesssim (pT)^{-1/2}\{\log(T)\}^{1/r_3}s$. So we have

$$\|\widehat{\mathbf{A}} - \mathbf{AH}\|_{\max} \lesssim T^{-1/2}\{\log(p)\}^{1/2}s.$$

(iii) The result follows from $T(\widehat{\mathbf{A}} - \mathbf{AH}^{-1}) = \mathbf{AH}^{-1}(\mathbf{HF}^\top - \widehat{\mathbf{F}}^\top)\widehat{\mathbf{F}} + \mathbf{U}(\widehat{\mathbf{F}} - \mathbf{FH}) + \mathbf{UFH}$.

□

Lemma A.3.9. *Under Conditions 2.2.1-2.2.3,*

$$\mathbb{E}\left\{\left\|\frac{1}{T}\mathbf{Y}\mathbf{Y}^\top - \frac{1}{T}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)\right\|_2\right\} \lesssim \sqrt{\frac{p}{T}}.$$

Proof. Recall that

$$T^{-1}\mathbf{Y}\mathbf{Y}^\top = T^{-1}\mathbf{A}\mathbf{F}^\top\mathbf{F}\mathbf{A}^\top + T^{-1}\mathbf{A}\mathbf{F}^\top\mathbf{U}^\top + T^{-1}\mathbf{U}\mathbf{F}\mathbf{A}^\top + T^{-1}\mathbf{U}\mathbf{U}^\top$$

and $T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top) = \mathbf{A}\mathbf{A}^\top + T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$. By Condition 2.2.3, with probability at least $1 - e^{-s}$,

$$\mathbb{E}\{\|T^{-1}\mathbf{A}\mathbf{F}^\top\mathbf{F}\mathbf{A}^\top - \mathbf{A}\mathbf{A}^\top\|_2\} \lesssim p^{1/2}T^{-1/2}s^{1/2}.$$

By Lemma A.3.1, with probability at least $1 - 4e^{-s}$,

$$\mathbb{E}(\|T^{-1}\mathbf{U}\mathbf{F}\mathbf{A}^\top\|_2) \lesssim p^{1/2}T^{-1/2}s.$$

By Theorem 1 in Koltchinskii and Lounici (2017), with probability at least $1 - e^{-s}$,

$$\mathbb{E}\{\|T^{-1}\mathbf{U}\mathbf{U}^\top - T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)\|_2\} \lesssim p^{1/2}T^{-1/2} + T^{-1/2}s^{1/2}.$$

The conclusion follows. □

Lemma A.3.10. *Under Conditions 2.2.1-2.3.1, for each $i = 1, \ldots, p$,*

$$\sup_x \left| \mathbb{P}\left\{ \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (\mathbf{P}_i \boldsymbol{y}_t)(\mathbf{Q}_i \boldsymbol{y}_t)^\top \right\|_2^2 \leqslant x \right\} - \mathbb{P}\left(\gamma_i \|\mathbf{Q}_i \boldsymbol{z}\|_2^2 \leqslant x\right) \right| \lesssim \frac{\{\log(T)\}^2 \log(p)}{T^{1/2}(1 + x^4)},$$

where γ_i is the eigenvalue of $\mathbf{\Gamma}_i = T^{-1}\sum_{t=1}^{T}(\mathbf{P}_i\boldsymbol{y}_t)(\mathbf{P}_i\boldsymbol{y}_t)^\top$, $\boldsymbol{z} \sim \mathcal{N}(\sigma_f^2\mathbf{A}\mathbf{A}^\top + \sigma_u^2\mathbf{I}_p)$ independent of $\mathbf{\Gamma}_i$, $\sigma_f^2 = T^{-1}\sum_{t=1}^{T}\mathbb{E}(f_{11}f_{t1})$ and $\sigma_u^2 = T^{-1}\sum_{t=1}^{T}\mathbb{E}(u_{11}u_{1t})$.

Proof. Recall that $\boldsymbol{y}_t = \mathbf{A}\boldsymbol{f}_t + \boldsymbol{u}_t$,

$$\begin{aligned}(\mathbf{P}_i\boldsymbol{y}_t)(\mathbf{Q}_i\boldsymbol{y}_t)^\top &= \mathbf{P}_i\boldsymbol{y}_t\boldsymbol{y}_t^\top\mathbf{Q}_i \\ &= \mathbf{P}_i\mathbf{A}\boldsymbol{f}_t\boldsymbol{f}_t^\top\mathbf{A}^\top\mathbf{Q}_i + \mathbf{P}_i\mathbf{A}\boldsymbol{f}_t\boldsymbol{u}_t^\top\mathbf{Q}_i + \mathbf{P}_i\boldsymbol{u}_t\boldsymbol{f}_t^\top\mathbf{A}^\top\mathbf{Q}_i + \mathbf{P}_i\boldsymbol{u}_t\boldsymbol{u}_t^\top\mathbf{Q}_i.\end{aligned}$$

By Condition 2.2.3 and Theorem 2 in Jirak (2016), for each $k = 1, \ldots, K$ and $i = 1, \ldots, p$,

$$\sup_x \left| \mathbb{P}\left(\frac{\sum_{t=1}^T f_{tk}}{\sqrt{T}\sigma_f} \leqslant x \right) - \mathbf{\Phi}(x) \right| \lesssim \frac{\{\log(T)\}^2}{T^{1/2}(1+x^4)}$$

and

$$\sup_x \left| \mathbb{P}\left(\frac{\sum_{t=1}^T u_{it}}{\sqrt{T}\sigma_u} \leqslant x \right) - \mathbf{\Phi}(x) \right| \lesssim \frac{\{\log(T)\}^2}{T^{1/2}(1+x^4)},$$

where $\sigma_f^2 = T^{-1}\sum_{t=1}^T \mathbb{E}(f_{11}f_{t1})$ and $\sigma_u^2 = T^{-1}\sum_{t=1}^T \mathbb{E}(u_{11}u_{1t})$. The conclusion follows from the proof of Theorem 4 in Koltchinskii and Lounici (2016). □

Lemma A.3.11. *Under Conditions 2.2.1-2.3.1, for each $i = 1, \ldots, p$, with probability at least $1 - e^{-s}$,*

$$\begin{aligned}\left| \|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2 - \mathbb{E}(\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2) \right| \lesssim & \frac{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}}{T^{9/8}e^{-s/4}} + \frac{B_i\sqrt{s}}{T} \\ & + \frac{\|T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)\|_2^2 s}{\bar{g}_i^2 T} + \frac{C_i\sqrt{s}}{T^{3/2}},\end{aligned}$$

where $\bar{g}_i = \min(\lambda_i - \lambda_{i+1}, \lambda_{i-1} - \lambda_i)$ for $i = 2, \ldots, p$ and $\bar{g}_1 = \lambda_1 - \lambda_2$.

Proof. By Lemma 1 in Koltchinskii and Lounici (2016), $\widehat{\mathbf{P}}_i - \mathbf{P}_i = \mathbf{L}_i + \mathbf{S}_i$, where $\mathbf{L}_i = \mathbf{Q}_i\mathbf{D}\mathbf{P}_i + \mathbf{P}_i\mathbf{D}\mathbf{Q}_i$, $\|\mathbf{S}_i\|_2 \lesssim \bar{g}_i^{-2}\|\mathbf{D}\|_2^2$, and $\mathbf{D} = T^{-1}\mathbf{Y}\mathbf{Y}^\top - T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)$. By Lemma A.3.9, with probability at least $1 - e^{-s}$,

$$\|\mathbf{D}\|_2 \lesssim p^{1/2}T^{-1/2} + T^{-1/2}s^{1/2}.$$

Notice $\mathbf{P}_i\mathbf{D}\mathbf{Q}_i = T^{-1}\sum_{t=1}^T (\mathbf{P}_i\boldsymbol{y}_t)(\mathbf{Q}_i\boldsymbol{y}_t)^\top$, we have

$$\begin{aligned}\|\mathbf{L}_i\|_2^2 &= \|\mathbf{Q}_i\mathbf{D}\mathbf{P}_i + \mathbf{P}_i\mathbf{D}\mathbf{Q}_i\|_2^2 \\ &= 2\|\mathbf{P}_i\mathbf{D}\mathbf{Q}_i\|_2^2 \\ &= \frac{2}{T}\left\| \frac{1}{\sqrt{T}}\sum_{t=1}^T (\mathbf{P}_i\boldsymbol{y}_t)(\mathbf{Q}_i\boldsymbol{y}_t)^\top \right\|_2^2.\end{aligned}$$

By Lemma A.3.10, with probability at least $1 - e^{-s}$,

$$\begin{aligned}&\left|\{\|\mathbf{L}_i\|_2^2 - \mathbb{E}(\|\mathbf{L}_i\|_2^2)\} - \{2^{-1}T\gamma_i\|\mathbf{Q}_i\boldsymbol{z}\|_2^2 - 2^{-1}T\mathbb{E}(\gamma_i\|\mathbf{Q}_i\boldsymbol{z}\|_2^2)\}\right| \\ \lesssim &\frac{\{\log(T)\}^{1/2}\{\log(p)\}^{1/4}}{T^{9/8}e^{-s/4}}.\end{aligned}$$

Following the proof of Theorem 4 in Koltchinskii and Lounici (2017), with probability at least $1 - e^{-s}$,

$$\left|\{2^{-1}T\gamma_i\|\mathbf{Q}_i\boldsymbol{z}\|_2^2 - 2^{-1}T\mathbb{E}(\gamma_i\|\mathbf{Q}_i\boldsymbol{z}\|_2^2)\}\right| \lesssim \frac{B_i\sqrt{s}}{T} + \frac{\|T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}')\|_\infty^2 s}{\bar{g}_i^2 T} + \frac{C_i\sqrt{s}}{T^{3/2}},$$

where B_i and C_i are given before proof of Theorem 2.4.6. The lemma is proved. □

Lemma A.3.12. *Under Conditions 2.2.1-2.3.1, for each $i = 1, \ldots, p$,*

$$\left|TB_i^{-1}\operatorname{Var}^{1/2}(\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2) - 1\right| \lesssim \frac{\log(T)\sqrt{\log(p)}}{T^{1/4}B_i} + \frac{1}{T} + \frac{C_i}{\sqrt{T}B_i}.$$

Proof. By Lemma A.3.10,

$$\begin{aligned}&\left|\operatorname{Var}(\|\mathbf{L}_i\|_2^2) - \frac{4}{T^2}\operatorname{Var}(\gamma_i\|\mathbf{Q}_i\boldsymbol{z}\|_2^2)\right| \\ \leqslant &\frac{4}{T^2}\int_0^\infty \left|\mathbb{P}\left\{\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T(\mathbf{P}_i\boldsymbol{y}_t)(\mathbf{Q}_i\boldsymbol{y}_t)^\top\right\|_2^2 > \sqrt{x}\right\} - \mathbb{P}\left(\gamma_i\|\mathbf{Q}_i\boldsymbol{z}\|_2^2 > \sqrt{x}\right)\right| dx \\ \lesssim &\frac{\{\log(T)\}^2\log(p)}{T^{5/2}}.\end{aligned}$$

Notice $4\operatorname{Var}(\gamma_i\|\mathbf{Q}_i\boldsymbol{z}\|_2^2) = \{(T+1)B_i^2 + 2C_i^2\}T^{-1}$ by the proof of Theorem 6 in Koltchinskii and Lounici (2017). Thus, we have

$$\operatorname{Var}(\|\mathbf{L}_i\|_2^2) \lesssim \frac{1}{T^2}\left[\frac{\{\log(T)\}^2\log(p)}{T^{1/2}} + \frac{T+1}{T}B_i^2 + \frac{2}{T}C_i^2\right]$$

and

$$\left|TB_i^{-1}\operatorname{Var}^{1/2}(\|\hat{\mathbf{P}}_i-\mathbf{P}_i\|_2^2)-1\right|\lesssim\frac{\log(T)\sqrt{\log(p)}}{T^{1/4}B_i}+\frac{1}{T}+\frac{C_i}{\sqrt{T}B_i}.$$

□

Lemma A.3.13. *Under Conditions 2.2.1-2.3.1, for each* $k=1,\ldots,K$,

$$\left|p\widetilde{B}_k^{-1}\operatorname{Var}^{1/2}(\|\bar{\mathbf{P}}_k-\widetilde{\mathbf{P}}_k\|_2^2)-1\right|\lesssim\frac{1}{p}+\frac{\widetilde{C}_k}{\sqrt{p}\widetilde{B}_k}.$$

Proof. By Lemma 1 in Koltchinskii and Lounici (2016), $\bar{\mathbf{P}}_k-\widetilde{\mathbf{P}}_k=\widetilde{\mathbf{L}}_k+\widetilde{\mathbf{S}}_k$, where $\widetilde{\mathbf{D}}=p^{-1}\mathbf{Y}^\top\mathbf{Y}-p^{-1}\mathbf{F}\mathbf{A}^\top\mathbf{A}\mathbf{F}^\top$, $\widetilde{\mathbf{L}}_k=\widetilde{\mathbf{Q}}_k\widetilde{\mathbf{D}}\widetilde{\mathbf{P}}_k+\widetilde{\mathbf{P}}_k\widetilde{\mathbf{D}}\widetilde{\mathbf{Q}}_k$ and $\|\widetilde{\mathbf{S}}_k\|_2\lesssim\widetilde{g}_k^{-2}\|\widetilde{\mathbf{D}}\|_2^2$, where

$$\widetilde{g}_k=\min\{\lambda_k(\mathbf{A}^\top\mathbf{A})-\lambda_{k+1}(\mathbf{A}^\top\mathbf{A}),\lambda_{k-1}(\mathbf{A}^\top\mathbf{A})-\lambda_k(\mathbf{A}^\top\mathbf{A})\}$$

for $k=2,\ldots,K$ and $\bar{g}_1=\lambda_1(\mathbf{A}^\top\mathbf{A})-\lambda_2(\mathbf{A}^\top\mathbf{A})$. By Lemma A.3.1, with probability at least $1-3e^{-s}$, $\|\mathbf{D}\|_2\lesssim p^{-2}T^{-1}s$. Similar to Lemmas A.3.11 and A.3.12 with $\widetilde{B}_k$ and $\widetilde{C}_k$ defined before the proof of Theorem 2.3.2, the lemma is proved. □

Appendix B

Supplemental materials for Chapter 3

This online supplementary material contains proofs of the main theorems, technical details, and extra results from numerical studies. Proofs of the main theorems are included in Section B.1. Technical details and discussions are reported in Section B.2. The theoretical validity of our procedure on selecting K and its numerical comparison with the competing eigenvalue-ratio based procedures are documented in Section B.3. Section B.4 displays extra results from simulation studies. We firs collect some notation throughout this supplementary files.

Notation. For a matrix $\mathbf{M} = (m_{ij})_{1\leqslant i,j\leqslant p} \in \mathbb{R}^{p\times p}$, write $||\mathbf{M}||_{\mathbb{F}} = (\sum_{i=1}^{p}\sum_{j=1}^{p} m_{ij}^2)^{1/2}$ to be the Frobenius norm, $\|\mathbf{M}\|_{\max} = \max_{i,j}|m_{ij}|$ to be the maximum norm and $\|\mathbf{M}\|_{\infty} = \max_i \sum_j |m_{ij}|$ to be the induced ℓ_∞ norm. The spectral norm of matrix $\mathbf{M}$ 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 $\boldsymbol{a} = (a_1, \ldots, a_p)'$ is defined by $||\boldsymbol{a}||_q = (\sum_{j=1}^{p}|a_j|^q)^{1/q}$ with $1 \leqslant q < \infty$. Denote the minimum and maximum eigenvalues of matrix $\mathbf{M}$ by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$, respectively. Let $\text{tr}(\mathbf{M}) = \sum_{j=1}^{p} m_{jj}$ be the trace of $\mathbf{M}$, $\text{vec}(\mathbf{M})$ be the vectorization of $\mathbf{M}$, and $\otimes$ be the Kronecker product. We write $\mathbf{I}$ for an identify 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 $a_n \leqslant Cb_n$ for some C independent of n; and $a_n \asymp b_n$ if and only if there exists C, D independent on n such that $C|b_n| \leqslant |a_n| \leqslant D|b_n|$. 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.

In this supplementary file, we constantly explore the tail probability of random variable X in the following sense: with probability at least $1 - \delta$, $X \lesssim \{1 + \log(1/\delta)\}$. Such an inequality is often proved with probability $1 - C\delta$ instead, where $C > 0$ is an absolute constant. In such cases, it is easy to show that the inequality still holds with the original probability. By replacing δ with

δ/C, we claim that with probability at least $1-\delta$, $X \lesssim \{1+\log(C/\delta)\} \lesssim \{1+\log(1/\delta)\}$. Thus, it will be said without further explanation that probability bound $1-C\delta$ can be replaced by $1-\delta$ via adjusting the constant. Also, note that $1+\log(1/\delta) = \log(e/\delta)$, so we can replace concentration bound $1+\log(1/\delta)$ by $\log(1/\delta)$. See Koltchinskii and Lounici (2017) for a similar discussion.

B.1 Proof of the main results

B.1.1 Invertibility of the projection matrix

Without loss of generality, we take $\mathcal{X}^d = [0,1]^d$. Consider coefficients $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^{Jd+1}$, where

$$\boldsymbol{a}_k = \left(a_0^{(k)}, a_{11}^{(k)}, \ldots, a_{J1}^{(k)}, \ldots, a_{1d}^{(k)}, \ldots, a_{Jd}^{(k)}\right)' \in \mathbb{R}^{Jd+1}, \quad k = 1, 2, \ldots$$

and define

$$\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle_n = \frac{1}{n}\sum_i \{a_0^{(1)} + \sum_j \sum_\ell a_{j\ell}^{(1)} \phi_j(X_{i\ell})\}\{a_0^{(2)} + \sum_j \sum_\ell a_{j\ell}^{(2)} \phi_j(X_{i\ell})\}. \tag{B.1.1}$$

In literature, conditions on the largest and smallest eigenvalues of $n^{-1}\boldsymbol{\Phi}'\boldsymbol{\Phi}$ are usually stated as an important assumption for theoretical guarantees (e.g. Fan et al., 2016). However, under standard nonparametric settings, we can establish it as following.

Lemma B.1.1. *Under Condition 3.3.2, whenever $J = o(\sqrt{n})$ and $d < J^{-1}n$, with probability at least $1-\delta$,*

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

where

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

as defined in Section 3.2.2.

Proof. From (B.1.1), $\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n = \boldsymbol{a}' \left(n^{-1} \boldsymbol{\Phi}' \boldsymbol{\Phi} \right) \boldsymbol{a}$ for any $\boldsymbol{a} \in \mathbb{R}^{Jd+1}$. For any $\delta > 0$, let

$$\mathcal{A}_\delta = \left\{ |\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n - \mathbb{E} \left(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n \right)| \gtrsim \frac{J}{n} \log(J^2/\delta) \mathbb{E}(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n) \right\}.$$

On $\mathcal{A}_\delta^c$, we have

$$\left\{ 1 - \frac{J}{n} \log(J^2/\delta) \right\} \mathbb{E} \left(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n \right) \lesssim \langle \boldsymbol{a}, \boldsymbol{a} \rangle_n \lesssim \left\{ 1 + \frac{J}{n} \log(J^2/\delta) \right\} \mathbb{E} \left(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n \right).$$

By Lemma B.2.4, $\mathbb{E} \left(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n \right) \asymp \| \boldsymbol{a} \|_2^2$. Thus, we have

$$\left\{ 1 - \frac{J}{n} \log(J^2/\delta) \right\} \| \boldsymbol{a} \|_2^2 \lesssim \boldsymbol{a}' \left(n^{-1} \boldsymbol{\Phi}' \boldsymbol{\Phi} \right) \boldsymbol{a} \lesssim \left\{ 1 + \frac{J}{n} \log(J^2/\delta) \right\} \| \boldsymbol{a} \|_2^2.$$

The conclusion follows from Lemma B.2.5, which implies $\mathbb{P}\{\mathcal{A}_\delta\} < \delta$. □

B.1.2 Proof of Theorems 3.3.1 and 3.3.3

Theorems 3.3.1 and 3.3.3 readily follow Propositions B.2.1–B.2.4.

B.1.3 Proof of Theorem 3.3.2

Recall that $\widehat{\mathbf{V}} = \widehat{\mathbf{G}} \mathcal{V}(\widehat{\boldsymbol{f}}_t) \widehat{\mathbf{G}}' + \widehat{\boldsymbol{\mathcal{D}}}$, similarly to the proof of Lemma B.2.17, we have

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

with probability at least $1 - \delta$. Then, by Lemma B.2.17, with probability at least $1 - \delta$,

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

By Lemmas B.2.2 and B.2.3, $\|\mathbb{Z}_0'(\mathbf{G}\bar{\boldsymbol{f}} + \bar{\boldsymbol{u}})\|_2 \lesssim \|\mathbb{Z}_0\|_{\mathbb{F}} T^{-1/2}\sqrt{\log(1/\delta)}$ with probability at least $1-\delta$. Thus, with probability at least $1-\delta$,

$$\begin{aligned}
\left\|\bar{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right\|_2 &\leqslant \|\{(\mathbb{Z}_0'\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1} - (\mathbb{Z}_0'\mathbf{V}^{-1}\mathbb{Z}_0)^{-1}\}\mathbb{Z}_0'\mathbf{V}^{-1}(\mathbf{G}\bar{\boldsymbol{f}} + \bar{\boldsymbol{u}})\|_2 \\
&\quad + \|(\mathbb{Z}_0'\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0'(\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1})(\mathbf{G}\bar{\boldsymbol{f}} + \bar{\boldsymbol{u}})\|_2 \\
&\leqslant \|(\mathbb{Z}_0'\mathbf{V}^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\|_2\|(\mathbb{Z}_0'\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\|_2\|\mathbf{V}^{-1}\|_2\|\mathbb{Z}_0'(\mathbf{G}\bar{\boldsymbol{f}} + \bar{\boldsymbol{u}})\|_2\|\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}\|_2 \\
&\quad + \|(\mathbb{Z}_0'\widehat{\mathbf{V}}^{-1}\mathbb{Z}_0)^{-1}\|_2\|\mathbb{Z}_0'(\mathbf{G}\bar{\boldsymbol{f}} + \bar{\boldsymbol{u}})\|_2\|\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}\|_2 \\
&\lesssim \frac{1}{\sqrt{nT}}\left\{\frac{\sqrt{J}}{n} + \frac{1}{\sqrt{n}} + \frac{1}{T} + \frac{p\sqrt{J}}{\sqrt{n^{1-2\alpha}T}} + \frac{1}{J^{(\kappa-1)/2}}\right\}\{1 + \sqrt{\log(1/\delta)}\}.
\end{aligned}$$

Therefore, for any $a > 0$,

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

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 $a = (nT)^{-1}C\vartheta_{n,T,J}^2\{1 + \log(21)\}$ gives

$$\mathbb{E}\left(\left\|\bar{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right\|_2^2\right) \leqslant \frac{2C\vartheta_{n,T,J}^2}{nT}.$$

For TOPE $\bar{\boldsymbol{\beta}}$ and the oracle generalized least squares (GLS) estimator $\widetilde{\boldsymbol{\beta}}$ whose jth components are denoted by $\bar{\beta}_j$ and $\widetilde{\beta}_j$ respectively, repeatedly employing Cauchy-Schwartz inequality to each of the (i, j) pair with $i, j = 1, \ldots, p$ leads to

$$|\operatorname{Cov}(\bar{\beta}_i, \bar{\beta}_j) - \operatorname{Cov}(\widetilde{\beta}_i, \widetilde{\beta}_j)|$$

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

which yields

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

B.1.4 Proof of Theorem 3.4.1

(i) For the oracle GLS estimator $\widetilde{\boldsymbol{\beta}}$, it holds

$$
\begin{aligned}
\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}=&(\mathbb{Z}_0'\mathbf{V}^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\mathbf{V}^{-1}\left(\mathbf{G}\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t+\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\right)\\
:=&\mathbf{A}\left(\mathbf{G}\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t+\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\right),
\end{aligned}
$$

where $\mathbf{A}=(\mathbb{Z}_0'\mathbf{V}^{-1}\mathbb{Z}_0)^{-1}\mathbb{Z}_0\mathbf{V}^{-1}$ and $\mathbf{V}=\mathbf{G}\,\mathrm{Var}(T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t)\mathbf{G}'+\mathrm{Var}(T^{-1}\sum_{t=1}^{T}u_{1t})\mathbf{I}_n$ as defined in (3.2.5) in the main paper. For any p-vector $\boldsymbol{c}$,

$$
(nT)^{1/2}\boldsymbol{c}'(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}):=\sum_{t=1}^{T}(W_{nt}+\widetilde{W}_{nt}),
$$

where $W_{nt}=n^{1/2}T^{-1/2}\boldsymbol{c}'\mathbf{A}\boldsymbol{u}_t$ and $\widetilde{W}_{nt}=n^{1/2}T^{-1/2}\boldsymbol{c}'\mathbf{A}\mathbf{G}\boldsymbol{f}_t$. Then

$$
\sum_{t=1}^{T}\mathbb{E}[|W_{nt}|^3]=n^{3/2}T^{-1/2}\|\boldsymbol{c}\|_2^3\mathbb{E}[\|\mathbf{A}\|_2^3\|]\mathbb{E}[\|\boldsymbol{u}_1\|_2^3]<\infty
$$

for any n, and since $\|\mathbf{A}\|_2 \leqslant \|\mathbf{A}\|_{\mathbb{F}} \leqslant p\|\mathbf{A}\|_2$ we have

$$\frac{\sum_{t=1}^{T}\mathbb{E}[|W_{nt}|^3]}{\left(\sum_{t=1}^{T}\mathbb{E}[W_{nt}^2]\right)^{3/2}} \leqslant \frac{T^{-1/2}\|\boldsymbol{c}\|_2^3\mathbb{E}\left[\|\mathbf{A}\|_2^3\right]\mathbb{E}\left[\|\boldsymbol{u}_1\|_2^3\right]}{\left\{\boldsymbol{c}'\mathbb{E}\left[\mathbf{A}\,\mathrm{Var}(\boldsymbol{u}_1)\mathbf{A}'\right]\boldsymbol{c}\right\}^{3/2}}$$
$$\leqslant \frac{T^{-1/2}\mathbb{E}\left[\|\mathbf{A}\|_2^3\right]\mathbb{E}\left[\|\boldsymbol{u}_1\|^3\right]}{\left\{\max_i \mathrm{Var}(u_{i1})\right\}^{3/2}\left\{\mathbb{E}\left[\|\mathbf{A}\|_{\mathbb{F}}^2\right]\right\}^{3/2}} \to 0$$

as T diverges to infinity. By Lyapunov central limit theorem, $\sum_{t=1}^{T} W_{nt}$ is hence asymptotically normal. Similarly, under Condition 3.3.4, 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 independent 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 $\boldsymbol{\Sigma} = \mathbb{E}\left[(\mathbb{Z}_0'\mathbf{V}^{-1}\mathbb{Z}_0)^{-1}\right]$. By Theorem 3.3.2, we have

$$\sqrt{nT}(\bar{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) \xrightarrow{p} \mathbf{0}. \tag{B.1.2}$$

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 $\mathbf{X}$, Lyapunov central limit theorem yields

$$(\mathbb{Z}_0'\mathbf{V}^{-1}\mathbb{Z}_0)^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p),$$

and Slusky's theorem leads to

$$(\mathbb{Z}_0'\mathbf{V}^{-1}\mathbb{Z}_0)^{1/2}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p).$$

B.2 Technical results

B.2.1 Preliminaries

Lemma B.2.1. *Under Condition 3.3.5,* $v_i(T) = T^{-1/2}\sum_{t=1}^{T} u_{it}$ *is sub-exponential for each* $i = 1, \ldots, n$.

Proof. Note that $\mathbb{E}(|u_{it}|^{4+\delta_1}) < \infty$ for any $t = 1, \ldots, T$, $i = 1, \ldots, n$ and $\delta_1 > 0$ and

$$\sum_{T=0}^{\infty} \alpha(T)^{1/3} < \sum_{t=0}^{\infty} \exp\{-CT^{r_1}/3\} < \infty.$$

By Theorem 4 in Tikhomirov (1981), $|\mathbb{P}\{v_i(T) < s\} - \mathbb{P}(W_i < s)| \leqslant C_1 T^{-1/2}(1+|s|)^{-4}\{\log(T)\}^3$ for each $i = 1, \ldots, n$ and any s, where $W \sim N(0, \sigma_i^2)$ and

$$\sigma_i^2 = \mathbb{E}(u_{i1}^2) + 2\sum_{t=2}^{\infty} \mathbb{E}(u_{i1}u_{it}).$$

Thus, we have

$$\mathbb{P}\{|v_i(T)| > s\} = \mathbb{P}\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_{it}\right| > s\right) \leqslant 2\exp\left\{-s^2/\left(2\sigma_i^2\right)\right\} + C_1 T^{-1/2}(1+s)^{-4}\{\log(T)\}^3$$

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

$$\begin{aligned}
& \mathbb{E}\{|v_i(T)|^k\} \\
= & \int_0^1 \mathbb{P}\{|v_i(T)| > s^{1/k}\}ds + \int_1^{\infty} \mathbb{P}\{|v_i(T)| > s^{1/k}\}ds \\
\leqslant & 1 + \int_1^{\infty} 2\exp\left\{-s^{2/k}/(2\sigma^2)\right\} ds + \int_1^{\infty} C_1 T^{-1/2}(1+s^{1/k})^{-4}\{\log(T)\}^3 ds \\
\leqslant & 1 + \int_0^{\infty} 2e^{-t}k(2\sigma_i^2)^{k/2}t^{k/2-1}dt + \int_1^{\infty} C_1 kT^{-1/2}\{\log(T)\}^3(1+t)^{-4}t^{k-1}dt \\
= & 1 + 2(2\sigma_i^2)^{k/2}k\Gamma(k/2) + C_1\pi T^{-1/2}\{\log(T)\}^3 k!,
\end{aligned}$$

so that

$$\begin{aligned}\mathbb{E}\left[\exp\{sv_i(T)\}\right] &\leqslant 1+\sum_{k=2}^{\infty}\frac{|s|^k\mathbb{E}\{|v_i(T)|^k\}}{k!}\\ &\leqslant 1+2(2+\sqrt{2\sigma_i^2s^2})\sum_{k=1}^{\infty}\frac{(\sigma_i^2s^2)^k k!}{(2k)!}+C_1\pi T^{-1/2}\{\log(T)\}^3\sum_{k=2}^{\infty}|s|^k\\ &\lesssim \exp\{2\sigma_i^2s^2+C_1\pi T^{-1/2}\{\log(T)\}^3\}\end{aligned}$$

for $|s|<\min\{1/\sigma_i,1\}$. The assertion follows from the definition of sub-exponential distributions.

□

Lemma B.2.2. *Under Conditions 3.2.1 and 3.3.5, for any $s>0$,*

$$\mathbb{P}\left\{\left\|\mathbf{A}\sum_{t=1}^{T}\boldsymbol{u}_t/T\right\|_2>s\|\mathbf{A}\|_{\mathbb{F}}/\sqrt{T}\right\}<2p\exp\left\{-\frac{s^2}{2\sigma^2}\right\}$$

for $p\times n$ matrix $\mathbf{A}$ and $\sigma^2=\max_i\sigma_i^2$ with σ_i^2 defined in Lemma B.2.1.

Proof. Write $\mathbf{A}=(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_p)'$, where $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_p$ are n-dimensional vectors. For each $m=1,\ldots,p$ and $w\geqslant 0$, by Conditions 3.2.1, 3.2.2, and 3.3.5, Lemma B.2.1, and Corollary 4 in Samson et al. (2000),

$$\begin{aligned}\mathbb{P}\left(\left|\boldsymbol{a}_m'\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\right|\geqslant s\right)&=\mathbb{P}\left(\left|\sum_{i=1}^{n}a_{mi}v_i(T)\right|\geqslant s\sqrt{T}\right)\\ &\leqslant 2\exp\left\{-\frac{s^2T}{2\sigma^2\|\boldsymbol{a}_m\|_2^2}\right\}.\end{aligned}$$

Hence

$$\mathbb{P}\left\{\left|\boldsymbol{a}_m'\sum_{t=1}^{T}\boldsymbol{u}_t/T\right|>\|\boldsymbol{a}_m\|_2 s/\sqrt{T}\right\}\leqslant 2\exp\left\{-\frac{s^2}{2\sigma^2}\right\}$$

for any $m=1,\ldots,p$ and

$$\mathbb{P}\left\{\left\|\mathbf{A}\sum_{t=1}^{T}\boldsymbol{u}_t/T\right\|_2>\|\mathbf{A}\|_{\mathbb{F}}s/\sqrt{T}\right\}\leqslant 2p\exp\left\{-\frac{s^2}{2\sigma^2}\right\}.$$

□

Conclusion in Lemma B.2.2 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 $\max_{j\leqslant n}\sum_{i=1}^{n}|\mathbb{E}(u_{it}u_{jt})|<C_2$, $\max_{i\leqslant n}\sum_{k=1}^{n}\sum_{m=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T}|\operatorname{cov}(u_{it}u_{kt},u_{is}u_{ms})|<C_2$, and $(nT)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T}|\mathbb{E}(u_{it}u_{js})|<C_2$ for some $C_2>0$, Corollary 4 in Samson et al. (2000) still applies in the above proof.

Lemma B.2.3. *For $p\times K$ matrix* $\mathbf{A}$*, under Condition 3.2.1 and 3.3.5,*

$$\mathbb{P}\left\{\left\|\mathbf{A}\sum_{t=1}^{T}\boldsymbol{f}_t/T\right\|_2>s\|\mathbf{A}\|_{\mathbb{F}}/\sqrt{T}\right\}\leqslant 2pC_3\exp(-C_4s^2/2)$$

for constants $C_3,C_4>0$.

Proof. The proof is similar to that of Lemma B.2.2. □

B.2.2 Some results for spline estimators

Lemma B.2.4. *Under Condition 3.3.2, there exist constants c_1, c_2 such that*

$$c_1\|\boldsymbol{a}\|_2^2\leqslant\mathbb{E}(\langle\boldsymbol{a},\boldsymbol{a}\rangle_n)\leqslant c_2\|\boldsymbol{a}\|_2^2.$$

Proof. It follows from Condition 3.3.2 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 $0<m_1\leqslant m_2<\infty$.

Denote $f_\ell(X_\ell)=\sum_j a_{j\ell}\phi_j(X_\ell)$, $\ell=1,\ldots,d$ and $f_0\equiv a_0$. Then, we have

$$\mathbb{E}(\langle\boldsymbol{a},\boldsymbol{a}\rangle_n)=\mathbb{E}\left[\left\{a_0+\sum_j\sum_l a_{j\ell}\phi_j(X_\ell)\right\}^2\right]=\mathbb{E}\left[\left\{a_0+\sum_{\ell=1}^{d}f_\ell(X_\ell)\right\}^2\right].$$

Since the basis functions are centralized,

$$\begin{aligned}\mathbb{E}(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n) &= \int_{\mathcal{X}} \left\{ a_0 + \sum_{\ell=1}^{d} f_\ell(X_\ell) \right\}^2 h(\mathbf{X}) d\mathbf{X} \asymp \int_{\mathcal{X}} \left\{ a_0 + \sum_{\ell=1}^{d} f_\ell(X_\ell) \right\}^2 d\mathbf{X} \\ &= a_0^2 + \int_{\mathcal{X}} \left\{ \sum_{\ell=1}^{d} f_\ell(X_\ell) \right\}^2 d\mathbf{X}\end{aligned}$$

By Lemma 1 of Stone (1985), we obtain

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

where $C_0 = 1 - (1 - m_1/m_2)^{1/2}$. Consequently, we have

$$\mathbb{E}(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n) \geqslant a_0^2 + \left(\frac{C_0}{2} \right)^{d-1} \left(\sum_{\ell} \sum_{j} a_{j\ell}^2 \right) \geqslant \min \left\{ 1, \left(\frac{C_0}{2} \right)^{d} \right\} \|\boldsymbol{a}\|^2.$$

Similarly, we can establish that

$$\int_0^1 \left\{ \sum_{\ell=1}^{d} f_\ell(X_\ell) \right\}^2 d\mathbf{X} \leqslant d^2 \sum_{\ell=1}^{d} \int_0^1 f_\ell^2(x) dx = d^2 \left(\sum_{\ell} \sum_{j} a_{j\ell}^2 \right)$$

and thus

$$\mathbb{E}(\langle \boldsymbol{a}, \boldsymbol{a} \rangle_n) \leqslant a_0^2 + d^2 \left(\sum_{\ell} \sum_{j} a_{j\ell}^2 \right) \leqslant (1 + d^2) \|\boldsymbol{a}\|^2.$$

□

Lemma B.2.5. *Under Condition 3.3.2, we have*

$$\mathbb{P} \left\{ \sup_{\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^{Jd+1}} \frac{|\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle_n - \mathbb{E}(\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle_n|)}{\sqrt{\mathbb{E}(\langle \boldsymbol{a}_1, \boldsymbol{a}_1 \rangle_n) \mathbb{E}(\langle \boldsymbol{a}_2, \boldsymbol{a}_2 \rangle_n)}} > s \right\} \leqslant C_1 J^2 \exp \left\{ -C_2 \frac{n}{J} \frac{s^2}{1+s} \right\}$$

for some constant $C_1, C_2 > 0$.

Proof. The proof is similar to that of Lemma A.2 in Huang et al. (2004). First, notice that

$$\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle_n - \mathbb{E}(\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle_n) = \frac{1}{n} \sum_{\ell,\ell'} \sum_{j,j'} a_{j\ell}^{(1)} a_{j'\ell'}^{(2)} \left\{ \boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'} - \mathbb{E}(\boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'}) \right\},$$

where $\boldsymbol{e}_{j\ell}$ is the $(J\ell + j + 1)$th natural basis of $\mathbb{R}^{Jd+1}$. Hence, we have $\boldsymbol{\Phi} \boldsymbol{e}_{j\ell} = \{\phi_j(X_{1\ell}), \ldots, \phi_j(X_{n\ell})\}'$. For any j, j', ℓ, ℓ',

$$\mathrm{Var}\left(\frac{1}{n} \boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'} \right) \leqslant \frac{1}{n^2} \sum_i \mathbb{E}\left\{ \phi_j^2(X_{i\ell}) \phi_{j'}^2(X_{i\ell'}) \right\} \lesssim \frac{1}{n}.$$

As $|\phi_j(X_\ell)| \leqslant M$ for each j, ℓ for some $M > 0$, Bernstein's inequality yields that for $s > 0$ and constants M_1, $M_2 > 0$

$$\mathbb{P}\left\{ \left| \frac{1}{n} \boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'} - \mathbb{E}\left(\frac{1}{n} \boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'} \right) \right| > c_2 s \right\} \leqslant \exp\left\{ -\frac{n s^2}{M_1 + M_2 s} \right\}.$$

By the union bound, for $s > 0$ and constants C_1, $C_2 > 0$

$$\mathbb{P}\left[\bigcup_{j,j',\ell,\ell'} \left\{ \left| \frac{1}{n} \boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'} - \mathbb{E}\left(\frac{1}{n} \boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'} \right) \right| > \frac{c_2 s}{Jd} \right\} \right] \leqslant C_1 J^2 \exp\left\{ -\frac{C_2 n s^2}{J^2 + sJ} \right\}.$$

Denote

$$\mathcal{B} = \bigcup_{j,j',\ell,\ell'} \left\{ \left| n^{-1} \boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'} - \mathbb{E}\left(n^{-1} \boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'} \right) \right| > c_2 s (Jd)^{-1} \right\},$$

so that $\mathbb{P}(\mathcal{B}) < C_1 J^2 \exp\{-C_2 n s^2 (J^2 + sJ)^{-1}\}$. For each $s > 0$, on $\mathcal{B}^c$,

$$\begin{aligned}
|\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle_n - \mathbb{E}(\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle_n)| &= \frac{1}{n} \sum_{\ell,\ell'} \sum_{j,j'} a_{j\ell}^{(1)} a_{j'\ell'}^{(2)} \left\{ \boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'} - \mathbb{E}(\boldsymbol{e}_{j\ell}' \boldsymbol{\Phi}' \boldsymbol{\Phi} \boldsymbol{e}_{j'\ell'}) \right\} \\
&\leqslant \sum_{j,\ell} \sum_{j',\ell'} |a_{j\ell}^{(1)}| |a_{j'\ell'}^{(2)}| \frac{c_2 s}{Jd} \\
&\leqslant \frac{c_2 s}{Jd} \left(Jd \sum_{j,\ell} |a_{j\ell}^{(1)}|^2 \right)^{1/2} \left(Jd \sum_{j',\ell'} |a_{j'\ell'}^{(2)}|^2 \right)^{1/2} \\
&= c_2 s \|\boldsymbol{a}_1\|_2 \|\boldsymbol{a}_2\|_2
\end{aligned}$$

$$\lesssim \quad s\sqrt{\mathbb{E}(\langle \boldsymbol{a}_1, \boldsymbol{a}_1 \rangle_n)\mathbb{E}(\langle \boldsymbol{a}_2, \boldsymbol{a}_2 \rangle_n)},$$

where the last inequality is due to Lemma B.2.4. The conclusion follows.

□

B.2.3 Technical results for the proof of Theorem 3.3.1

Lemma B.2.6. *Under Conditions 3.3.1 and 3.3.3, for each n, with probability at least $1-\delta$,*

$$1 - n^{-1}\log(1/\delta) \lesssim \lambda_{\min}\left(\frac{1}{n}\mathbf{G}'\mathbf{P}\mathbf{G}\right) < \lambda_{\max}\left(\frac{1}{n}\mathbf{G}'\mathbf{P}\mathbf{G}\right) \lesssim 1 + n^{-1}\log(1/\delta).$$

Proof. Denote $\mathbf{R} = \mathbf{G} - \mathbf{P}\mathbf{G}$, and we have $\mathbf{G}'\mathbf{P}\mathbf{G} = \mathbf{G}'\mathbf{G} - \mathbf{G}'\mathbf{R}$. Thus we have

$$\begin{aligned}
\lambda_{\min}\left(\frac{1}{n}\mathbf{G}'\mathbf{P}\mathbf{G}\right) &\geqslant \lambda_{\min}\left(\frac{1}{n}\mathbf{G}'\mathbf{G}\right) + \lambda_{\min}\left(-\frac{1}{n}\mathbf{G}'\mathbf{R}\right), \\
\lambda_{\max}\left(\frac{1}{n}\mathbf{G}'\mathbf{P}\mathbf{G}\right) &\leqslant \lambda_{\max}\left(\frac{1}{n}\mathbf{G}'\mathbf{G}\right) + \lambda_{\max}\left(-\frac{1}{n}\mathbf{G}'\mathbf{R}\right).
\end{aligned}$$

Note that $\|\mathbf{R}\|_{\mathbb{F}}^2 \lesssim nJ^{-\kappa}$ by Condition 3.3.3. Thus, combining Condition 3.3.1, it holds that, with probability at least $1-\delta$,

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

Thus, with probability at least $1-\delta$, $|\lambda(\mathbf{G}'\mathbf{R}/n)| \lesssim J^{-\kappa}\{1 + n^{-1}\log(1/\delta)\}$. By Condition 3.3.1, with probability at least $1-\delta$,

$$1 - n^{-1}\log(1/\delta) \lesssim \lambda_{\min}\left(\frac{1}{n}\mathbf{G}'\mathbf{G}\right) < \lambda_{\max}\left(\frac{1}{n}\mathbf{G}'\mathbf{G}\right) \lesssim 1 + n^{-1}\log(1/\delta).$$

The conclusion follows. □

Lemma B.2.7. *Under Conditions 3.2.1 and 3.3.2-3.3.5, for $\widetilde{\mathbf{U}} = \mathbf{U} + \mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\}$ defined in Section 3.2.2 in the main paper and $\widehat{\boldsymbol{\beta}}$ the OLS estimator in Proposition B.2.5,*

(i) $\mathbb{E}(\|\mathbf{F}'\widetilde{\mathbf{U}}'\|_{\mathbb{F}}^2) = O((n+p^2n^{2\alpha})T)$, $\mathbb{E}(\|\widetilde{\mathbf{U}}'\mathbf{\Phi}\|_{\mathbb{F}}^2) = O(nJ(T+p^2n^{2\alpha}))$ *and* $\mathbb{E}(\|\mathbf{\Phi}'\widetilde{\mathbf{U}}\mathbf{F}\|_{\mathbb{F}}^2) = O(p^2n^{1+2\alpha}TJ)$.

(ii) With probability at least $1-3\delta$, $\|\mathbf{F}'\widetilde{\mathbf{U}}'\|_{\mathbb{F}} \lesssim \{(n+p^2)T\}^{1/2}\{1+\sqrt{\log(1/\delta)}\}$, $\|\widetilde{\mathbf{U}}'\mathbf{\Phi}\|_{\mathbb{F}} \lesssim \{nJ(T+p^2)\}^{1/2}\{1+\sqrt{\log(1/\delta)}\}$ *and* $\|\mathbf{\Phi}'\widetilde{\mathbf{U}}\mathbf{F}\|_{\mathbb{F}} \lesssim (p^2nTJ)^{1/2}\{1+\sqrt{\log(1/\delta)}\}$.

(iii) With probability at least $1-4\delta$,

$$\|\mathbf{P}\widetilde{\mathbf{U}}\|_{\mathbb{F}} \lesssim \sqrt{J(T+p^2n^{2\alpha})}\{1+n^{-1}J\log(J^2/\delta)\}^{3/2}\{1+\sqrt{\log(1/\delta)}\}.$$

Proof. (i) By Lemma B.1 of Fan et al. (2016), $\mathbb{E}(\|\mathbf{F}'\mathbf{U}'\|_{\mathbb{F}}^2) = O(nT)$, $\mathbb{E}(\|\mathbf{U}'\mathbf{\Phi}\|_{\mathbb{F}}^2) = O(nJT)$, $\mathbb{E}(\|\mathbf{\Phi}'\mathbf{U}\mathbf{F}\|_{\mathbb{F}}^2) = O(nTJ)$, and $\mathbb{E}(\|\mathbf{P}\mathbf{U}\|_{\mathbb{F}}^2) = O(JT)$. Thus, it suffices to show

$$\mathbb{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\mathbf{F}\|_{\mathbb{F}}^2] = O(p^2T), \tag{B.2.1}$$

$$\mathbb{E}[\|\mathbf{\Phi}'\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\|_{\mathbb{F}}^2] = O(p^2nJ), \tag{B.2.2}$$

$$\mathbb{E}[\|\mathbf{\Phi}'\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\mathbf{F}\|_{\mathbb{F}}^2] = O(p^2nTJ). \tag{B.2.3}$$

By Proposition B.2.5, $\mathbb{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\|_{\mathbb{F}}^2] \leqslant \mathbb{E}(\|\mathbb{Z}\|_{\mathbb{F}}^2)\|\mathbf{I}_T\|_2^2\mathbb{E}(\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_{\mathbb{F}}^2) = O(p^2n^{2\alpha})$. Then (B.2.1) follows from Cauchy-Schwartz inequality that

$$\mathbb{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\mathbf{F}\|_{\mathbb{F}}^2] \leqslant \mathbb{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\|_{\mathbb{F}}^2]\mathbb{E}(\|\mathbf{F}\|_{\mathbb{F}}^2) = O\left(p^2n^{2\alpha}T\right).$$

As a consequence of Lemma B.2.4, we have $\mathbb{E}(\|\mathbf{\Phi}\|_2^2) = O(n)$, and consequently $\mathbb{E}(\|\mathbf{\Phi}\|_{\mathbb{F}}^2) \leqslant (Jd+1)\mathbb{E}(\|\mathbf{\Phi}\|_2^2) = O(nJ)$, and (B.2.2) holds since

$$\mathbb{E}[\|\mathbf{\Phi}'\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\|_{\mathbb{F}}^2] \leqslant \mathbb{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\|_{\mathbb{F}}^2]\mathbb{E}(\|\mathbf{\Phi}\|_{\mathbb{F}}^2) = O\left(p^2n^{1+2\alpha}J\right).$$

Applying Cauchy-Schwartz inequality, (B.2.3) follows

$$\mathbb{E}[\|\mathbf{\Phi}'\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\mathbf{F}\|_{\mathbb{F}}^2] \leqslant \mathbb{E}[\|\mathbb{Z}\{\mathbf{I}_T \otimes (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\|_{\mathbb{F}}^2]\mathbb{E}(\|\mathbf{\Phi}\|_{\mathbb{F}}^2)\mathbb{E}(\|\mathbf{F}\|_{\mathbb{F}}^2) = O\left(p^2n^{1+2\alpha}TJ\right).$$

(ii) Since $\mathbb{E}(\|\mathbf{F}'\widetilde{\mathbf{U}}'\|_{\mathbb{F}}^2) \leqslant C_0(n+p^2n^{2\alpha})T$ for some $C_0 > 0$, for $s > 0$ we have

$$\begin{aligned}
& \mathbb{P}(\|\mathbf{F}'\widetilde{\mathbf{U}}'\|_{\mathbb{F}}/\sqrt{C_0(n+p^2n^{2\alpha})T} > M) \\
\leqslant\ & \exp(-sM)\mathbb{E}[\exp\{s\|\mathbf{F}'\widetilde{\mathbf{U}}'\|_{\mathbb{F}}/\sqrt{C_0(n+p^2n^{2\alpha})T}\}] \\
\leqslant\ & \exp(-sM)\mathbb{E}\left[1 + s\|\mathbf{F}'\widetilde{\mathbf{U}}'\|_{\mathbb{F}}/\sqrt{C_0(n+p^2n^{2\alpha})T}\right. \\
& \left. + s^2\|\mathbf{F}'\widetilde{\mathbf{U}}'\|_{\mathbb{F}}^2/\{2C_0(n+p^2n^{2\alpha})T\} + o(s^2\|\mathbf{F}'\widetilde{\mathbf{U}}'\|_{\mathbb{F}}^2/\{2C_0(n+p^2n^{2\alpha})T\})\right] \\
\leqslant\ & \exp(-sM + s + s^2/2 + o(s^2)).
\end{aligned}$$

The minimum of the right hand side is $\exp\{-(M-1)^2/2\}$. Letting $\delta = \exp\{-(M-1)^2/2\}$, we have with probability at least $1-\delta$,

$$\|\mathbf{F}'\widetilde{\mathbf{U}}'\|_{\mathbb{F}} \lesssim \sqrt{(n+p^2n^{2\alpha})T}\{1+\sqrt{\log(1/\delta)}\}.$$

The remaining two bounds follows similarly.

(iii) By Lemma B.1.1, with probability at least $1-\delta$, we have

$$\|\mathbf{\Phi}\|_2^2 = \lambda_{\max}(\mathbf{\Phi}'\mathbf{\Phi}) \lesssim n\left\{1 + \frac{J}{n}\log(J^2/\delta)\right\}$$

and

$$\begin{aligned}
\|(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\|_2 = \lambda_{\min}^{-1}(\mathbf{\Phi}'\mathbf{\Phi}) &\lesssim \left[n\left\{1 - \frac{J}{n}\log(J^2/\delta)\right\}\right]^{-1} \\
&\lesssim n^{-1}\left[1 + \frac{J}{n}\log(J^2/\delta) + o\left\{\frac{J}{n}\log(J^2/\delta)\right\}\right] \\
&\lesssim n^{-1}\left\{1 + \frac{J}{n}\log(J^2/\delta)\right\}.
\end{aligned}$$

Hence, with probability at least $1-4\delta$,

$$\|\mathbf{P}\widetilde{\mathbf{U}}\|_{\mathbb{F}} \leqslant \|\mathbf{\Phi}\|_2\|(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\|_2\|\mathbf{\Phi}'\widetilde{\mathbf{U}}\|_{\mathbb{F}} \lesssim \sqrt{J(T+p^2n^{2\alpha})}\left\{1 + \frac{J}{n}\log(J^2/\delta)\right\}^{3/2}\{1+\sqrt{\log(1/\delta)}\}.$$

□

Lemma B.2.8. *With probability at least* $1-2\delta$*,*

(i) $\|\widetilde{\mathbf{U}}'\mathbf{\Phi}\mathbf{B}\|_{\mathbb{F}} \lesssim \sqrt{n(T+p^2n^{2\alpha})}\{1+\sqrt{\log(1/\delta)}\}$,

(ii) $\|\mathbf{B}'\mathbf{\Phi}'\widetilde{\mathbf{U}}\mathbf{F}\|_{\mathbb{F}} \lesssim p\sqrt{n^{1+2\alpha}T}\{1+\sqrt{\log(1/\delta)}\}^2\{1+\sqrt{\log(1/\delta)}\}$.

Proof. By Proposition B.2.5,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}'\mathbf{\Phi}\mathbf{B}\|_{\mathbb{F}}^2\right] &\leqslant \mathbb{E}\left[\|\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\mathbf{G}\|_{\mathbb{F}}^2\right] \\
&\quad + \mathbb{E}\left[\|\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\mathbf{R}\|_{\mathbb{F}}^2\right] = O(p^2n^{1+2\alpha}), \\
\mathbb{E}\left[\|\mathbf{B}'\mathbf{\Phi}'\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\mathbf{F}\|_{\mathbb{F}}^2\right] &\leqslant \mathbb{E}\left[\|\mathbf{G}'\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\mathbf{F}\|_{\mathbb{F}}^2\right] \\
&\quad + \mathbb{E}\left[\|\mathbf{R}'\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\mathbf{F}\|_{\mathbb{F}}^2\right] = O(p^2n^{1+2\alpha}T).
\end{aligned}$$

By Lemma C.6 in Fan et al. (2016), $\mathbb{E}\left(\|\mathbf{U}\mathbf{\Phi}\mathbf{B}\|_{\mathbb{F}}^2\right) = O(nT)$ and $\mathbb{E}\left(\|\mathbf{B}\mathbf{\Phi}'\mathbf{U}\mathbf{F}\|_{\mathbb{F}}^2\right) = O(nT)$. So similar to the proof of Lemma B.2.7, with probability at least $1-4\delta$,

$$\begin{aligned}
\|\widetilde{\mathbf{U}}\mathbf{\Phi}\mathbf{B}\|_{\mathbb{F}} &\leqslant \|\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\mathbf{B}\}\mathbf{\Phi}\|_{\mathbb{F}} + \|\mathbf{U}\mathbf{\Phi}\mathbf{B}\|_{\mathbb{F}} \lesssim \sqrt{n(T+p^2n^{2\alpha})}\{1+\sqrt{\log(1/\delta)}\}, \\
\|\mathbf{B}'\mathbf{\Phi}'\widetilde{\mathbf{U}}\mathbf{F}\|_{\mathbb{F}} &\leqslant \|\mathbf{B}'\mathbf{\Phi}'\mathbb{Z}\{\mathbf{I}_T\otimes(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\}\mathbf{F}\|_{\mathbb{F}} + \|\mathbf{B}'\mathbf{\Phi}'\mathbf{U}\mathbf{F}\|_{\mathbb{F}} \lesssim p\sqrt{n^{1+2\alpha}T}\{1+\sqrt{\log(1/\delta)}\}.
\end{aligned}$$

□

Denote $\mathbf{K}$ a $K\times K$ diagonal matrix whose diagonals are the first K eigenvalues of $(nT)^{-1}\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}$. Then $(nT)^{-1}\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}\widehat{\mathbf{F}} = \widehat{\mathbf{F}}\mathbf{K}$. Let

$$\mathbf{H} = \frac{1}{nT}\mathbf{B}'\mathbf{\Phi}'\mathbf{\Phi}\mathbf{B}\mathbf{F}'\widehat{\mathbf{F}}\mathbf{K}^{-1}$$

Substituting (3.2.4), we have

$$\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H} = \left(\sum_{i=1}^{8}\mathbf{A}_i\right)\mathbf{K}^{-1}$$

where

$$\mathbf{A}_1 = \frac{1}{nT}\mathbf{F}\mathbf{B}'\mathbf{\Phi}'\widetilde{\mathbf{U}}\widehat{\mathbf{F}}, \quad \mathbf{A}_2 = \frac{1}{nT}\widetilde{\mathbf{U}}'\mathbf{\Phi}\mathbf{B}\mathbf{F}'\widehat{\mathbf{F}}, \quad \mathbf{A}_3 = \frac{1}{nT}\widetilde{\mathbf{U}}'\mathbf{P}\widetilde{\mathbf{U}}\widehat{\mathbf{F}},$$
$$\mathbf{A}_4 = \frac{1}{nT}\mathbf{F}\mathbf{B}'\mathbf{\Phi}'\mathbf{R}\mathbf{F}'\widehat{\mathbf{F}}, \quad \mathbf{A}_5 = \frac{1}{nT}\mathbf{F}\mathbf{R}'\mathbf{\Phi}\mathbf{B}\mathbf{F}'\widehat{\mathbf{F}},$$
$$\mathbf{A}_6 = \frac{1}{nT}\mathbf{F}\mathbf{R}'\mathbf{P}\mathbf{R}\mathbf{F}'\widehat{\mathbf{F}}, \quad \mathbf{A}_7 = \frac{1}{nT}\mathbf{F}\mathbf{R}'\mathbf{P}\widetilde{\mathbf{U}}\widehat{\mathbf{F}}, \quad \mathbf{A}_8 = \frac{1}{nT}\widetilde{\mathbf{U}}'\mathbf{P}\mathbf{R}\mathbf{F}'\widehat{\mathbf{F}}.$$

Next, in Lemmas B.2.9-B.2.13, we will provide a bound on $\|\mathbf{H} - \mathbf{I}\|_{\mathbb{F}}$ in probability.

Lemma B.2.9. *With probability at least* $1 - 5\delta$, $\|\mathbf{K}^{-1}\|_2 \lesssim 1 + n^{-1}\log(1/\delta)$.

Proof. The K largest eigenvalues of $(nT)^{-1}\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}$ are the same as those of

$$\mathbf{W} = (nT)^{-1}(\mathbf{\Phi}'\mathbf{\Phi})^{-1/2}\mathbf{\Phi}'\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}'\mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1/2}.$$

Substituting $\widetilde{\mathbf{Y}} = \mathbf{G}\mathbf{F}' + \widetilde{\mathbf{U}}$ and $T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K$, we have $\mathbf{W} = \sum_{i=1}^4 \mathbf{W}_i$ where

$$\begin{aligned}
\mathbf{W}_1 &= \frac{1}{n}(\mathbf{\Phi}'\mathbf{\Phi})^{-1/2}\mathbf{\Phi}'\mathbf{G}\mathbf{G}'\mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1/2}, \\
\mathbf{W}_2 &= \frac{1}{nT}(\mathbf{\Phi}'\mathbf{\Phi})^{-1/2}\mathbf{\Phi}'\mathbf{G}\mathbf{F}'\widetilde{\mathbf{U}}'\mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1/2}, \\
\mathbf{W}_3 &= \mathbf{W}_2', \\
\mathbf{W}_4 &= \frac{1}{nT}(\mathbf{\Phi}'\mathbf{\Phi})^{-1/2}\mathbf{\Phi}'\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}'\mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1/2}.
\end{aligned}$$

By Lemma B.1.1, with probability at least $1 - \delta$, we have

$$\|\mathbf{\Phi}\|_2^2 = \lambda_{\max}(\mathbf{\Phi}'\mathbf{\Phi}) \lesssim n\left\{1 + \frac{J}{n}\log(J^2/\delta)\right\}$$

and

$$\begin{aligned}
\|(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\|_2 = \lambda_{\min}^{-1}(\mathbf{\Phi}'\mathbf{\Phi}) &\lesssim \left[n\left\{1 - \frac{J}{n}\log(J^2/\delta)\right\}\right]^{-1} \\
&\lesssim n^{-1}\left[1 + \frac{J}{n}\log(J^2/\delta) + o\left\{\frac{J}{n}\log(J^2/\delta)\right\}\right]
\end{aligned}$$

$$\lesssim n^{-1}\left\{1+\frac{J}{n}\log(J^2/\delta)\right\}.$$

By Lemma B.2.6, with probability at least $1-\delta$, $\|\mathbf{PG}\|_2^2 = \lambda_{\max}(\mathbf{G}'\mathbf{PG}) \lesssim n(1+J^{-\kappa})\{1+n^{-1}\log(1/\delta)\}$. Hence, with probability at least $1-5\delta$,

$$\begin{aligned}\|\mathbf{W}_2\|_2 &\leqslant \frac{1}{n}\|(\mathbf{\Phi}'\mathbf{\Phi})^{-1/2}\|_2^2\|\mathbf{\Phi}\|_2\|\mathbf{PG}\|_2\left\|\frac{1}{T}\mathbf{F}'\widetilde{\mathbf{U}}'\mathbf{\Phi}\right\|_{\mathbb{F}}\\ &\lesssim p\sqrt{\frac{J}{n^{1-2\alpha}T}}(1+J^{-\kappa})\{1+n^{-1}J\log(J^2/\delta)\}^{3/2}\{1+\sqrt{\log(1/\delta)}\}\{1+n^{-1}\log(1/\delta)\}\end{aligned}$$

and by Lemma B.2.7, with probability at least $1-4\delta$,

$$\begin{aligned}\|\mathbf{W}_4\|_2 &\leqslant \frac{1}{nT}\|(\mathbf{\Phi}'\mathbf{\Phi})^{-1/2}\|_2^2\|\mathbf{\Phi}'\widetilde{\mathbf{U}}\|_{\mathbb{F}}^2\\ &\lesssim \frac{J(T+p^2n^{2\alpha})}{nT}\{1+J\log(J^2/\delta)/n\}\{1+\sqrt{\log(1/\delta)}\}.\end{aligned}$$

By Weyl's Theorem, $|\lambda_k(\mathbf{W})-\lambda_k(\mathbf{W}_1)| \leqslant \|\mathbf{W}-\mathbf{W}_1\|_2$ for each $k=1,\ldots,K$, which implies, with probability at least $1-5\delta$,

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

Note that the K largest eigenvalues of $\mathbf{W}_1$ is also the K largest eigenvalues of $n^{-1}\mathbf{G}'\mathbf{PG}$. Thus, by Lemma B.2.6, with probability at least $1-5\delta$, $\|\mathbf{K}^{-1}\|_2 \lesssim 1+n^{-1}\log(1/\delta)$. □

Lemma B.2.10. *With probability at least* $1-7\delta$*,*

(i) $\|\mathbf{A}_1\|_{\mathbb{F}}, \|\mathbf{A}_2\|_{\mathbb{F}} \lesssim \sqrt{n^{-1}(T+p^2n^{2\alpha})}\{1+\sqrt{\log(1/\delta)}\}$,

(ii) $\|\mathbf{A}_3\|_{\mathbb{F}} \lesssim n^{-1}T^{-1/2}J(T+p^2n^{2\alpha})\{1+\sqrt{\log(1/\delta)}\}$,

(iii) $\|\mathbf{A}_4\|_{\mathbb{F}}, \|\mathbf{A}_5\|_{\mathbb{F}} \lesssim (J^{-\kappa/2}\sqrt{T})\{1+\sqrt{\log(1/\delta)}\}$,

(iv) $\|\mathbf{A}_7\|_{\mathbb{F}}, \|\mathbf{A}_8\|_{\mathbb{F}} \lesssim \sqrt{(T+p^2n^{2\alpha})(nJ^{\kappa-1})^{-1}}\{1+\sqrt{\log(1/\delta)}\}$*;*

and $\|\mathbf{A}_6\|_{\mathbb{F}} \lesssim J^{-\kappa}\sqrt{T}$.

Proof. Notice that $\|\mathbf{F}\|_{\mathbb{F}} = \sqrt{KT}$ with probability 1 and $\|\widehat{\mathbf{F}}\|_{\mathbb{F}} = \sqrt{KT}$. Then, (i) follows from Lemma B.2.8 and (ii) follows from Lemma B.2.7. By Condition 3.3.3 that $\|\mathbf{R}\|_{\mathbb{F}}^2 \lesssim nJ^{-\kappa}$, (iii) follows from Lemma B.2.6 and $\mathbf{\Phi B} = \mathbf{PG}$. Part (iv) follows from Lemma B.2.7 and $\|\mathbf{R}\|_{\mathbb{F}}^2 \lesssim nJ^{-\kappa}$. Result on $\mathbf{A}_6$ follows similarly to (iii) given $\|\mathbf{P}\|_2 = 1$.

□

Lemma B.2.11. *With probability at least* $1 - 3\delta$*,*

(i) $\|\mathbf{A}_1\|_{\max}, \|\mathbf{A}_2\|_{\max} \lesssim n^{-1/2}T^{-1}\sqrt{T + p^2n^{2\alpha}}\{\log(T)\}^{2/r_2}\{1 + \log(1/\delta)\}$,

(ii) $\|\mathbf{A}_3\|_{\max} \lesssim n^{-1/2}T^{-1}\sqrt{T + p^2n^{2\alpha}}\{\log(T)\}^{1/r_2}\{1 + \log(1/\delta)\}$,

(iii) $\|\mathbf{A}_4\|_{\max}, \|\mathbf{A}_5\|_{\max} \lesssim n^{-1}T^{-1}\{\log(T)\}^{3/r_2}J^{-\kappa}\{1 + \log(1/\delta)\}$,

(iv) $\|\mathbf{A}_7\|_{\max}, \|\mathbf{A}_8\|_{\max} \lesssim (nT)^{-1}J^{-\kappa}\sqrt{J(T + p^2n^{2\alpha})}\{\log(T)\}^{2/r_2}\{1 + \log(1/\delta)\}$*;*

and $\|\mathbf{A}_6\|_{\max} \lesssim (nT)^{-1}J^{-2\kappa}\{\log(T)\}^{3/r_2}$.

Proof. By Lemma B.1 in Fan et al. (2011), with probability at least $1 - 1\delta$, $\|\widetilde{\mathbf{U}}\mathbf{P}\widetilde{\mathbf{U}}\|_{\max} \lesssim \sqrt{n}(T + p^2n^{2\alpha})\{1 + \log(1/\delta)\}$. Also, the proof of Lemma D.2 in Wang and Fan (2017) implies that $\|\mathbf{U}'\mathbf{\Phi B}\|_\infty \lesssim \sqrt{n}T$. Hence, with probability at least $1 - \delta$, $\|\widetilde{\mathbf{U}}'\mathbf{\Phi B}\|_\infty \lesssim \sqrt{n}(T + p^2n^{2\alpha})\{1 + \log(1/\delta)\}$ by Lemma B.2.8. Then, the results follow from that $\|\mathbf{F}\|_{\max} \lesssim \{\log(T) + \log(1/\delta)\}^{1/r_2}$ with probability at least $1 - \delta$. □

Proposition B.2.1. *Given* $J = o(\sqrt{n})$ *and* $\kappa \geqslant 1$*,*

(i) With probability at least $1 - 12\delta$*,*

$$\frac{1}{T}\|\widehat{\mathbf{F}} - \mathbf{FH}\|_{\mathbb{F}}^2 \lesssim \left(\frac{1}{n} + \frac{p^2}{n^{1-2\alpha}T} + \frac{1}{J^\kappa}\right)\{1 + \sqrt{\log(1/\delta)}\}^2\{1 + n^{-1}\log(1/\delta)\}. \quad \text{(B.2.4)}$$

(ii) With probability at least $1 - 8\delta$*,*

$$\|\widehat{\mathbf{F}} - \mathbf{FH}\|_{\max} \lesssim \left(\frac{1}{\sqrt{n}} + \frac{p}{\sqrt{n^{1-2\alpha}T}}\right)\{\log(T)\}^{2/r_2}\{1 + \log(1/\delta)\}. \quad \text{(B.2.5)}$$

Proof. By Lemma B.2.9, $\|\mathbf{K}^{-1}\|_2 \lesssim 1 + n^{-1}\log(1/\delta)$ with probability at least $1 - 5\delta$. The result follows from Lemmas B.2.10 and B.2.11. □

Lemma B.2.12. *With probability at least* $1 - 20\delta$*,*

(i) $T^{-1}\|\mathbf{A}_1\|_{\mathbb{F}}^2 \lesssim \{n^{-2} + n^{-1+2\alpha}T^{-1}p^2 + (nTJ^{\kappa})^{-1}(T + p^2n^{2\alpha})\}\{1 + \sqrt{\log(1/\delta)}\}^2\{1 + n^{-1}\log(1/\delta)\}$,

(ii) $T^{-2}\|\mathbf{F}'\mathbf{A}_2\|_{\mathbb{F}}^2 \lesssim n^{-1+2\alpha}T^{-1}p^2\{1 + \sqrt{\log(1/\delta)}\}^2$,

(iii) $T^{-2}\|\mathbf{F}'(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\mathbb{F}}^2 \lesssim \{n^{-2} + n^{-1+2\alpha}T^{-1}p^2 + J^{-\kappa}\}\{1 + \sqrt{\log(1/\delta)}\}^2$,

(iv) $T^{-2}\|\widehat{\mathbf{F}}'(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\mathbb{F}}^2 \lesssim \{n^{-2} + n^{-1+2\alpha}T^{-1}p^2 + J^{-\kappa}\}\{1 + \sqrt{\log(1/\delta)}\}^2$.

Proof. (i) First, by lemmas B.2.6 and B.2.9, with probability at least $1 - 6\delta$,

$$\|\mathbf{H}\|_2 \leqslant \frac{1}{nT}\|\mathbf{P}\mathbf{G}\|_{\mathbb{F}}^2\|\mathbf{F}\|_{\mathbb{F}}\|\widehat{\mathbf{F}}\|_{\mathbb{F}}\|\mathbf{K}^{-1}\|_2 \lesssim 1 + n^{-1}\log(1/\delta).$$

Then, by Lemma B.2.8 and Proposition B.2.1, with probability at least $1 - 20\delta$,

$$\begin{aligned}
&\|\mathbf{B}'\mathbf{\Phi}'\widetilde{\mathbf{U}}\widehat{\mathbf{F}}\|_{\mathbb{F}}^2 \\
\leqslant\ & 2\|\mathbf{B}'\mathbf{\Phi}'\widetilde{\mathbf{U}}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\mathbb{F}}^2 + 2\|\mathbf{B}'\mathbf{\Phi}'\widetilde{\mathbf{U}}\mathbf{F}\mathbf{H}\|_{\mathbb{F}}^2 \\
\lesssim\ & \left\{n(T + p^2n^{2\alpha})\left(\frac{T}{n} + \frac{p^2}{n^{1-2\alpha}} + \frac{T}{J^{\kappa}}\right) + p^2n^{1+2\alpha}T\right\} \\
& \{1 + \sqrt{\log(1/\delta)}\}^2\{1 + n^{-1}\log(1/\delta)\} \\
\lesssim\ & \left\{T^2 + p^2n^{2\alpha}T + p^4n^{4\alpha} + nT(T + p^2n^{2\alpha})/J^{\kappa}\right\} \\
& \{1 + \sqrt{\log(1/\delta)}\}^2\{1 + n^{-1}\log(1/\delta)\}.
\end{aligned}$$

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

(ii) By Lemma B.2.8, with probability at least $1 - 4\delta$,

$$\frac{1}{T^2}\|\mathbf{F}'\mathbf{A}_2\|_{\mathbb{F}}^2 \leqslant \frac{1}{n^2T^4}\|\mathbf{F}'\widetilde{\mathbf{U}}'\mathbf{\Phi}\mathbf{B}\|_{\mathbb{F}}^2\|\mathbf{F}\|_{\mathbb{F}}^2\|\widehat{\mathbf{F}}\|_{\mathbb{F}}^2 \lesssim \frac{p^2}{nT}\{1 + \sqrt{\log(1/\delta)}\}^2.$$

(iii) Combining (i) and (ii), the result follows from Lemma B.2.10.

(iv) The result follows from

$$\frac{1}{T}\|\widehat{\mathbf{F}}'(\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H})\|_{\mathbb{F}} \leqslant \frac{1}{T}\|\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H}\|_{\mathbb{F}}^2 + \frac{1}{T}\|\mathbf{H}'\mathbf{F}'(\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H})\|_{\mathbb{F}}.$$

□

Lemma B.2.13. *With probability at least* $1-20\delta$*,*

$$\|\mathbf{H}'\mathbf{H}-\mathbf{I}_K\|_{\mathbb{F}}^2 \lesssim \left(\frac{1}{n^2}+\frac{p^2}{n^{1-2\alpha}T}+\frac{1}{J^{\kappa}}\right)\{1+\sqrt{\log(1/\delta)}\}^2\{1+n^{-1}\log(1/\delta)\}.$$

Proof. By Condition 3.2.2, $\mathbf{F}'\mathbf{F}=T\mathbf{I}_K$ with probability 1 and $\widehat{\mathbf{F}}'\widehat{\mathbf{F}}=T\mathbf{I}_K$. So

$$\mathbf{H}'\mathbf{H}=\frac{1}{T}(\mathbf{F}\mathbf{H})'\mathbf{F}\mathbf{H}=\frac{1}{T}(\mathbf{F}\mathbf{H}-\widehat{\mathbf{F}})'\mathbf{F}\mathbf{H}+\frac{1}{T}\widehat{\mathbf{F}}'(\mathbf{F}\mathbf{H}-\widehat{\mathbf{F}})+\mathbf{I}_K$$

and $\|\mathbf{H}'\mathbf{H}-\mathbf{I}_K\|_{\mathbb{F}} \leqslant T^{-1}\|(\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H})'\mathbf{F}\|_{\mathbb{F}}\|\mathbf{H}\|_2 + T^{-1}\|\mathbf{F}'(\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H})\|_{\mathbb{F}}$, which gives the desired result. □

Define $\widehat{\mathbf{B}}=T^{-1}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\widetilde{\mathbf{Y}}\widehat{\mathbf{F}}$ so that $\widehat{\mathbf{G}}=T^{-1}\mathbf{P}\widetilde{\mathbf{Y}}\widehat{\mathbf{F}}=\mathbf{\Phi}\widehat{\mathbf{B}}$, we have

$$\widehat{\mathbf{B}}-\mathbf{B}\mathbf{H}=\sum_{i=1}^{4}\mathbf{C}_i$$

where

$$\mathbf{C}_1=\frac{1}{T}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{R}\mathbf{F}'\widehat{\mathbf{F}}, \quad \mathbf{C}_2=\frac{1}{T}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\widetilde{\mathbf{U}}\mathbf{F}\mathbf{H},$$
$$\mathbf{C}_3=\frac{1}{T}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\widetilde{\mathbf{U}}(\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H}), \quad \mathbf{C}_4=\frac{1}{T}\mathbf{B}\mathbf{F}'(\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H}).$$

Proposition B.2.2. *With probability at least* $1-20\delta$*,*

(i) $\|\widehat{\mathbf{B}}-\mathbf{B}\mathbf{H}\|_{\mathbb{F}}^2 \lesssim \{n^{-2}J+n^{-1+2\alpha}T^{-1}p^2J+n^{-2+4\alpha}T^{-2}p^4J+J^{-\kappa+1}\}\{1+J\log(J^2/\delta)/n\}^3$ $\{1+\sqrt{\log(1/\delta)}\}^4$*,*

(ii) $n^{-1}\|\widehat{\mathbf{G}}-\mathbf{G}\mathbf{H}\|_{\mathbb{F}}^2 \lesssim (n^{-2}J + n^{-1+2\alpha}T^{-1}p^2J + n^{-2+4\alpha}T^{-2}p^4J + J^{-\kappa+1})\{1+J\log(J^2/\delta)/n\}^4\{1+\sqrt{\log(1/\delta)}\}^4$.

Proof. (i) By Lemmas B.1.1, B.2.7, B.2.8 and B.2.12, with probability at least $1-20\delta$,

$$\begin{aligned}
\|\mathbf{C}_1\|_{\mathbb{F}}^2 &\lesssim \frac{1}{J^\kappa}\{1+J\log(J^2/\delta)/n\}^3,\\
\|\mathbf{C}_2\|_{\mathbb{F}}^2 &\lesssim \frac{p^2J}{n^{2\alpha}T}\{1+J\log(J^2/\delta)/n\}^2\{1+\sqrt{\log(1/\delta)}\}^2,\\
\|\mathbf{C}_3\|_{\mathbb{F}}^2 &\lesssim \left(\frac{J}{n^2}+\frac{p^2J}{n^{2-2\alpha}T}+\frac{p^4J}{n^{2-4\alpha}T^2}+\frac{T+p^2}{nTJ^{\kappa-1}}\right)\{1+J\log(J^2/\delta)/n\}^2\\
&\qquad \{1+\sqrt{\log(1/\delta)}\}^4,\\
\|\mathbf{C}_4\|_{\mathbb{F}}^2 &\lesssim \left(\frac{J}{n^2}+\frac{p^2J}{n^{1-2\alpha}T}+\frac{1}{J^{\kappa-1}}\right)\{1+J\log(J^2/\delta)/n\}^3\{1+\sqrt{\log(1/\delta)}\}^2.
\end{aligned}$$

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

(ii) The result follows from

$$\frac{1}{n}\|\widehat{\mathbf{G}}-\mathbf{G}\mathbf{H}\|_{\mathbb{F}}^2 \leqslant \frac{2}{n}\|\mathbf{\Phi}(\widehat{\mathbf{B}}-\mathbf{B}\mathbf{H})\|_{\mathbb{F}}^2 + \frac{2}{n}\|\mathbf{R}\mathbf{H}\|_{\mathbb{F}}^2.$$

□

Proposition B.2.3. *With probability at least* $1-20\delta$*,*

(i) $\|\widehat{\mathbf{B}}-\mathbf{B}\mathbf{H}\|_{\max} \lesssim n^{-1/2}T^{-1}\{(T+p^2n^{2\alpha})\log(n)\}^{1/2}\{1+\log(1/\delta)\}$,

(ii) $\|\widehat{\mathbf{G}}-\mathbf{G}\mathbf{H}\|_{\max} \lesssim T^{-1}\{(T+p^2n^{2\alpha})\log(n)\}^{1/2}\{1+\log(1/\delta)\}$,

(iii) $\|\widehat{\mathbf{G}}-\mathbf{G}\mathbf{H}^{-1}\|_{\max} \lesssim T^{-1}\{(T+p^2n^{2\alpha})\log(n)\}^{1/2}\{1+\log(1/\delta)\}$.

Proof. (i) By Lemma B.1 in Fan et al. (2011), with probability at least $1-\delta$, $\|\mathbf{F}\widetilde{\mathbf{U}}\|_{\max} \lesssim \sqrt{(T+p^2)\log(n)}\{1+\log(1/\delta)\}$. Then, by Lemmas B.1.1, B.2.7, B.2.8 and B.2.12, with

probability at least $1 - 20\delta$,

$$\begin{aligned}
\|\mathbf{C}_1\|_{\max} &\lesssim \frac{\{\log(T)\}^{2/r_2}}{\sqrt{n}TJ^{\kappa}}\{1+\log(1/\delta)\},\\
\|\mathbf{C}_2\|_{\max} &\lesssim \frac{\sqrt{(T+p^2n^{2\alpha})\log(n)}}{\sqrt{n}T}\{1+\log(1/\delta)\},\\
\|\mathbf{C}_3\|_{\max} &\lesssim \left(\frac{T+p^2n^{2\alpha}}{nT^2}\right)\{\log(T)\}^{2/r_2}\{1+\log(1/\delta)\},\\
\|\mathbf{C}_4\|_{\max} &\lesssim \left(\frac{T+p^2n^{2\alpha}}{\sqrt{n}T^2J^{\kappa}}\right)\{\log(T)\}^{2/r_2}\{1+\log(1/\delta)\}.
\end{aligned}$$

So $\|\widehat{\mathbf{B}}-\mathbf{BH}\|_{\max} \lesssim n^{-1/2}T^{-1}\{(T+p^2n^{2\alpha})\log(n)\}^{1/2}\{1+\log(1/\delta)\}$.

(ii) The result follows from

$$\|\widehat{\mathbf{G}}-\mathbf{GH}\|_{\max} \leqslant \frac{2}{n}\|\mathbf{\Phi}(\widehat{\mathbf{B}}-\mathbf{BH})\|_{\max} + \frac{2}{n}\|\mathbf{RH}\|_{\max}.$$

(iii) The result follows from

$$\widehat{\mathbf{G}}-\mathbf{GH}^{-1} = \frac{1}{T}\mathbf{GH}^{-1}(\mathbf{HF}'-\widehat{\mathbf{F}}')\widehat{\mathbf{F}} + \frac{1}{T}\mathbf{P}\widetilde{\mathbf{U}}(\widehat{\mathbf{F}}-\mathbf{FH}) + \frac{1}{T}\mathbf{P}\widehat{\mathbf{U}}\mathbf{FH}.$$

□

Proposition B.2.4. *With probability at least* $1-20\delta$*,*

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

Proof. Note that

$$\mathbf{HK} = \frac{1}{n}\mathbf{B}'\mathbf{\Phi}'\mathbf{\Phi}\mathbf{B}\left(\frac{1}{T}\mathbf{F}'\widehat{\mathbf{F}}-\mathbf{H}\right) + \frac{1}{n}\mathbf{B}'\mathbf{\Phi}'\mathbf{\Phi}\mathbf{BH}.$$

By Lemma B.2.12, with probability at least $1-20\delta$,

$$\left\|\frac{1}{n}\mathbf{B}'\mathbf{\Phi}'\mathbf{\Phi}\mathbf{B}\left(\frac{1}{T}\mathbf{F}'\widehat{\mathbf{F}}-\mathbf{H}\right)\right\|_{\mathbb{F}}$$

$$
\begin{aligned}
&\leqslant \frac{1}{n}\|\mathbf{\Phi B}\|_{\mathbb{F}}^2\frac{1}{T}\|\mathbf{F}'(\widehat{\mathbf{F}}-\mathbf{FH})\|_{\mathbb{F}} \\
&\lesssim \left(\frac{1}{n}+\frac{p}{\sqrt{n^{1-2\alpha}T}}+\frac{1}{J^{\kappa/2}}\right)\{1+\sqrt{\log(1/\delta)}\}\sqrt{1+n^{-1}\log(1/\delta)}.
\end{aligned}
$$

In addition, by Conditions 3.3.1 and 3.3.3, $\|\mathbf{G}'\mathbf{G}-\mathbf{B}'\mathbf{\Phi}'\mathbf{\Phi B}\|_{\mathbb{F}} \lesssim nJ^{-\kappa/2}$. Therefore, with probability at least $1-20\delta$,

$$
\left\|\frac{1}{n}\mathbf{G}'\mathbf{GH}-\mathbf{HK}\right\|_{\mathbb{F}} \lesssim \left(\frac{1}{n}+\frac{p}{\sqrt{n^{1-2\alpha}T}}+\frac{1}{J^{\kappa/2}}\right)\{1+\sqrt{\log(1/\delta)}\}\sqrt{1+n^{-1}\log(1/\delta)}.
$$

This implies that with probability at least $1-20\delta$, $\mathbf{H}$ (up to an error term) is a matrix consisting of eigenvectors of $n^{-1}\mathbf{G}'\mathbf{G}$. By Condition 3.2.2, $\mathbf{G}'\mathbf{G}$ 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 $1-20\delta$,

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

for some diagonal matrix $\mathbf{D}$. By Lemma B.2.13, with probability at least $1-20\delta$, for each $i=1,\ldots,K$,

$$
|\lambda(\mathbf{H})-\eta| \lesssim \left(\frac{1}{n}+\frac{p}{\sqrt{n^{1-2\alpha}T}}+\frac{1}{J^{\kappa/2}}\right)\{1+\sqrt{\log(1/\delta)}\}\sqrt{1+n^{-1}\log(1/\delta)}
$$

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 $1-20\delta$,

$$
\begin{aligned}
\|\mathbf{H}-\mathbf{I}_K\|_{\mathbb{F}}^2 &= \sum_{i\neq j}h_{ij}^2+\sum_{i=1}^{K}(h_{ii}-1)^2 \\
&\lesssim \left(\frac{1}{n^2}+\frac{p^2}{n^{1-2\alpha}T}+\frac{1}{J^{\kappa}}\right)\{1+\sqrt{\log(1/\delta)}\}^2\{1+n^{-1}\log(1/\delta)\}.
\end{aligned}
$$

□

B.2.4 Technical results for the proof of Theorem 3.3.2

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 3.2.2 in the main paper, where

$$\widehat{\boldsymbol{\Sigma}}_f(s) = \frac{1}{T-s}\sum_{t=1}^{T-s}(\boldsymbol{f}_t - \bar{\boldsymbol{f}})(\boldsymbol{f}_{t+s} - \bar{\boldsymbol{f}})'$$

and

$$\widehat{\boldsymbol{\Sigma}}_f(-s) = \frac{1}{T-s}\sum_{t=s}^{T}(\boldsymbol{f}_{t-s} - \bar{\boldsymbol{f}})(\boldsymbol{f}_t - \bar{\boldsymbol{f}})'$$

for $s \geqslant 0$, respectively.

Lemma B.2.14. *Under Condition 3.2.2, with probability at least* $1-\delta$*,*

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

Proof. Note that

$$\mathcal{V}(\boldsymbol{f}_t) = \frac{1}{T^2}\sum_{t,s=1}^{T}(\boldsymbol{f}_t - \bar{\boldsymbol{f}})(\boldsymbol{f}_s - \bar{\boldsymbol{f}})' = \frac{1}{T^2}\mathbf{F}'\mathbf{P}_1\mathbf{F},$$

where $\mathbf{P}_1$ is the projection matrix onto $(1,\ldots,1)' \in \mathbb{R}^T$. Thus, by Theorem 3.3.1

$$\begin{aligned}
\left\|\mathcal{V}(\widehat{\boldsymbol{f}}_t) - \mathcal{V}(\boldsymbol{f}_t)\right\|_{\mathbb{F}}^2 &= \frac{1}{T^4}\left\|\widehat{\mathbf{F}}\mathbf{P}_1\widehat{\mathbf{F}}' - \mathbf{F}\mathbf{P}_1\mathbf{F}'\right\|_{\mathbb{F}}^2 \\
&\leqslant \frac{1}{T^4}\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2\{\|\mathbf{P}_1\widehat{\mathbf{F}}\|_{\mathbb{F}}^2 + \|\mathbf{P}_1\mathbf{F}\|_{\mathbb{F}}^2\} \\
&\lesssim \frac{1}{T^2}\left(\frac{1}{n} + \frac{p^2}{n^{1-2\alpha}T} + \frac{1}{J^{\kappa}}\right)\{1+\sqrt{\log(20/\delta)}\}^2.
\end{aligned}$$

The conclusion follows. □

Lemma B.2.15. *Under Conditions 3.2.1, 3.2.2, and 3.3.5,*

$$\left\| \mathcal{V}\left(\boldsymbol{f}_t\right) - \operatorname{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t\right)\right\|_{\mathbb{F}} \lesssim \frac{1}{T^2}$$

Proof. Recall that

$$\operatorname{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t\right) = \frac{1}{T^2}\sum_{t,s}\operatorname{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_s)$$

and

$$\mathcal{V}\left(\boldsymbol{f}_t\right) = \frac{1}{T^2}\sum_{t,s}(\boldsymbol{f}_t - \bar{\boldsymbol{f}})(\boldsymbol{f}_s - \bar{\boldsymbol{f}})'.$$

By Davydov's inequality (Athreya and Lahiri, 2006), for each $k = 1, \ldots, K$ and $t, s = 1, \ldots, T$, $|\mathbb{E}(f_{tk}f_{sk})^2| \lesssim \{\alpha(|t-s|)\}^{1/r_1}\{\mathbb{E}(|f_{tk}|^{2q_1})\}^{1/q_1}\{\mathbb{E}(|f_{sk}|^{2q_2})\}^{1/q_2}$, 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 3.3.5, $\mathbb{E}(|f_{tk}|^{q_1})$ and $\mathbb{E}(|f_{sk'}|^{q_2})$ exist for each $t = 1, \ldots, T$ and $\alpha(|t-s|) < \exp(-C_1|t-s|^{r_1})$, so $|\mathbb{E}(f_{tk}f_{sk})^2| \lesssim \exp(-|t-s|)$. Thus,

$$\|\operatorname{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_s)\|_{\mathbb{F}} \lesssim \exp(-|t-s|)$$

and

$$\left\| \mathcal{V}\left(\boldsymbol{f}_t\right) - \operatorname{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t\right)\right\|_{\mathbb{F}} = \left\|\frac{1}{T^2}\sum_{t,s}\operatorname{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_s)\right\|_{\mathbb{F}} \lesssim \frac{1}{T^2}\sum_{t=1}^{T}\exp(-t) \lesssim \frac{1}{T^2}.$$

□

Lemma B.2.16. *For each $i = 1, \ldots, n$, with probability at least $1 - \delta$,*

$$\left|\mathcal{V}\left(\hat{u}_{it}\right) - \operatorname{Var}\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}\right)\right|$$

$$\lesssim \frac{1}{T}\left[\frac{1}{\sqrt{n}T}+\frac{1}{n^{3/2-\alpha}}+\frac{p}{T^{1/2}n^{3/2}}+\frac{\{(T+p^2)\log(n)\}^{1/4}}{\sqrt{n^2T}}+\frac{1}{nJ^{\kappa/2}}\right]\{1+\sqrt{\log(21/\delta)}\},$$

where $\mathcal{V}(\widehat{u}_{1t})$ is defined in Section 3.2.2 of the main paper.

Proof. Denote $\widehat{\mathbf{U}}=\{\widehat{u}_{it}\}_{i=1,t=1}^{n,T}$. Note that

$$\mathbf{U}-\widehat{\mathbf{U}}=(\widehat{\mathbf{G}}-\mathbf{G}\mathbf{H}^{-1})(\widehat{\mathbf{F}}'-\mathbf{H}\mathbf{F}')+\mathbf{G}\mathbf{H}^{-1}(\widehat{\mathbf{F}}'-\mathbf{H}\mathbf{F}')+(\widehat{\mathbf{G}}-\mathbf{G}\mathbf{H}^{-1})\mathbf{H}\mathbf{F}'.$$

Then by Propositions B.2.1 and B.2.3 and Cauchy-Schwartz inequality, with probability at least $1-20\delta$,

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

Thus, similarly to the proof of Lemmas B.2.14 and B.2.15, with probability at least $1-21\delta$,

$$|\mathcal{V}(\widehat{u}_{it})-\mathcal{V}(u_{it})|\lesssim\frac{1}{nT}\left[\frac{1}{\sqrt{n}}+\frac{p}{\sqrt{n^{1-2\alpha}T}}+\frac{\{(T+p^2n^{2\alpha})\log(n)\}^{1/4}}{\sqrt{T}}+\frac{1}{J^{\kappa/2}}\right]\{1+\sqrt{\log(20/\delta)}\}$$

and

$$\left|\mathcal{V}(u_{it})-\operatorname{Var}\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}\right)\right|\lesssim\frac{1}{\sqrt{n}T^2}\{1+\sqrt{\log(1/\delta)}\}.$$

The conclusion follows. □

Lemma B.2.17. *With probability at least $1-\delta$,*

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

Proof. Recall that $\mathbf{V}=\mathbf{G}\operatorname{Var}\left(T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t\right)\mathbf{G}'+\boldsymbol{\mathcal{D}}$, so

$$\lambda_{\min}(\mathbf{V})\geqslant\lambda_{\min}\left\{\mathbf{G}\operatorname{Var}\left(T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_t\right)\mathbf{G}'\right\}+\lambda_{\min}(\mathbf{D})\gtrsim T^{-1}.$$

Note that

$$\begin{aligned}\widehat{\mathbf{V}} - \mathbf{V} =& \mathbf{G}\left\{\mathcal{V}(\widehat{\boldsymbol{f}}_t) - \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t\right)\right\}\mathbf{G}' + (\widehat{\mathbf{G}} - \mathbf{G})\mathcal{V}(\widehat{\boldsymbol{f}}_t)\widehat{\mathbf{G}}' \\ &+ \mathbf{G}\mathcal{V}(\widehat{\boldsymbol{f}}_t)(\widehat{\mathbf{G}} - \mathbf{G})' + (\widehat{\boldsymbol{\mathcal{D}}} - \boldsymbol{\mathcal{D}}).\end{aligned}$$

In addition, by the proof of Theorem 2 in Fan et al. (2008), $\|\mathbf{G}\mathbf{V}^{-1}\mathbf{G}\|_2 = O(T)$. Thus,

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

From Lemmas B.2.14 and B.2.15, with probability at least $1 - \delta$,

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

and

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

which leads to the desired assertion by lemma B.2.16 and Theorem 3.3.1. □

As a straightforward corollary to Lemma B.2.17, with probability at least $1 - \delta$,

$$\left\|\widehat{\mathbf{V}} - \mathbf{V}\right\|_{\mathbf{V},\mathbb{F}} \lesssim \left\{\frac{\sqrt{J}}{n} + \frac{1}{\sqrt{n}} + \frac{1}{T} + \frac{p\sqrt{J}}{\sqrt{n^{1-2\alpha}T}} + \frac{1}{J^{(\kappa-1)/2}}\right\}\sqrt{\log(1/\delta)},$$

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 $\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\mathbf{V},\mathbb{F}} \lesssim \{n^{-1}\sqrt{J} + p\sqrt{J}n^{-1/2+\alpha}T^{-1/2} + J^{-(\kappa-1)/2}\}\sqrt{\log(1/\delta)}$, which mimics the optimal rate from Fan et al. (2013) and Wang and Fan (2017).

B.2.5 Some legitimate preliminary estimators

In this section, we will discuss some preliminary estimators $\widehat{\boldsymbol{\beta}}^0$ that satisfy the condition of TOPE. That is, $\|\widehat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}\|_2 = O_P(n^{-1/2+\alpha}T^{-1/2})$ for $\alpha \in [0, 1/2)$ as in Section 3.2.2. We start with the ordinary least squares (OLS) estimator based on an average version of model (3.1.1) or (3.2.2) over time,

$$\widehat{\boldsymbol{\beta}}^{\text{OLS}} = (\mathbb{Z}_0'\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\bar{\boldsymbol{y}}. \tag{B.2.6}$$

Proposition B.2.5. *Under Conditions 3.2.1, 3.2.2, and 3.3.4, with probability at least* $1-\delta$*,*

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

Proof. Combining (B.2.6) and $\bar{\boldsymbol{y}} = \mathbb{Z}_0'\boldsymbol{\beta} + \mathbf{G}T^{-1}\sum_{t=1}^T \boldsymbol{f}_t + T^{-1}\sum_{t=1}^T \boldsymbol{u}_t$, we have

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{\text{OLS}} &= (\mathbb{Z}_0'\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\left\{\frac{1}{T}\sum_{t=1}^T(\mathbf{Z}_t\boldsymbol{\beta} + \mathbf{G}\boldsymbol{f}_t + \boldsymbol{u}_t)\right\} \\
&= \boldsymbol{\beta} + (\mathbb{Z}_0'\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\mathbf{G}\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}_t\right) + (\mathbb{Z}_0'\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{u}_t\right) \\
&\equiv \boldsymbol{\beta} + (\mathrm{I}) + (\mathrm{II}).
\end{aligned}$$

By Condition 3.3.4, with probability 1, $\|\mathbf{P}_Z\mathbf{G}\|_{\mathbb{F}}^2 \lesssim n^{2\alpha}$. In addition, eigenvalues of $n^{-1}\mathbb{Z}_0'\mathbb{Z}_0$ is bounded away from 0 and infinity almost surely by Condition 3.3.4 (i). Thus, eigenvalues of $(n^{-1}\mathbb{Z}_0'\mathbb{Z}_0)^{-1}$ is bounded away from 0 and infinity almost surely. That is,

$$\left\|(\mathbb{Z}_0'\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\right\|_{\mathbb{F}}^2 = \mathrm{tr}\{(\mathbb{Z}_0'\mathbb{Z}_0)^{-1}\} \lesssim \frac{p}{n},$$

and thus

$$\|\,(\mathbb{Z}_0'\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\mathbf{G}\|_{\mathbb{F}}^2 \leqslant \|\,(\mathbb{Z}_0'\mathbb{Z}_0)^{-1}\mathbb{Z}_0'\|_{\mathbb{F}}^2\|\mathbf{P}_Z\mathbf{G}\|_{\mathbb{F}}^2 \lesssim \frac{p^2}{n^{1-2\alpha}}$$

by Cauchy-Schwarz inequality. In light of Lemma B.2.3, we have

$$\mathbb{P}\left\{\|(\mathrm{I})\|_2 > sT^{-1/2}\left\|\left(\mathbb{Z}_0'\mathbb{Z}_0\right)^{-1}\mathbb{Z}_0'\mathbf{G}\right\|_{\mathbb{F}}\right\} < C_1\exp(-C_2s^2).$$

By Lemma B.2.2, it holds

$$\mathbb{P}\left\{\|(\mathrm{II})\|_2 > sT^{-1/2}\left\|\left(\mathbb{Z}_0'\mathbb{Z}_0\right)^{-1}\mathbb{Z}_0'\right\|_{\mathbb{F}}\right\} < C_1\exp(-C_2s^2).$$

Thus, we have

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}-\boldsymbol{\beta}\|_2 > sT^{-1/2}\left\{\left\|\left(\mathbb{Z}_0'\mathbb{Z}_0\right)^{-1}\mathbb{Z}_0'\mathbf{G}\right\|_{\mathbb{F}}+\left\|\left(\mathbb{Z}_0'\mathbb{Z}_0\right)^{-1}\mathbb{Z}_0'\right\|_{\mathbb{F}}\right\}\right\} < 2C_1\exp(-C_2s^2).$$

□

Hence, $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ is a legitimate preliminary estimator for the TOPE. In addition to the OLS estimator, one may consider the following preliminary estimator. As discussed in Section 3.2.1 in the main paper, $\boldsymbol{z}_{it}$ and $\boldsymbol{g}(\boldsymbol{x}_i)$ are allowed to be dependent so that we can 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 (3.1.1) in the main paper can be rewritten as

$$y_{it} = \boldsymbol{z}_{it}'\boldsymbol{\beta} + \boldsymbol{z}_{i\cdot}'\boldsymbol{\eta}_t + \boldsymbol{g}_0(\boldsymbol{x}_i)'\boldsymbol{f}_t + u_{it},$$

where $\boldsymbol{\eta}_t = \mathbf{A}'\boldsymbol{f}_t$. Under Condition 3.2.1, $\boldsymbol{g}_0(\boldsymbol{x}_i)'\boldsymbol{f}_t + u_{it}$ is uncorrelated with the regressors $\boldsymbol{z}_{it}$. Hence, we can use the following random-effects GLS (Schmidheiny and Basel, 2011) to estimate $(\boldsymbol{\beta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)$ by

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\eta}}_1\\ \vdots\\ \widehat{\boldsymbol{\eta}}_T\end{bmatrix} = (\mathbf{W}'\widehat{\boldsymbol{\Sigma}}_R^{-1}\mathbf{W})^{-1}\mathbf{W}'\widehat{\boldsymbol{\Sigma}}_R^{-1}\mathbf{y},$$

where $\mathbf{y} = (y_{11}, \ldots, y_{n1}, \ldots, y_{1T}, \ldots, y_{nT})'$,

$$\mathbf{W} = \begin{bmatrix} \boldsymbol{z}_{11}' & \boldsymbol{z}_{1\cdot}' & & \\ \vdots & \vdots & & \\ \boldsymbol{z}_{n1}' & \boldsymbol{z}_{n\cdot}' & & \\ \vdots & & \ddots & \\ \boldsymbol{z}_{1T}' & & & \boldsymbol{z}_{1\cdot}' \\ \vdots & & & \vdots \\ \boldsymbol{z}_{nT}' & & & \boldsymbol{z}_{n\cdot}' \end{bmatrix}$$

and $\widehat{\boldsymbol{\Sigma}}_R$ is an estimator of $\boldsymbol{\Sigma}_R$, the covariance matrix of $\mathbf{v} = (\boldsymbol{g}_0(\boldsymbol{x}_1)'\boldsymbol{f}_1 + u_{11}, \ldots, \boldsymbol{g}_0(\boldsymbol{x}_n)'\boldsymbol{f}_1 + u_{n1}, \ldots, \boldsymbol{g}_0(\boldsymbol{x}_1)'\boldsymbol{f}_T + u_{1T}, \ldots, \boldsymbol{g}_0(\boldsymbol{x}_n)'\boldsymbol{f}_T + u_{nT})'$. Under Condition 3.2.1 in the main paper, $\boldsymbol{\Sigma}_R$ is a block diagonal matrix $\operatorname{diag}(\boldsymbol{\Sigma}_{R,1}, \ldots, \boldsymbol{\Sigma}_{R,T})$ with

$$\boldsymbol{\Sigma}_{R,t} = \mathbb{E}\left\{ \begin{bmatrix} \boldsymbol{g}_0(\boldsymbol{x}_1)' \\ \vdots \\ \boldsymbol{g}_0(\boldsymbol{x}_n)' \end{bmatrix} \begin{bmatrix} \boldsymbol{g}_0(\boldsymbol{x}_1) & \ldots & \boldsymbol{g}_0(\boldsymbol{x}_n) \end{bmatrix} \right\} + \sigma_u^2 \mathbf{I}_n$$

for each $t = 1, \ldots, T$, where $\operatorname{var}(u_{it}) = \sigma_u^2$. There are a variety of estimators of $\widehat{\boldsymbol{\Sigma}}_{R,1}$. For instance, Bai (2009b) and Schmidheiny and Basel (2011) 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 (Bai, 2009b; Lam and Yao, 2012; Leek and Storey, 2007) and can be extended to the iterative feasible GLS estimator (Bai, 2009b; Phillips, 2010). That is, we can update $\widehat{\boldsymbol{\Sigma}}_{R,1}^{\text{new}}$ using $(\widehat{\boldsymbol{\beta}}^{\text{old}}, \widehat{\boldsymbol{\eta}}_1^{\text{old}}, \ldots, \widehat{\boldsymbol{\eta}}_T^{\text{old}})$ from the previous step and iteratively update $(\widehat{\boldsymbol{\beta}}^{\text{new}}, \widehat{\boldsymbol{\eta}}_1^{\text{new}}, \ldots, \widehat{\boldsymbol{\eta}}_T^{\text{new}})$ using the update $\widehat{\boldsymbol{\Sigma}}_{R,1}^{\text{new}}$. The update $(\widehat{\boldsymbol{\beta}}^{\text{new}}, \widehat{\boldsymbol{\eta}}_1^{\text{new}}, \ldots, \widehat{\boldsymbol{\eta}}_T^{\text{new}})$ admits the following shrinkage of errors.

Proposition B.2.6 (Lemma 1 in Phillips (2010))**.** *Under Conditions C1 to C3 in Phillips (2010), if* $T \geqslant p+1$ *and* $\mathbf{A}_0 = \mathbb{E}(\mathbf{W}'\mathbf{\Sigma}_R^{-1}\mathbf{W})$ *is nonsingular,*

$$\begin{aligned}&\sqrt{n}\left\{(\widehat{\boldsymbol{\beta}}^{new'}, \widehat{\boldsymbol{\eta}}_1^{new'}, \ldots, \widehat{\boldsymbol{\eta}}_T^{new'})' - (\boldsymbol{\beta}', \boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_T')'\right\}\\ =&\frac{2T}{T-1}\sqrt{n}\left\{(\widehat{\boldsymbol{\beta}}^{old'}, \widehat{\boldsymbol{\eta}}_1^{old'}, \ldots, \widehat{\boldsymbol{\eta}}_T^{old'})' - (\boldsymbol{\beta}', \boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_T')'\right\}\boldsymbol{\psi}\mathbf{A}_0^{-1}\boldsymbol{\psi} + o_P(1),\end{aligned}$$

where $\boldsymbol{\psi}$ *is given in Phillips (2010).*

Together along with $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leqslant \|(\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\eta}}_1', \ldots, \widehat{\boldsymbol{\eta}}_T')' - (\boldsymbol{\beta}', \boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_T')'\|_2$, Proposition B.2.6 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 the TOPE.

B.3 Technical results for Section 3.5.1

B.3.1 Proof of Theorem 3.5.1

Recall the notation from Section 3.5.1 that $\boldsymbol{\gamma}_i$ and $\widehat{\boldsymbol{\gamma}}_i$ denote the eigenvectors corresponding to the ith largest eigenvalues of $\mathbf{V}$ and $\widetilde{\mathbf{V}}$, respectively; and $\theta_{it} = \boldsymbol{\gamma}_i'\widetilde{\boldsymbol{y}}_t$, $\widehat{\theta}_{it} = \widehat{\boldsymbol{\gamma}}_i'\widetilde{\boldsymbol{y}}_t$, and $\widetilde{\boldsymbol{w}}_t = (\widehat{\boldsymbol{\gamma}}_{K_0+1}, \ldots, \widehat{\boldsymbol{\gamma}}_n)'\widetilde{\boldsymbol{y}}_t$ for $\widetilde{\boldsymbol{y}}_t = \boldsymbol{y}_t - \mathbf{Z}_t\widehat{\boldsymbol{\beta}}$ discussed in Section 3.5.1. Let $\widehat{\boldsymbol{\mu}} = (\mathbf{I}_S \otimes \widehat{\mathbf{\Gamma}})[\text{vec}\{\widehat{\mathbf{\Sigma}}(1)\}', \ldots, \text{vec}\{\widehat{\mathbf{\Sigma}}(S)\}']'$, and $\boldsymbol{\mu} = (\mathbf{I}_S \otimes \mathbf{\Gamma})[\text{vec}\{\widehat{\mathbf{\Sigma}}(1)\}', \ldots, \text{vec}\{\widehat{\mathbf{\Sigma}}(S)\}']'$, where $\widehat{\mathbf{\Gamma}} = \text{diag}\{\widehat{\mathbf{\Sigma}}(0)\}^{-1/2} \otimes \text{diag}\{\widehat{\mathbf{\Sigma}}(0)\}^{-1/2}$, $\mathbf{\Gamma} = \text{diag}\{\mathbf{\Sigma}(0)\}^{-1/2} \otimes \text{diag}\{\mathbf{\Sigma}(0)\}^{-1/2}$, $\widehat{\mathbf{\Sigma}}(s) = \sum_{t=1}^{T-s} \widetilde{\boldsymbol{w}}_{t+s}\widetilde{\boldsymbol{w}}_t'/T$, and $\mathbf{\Sigma}(s) = \sum_{t=1}^{T} \boldsymbol{w}_{t+s}\boldsymbol{w}_t'/T$ for each s. Then, consider

$$\widehat{\psi} = \sqrt{T}\max_{1\leqslant l\leqslant (n-K_0)^2S}\widehat{\mu}_l,\ \psi = \sqrt{T}\max_{1\leqslant l\leqslant (n-K_0)^2S}\mu_l,\ G^0 = \max_{1\leqslant l\leqslant (n-K_0)^2S}G_l,$$

where $\boldsymbol{\mathcal{G}} = (G_1, \ldots, G_l, \ldots, G_{(n-K_0)^2S}) \sim N(0, \mathbf{\Xi}_T)$, $\mathbf{\Xi}_T = (\mathbf{I}_S \otimes \mathbf{\Gamma})\mathbb{E}(\boldsymbol{\xi}_T\boldsymbol{\xi}_T')(\mathbf{I}_S \otimes \mathbf{\Gamma})$, and $\boldsymbol{\xi}_T = \sqrt{T}[\text{vec}\{\widehat{\mathbf{\Sigma}}(1)\}', \ldots, \text{vec}\{\widehat{\mathbf{\Sigma}}(S)\}']'$ as defined in Section 3.5.1 in the main paper.

Similar to arguments in the proof of Proposition 1 in the appendix to Chang et al. (2017), Theorem 3.5.1 follows from $\sup_{1\leqslant K_0\leqslant K}\sup_s|\mathbb{P}(\widehat{\psi} \leqslant s) - \mathbb{P}(G^0 \leqslant s)| = o(1)$. To that end,

first, recall that for each K_0, $\widehat{\mathbf{\Sigma}}(0) = \sum_{t=1}^{T} \widetilde{\boldsymbol{w}}_t \widetilde{\boldsymbol{w}}_t'/T$ and $\mathbf{\Sigma}(0) = \sum_{t=1}^{T} \boldsymbol{w}_t \boldsymbol{w}_t'/T$, where $\widetilde{\boldsymbol{w}}_t = (\widehat{w}_{K_0+1,t}, \ldots, \widehat{w}_{nt})'$ and $\boldsymbol{w}_t = (w_{K_0+1,t}, \ldots, w_{nt})'$. Lemma B.3.1, together with Proposition B.2.5 and Condition 3.3.5, implies that $\sup_{1 \leqslant i \leqslant K} \sup_{1 \leqslant t \leqslant T} \mathbb{P}(|\widehat{w}_{it}| > s) \lesssim \exp(-s^r)$ for some $r > 0$. Then, for $\operatorname{diag}\{\widehat{\mathbf{\Sigma}}(0)\}^{-1/2} = (\widehat{\sigma}_1^{(0)}, \ldots, \widehat{\sigma}_{n-K_0}^{(0)})$ and $\operatorname{diag}\{\mathbf{\Sigma}(0)\}^{-1/2} = (\sigma_1^{(0)}, \ldots, \sigma_{n-K_0}^{(0)})$, Lemma B.3.1 and Theorem 1 in Merlevède et al. (2011) imply that

$$\sup_{1 \leqslant K_0 \leqslant K} \sup_{1 \leqslant i \leqslant n-K_0} \mathbb{P}(|\widehat{\sigma}_i^{(0)} - \sigma_i^{(0)}| > \varepsilon) \lesssim nT \exp(-T^{\iota}\varepsilon^{\iota}) + n \exp(-T\varepsilon^2)$$

for some $0 < \iota < 1$ and any $\epsilon > 0$. Hence, for each $K_0 = 1, \ldots, K$ and any $\epsilon > 0$,

$$\mathbb{P}(|\widehat{\psi} - \psi| > \varepsilon) \lesssim nT \exp(-T^{\iota}\varepsilon^{\iota}) + n \exp(-T\varepsilon^2). \tag{B.3.1}$$

Along Lemma A.4 in Chang et al. (2017) and anti-concentration inequality of Gaussian random variables, (B.3.1) implies

$$\begin{aligned} & \sup_{1 \leqslant K_0 \leqslant K} \sup_{s} |\mathbb{P}(\widehat{\psi} \leqslant s) - \mathbb{P}(G^0 \leqslant s)| \\ \leqslant & \sup_{1 \leqslant K_0 \leqslant K} \sup_{s} |\mathbb{P}(\widehat{\psi} \leqslant s) - \mathbb{P}(\psi \leqslant s)| + \sup_{s} |\mathbb{P}(\psi \leqslant s + \varepsilon) - \mathbb{P}(G^0 \leqslant s + \varepsilon)| \\ & + \sup_{s} |\mathbb{P}(s < G^0 \leqslant s + \varepsilon)| = o(1). \end{aligned}$$

This completes the proof of Theorem 3.5.1.

B.3.2 Proof of Theorem 3.5.3

For each $k > 1$, rejecting $H_0(k)$ leads to the rejection of $H_0(k-1)$ as well. That is, $\{\text{Reject } H_0(k-1)\} \supset \{\text{Reject } H_0(k)\}$ so that $\mathbb{P}\left[\{\text{Reject } H_0(k-1)\} \cup \{\text{FTR } H_0(k)\}\right] = 1$, where FTR stands for failing to reject. Hence, for each n, Theorem 3.5.2 implies that

$$\begin{aligned} \mathbb{P}(\widehat{K} = K) &= \mathbb{P}([\cap_{k=1}^{K-1} \{\text{Reject } H_0(k)\}] \cap \{\text{FTR } H_0(K)\}) \\ &= \mathbb{P}\left[\{\text{Reject } H_0(K-1)\} \cap \{\text{FTR } H_0(K)\}\right] \end{aligned}$$

$$
\begin{aligned}
&= -\mathbb{P}\left[\{\text{Reject } H_0(K-1)\} \cup \{\text{FTR } H_0(K)\}\right] + \mathbb{P}\left\{\text{Reject } H_0(K-1)\right\} \\
&\quad +\mathbb{P}\left\{\text{FTR } H_0(K)\right\} \\
&\to 1-\alpha_n.
\end{aligned}
$$

as T goes to infinity. Therefore, $\inf_n \mathbb{P}(\widehat{K} = K) \to 1$ as T diverges to infinity with α_n given in Theorem 3.5.2.

B.3.3 Technical results for Section B.3.1

For $r \times n$ and $n \times r$ half orthogonal matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ satisfying $\mathbf{H}_1'\mathbf{H}_1 = \mathbf{H}_2'\mathbf{H}_2 = \mathbf{I}_r$, as Chang et al. (2018) we define

$$
D(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)) = \{1 - \text{tr}(\mathbf{H}_1\mathbf{H}_1'\mathbf{H}_2\mathbf{H}_2')/r\}^{1/2}
$$

where $\mathcal{M}(\mathbf{H}_1)$ and $\mathcal{M}(\mathbf{H}_2)$ are the column spaces of $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively.

Lemma B.3.1. *Under Conditions 3.2.1(b) and 3.3.5 in the main paper, for $\widehat{\theta}_{it}$ and θ_{it} defined in Section B.3.1,*

$$
\sup_{1\leqslant i\leqslant K} \sup_{1\leqslant t\leqslant T} |\widehat{\theta}_{it} - \theta_{it}| = O_p(\|\widetilde{\mathbf{V}} - \mathbf{V}\|_{\mathbb{F}}/\nu) + O_p(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2)
$$

where $\nu = \min_{1\leqslant k\leqslant K} \lambda_k(\mathbf{V}) - \lambda_{k+1}(\mathbf{V})$.

Proof. For each $i = 1, \ldots, K$, denoting $\mathbf{H}_1 = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_i)$ and $\widehat{\mathbf{H}}_1 = (\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_i)$, by the remark after Lemma 1 in Chang et al. (2018), we have

$$
D(\mathcal{M}(\widehat{\mathbf{H}}_1), \mathcal{M}(\mathbf{H}_1)) = O_p(\|\widehat{\mathbf{H}}_1 - \mathbf{H}_1\|_2) = O_p(\|\widetilde{\mathbf{V}} - \mathbf{V}\|_2/\nu_i),
$$

where $\nu_i = \lambda_i(\mathbf{V}) - \lambda_{i+1}(\mathbf{V})$. Thus, there exists some orthogonal matrix $\mathbf{Q}_i$ such that,

$$
\sup_{1\leqslant i\leqslant K} \|\widehat{\boldsymbol{\gamma}}_i - \mathbf{Q}_i\boldsymbol{\gamma}_i\|_2 = O_p(D(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2))) = O_p(\|\widetilde{\mathbf{V}} - \mathbf{V}\|_2/\nu),
$$

where $\nu = \min_{1 \leqslant k \leqslant K} \lambda_k(\mathbf{V}) - \lambda_{k+1}(\mathbf{V})$. The conclusion follows from

$$|\widehat{\theta}_{it} - \theta_{it}| \leqslant |(\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)'(\boldsymbol{y}_t - \mathbf{Z}_t\boldsymbol{\beta})| + |\widehat{\boldsymbol{\gamma}}_i'\mathbf{Z}_t(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})|.$$

□

Here ν is the smallest eigenvalue of $\mathbf{G}\mathbf{G}'$ and determines the strength of latent factors. Lemma B.3.1 also shows that

$$D(\mathcal{M}(\widehat{\boldsymbol{\gamma}}_{K+1}, \ldots, \widehat{\boldsymbol{\gamma}}_n), \mathcal{M}(\boldsymbol{\gamma}_{K+1}, \ldots, \boldsymbol{\gamma}_n)) = O_p(\|\widetilde{\mathbf{V}} - \mathbf{V}\|_{\mathbb{F}}/\nu) + O_p(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2)$$

so that

$$\|(\widehat{w}_{K+1,t}, \ldots, \widehat{w}_{nt})' - \mathbf{Q}(w_{K+1,t}, \ldots, w_{nt})'\|_2 = O_p(\|\widetilde{\mathbf{V}} - \mathbf{V}\|_{\mathbb{F}}/\nu) + O_p(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2),$$

for some orthogonal matrix $\mathbf{Q}$.

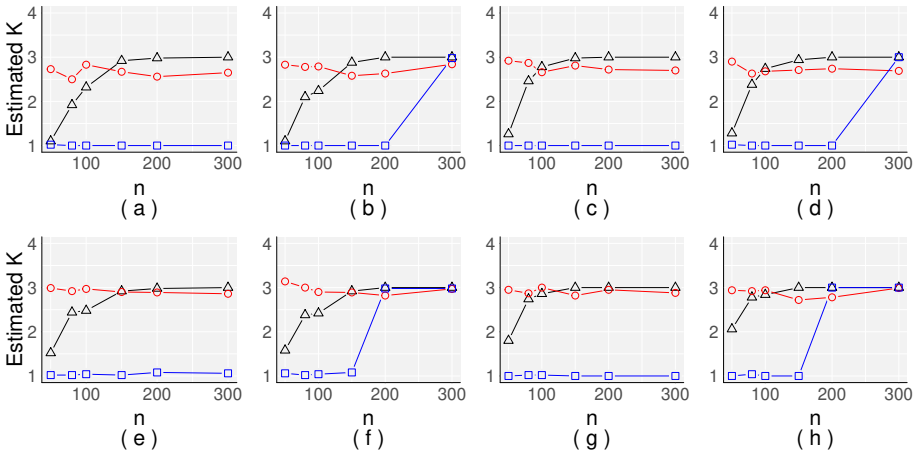

Figure B.1: Comparisons of the empirical mean of $\widehat{K}$ using HDWN testing-based procedure ("–○–") along with those of ALT_1 ("–△–") and ALT_2 (" –▫–"). In the simulation, $rt = 10$. In the first row, $T = 50$ and in the second row, $T = 100$. In (a) and (e), $\boldsymbol{f}_k$ follows $\mathrm{AR}(3)$ for each k and u_{it} is temporally independent for each i. In (b) and (f), $\boldsymbol{f}_k$ follows $\mathrm{AR}(3)$ and $\boldsymbol{u}_i$ follows $\mathrm{ARCH}(1)$ model. In (c) and (g), $\boldsymbol{f}_k$ follows $\mathrm{GARCH}(2,2)$ for each k and u_{it}'s are temporally independent. In (d) and (h), $\boldsymbol{f}_k$ follows $\mathrm{GARCH}(2,2)$ for each k and $\boldsymbol{u}_i$ follows $\mathrm{ARCH}(1)$ for each i.

B.3.4 Simulation resutls

In this section, we demonstrate the performance of proposed HDWN testing-based procedure for determining dimension K of latent process $\boldsymbol{f}_t$. For comparison, we consider the eigenvalue-ratio procedures using projected data $\mathbf{P}\widetilde{\mathbf{Y}}$ (ALT_1) (Fan et al., 2016) and original data $\widetilde{\mathbf{Y}}$ (ALT_2) (Ahn and Horenstein, 2013; Lam and Yao, 2012).

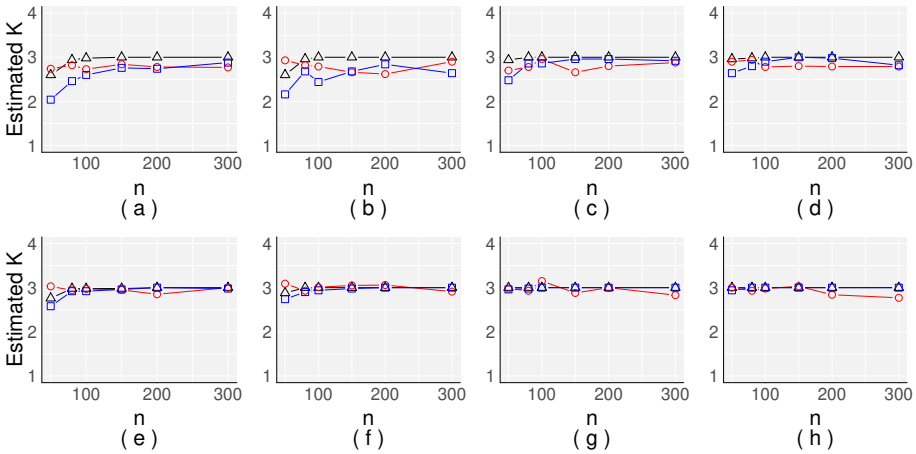

Figure B.2: Comparisons of the empirical mean of $\widehat{K}$ using HDWN testing-based procedure ("–○–") along with those of ALT_1 ("–△–"), and ALT_2 (" –▫–"). In the simulation, $rt = 5$. In the first row, $T = 50$ and in the second row, $T = 100$. In (a) and (e), $\boldsymbol{f}_k$ follows $\mathrm{AR}(3)$ for each k and u_{it} is temporally independent for each i. In (b) and (f), $\boldsymbol{f}_k$ follows $\mathrm{AR}(3)$ and $\boldsymbol{u}_i$ follows $\mathrm{ARCH}(1)$ model. In (c) and (g), $\boldsymbol{f}_k$ follows $\mathrm{GARCH}(2,2)$ for each k and u_{it}'s are temporally independent. In (d) and (h), $\boldsymbol{f}_k$ follows $\mathrm{GARCH}(2,2)$ for each k and $\boldsymbol{u}_i$ follows $\mathrm{ARCH}(1)$ for each i.

For numerical studies, we set $n = 50, 80, 100, 150, 200, 300$ and $T = 50, 100$. From model (3.1.1), data are generated using the same setting in Section 3.6.1 with $K = 3$ except that: 1) the three independent and identically distributed component series in $\boldsymbol{f}_t$ either follow $\mathrm{AR}(3)$ with autoregressive coefficient $\boldsymbol{\rho} = (0, 0, 0.5)$ or $\mathrm{GARCH}(2,2)$ with autoregressive coefficient $\boldsymbol{\alpha} = (0.12, 0.04)$ and variance coefficient $\boldsymbol{\sigma} = (0.4, 0.08)$, and standard normal innovations are used; 2) we rescale $\mathbf{G}$ such that $n^{-1/2}\mathbf{G}'\mathbf{G}$ is a 3×3 diagonal matrix with diagonals $5 \cdot rt, 5, 3$ and $rt \in \{2, 5, 10\}$; 3) finally, u_{it} are either *i.i.d.* standard normal or the n component series in $\boldsymbol{u}_t$ are independent $\mathrm{ARCH}(1)$ series with autoregressive coefficient $\alpha = 0.2$ and standard normal innovation. For each setting, 100 experiments are conducted. For the proposed HDWN testing-

based procedure, OLS $\widehat{\boldsymbol{\beta}}^{\text{OLS}}$ and thresholding estimator for $\widetilde{\mathbf{V}}$ (Bickel and Levina, 2008b) are employed and the significance level is set as $\alpha_n = 0.5n^{-1/5}$. The mean of estimated K is reported for comparison.

For the proposed HDWN-testing based procedure, committing Type I error results in large $\widehat{K}$ compared to the true K while committing Type II error leads to a smaller $\widehat{K}$. In practice, we are more keen on a slightly over-complicate model than an under-complicate one. Thus, we can set the significance level $\alpha = 0.1$ or 0.2, which provides greater powers. That is, we decrease the probability of choosing a smaller K and increase the probability of choosing a greater K. On the other hand, the proposed procedure has smaller chance of selecting wrong K when power is approaching to 1 and significance level is close to 0. Therefore, we can choose a smaller significance level for large n and T. That is, we can let $\alpha_n \to 0$ as n diverges as seen in Theorem 3.5.2.

Given small n and large ratio between the largest and second largest eigenvalues of $\mathbf{G}\mathbf{G}'$ (panels (a) and (e) in Figure B.1, for example), the proposed method outperforms eigenvalue-ratio procedures. As discussed in Section 3.5.1, the empirical eigenvalues of $\widetilde{\mathbf{Y}}\mathbf{P}\widetilde{\mathbf{Y}}$ corresponding to the nonzero counterparts diverge in n while the remaining stays in constant order. Thus, the performance of eigenvalue-ratio procedures becomes satisfactory only when n is large enough (ALT_1). On the other hand, as expected, it is observed in Figure B.1 that using the projected data (ALT_1) in the eigenvalue ratio procedure provides better estimates on K compared to that using the original data (ALT_2). When the ratio between the largest and second largest eigenvalue of $\mathbf{G}\mathbf{G}'$ is mild (Ahn and Horenstein, 2013; Fan et al., 2016; Lam and Yao, 2012), such as 5 or 2 in Figures B.2 and B.3, the performance of eigenvalue-ratio procedures improves substantially while the performance of proposed method remains satisfactory.

B.4 Additional simulation studies

In this section, Figures B.4 to B.2 display additional results from the simulation studies in the main paper. Section B.4.1 displays the mean squared error (MSE) for estimating $\boldsymbol{\beta}$, Section B.4.2 reports comparisons of the empirical coverage probability (ECP) and maximum marginal length

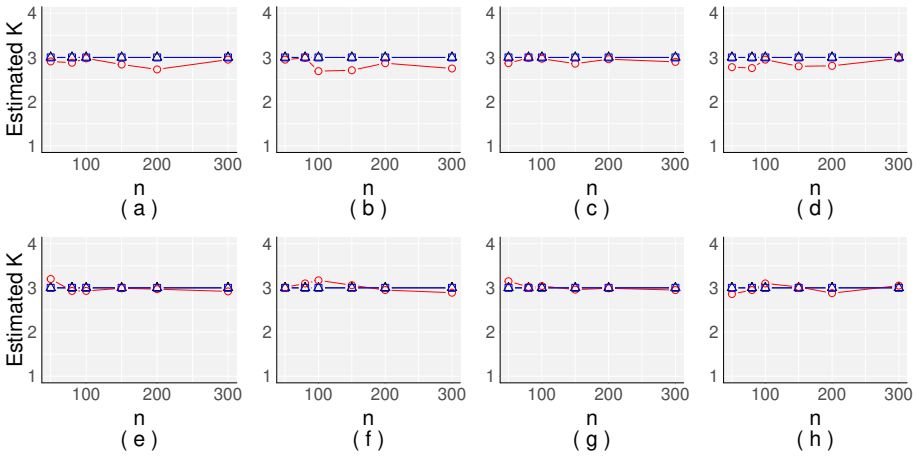

Figure B.3: Comparisons of the empirical mean of $\widehat{K}$ using HDWN testing-based procedure ("–○–") along with those of ALT_1 ("–△–"), and ALT_2 (" –▫–"). In the simulation, $rt = 2$. In the first row, $T = 50$ and in the second row, $T = 100$. In (a) and (e), $\boldsymbol{f}_k$ follows $\mathrm{AR}(3)$ for each k and u_{it} is temporally independent for each i. In (b) and (f), $\boldsymbol{f}_k$ follows $\mathrm{AR}(3)$ and $\boldsymbol{u}_i$ follows $\mathrm{ARCH}(1)$ model. In (c) and (g), $\boldsymbol{f}_k$ follows $\mathrm{GARCH}(2,2)$ for each k and u_{it}'s are temporally independent. In (d) and (h), $\boldsymbol{f}_k$ follows $\mathrm{GARCH}(2,2)$ for each k and $\boldsymbol{u}_i$ follows $\mathrm{ARCH}(1)$ for each i.

(MML) of 95% confidence regions for different estimators as considered in Section 3.6 in the main paper. Figures B.25 and B.26 in Section B.4.3 includes more plots from the real data study.

B.4.1 Mean squared error for estimating $\boldsymbol{\beta}$

This section displays the logarithm of MSE for estimating $\boldsymbol{\beta}$ with respect to the logarithm of nT with different choices of n, T, and the dependence across t in $\boldsymbol{f}_k = (f_{k1}, \ldots, f_{kt}, \ldots, f_{kT})$ for $k = 1, 2, 3$ and that in $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{it}, \ldots, u_{iT})$ for each $i = 1, \ldots, n$.

- Figures B.4 and B.5 are about independent f_{kt} for each k and t with $T = 100$ and 500, respectively.
- In Figures B.6 and B.7, $\boldsymbol{f}_k$ follows $\mathrm{ARMA}(1,1)$ model with normal or t_8 innovations for each $k = 1, 2, 3$, and $T = 100$ and 500.
- Finally, in Figures B.8–B.10, $\boldsymbol{f}_k$ follows $\mathrm{AR}(1)$ model with standard normal or centered χ^2_5 innovations for each $k = 1, 2, 3$, and $T = 20, 100, 500$.

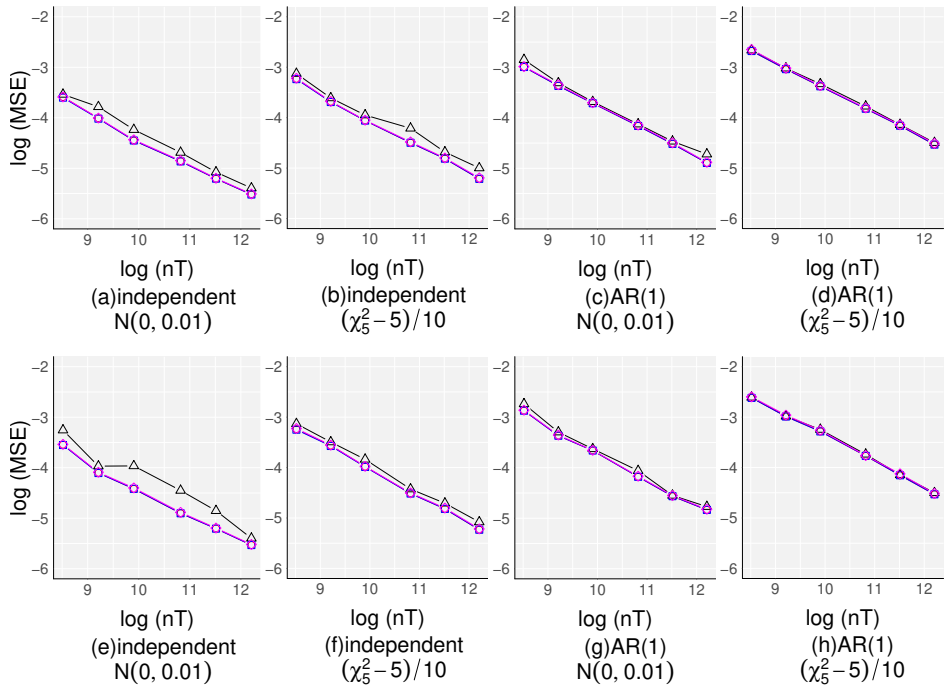

Figure B.4: Comparisons of the logarithm of MSE for estimating $\boldsymbol{\beta}$ by TOPE ("–○–") along those of the oracle estimator (" –▫–"), the GLS estimator (" –◇–"), and the OLS ("–△–"). Results are about $T = 100$. In the first row, $f_{kt} \sim N(0, 1)$ are independent in k, t. In the second row, $f_{kt} \sim t_8$ are independent in k, t. Distributions and serial correlations of $\boldsymbol{u}_i$ are displayed in the plots. Results are based on 500 replications.

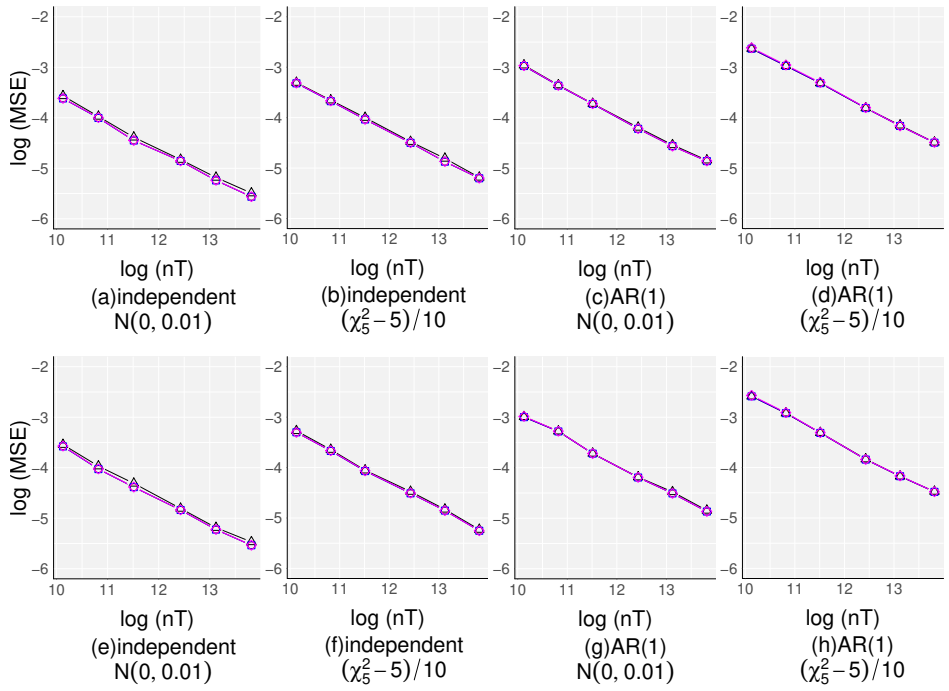

Figure B.5: Comparisons of the logarithm of MSE for estimating $\boldsymbol{\beta}$ by TOPE ("–○–") along those of the oracle estimator (" –▫–"), the GLS estimator (" –◇–"), and the OLS ("–△–"). Results are about $T = 500$. In the first row, $f_{kt} \sim N(0, 1)$ are independent in k, t. In the second row, $f_{kt} \sim t_8$ are independent in k, t. Distributions and serial correlations of $\boldsymbol{u}_i$ are displayed in the plots. Results are based on 500 replications.

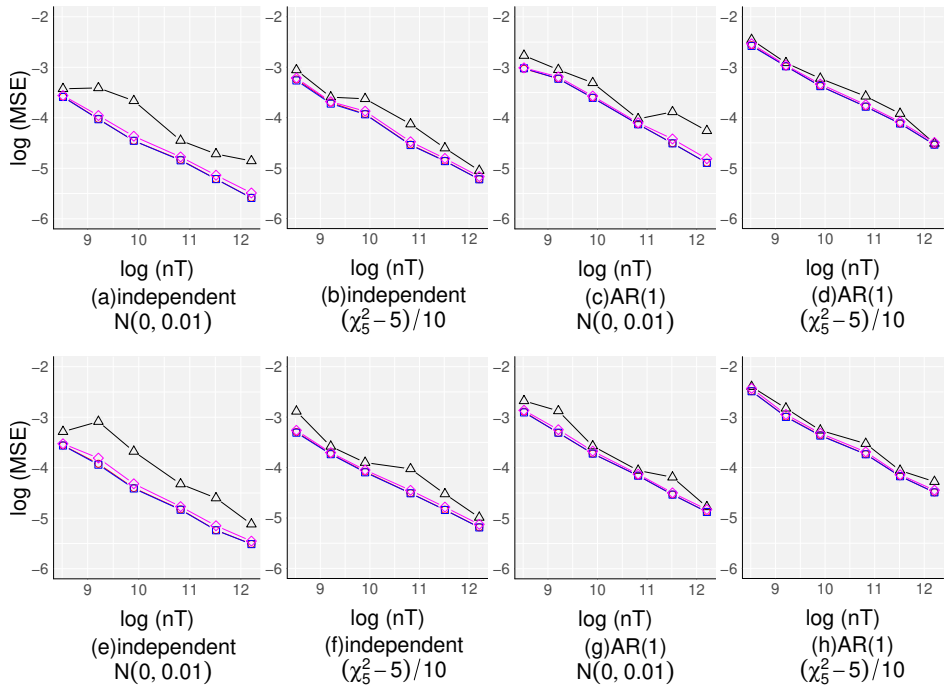

Figure B.6: Comparisons of the logarithm of MSE for estimating $\boldsymbol{\beta}$ by TOPE ("–○–") along those of the oracle estimator (" –▫–"), the GLS estimator (" –◇–"), and the OLS ("–△–"). Results are about $T = 100$. In the first row, $\boldsymbol{f}_k$ follows ARMA$(1, 1)$ with $N(0, 1)$ innovation for each k; in the second row, $\boldsymbol{f}_k$ follows ARMA$(1, 1)$ with t_8 innovation for each k. Distributions and serial correlations of $\boldsymbol{u}_i$ are displayed in the plots. Results are based on 500 replications.

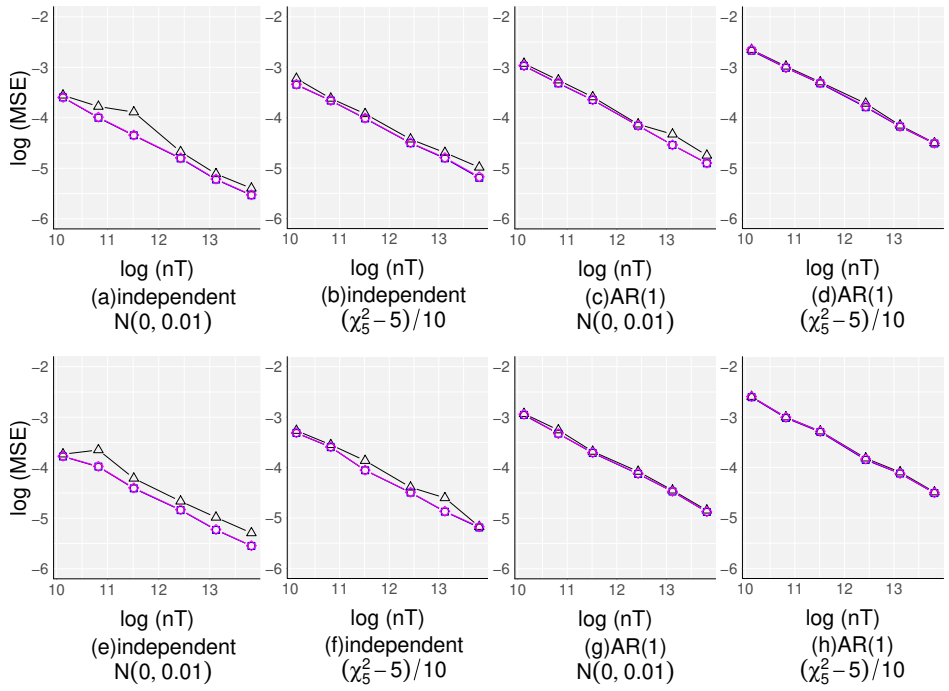

Figure B.7: Comparisons of the logarithm of MSE for estimating $\boldsymbol{\beta}$ by TOPE ("–○–") along those of the oracle estimator (" –▫–"), the GLS estimator (" –◇–"), and the OLS ("–△–"). Results are about $T = 500$. In the first row, $\boldsymbol{f}_k$ follows ARMA$(1, 1)$ with $N(0, 1)$ innovation for each k; in the second row, $\boldsymbol{f}_k$ follows ARMA$(1, 1)$ with t_8 innovation for each k. Distributions and serial correlations of $\boldsymbol{u}_i$ are displayed in the plots. Results are based on 500 replications.

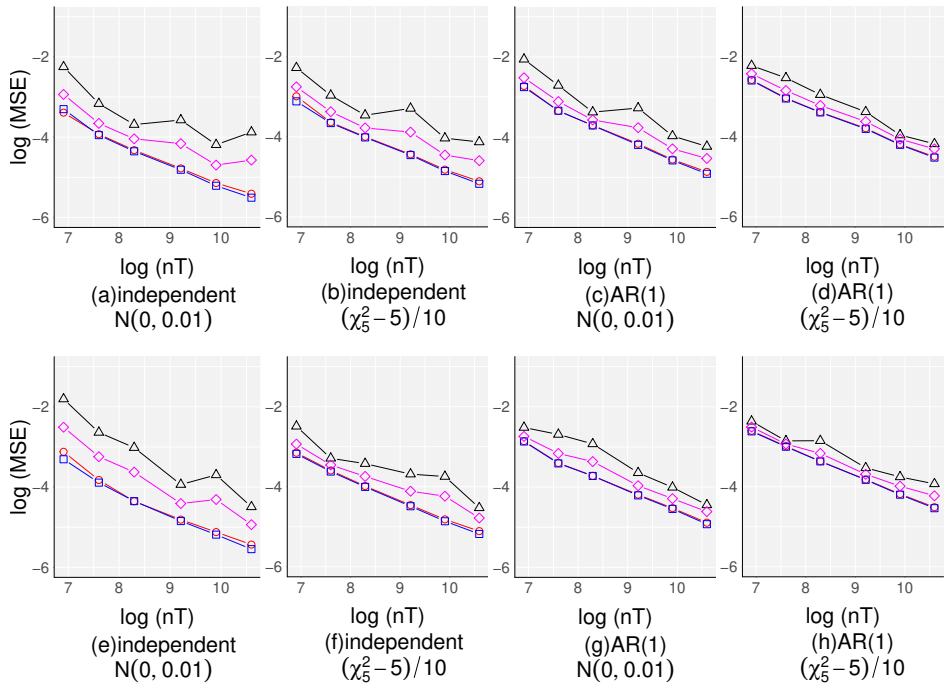

Figure B.8: Comparisons of the logarithm of MSE for estimating $\boldsymbol{\beta}$ by TOPE ("–○–") along those of the oracle estimator (" –□–"), the GLS estimator (" –◇–"), and the OLS ("–△–"). Results are about $T = 20$. For each k, $\boldsymbol{f}_k$ follows AR(1) with $N(0, 1)$ innovation in the first row; in the second row, $\boldsymbol{f}_k$ follows AR(1) with centered χ^2_5 innovation. Distributions and serial correlations of $\boldsymbol{u}_i$ are displayed in the plots. Results are based on 500 replications.

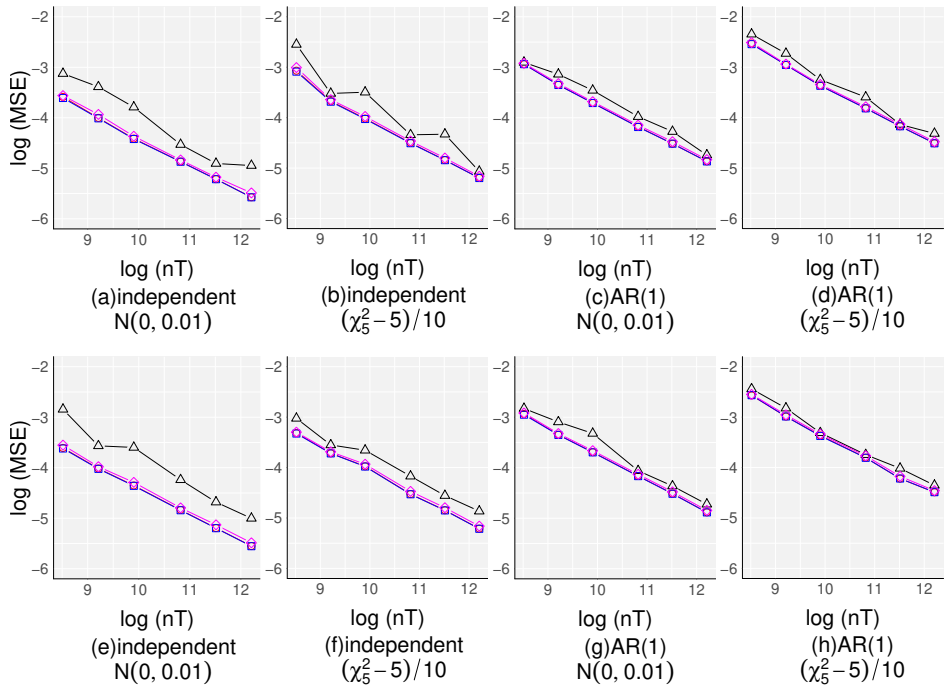

Figure B.9: Comparisons of the logarithm of MSE for estimating $\boldsymbol{\beta}$ by TOPE ("–○–") along those of the oracle estimator (" –▫–"), the GLS estimator (" –◇–"), and the OLS ("–△–"). Results are about $T = 100$. For each k, $\boldsymbol{f}_k$ follows AR(1) with $N(0, 1)$ innovation in the first row; in the second row, $\boldsymbol{f}_k$ follows AR(1) with centered χ^2_5 innovation. Distributions and serial correlations of $\boldsymbol{u}_i$ are displayed in the plots. Results are based on 500 replications.

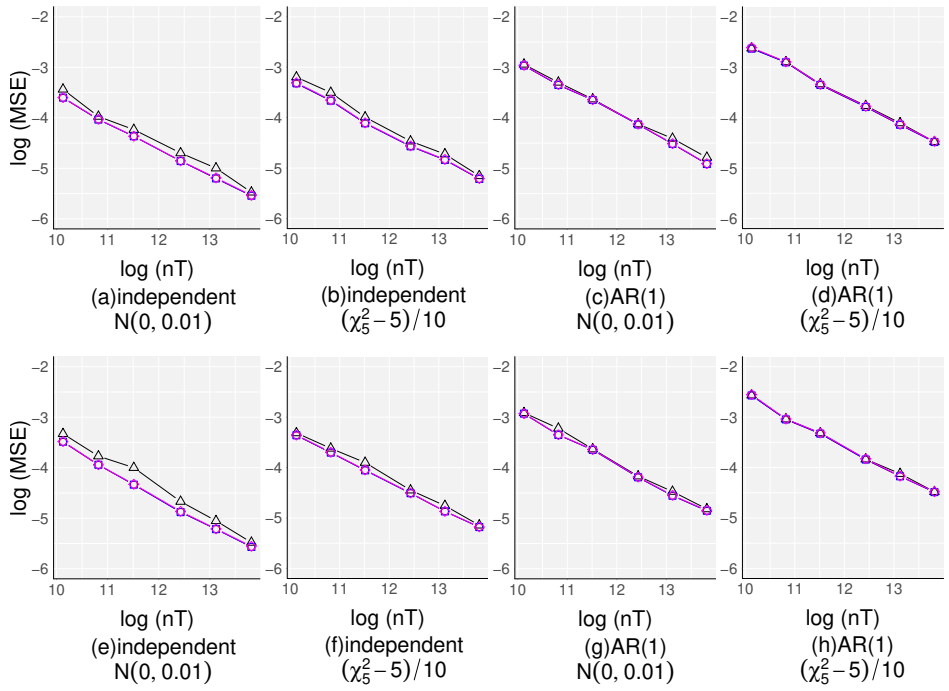

Figure B.10: Comparisons of the logarithm of MSE for estimating $\boldsymbol{\beta}$ by TOPE ("–○–") along those of the oracle estimator (" –▫–"), the GLS estimator (" –◇–"), and the OLS ("–△–"). Results are about $T = 500$. For each k, $\boldsymbol{f}_k$ follows AR(1) with $N(0, 1)$ innovation in the first row; in the second row, $\boldsymbol{f}_k$ follows AR(1) with centered χ^2_5 innovation. Distributions and serial correlations of $\boldsymbol{u}_i$ are displayed in the plots. Results are based on 500 replications.

B.4.2 Plots of empirical covering probability and maximum marginal length

This section displays the ECP and MML, which are defined in the main paper as the empirical frequency of the confidence region covering the true regression coefficients and the maximum width along the p directions of the confidence region, respectively. We display the ECP and MML along varying T for combinations of different methods, n, and dependence across t in $\boldsymbol{f}_k = (f_{k1}, \ldots, f_{kt}, \ldots, f_{kT})$ for $k = 1, 2, 3$ and that in $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{it}, \ldots, u_{iT})$ for each $i = 1, \ldots, n$.

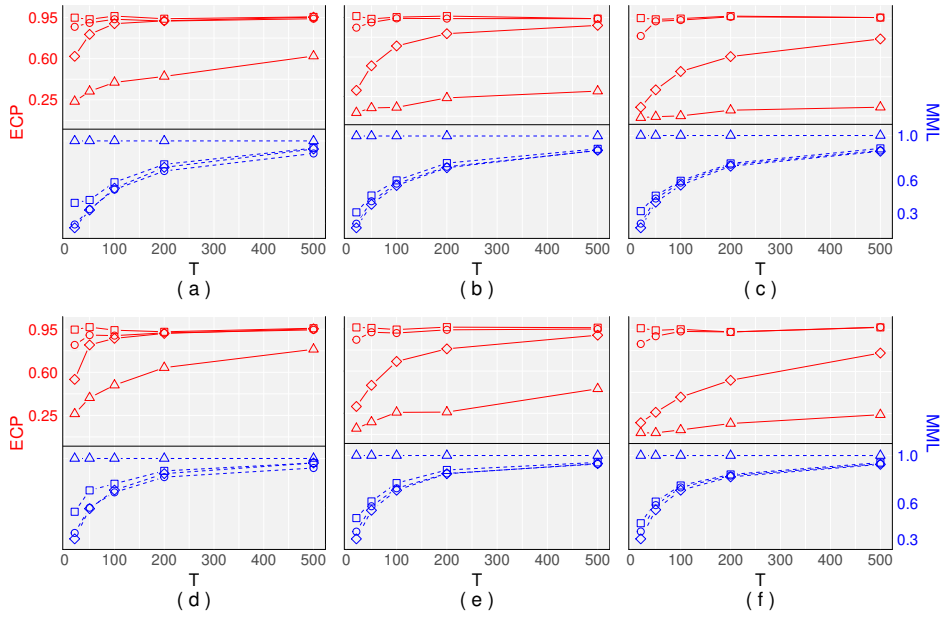

Figure B.11: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $f_{kt} \sim (\chi^2_5 - 5)$ are independent in k, t; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $u_{it} \sim N(0, 0.01)$ are independent in i, t. In the second row, $u_{it} \sim (\chi^2_5 - 5)/10$ are independent in i, t. Results are based on 500 replications.

- Figures B.11–B.14 displays results for independent f_{kt} following the centered χ^2_5 or t_8 distributions, and it is either that u_{it} are independent in i, t or $\boldsymbol{u}_i$ follows AR(1) model with different innovations.

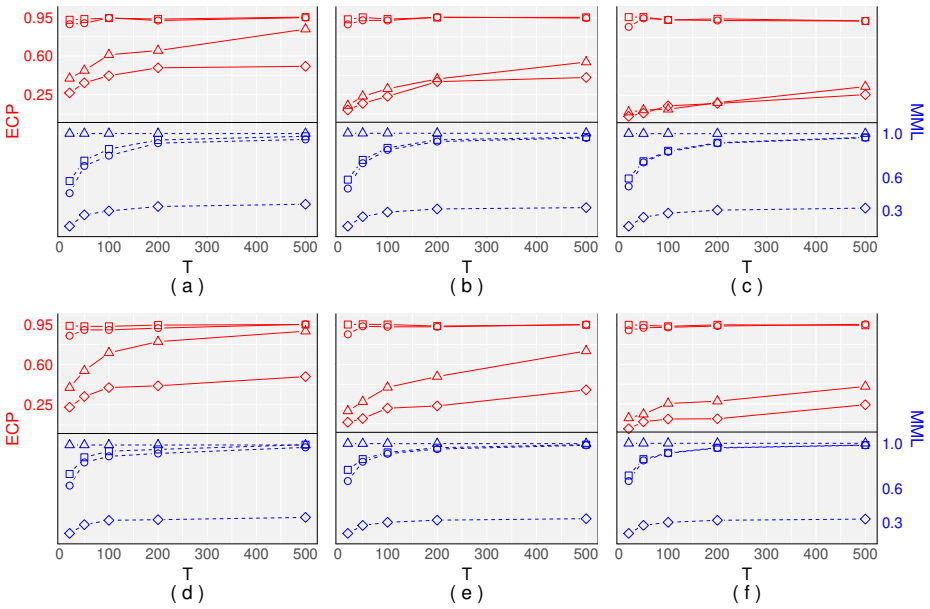

Figure B.12: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $f_{kt} \sim (\chi^2_5 - 5)$ are independent in k, t; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $\boldsymbol{u}_i$ follows the AR(1) model with $N(0, 0.01)$ innovation while same model is used for $\boldsymbol{u}_i$ in the second row with $(\chi^2_5 - 5)/10$ innovation. Results are based on 500 replications.

- In Figures B.15–B.18, $\boldsymbol{f}_k$ follows AR(1) model with different innovations while it is either that u_{it} are independent in i, t or $\boldsymbol{u}_i$ follows AR(1) model with different innovations (such as normal or centered χ^2_5).

- Figures B.19–B.24 are about results when $\boldsymbol{f}_k$ follows the ARMA$(1, 1)$ model for each k with normal, centered χ^2_5 and t_8 innovations, respectively. Residual u_{it} either are independent in i, t or $\boldsymbol{u}_i$ is AR(1) processes for each i with different innovation.

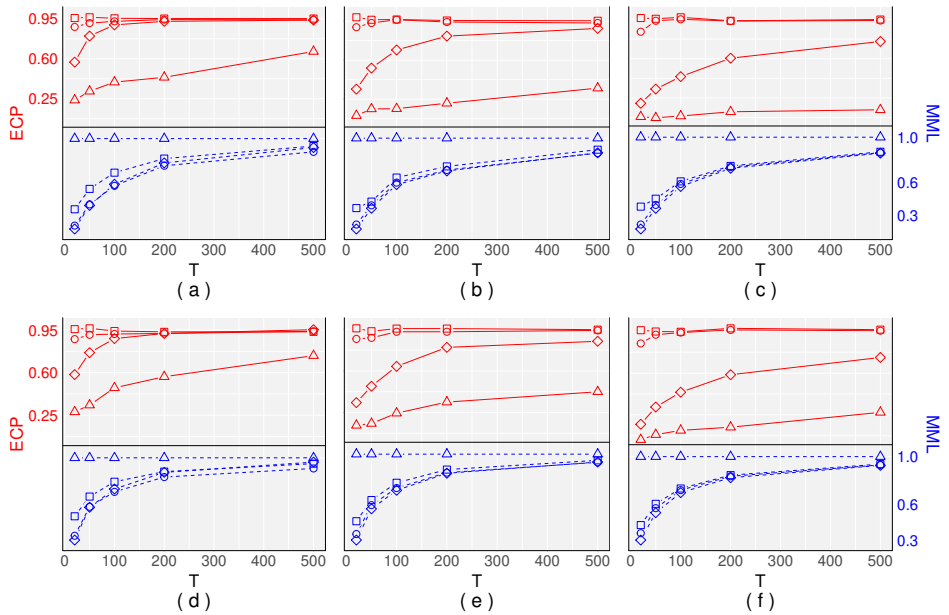

Figure B.13: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $f_{kt} \sim t_8$ are independent in k, t; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $u_{it} \sim N(0, 0.01)$ are independent in i, t. In the second row, $u_{it} \sim (\chi^2_5 - 5)/10$ are independent in i, t. Results are based on 500 replications.

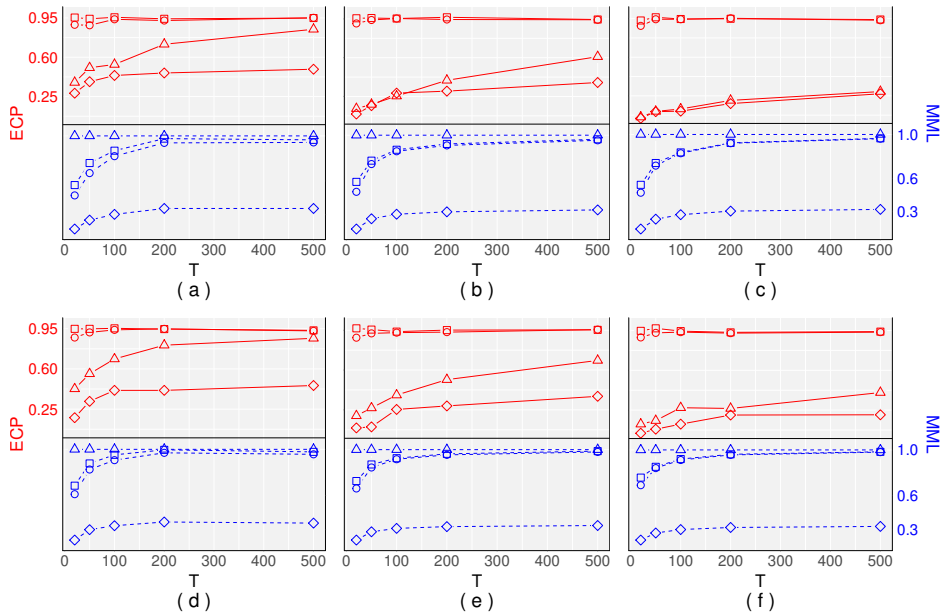

Figure B.14: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $f_{kt} \sim t_8$ are independent in k, t; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $\boldsymbol{u}_i$ follows the AR(1) model with $N(0, 0.01)$ innovation while same model is used for $\boldsymbol{u}_i$ in the second row with $(\chi^2_5 - 5)/10$ innovation. Results are based on 500 replications.

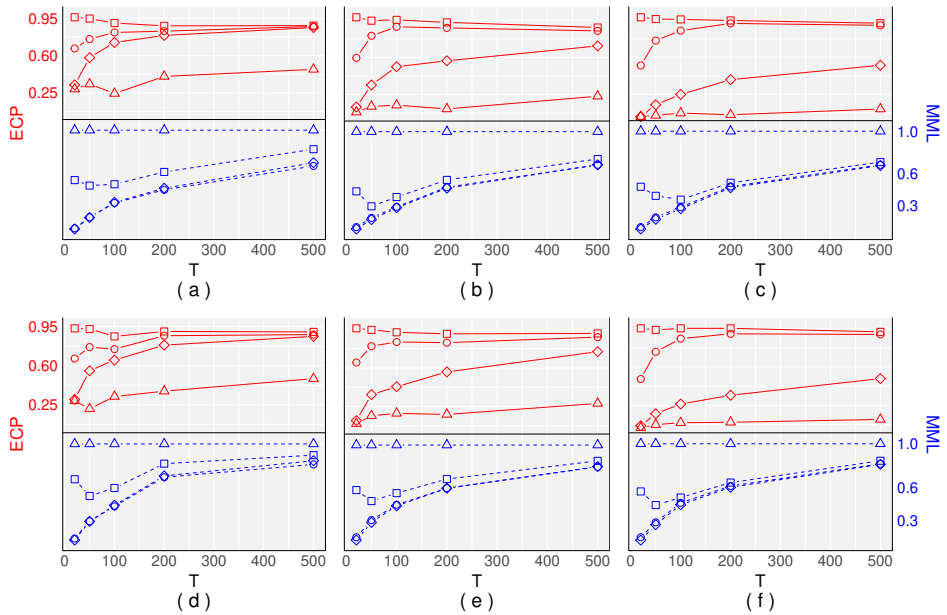

Figure B.15: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows AR(1) with $N(0,1)$ innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $u_{it} \sim N(0, 0.01)$ are independent in i, t. In the second row, $u_{it} \sim (\chi^2_5 - 5)/10$ are independent in i, t. Results are based on 500 replications.

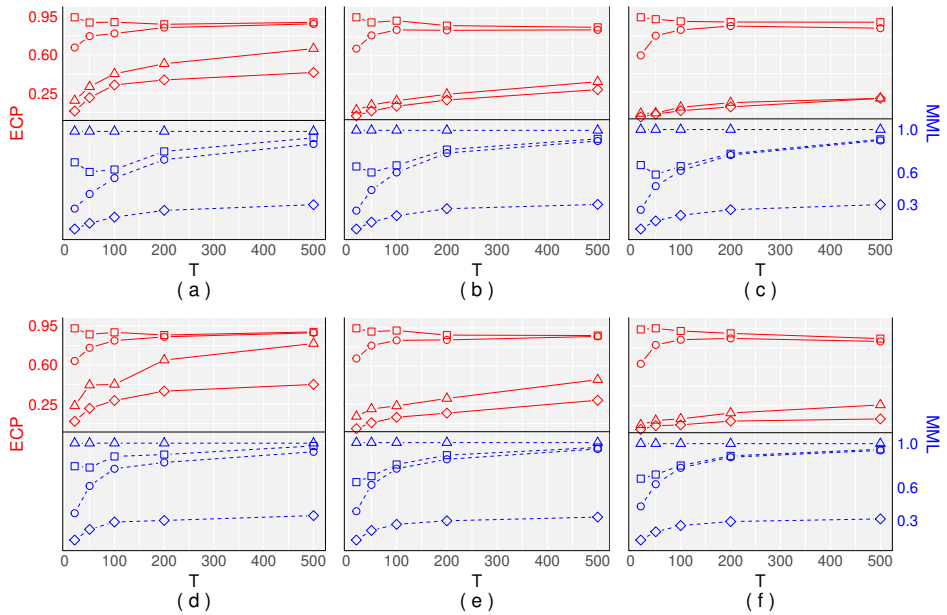

Figure B.16: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –▫–" for ECP and " - -▫- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows AR(1) with $N(0,1)$ innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $\boldsymbol{u}_i$ follows the AR(1) model with $N(0, 0.01)$ innovation while same model is used for $\boldsymbol{u}_i$ in the second row with $(\chi^2_5 - 5)/10$ innovation. Results are based on 500 replications.

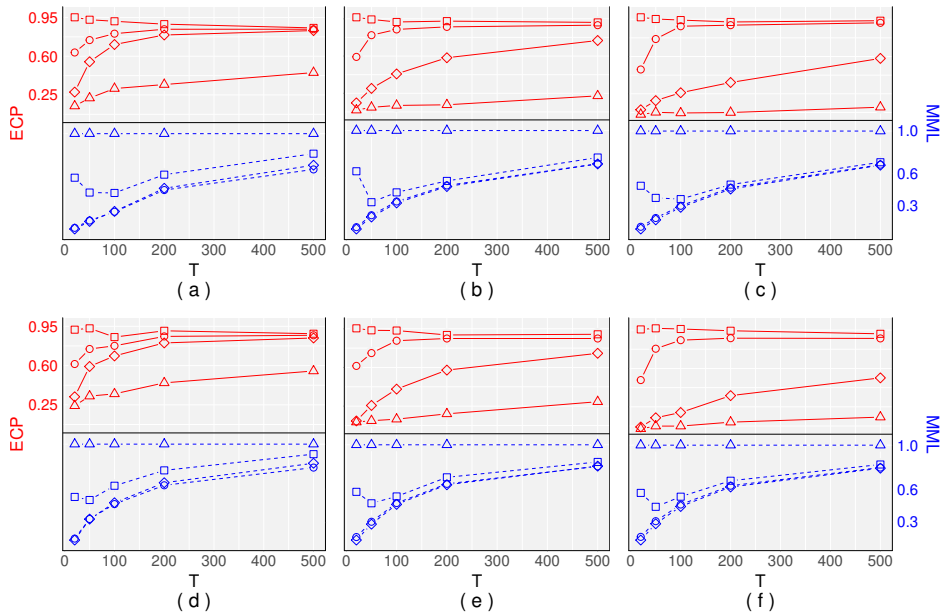

Figure B.17: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows AR(1) with centered χ^2_5 innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $u_{it} \sim N(0, 0.01)$ are independent in i, t. In the second row, $u_{it} \sim (\chi^2_5 - 5)/10$ are independent in i, t. Results are based on 500 replications.

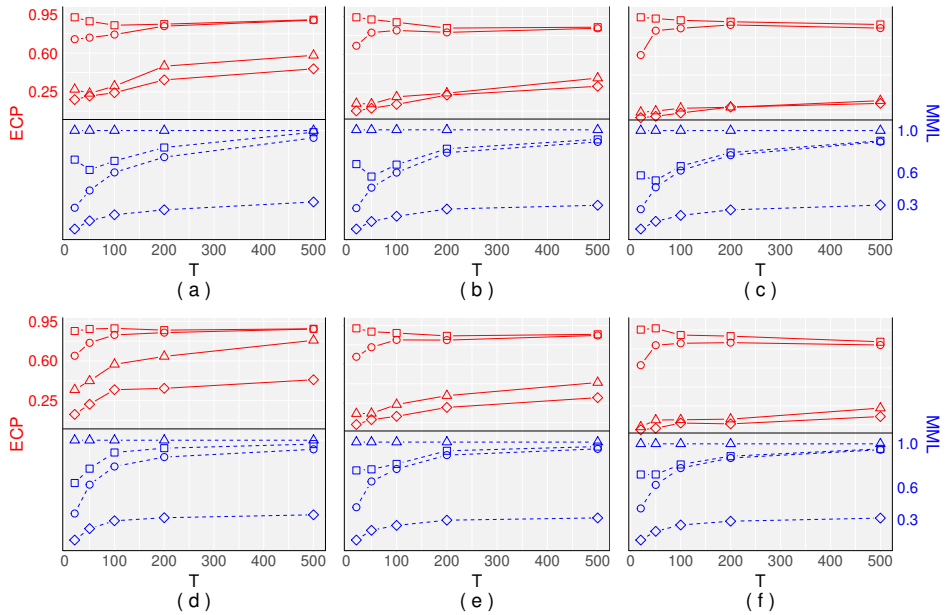

Figure B.18: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows AR(1) with centered χ^2_5 innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $\boldsymbol{u}_i$ follows the AR(1) model with $N(0, 0.01)$ innovation while same model is used for $\boldsymbol{u}_i$ in the second row with $(\chi^2_5 - 5)/10$ innovation. Results are based on 500 replications.

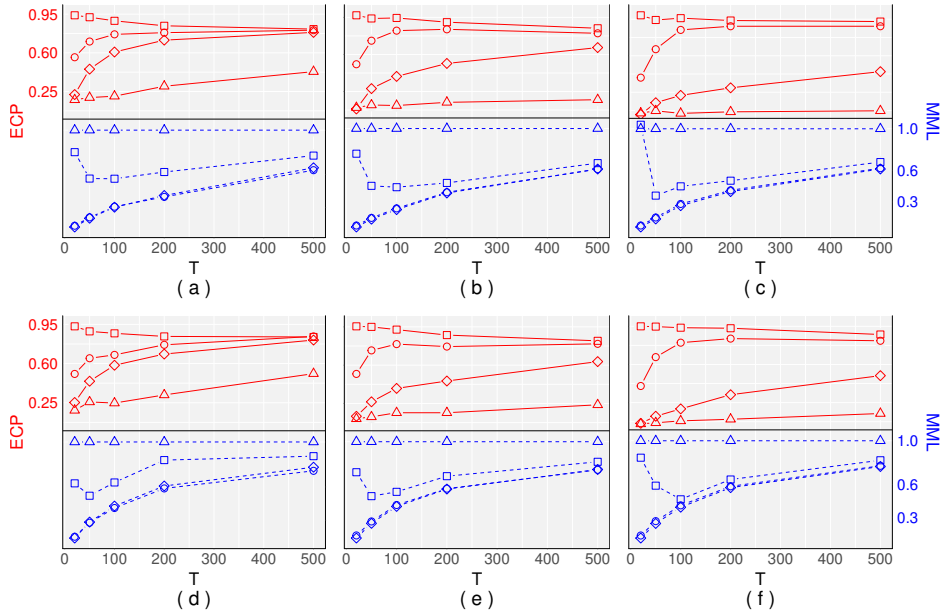

Figure B.19: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows ARMA$(1,1)$ with $N(0,1)$ innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $u_{it} \sim N(0, 0.01)$ are independent in i, t. In the second row, $u_{it} \sim (\chi^2_5 - 5)/10$ are independent in i, t. Results are based on 500 replications.

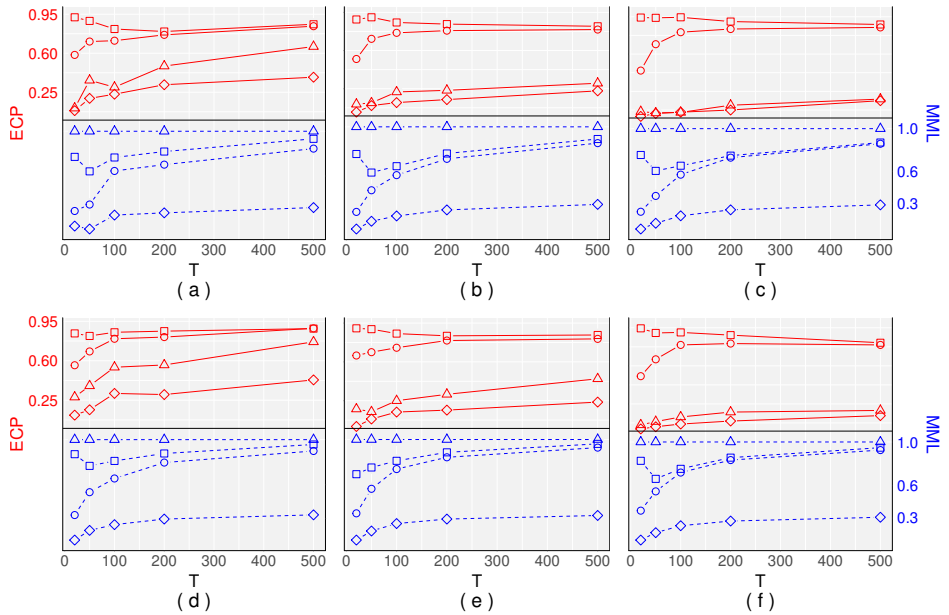

Figure B.20: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows ARMA$(1,1)$ with $N(0,1)$ innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $\boldsymbol{u}_i$ follows the AR(1) model with $N(0, 0.01)$ innovation while same model is used for $\boldsymbol{u}_i$ in the second row with $(\chi^2_5 - 5)/10$ innovation. Results are based on 500 replications.

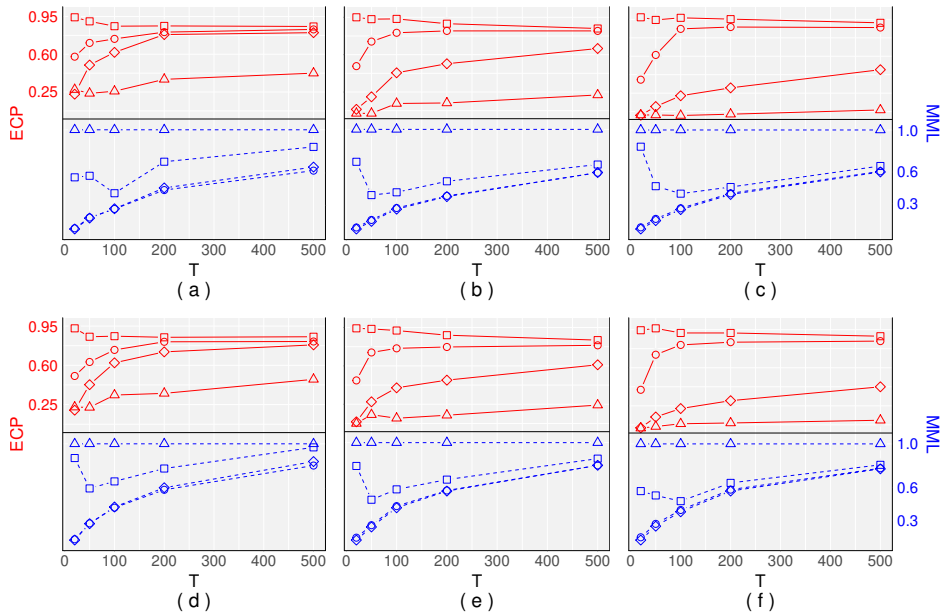

Figure B.21: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows ARMA$(1,1)$ with centered χ^2_5 innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $u_{it} \sim N(0, 0.01)$ are independent in i, t. In the second row, $u_{it} \sim (\chi^2_5 - 5)/10$ are independent in i, t. Results are based on 500 replications.

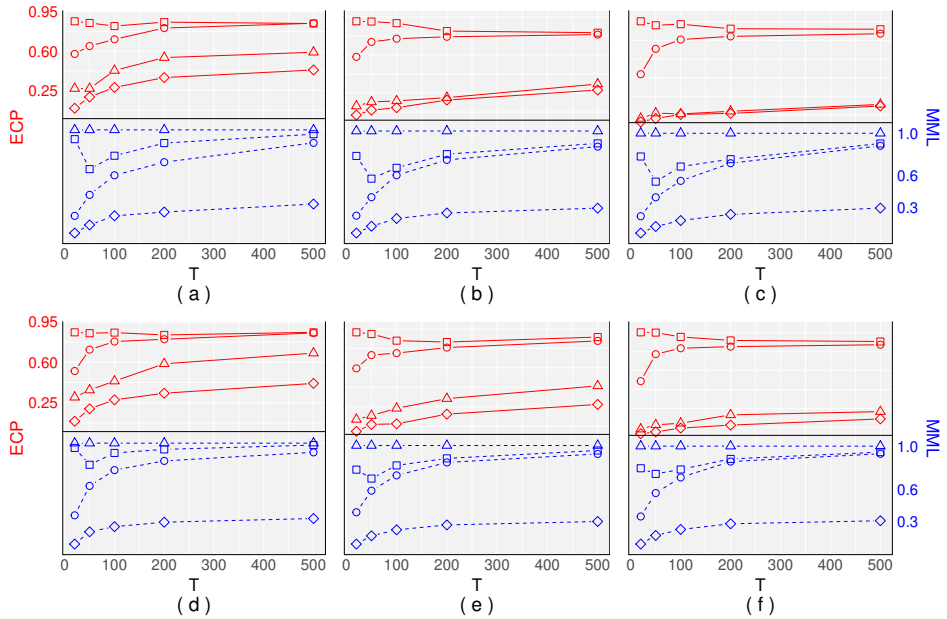

Figure B.22: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows ARMA$(1,1)$ with centered χ^2_5 innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $\boldsymbol{u}_i$ follows the AR(1) model with $N(0, 0.01)$ innovation while same model is used for $\boldsymbol{u}_i$ in the second row with $(\chi^2_5 - 5)/10$ innovation. Results are based on 500 replications.

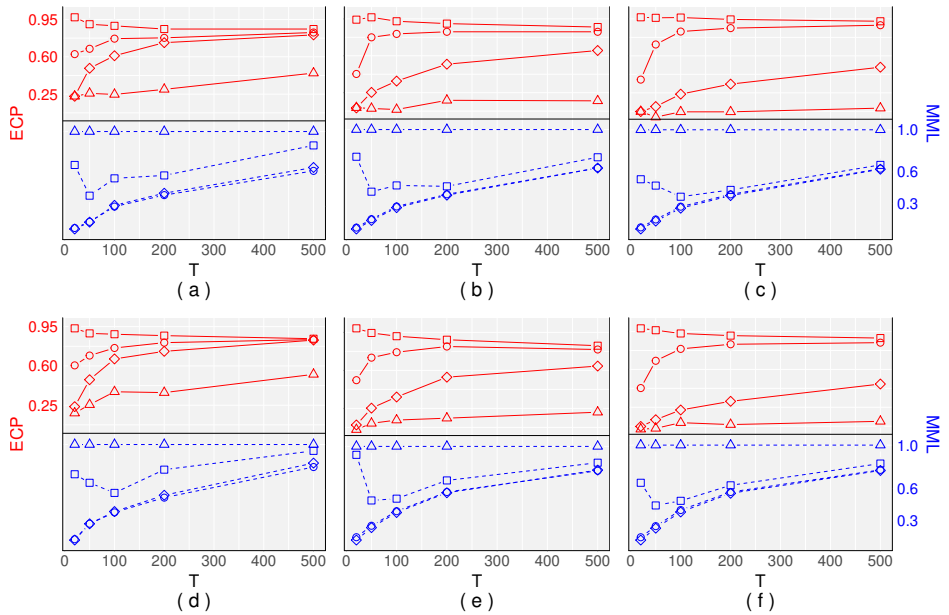

Figure B.23: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows ARMA$(1,1)$ with centered t_8 innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $u_{it} \sim N(0, 0.01)$ are independent in i, t. In the second row, $u_{it} \sim (\chi^2_5 - 5)/10$ are independent in i, t. Results are based on 500 replications.

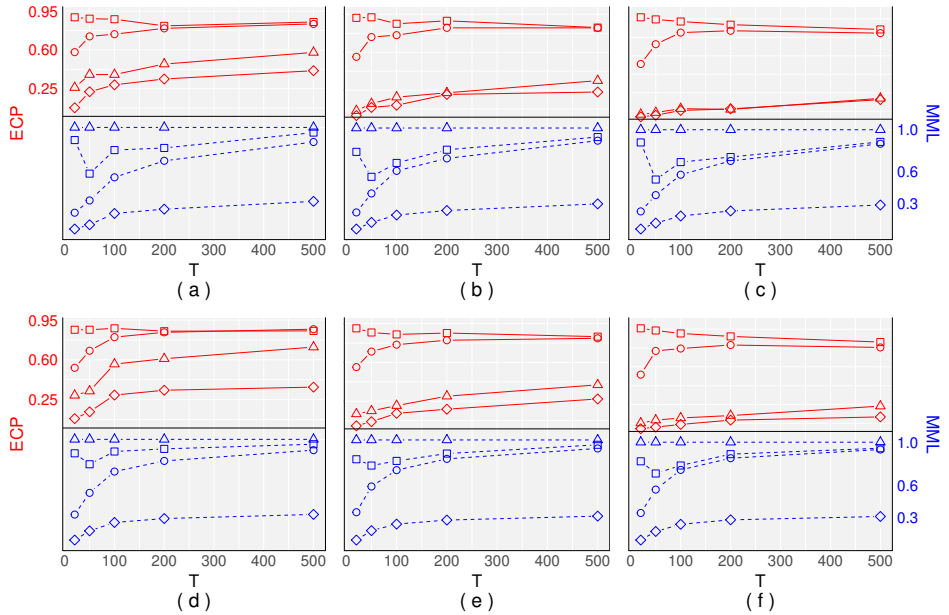

Figure B.24: Comparisons of the ECP and MML of 95% confidence region of TOPE ("–○–" for ECP and '- -○- -" for MML) along those of the oracle estimator (" –□–" for ECP and " - -□- -" for MML), the GLS estimator (" –◇–" for ECP and "- -◇- -" for MML), and the OLS (" –△–" for ECP and "- -△- -" for MML). In simulations, $\boldsymbol{f}_k$ follows ARMA$(1,1)$ with centered t_8 innovation for each k; $n = 100, 500, 2000$ for the first, second, and third column, respectively. In the first row, $\boldsymbol{u}_i$ follows the AR(1) model with $N(0, 0.01)$ innovation while same model is used for $\boldsymbol{u}_i$ in the second row with $(\chi^2_5 - 5)/10$ innovation. Results are based on 500 replications.

B.4.3 Extra displays from real data analysis

In this section, extra plots are displayed for the real data analysis conducted in Section 3.7 in the main paper.

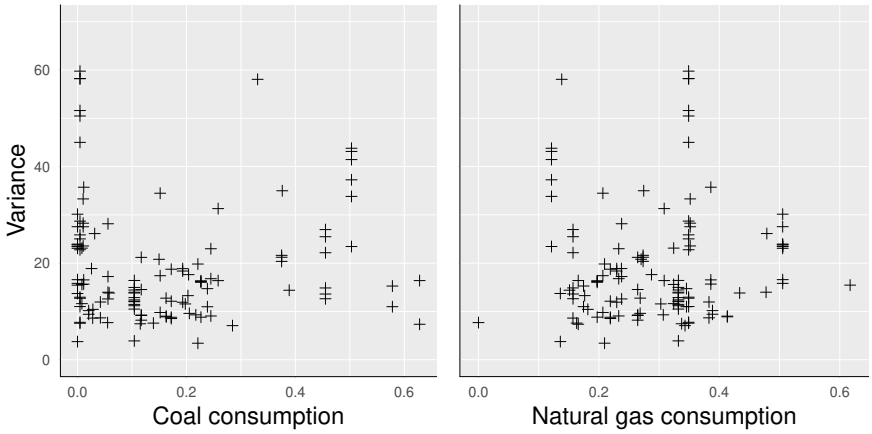

Figure B.25: Variances of the mean PM2.5 concentration at 129 monitoring sites versus coal and natural gas consumption of the states, in which the monitoring sites reside.

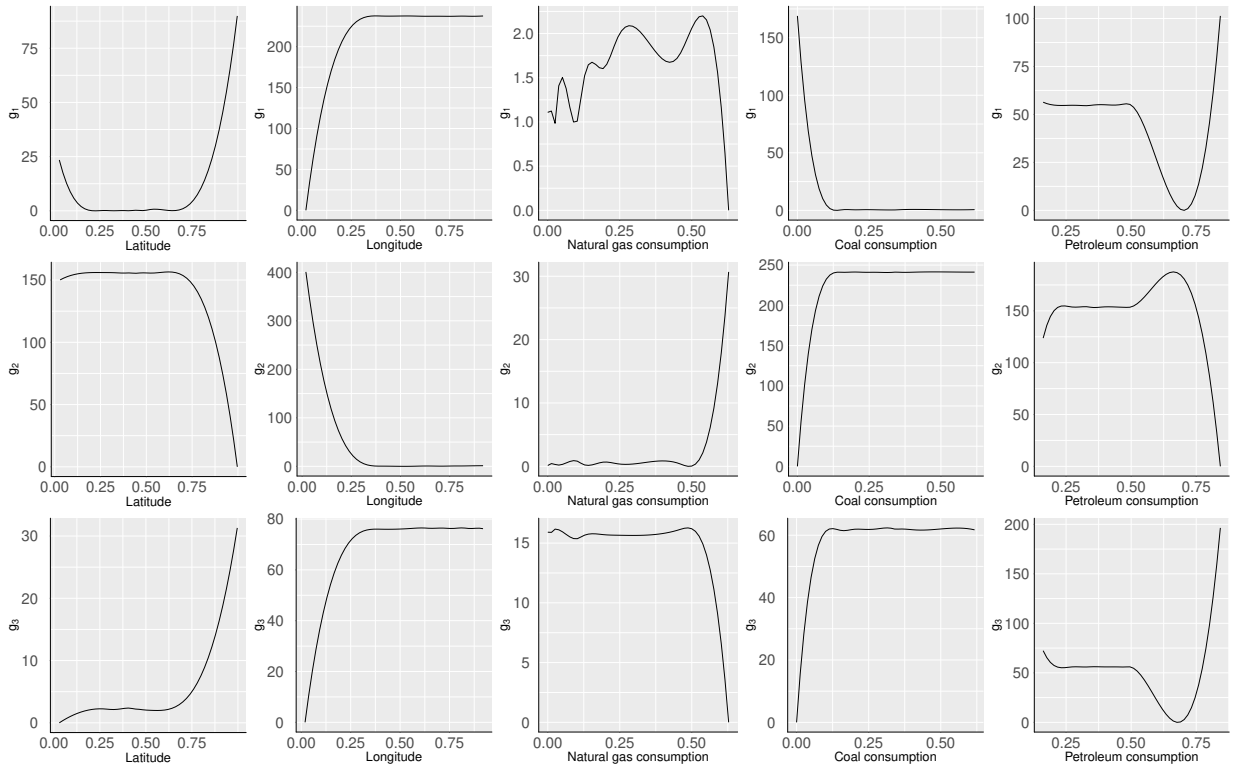

Figure B.26: Recovered nonparametric loading functions $\widehat{g}_k(x)$ for latitude, longitude, energy consumption proportion of natural gas, coal, and petroleum, for each $k = 1, 2, 3$.

Appendix C

Supplemental materials for Chapter 4

This supplementary material contains technical results used for the main paper.

C.1 Technical Results

Proof of Theorem 4.3.1. For simplicity, in the proof of Theorem 4.3.1, we assume $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$ are temporally independent, and $\mathrm{Var}(\boldsymbol{u}_t) = \sigma_u^2 \mathbf{I}_p$.

Step 1. In this step, we construct a series of "candidate cluster assignments" and their corresponding loading matrices and prove that they belong to the subset of cluster assignments and loading matrices in the theorem statement. Denote $\mathcal{S}_m$ as the set of all permutations of $\{1, \ldots, m\}$. Recall that in Section 4.3, we consider the $0-1$ loss function for estimated labels $\widehat{\boldsymbol{z}}$ as

$$L(\widehat{\boldsymbol{z}}, \boldsymbol{z}) = \inf_{\boldsymbol{\Pi} \in \mathcal{S}_m} \left[\frac{1}{p} \sum_{i=1}^{p} \mathcal{I}\{\boldsymbol{\Pi}(\widehat{z}_i) \neq z_i\} \right].$$

In model (4.2.2), we do not distinguish for cluster label switching. Thus, we introduce the permutation $\boldsymbol{\Pi}$ to avoid the error from label switching. Similar setting can be seen in Lu and Zhou (2016) and Gao et al. (2018). Now we impose the assumption that $p > 8$ and $p/(4m^2) > 1$. For each $j = 1, \ldots, m$ and $\boldsymbol{z} = (z_1, \ldots, z_p)$, define $p_j(\boldsymbol{z}) = \sum_{i=1}^{p} \mathcal{I}(i : z_i = j)$. Without loss of generality, let $\boldsymbol{z}^* \in \{1, \ldots, m\}^p$ satisfy that $p_1(\boldsymbol{z}^*) \leqslant p_2(\boldsymbol{z}^*) \leqslant \cdots \leqslant p_m(\boldsymbol{z}^*)$ and $p_1(\boldsymbol{z}^*) = p_2(\boldsymbol{z}^*) = [p/m]$. As suggested by Lu and Zhou (2016) and Gao et al. (2018), for each $j = 1, \ldots, m$, let $\mathcal{T}_j$ be a subset of $\{i : z_i = j\}$ with cardinality $[p_j(\boldsymbol{z}^*) - p/(4m^2)]$, $\mathcal{T} = \cup_{j=1}^{m} \mathcal{S}_j$ and

$$\mathcal{Z}^* = \{\boldsymbol{z} \in \{1, \ldots, m\}^p : z_i = z_i^* \text{ for all } i \in \mathcal{T}\}.$$

Different from the approach of Lu and Zhou (2016) and Gao et al. (2018) that $m/4$ cluster assignments are constructed by filling $\{1, \ldots, p\} \backslash \mathcal{T}$ with the same assignments, we fill it with different

assignments to get a denser covering. That is, we construct a series of candidate cluster assignments by letting $n = p/4$ and

$$\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)}\} = \{\boldsymbol{z} \in \mathcal{Z}^* : z_i = z_i^* \text{ for all } i \in \mathcal{T}\}.$$

Since the number of covering does not depend on m, this construction works when m is finite. Note that, for each $\ell = 1, \ldots, n$ and $j = 1, \ldots, m$, $p_j(\boldsymbol{z}^{(\ell)}) \geqslant [p/m - p/(4m^2)] \asymp p$. Thus, for each $\ell = 1, \ldots, n$, $\boldsymbol{z}^{(\ell)}$ satisfies Condition 4.2.1, so $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)} \in \mathcal{Z}$. Next, we construct a series of candidate loading matrices corresponding to the series of cluster assignments above. Specially, we let

$$\mathbf{C}_0(\boldsymbol{z}) = \begin{bmatrix} D_A\mathbf{A}_1^{(0)} & d_B\mathbf{B}_1^{(0)} & & \\ \vdots & & \ddots & \\ D_A\mathbf{A}_m^{(0)} & & & d_B\mathbf{B}_m^{(0)} \end{bmatrix}, \tag{C.1.1}$$

and

$$(\mathbf{A}_j^{(0)}, \mathbf{B}_j^{(0)}) = \begin{bmatrix} \mathbf{D}_j^{(1)} \\ \mathbf{D}_j^{(2)} \end{bmatrix},$$

where

$$\mathbf{D}_j^{(1)} = \begin{bmatrix} 1 & 1 & \ldots & 1 & 1 \\ -1 & 1 & \ldots & 1 & 1 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & -r_j + 1 & 1 \end{bmatrix}$$

is a $r_j \times (r_0 + r_j)$ matrix and

$$\mathbf{D}_j^{(2)} = \begin{bmatrix} \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}$$

is a $(p_j(\boldsymbol{z}) - r_j) \times (r_0 + r_j)$ block diagonal matrix consisting of $r_0 + r_j$ vectors of $d_0\mathbf{1}$ for $j = 1, \ldots, m$. For simplicity, we assume that $(p_j(\boldsymbol{z}) - r_j)/(r_0 + r_j)$ is an integer for each $j = 1, \ldots, m$ and the vectors $\mathbf{1}$'s in $\mathbf{D}_j^{(2)}$ are of same length $(p_j(\boldsymbol{z}) - r_j)/(r_0 + r_j)$. It is easy to see that for each $\ell = 1, \ldots, n$, $\mathbf{C}_0(\boldsymbol{z}^{(\ell)})$ has the form in (4.2.5) and satisfies Conditions 4.2.6 and 4.2.7, so $\mathbf{C}_0(\boldsymbol{z}^{(1)}), \ldots, \mathbf{C}_0(\boldsymbol{z}^{(n)}) \in \mathcal{C}$. Therefore, $(\boldsymbol{z}^{(1)}, \mathbf{C}_0(\boldsymbol{z}^{(1)})), \ldots, (\boldsymbol{z}^{(n)}, \mathbf{C}_0(\boldsymbol{z}^{(n)}))$ truly belongs to the class in the statement of Theorem 4.3.1:

$$\left(\boldsymbol{z}^{(1)}, \mathbf{C}_0(\boldsymbol{z}^{(1)})\right), \ldots, \left(\boldsymbol{z}^{(n)}, \mathbf{C}_0(\boldsymbol{z}^{(n)})\right) \subset (\mathcal{Z}, \mathcal{C}).$$

Step 2. Next for each $\ell \neq \ell'$, we prove that $(\boldsymbol{z}^{(\ell)}, \mathbf{C}_0(\boldsymbol{z}^{(\ell)}))$ and $(\boldsymbol{z}^{(\ell')}, \mathbf{C}_0(\boldsymbol{z}^{(\ell')}))$ are well-separated and the Kullback-Leibler (K-L) divergence between $(\boldsymbol{z}^{(\ell)}, \mathbf{C}_0(\boldsymbol{z}^{(\ell)}))$ and $(\boldsymbol{z}^{(\ell')}, \mathbf{C}_0(\boldsymbol{z}^{(\ell')}))$ are bounded. By the definition of $\mathcal{T}$, for each $\ell \neq \ell'$, we have

$$L(\boldsymbol{z}^{(\ell)}, \boldsymbol{z}^{(\ell')}) = \frac{1}{4p}. \tag{C.1.2}$$

Next, we consider the Kullback-Leibler divergence between $(\boldsymbol{z}^{(\ell)}, \mathbf{C}_0(\boldsymbol{z}^{(\ell)}))$ and $(\boldsymbol{z}^{(\ell')}, \mathbf{C}_0(\boldsymbol{z}^{(\ell')}))$ for each $\ell \neq \ell'$. Note that the covariance matrix for model (4.2.2) and group assignments $\boldsymbol{z}$ is

$$\boldsymbol{\Sigma}(\boldsymbol{z}) = \mathbf{C}_0(\boldsymbol{z})\mathbf{C}_0(\boldsymbol{z})^\top + \sigma_u^2 \mathbf{I}_p. \tag{C.1.3}$$

By the following fact on the Kullback-Leibler divergence between multivariate Gaussians: denoting $\mathbb{P}_1$ and $\mathbb{P}_2$ as the probability measure corresponding to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_2)$, respectively, if $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are non-degenerating, then

$$\mathbb{KL}\left(\mathbb{P}_1, \mathbb{P}_2\right) = \frac{T}{2}\left\{\mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\right) - p + \log\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)\right\}.$$

By the definition of $\boldsymbol{\Sigma}(\boldsymbol{z})$ in (C.1.3), $\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell)})$ is non-degenerating for each $\ell = 1, \ldots, n$. Thus, the K-L divergence between $\mathbb{P}_{\boldsymbol{z}^{(\ell)}, \mathbf{C}_0(\boldsymbol{z}^{(\ell)})}$ and $\mathbb{P}_{\boldsymbol{z}^{(\ell')}, \mathbf{C}_0(\boldsymbol{z}^{(\ell')})}$ is

$$
\begin{aligned}
&\mathbb{KL}\left(\mathbb{P}_{\boldsymbol{z}^{(\ell)}, \mathbf{C}_0(\boldsymbol{z}^{(\ell)})}, \mathbb{P}_{\boldsymbol{z}^{(\ell')}, \mathbf{C}_0(\boldsymbol{z}^{(\ell')})}\right) \\
=&\frac{T}{2}\left\{\operatorname{tr}\left(\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell)})^{-1}\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell')})\right) - p + \log\left(\frac{|\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell)})|}{|\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell')})|}\right)\right\}.
\end{aligned}
$$

By Lemmas C.2.17 and C.2.18, we have

$$
\begin{aligned}
&\mathbb{KL}\left(\mathbb{P}_{\boldsymbol{z}^{(\ell)}, \mathbf{C}_0(\boldsymbol{z}^{(\ell)})}, \mathbb{P}_{\boldsymbol{z}^{(\ell')}, \mathbf{C}_0(\boldsymbol{z}^{(\ell')})}\right) \\
=&\frac{r_0 T}{2}\log\left(\frac{D_A^2\sum_{k=1}^m \frac{p_k(\boldsymbol{z}^{(\ell)})}{r_0+r_k} + \sigma_u^2}{D_A^2\sum_{k=1}^m \frac{p_k(\boldsymbol{z}^{(\ell')})}{r_0+r_k} + \sigma_u^2}\right) + \frac{T}{2}\sum_{k=1}^m r_k \log\left(\frac{\frac{d_B^2 p_k(\boldsymbol{z}^{(\ell)})}{r_0+r_k} + \sigma_u^2}{\frac{d_B^2 p_k(\boldsymbol{z}^{(\ell')})}{r_0+r_k} + \sigma_u^2}\right) \\
=&\frac{r_0 T}{2}\log\left(1 + \frac{D_A^2\left(\frac{1}{r_0+r_{k_1}} - \frac{1}{r_0+r_{k_2}}\right)}{D_A^2\sum_{k=1}^m \frac{p_k(\boldsymbol{z}^{(\ell)})}{r_0+r_k} + \sigma_u^2}\right) + \frac{r_{k_1} T}{2}\log\left(1 + \frac{\frac{d_B^2}{r_0+r_{k_1}}}{\frac{d_B^2 p_{k_1}(\boldsymbol{z}^{(\ell')})}{r_0+r_{k_1}} + \sigma_u^2}\right) \\
&+ \frac{r_{k_2} T}{2}\log\left(1 - \frac{\frac{d_B^2}{r_0+r_{k_2}}}{\frac{d_B^2 p_{k_2}(\boldsymbol{z}^{(\ell')})}{r_0+r_{k_2}} + \sigma_u^2}\right),
\end{aligned}
$$

for some $k_1 \neq k_2$. By the definition of $\mathcal{Z}^*$, $p_j(\boldsymbol{z}) \asymp p/m$ for each $j = 1, \ldots, m$ and any $\boldsymbol{z} \in \mathcal{Z}^*$. Thus, by the well known fact that $\log(1+x) \leqslant x$,

$$
\begin{aligned}
&\mathbb{KL}\left(\mathbb{P}_{\boldsymbol{z}^{(j)}, \mathbf{C}_0(\boldsymbol{z}^{(j)})}, \mathbb{P}_{\boldsymbol{z}^{(\ell)}, \mathbf{C}_0(\boldsymbol{z}^{(\ell)})}\right) \qquad \text{(C.1.4)} \\
\leqslant&\frac{T}{2}\left\{\frac{D_A^2 r_0\left(\frac{1}{r_0+r_{k_1}} - \frac{1}{r_0+r_{k_2}}\right)}{D_A^2\sum_{k=1}^m \frac{p_k(\boldsymbol{z}^{(j)})}{r_0+r_k} + \sigma_u^2} + \frac{\frac{d_B^2 r_{k_1}}{r_0+r_{k_1}}}{\frac{d_B^2 p_{k_1}(\boldsymbol{z}^{(\ell)})}{r_0+r_{k_1}} + \sigma_u^2} - \frac{\frac{d_B^2 r_{k_2}}{r_0+r_{k_2}}}{\frac{d_B^2 p_{k_2}(\boldsymbol{z}^{(\ell)})}{r_0+r_{k_2}} + \sigma_u^2}\right\} \\
\leqslant&\frac{d_B^2 r_0 \max_j r_j m T}{2(D_A^2 r_0 + d_B^2 \min_j r_j)p}.
\end{aligned}
$$

Step 3. We finalize the proof by the generalized FanoâĂŹs lemma. Specifically by (C.1.2), (C.1.4) and Lemma in (Yu, 1997), we have

$$\inf_{\hat{\boldsymbol{z}}} \sup_{\boldsymbol{z}\in\mathcal{Z},\mathbf{C}\in\mathcal{C}} \mathbb{E}\{L(\hat{\boldsymbol{z}},\boldsymbol{z})\} \geqslant \frac{1}{8p}\left\{1 - \frac{\frac{d_B^2 r_0 \max_j r_j mT}{2(D_A^2 r_0 + d_B^2 \min_j r_j)p} + \log(2)}{\log(p/4)}\right\}.$$

Letting $\theta = (D_A^2 r_0 + d_B^2 \min_j r_j)^{-1} d_B^2 r_0 \max_j r_j$ and

$$p = \frac{\theta mT}{2\{\varepsilon \log(p/4) - \log(2)\}} \vee 8$$

for some $\varepsilon \in (\log(2)/\log(p/4), 1)$, we have

$$\inf_{\hat{\boldsymbol{z}}} \sup_{\boldsymbol{z}\in\mathcal{Z},\mathbf{C}\in\mathcal{C}} \mathbb{E}\{L(\hat{\boldsymbol{z}},\boldsymbol{z})\} \geqslant \frac{(D_A^2 r_0 + d_B^2 \min_j r_j)\{\varepsilon\log(p/4) - \log(2)\}(1-\varepsilon)m}{16 d_B^2 r_0 \max_j r_j pT} \wedge \frac{1}{64}.$$

□

Proof of Corollary 4.3.1. **Step 1.** In this step, we construct a series of "candidate cluster assignments" and their corresponding loading matrices and prove that they belong to the subset of cluster assignments and loading matrices in the theorem statement. Denote $\mathcal{S}_m$ as the set of all permutations of $\{1,\ldots,m\}$. Recall that in Section 4.3, we consider the $0-1$ loss function for estimated labels $\hat{\boldsymbol{z}}$ as

$$L(\hat{\boldsymbol{z}},\boldsymbol{z}) = \inf_{\boldsymbol{\Pi}\in\mathcal{S}_m}\left[\frac{1}{p}\sum_{i=1}^{p}\mathcal{I}\{\boldsymbol{\Pi}(\hat{z}_i)\neq z_i\}\right].$$

Similar as the proof of Theorem 4.3.1, for each $j = 1,\ldots,m$, we let $\mathcal{T}_j$ be a subset of $\{i : z_i = j\}$ with cardinality $[p_j(\boldsymbol{z}^*) - p/(4m^2)]$, $\mathcal{T} = \cup_{j=1}^m \mathcal{S}_j$ and

$$\mathcal{Z}^* = \{\boldsymbol{z}\in\{1,\ldots,m\}^p : z_i = z_i^* \text{ for all } i\in\mathcal{T}\}.$$

Then, we construct a series of candidate cluster assignments by letting $n = p/4$ and

$$\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)}\} = \{\boldsymbol{z} \in \mathcal{Z}^* : z_i = z_i^* \text{ for all } i \in \mathcal{T}\}.$$

Note that, for each $\ell = 1, \ldots, n$ and $j = 1, \ldots, m$, $p_j(\boldsymbol{z}^{(\ell)}) \geqslant [p/m - p/(4m^2)] \asymp p$. Thus, for each $\ell = 1, \ldots, n$, $\boldsymbol{z}^{(\ell)}$ satisfies Condition 4.2.1, so $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)} \in \mathcal{Z}$. Then, the membership matrix relative to $\boldsymbol{z}$ is

$$\boldsymbol{\Gamma}(\boldsymbol{z}) = \begin{bmatrix} \mathbf{1}_{p_1(\boldsymbol{z})} & & \\ & \ddots & \\ & & \mathbf{1}_{p_m(\boldsymbol{z})} \end{bmatrix}.$$

Next, we construct a series of candidate loading matrices corresponding to the series of cluster assignments above. Specially, we let $\boldsymbol{\Psi} = \boldsymbol{\sigma}_u^2 \mathbf{I}$ and

$$\mathbf{V}_0 = \begin{bmatrix} b & a & \ldots & a \\ a & b & \ldots & a \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \ldots & b \end{bmatrix}$$

where the diagonal elements are all b, off-diagonal elements are all a and $0 < a < b$. It is easy to see that $\mathbf{V}_0$ is a symmetric matrix. Therefore, $(\boldsymbol{z}^{(1)}, \mathbf{V}_0), \ldots, (\boldsymbol{z}^{(n)}, \mathbf{V}_0)$ truly belongs to the class in the statement of Theorem 4.3.1:

$$\big(\boldsymbol{z}^{(1)}, \mathbf{V}_0), \ldots, (\boldsymbol{z}^{(n)}, \mathbf{V}_0)\big) \subset (\mathcal{Z}, \mathcal{V}).$$

Step 2. Next for each $\ell \neq \ell'$, we prove that $(\boldsymbol{z}^{(\ell)}, \mathbf{V}_0)$ and $(\boldsymbol{z}^{(\ell')}, \mathbf{V}_0)$ are well-separated and the Kullback-Leibler (K-L) divergence between $(\boldsymbol{z}^{(\ell)}, \mathbf{V}_0)$ and $(\boldsymbol{z}^{(\ell')}, \mathbf{V}_0)$ are bounded. By the

definition of $\mathcal{T}$, for each $\ell \neq \ell'$, we have

$$L(\boldsymbol{z}^{(\ell)}, \boldsymbol{z}^{(\ell')}) = \frac{1}{4p}. \tag{C.1.5}$$

Next, we consider the Kullback-Leibler divergence between $(\boldsymbol{z}^{(\ell)}, \mathbf{V}_0)$ and $(\boldsymbol{z}^{(\ell')}, \mathbf{V}_0)$ for each $\ell \neq \ell'$. Note that the covariance matrix for model (4.2.10) and group assignments $\boldsymbol{z}$ is

$$\boldsymbol{\Sigma}_G(\boldsymbol{z}) = \boldsymbol{\Gamma}(\boldsymbol{z})\mathbf{V}_0\boldsymbol{\Gamma}(\boldsymbol{z})^\top + \mathbf{I}_p. \tag{C.1.6}$$

By the following fact on the Kullback-Leibler divergence between multivariate Gaussians: denoting $\mathbb{P}_1$ and $\mathbb{P}_2$ as the probability measure corresponding to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_2)$, respectively, if $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are non-degenerating, then

$$\mathbb{KL}\left(\mathbb{P}_1, \mathbb{P}_2\right) = \frac{T}{2}\left\{\mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\right) - p + \log\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)\right\}.$$

By the definition of $\boldsymbol{\Sigma}_G(\boldsymbol{z})$ in (C.1.6), $\boldsymbol{\Sigma}_G(\boldsymbol{z}^{(\ell)})$ is non-degenerating for each $\ell = 1, \ldots, n$. Thus, the K-L divergence between $\mathbb{P}_{\boldsymbol{z}^{(\ell)}, \mathbf{V}_0)}$ and $\mathbb{P}_{\boldsymbol{z}^{(\ell')}, \mathbf{V}_0)}$ is

$$\begin{aligned}
&\mathbb{KL}\left(\mathbb{P}_{\boldsymbol{z}^{(\ell)}, \mathbf{V}_0)}, \mathbb{P}_{\boldsymbol{z}^{(\ell')}, \mathbf{V}_0)}\right) \\
=&\frac{T}{2}\left\{\mathrm{tr}\left(\boldsymbol{\Sigma}_G(\boldsymbol{z}^{(\ell)})^{-1}\boldsymbol{\Sigma}_G(\boldsymbol{z}^{(\ell')})\right) - p + \log\left(\frac{|\boldsymbol{\Sigma}_G(\boldsymbol{z}^{(\ell)})|}{|\boldsymbol{\Sigma}_G(\boldsymbol{z}^{(\ell')})|}\right)\right\}.
\end{aligned}$$

With similar arguments as Lemmas C.2.17 and C.2.18, we have $\mathrm{tr}\left(\boldsymbol{\Sigma}_G(\boldsymbol{z}^{(\ell)})^{-1}\boldsymbol{\Sigma}_G(\boldsymbol{z}^{(\ell')})\right) = p$ and

$$\log\left(\frac{|\boldsymbol{\Sigma}_G(\boldsymbol{z}^{(\ell)})|}{|\boldsymbol{\Sigma}_G(\boldsymbol{z}^{(\ell')})|}\right) = \sum_{j=1}^{m}\log\left(\frac{(b-a)p_j(\boldsymbol{z}^{(\ell)}) + \sigma_u^2}{(b-a)p_j(\boldsymbol{z}^{(\ell')}) + \sigma_u^2}\right)$$

By the definition of $\mathcal{Z}^*$, $p_j(\boldsymbol{z}) \asymp p/m$ for each $j = 1, \ldots, m$ and any $\boldsymbol{z} \in \mathcal{Z}^*$. Thus, by the well known fact that $\log(1+x) \leqslant x$, for some $k_1 \neq k_2$,

$$\mathbb{KL}\left(\mathbb{P}_{\boldsymbol{z}^{(j)}, \mathbf{V}_0}, \mathbb{P}_{\boldsymbol{z}^{(\ell)}, \mathbf{V}_0}\right) \tag{C.1.7}$$

$$\leqslant \frac{T}{2}\left\{\frac{1}{(b-a)p_{k_1}(\boldsymbol{z}^{(\ell)})+\sigma_u^2}-\frac{1}{(b-a)p_{k_2}(\boldsymbol{z}^{(\ell)})+\sigma_u^2}\right\}$$
$$\leqslant \frac{mT}{2(b-a)p}.$$

Step 3. We finalize the proof by the generalized FanoâĂŹs lemma. Specifically by (C.1.5), (C.1.7) and Lemma in (Yu, 1997), we have

$$\inf_{\hat{\boldsymbol{z}}}\sup_{\boldsymbol{z}\in\mathcal{Z},\mathbf{C}\in\mathcal{C}}\mathbb{E}\{L(\hat{\boldsymbol{z}},\boldsymbol{z})\}\geqslant\frac{1}{8p}\left\{1-\frac{\frac{mT}{2(b-a)p}+\log(2)}{\log(p/4)}\right\}.$$

Letting

$$p=\frac{(b-a)mT}{2\{\varepsilon\log(p/4)-\log(2)\}}\vee 8$$

for some $\varepsilon\in(\log(2)/\log(p/4),1)$, we have

$$\inf_{\hat{\boldsymbol{z}}}\sup_{\boldsymbol{z}\in\mathcal{Z},\mathbf{C}\in\mathcal{C}}\mathbb{E}\{L(\hat{\boldsymbol{z}},\boldsymbol{z})\}\geqslant\frac{(b-a)\{\varepsilon\log(p/4)-\log(2)\}(1-\varepsilon)m}{16pT}\wedge\frac{1}{64}.$$

□

Proof of Theorem 4.4.1. The conclusion follows from Lemmas C.2.5, C.2.8 and C.2.9. □

Proof of Theorem 4.4.2. The conclusion follows by applying similar discussion as Theorem 4.4.1 to the first r_0+r_j largest eigenvalues of $T^{-1}\mathbf{Y}_j^\top\mathbf{Y}_j$ and that

$$\frac{1}{T}\|\widehat{\mathbf{F}}_j-\mathbf{F}_j\|_{\mathbb{F}}^2\leqslant\frac{1}{T}\|(\widehat{\mathbf{F}}_0,\widehat{\mathbf{F}}_j)-(\mathbf{F}_0,\mathbf{F}_j)\|_{\mathbb{F}}^2\lesssim\frac{m}{p}s^3,$$
$$\frac{m}{p}\|\widehat{\mathbf{B}}_j-\mathbf{B}_j\|_{\mathbb{F}}^2\leqslant\frac{1}{p}\|(\widehat{\mathbf{A}}_j,\widehat{\mathbf{B}}_j)-(\mathbf{A}_j,\mathbf{B}_j)\|_{\mathbb{F}}^2\lesssim\left(\frac{m}{p}+\frac{1}{T}\right)s^4,$$

for each $j=1,\ldots,m$. □

Proof of Theorem 4.4.3. The conclusion follows from Lemmas C.2.11, C.2.14 and C.2.15. □

Proof of Corollary 4.4.1. (i) Recall that $\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2 = \left(\sum_{i=1}^3 \mathbf{N}_i\right)\mathbf{K}_2^{-1}$. By Lemma C.2.10, $\|\mathbf{K}_2^{-1}\|_2 \lesssim d_B^{-1}(\min_j r_j)^{-1/2}(1 + D_A\sqrt{r_0}T^{-1/2}\sqrt{s})$ with probability at least $1-7e^{-s}$. Also, by Lemma C.2.16, $\sum_{i=1}^3 \|\mathbf{N}_i\|_{\max} \lesssim D_A\sqrt{r_0}p^{-1/2}\{\log(T)\}^{2/r_2}s$. Thus, with probability at least $1 - 10e^{-s}$,

$$\begin{aligned}\|\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2\|_{\max} &= \left(\sum_{i=1}^3 \|\mathbf{N}_i\|_{\max}\right)\|\mathbf{K}_2^{-1}\|_2 \\ &\lesssim D_A\sqrt{r_0}d_B^{-1}(\min_j r_j)^{-1/2}p^{-1/2}\{\log(T)\}^{2/r_2}s.\end{aligned}$$

(ii) Note that $\widehat{\mathbf{C}} - \mathbf{C}\mathbf{H}_2^{-1} = T^{-1}\mathbf{C}\mathbf{H}_2^{-1}(\mathbf{H}_2\mathbf{F}^\top - \widehat{\mathbf{F}}^\top)\widehat{\mathbf{F}} + T^{-1}\mathbf{U}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2) + T^{-1}\mathbf{U}\mathbf{F}\mathbf{H}_2$. By Lemma B.1 in Fan et al. (2011), with probability at least $1 - e^{-s}$, $\|\mathbf{U}\mathbf{F}\|_{\max} \lesssim \sqrt{T\log(p)}s$. Thus, by Lemma C.2.14, with probability at least $1 - 10e^{-s}$,

$$\begin{aligned}&\|\widehat{\mathbf{C}} - \mathbf{C}\mathbf{H}_2^{-1}\|_{\max} \\ =&\frac{1}{T}\|\mathbf{C}\|_{\max}\|\mathbf{H}_2^{-1}\|_2\|\mathbf{H}_2\mathbf{F}^\top - \widehat{\mathbf{F}}^\top\|_{\max}\|\widehat{\mathbf{F}}\|_{\mathbb{F}} + \frac{1}{T}\|\mathbf{U}\|_2\|\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2\|_{\max} + \frac{1}{T}\|\mathbf{U}\mathbf{F}\|_{\max}\|\mathbf{H}_2\|_2 \\ \lesssim&\frac{D_A^2 r_0\{\log(T)\}^{2/r_2}s}{d_B^2\min_j r_j\sqrt{pT}} + \frac{D_A\sqrt{r_0}\{\log(T)\}^{2/r_2}s}{d_B\sqrt{\min_j r_j}T} + \frac{D_A\sqrt{r_0}\sqrt{\log(p)}s}{d_B\sqrt{\min_j r_j}\sqrt{T}} \\ \lesssim&\frac{D_A\sqrt{r_0}\sqrt{\log(p)}s}{d_B\sqrt{\min_j r_j}\sqrt{T}}.\end{aligned}$$

(iii) The conclusion follows from the result above and Lemma C.2.14.

□

Proof of Theorem 4.4.4. In the proof of Theorem 4.4.4, we assume that the number of common factors r_0 and number of all factors K are correctly estimated. Then, we consider a block diagonal matrix $\mathbf{B} = \mathrm{diag}(\mathbf{B}_1, \ldots, \mathbf{B}_m)$, and its estimate $\widehat{\mathbf{B}} = \mathbf{B} + \mathbf{E}$. Denote $\mathbf{B} = \{b_{ik}\}_{i=1,k=1}^{p,K-r_0}$, $\widehat{\mathbf{B}} = \{\widehat{b}_{ik}\}_{i=1,k=1}^{p,K-r_0}$ and $\mathbf{E} = \{e_{ik}\}_{i=1,k=1}^{p,K-r_0}$. By Corollary 4.4.1, with probability at least $1 - 10e^{-s}$,

$$\max_{i,k}|e_{ik}| \lesssim \frac{D_A\sqrt{r_0}}{d_B\sqrt{\min_j r_j}}\left\{\sqrt{\frac{\log(p)}{T}} + \frac{1}{\sqrt{p}}\right\}s.$$

As Algorithm 4 suggests, we let $\tau = \delta \log(pT/m)(T^{-1/2}\sqrt{\log(p)} + p^{-1/2})$. Then, for each $i = 1, \ldots, p$ and $k = 1, \ldots, K - r_0$, we have

$$\mathbb{P}(|\widehat{b}_{ik}| > \tau, b_{ik} = 0) = \mathbb{P}(|e_{ik}| > \tau) \leqslant 10\{m^{-1}\log^{-1}(p)pT\}^{-\frac{\delta d_B \sqrt{\min_j r_j}}{C D_A \sqrt{r_0}}},$$

for some positive constant C. Similarly, for each $i = 1, \ldots, p$ and $k = 1, \ldots, K - r_0$, we have

$$\begin{aligned}
&\mathbb{P}(|\widehat{b}_{ik}| \leqslant \tau, b_{ik} \neq 0\}) = \mathbb{P}(|e_{ik}| \geqslant |b_{ik}| - \tau) \\
\leqslant &10 \exp\left\{-\frac{|b_{ik}|}{\delta \log\{m^{-1}\log^{-1}(p)pT\}(T^{-1/2}\sqrt{\log(p)} + p^{-1/2})}\right\} \{m^{-1}\log^{-1}(p)pT\}^{-\frac{C D_A \sqrt{r_0}}{\delta d_B \sqrt{\min_j r_j}}}.
\end{aligned}$$

For each $k = 1, \ldots, K - r_0$, let $\boldsymbol{i}_k = (\mathcal{I}(b_{1k} \neq 0), \ldots, \mathcal{I}(b_{pk} \neq 0))^\top$, $\widehat{\boldsymbol{i}}_k = (\mathcal{I}(|\widehat{b}_{1k}| > \tau), \ldots, \mathcal{I}(|\widehat{b}_{pk}| > \tau)^\top$, $\widehat{\mathbf{I}} = (\widehat{\boldsymbol{i}}_1, \ldots, \widehat{\boldsymbol{i}}_{K-r_0})$ and $\mathbf{I} = (\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{K-r_0})$. Then, by letting $\delta = (d_B\sqrt{\min_j r_j})^{-1} C D_A \sqrt{r_0}$, we have

$$\begin{aligned}
\mathbb{P}(\widehat{\mathbf{I}} \neq \mathbf{I}\}) =&\mathbb{P}\left\{\bigcup_{i,k=1}^{p,K-r_0} (|\widehat{b}_{ik}| > \tau, b_{ik} = 0) \cup (|\widehat{b}_{ik}| \leqslant \tau, b_{ik} \neq 0\})\right\} \\
\leqslant &10 \exp\left\{-\frac{\min_{b_{ik}\neq 0}|b_{ik}|}{\delta \log\{m^{-1}\log^{-1}(p)pT\}(T^{-1/2}\sqrt{\log(p)} + p^{-1/2})}\right\} \{m^{-1}\log^{-1}(p)pT\}^{-1}.
\end{aligned}$$

Note that by the definition of $\widetilde{\mathbf{C}}$ in (4.2.5) , there exist a $p \times p$ permutation matrix $\mathbf{\Pi}$ satisfying that $\mathbf{I} = \mathbf{\Pi}\mathbf{I}_{\widetilde{\mathbf{C}}}$, where $\mathbf{I}_{\widetilde{\mathbf{C}}} = \{\mathcal{I}(\widetilde{c}_{ik} = 0)\}_{i,k=1}^{p,K}$ and $\widetilde{c}_{ik}$ is element of $\widetilde{\mathbf{C}}$. By the definition of $\widehat{\boldsymbol{z}}(\mathbf{I})$ in Algorithm 4, row switching of $\mathbf{I}$ will not affect $\widehat{\boldsymbol{z}}(\mathbf{I})$, that is, $\widehat{\boldsymbol{z}}(\mathbf{I}) = \widehat{\boldsymbol{z}}(\mathbf{\Pi}\mathbf{I}_{\widetilde{\mathbf{C}}}) = \widehat{\boldsymbol{z}}(\mathbf{I}_{\widetilde{\mathbf{C}}})$. Under Condition 4.2.6, for each $j = 1, \ldots, m$, there exists $i_j \in \{i : z_i = j\}$ satisfying that $\sum_{\ell=1}^{r_j} \mathcal{I}(b_{i\ell}^{(j)} \neq 0) = r_j$. Thus, for each $i' \in \{i : z_i = j\}$, there exists some k satisfying that $\mathbf{i}_{i_j}(k) = \mathbf{i}_{i'}(k) = 1$ so $\widehat{z}_{i_j}(\mathbf{I}_{\widetilde{\mathbf{C}}}) = \widehat{z}_{i'}(\mathbf{I}_{\widetilde{\mathbf{C}}})$. Similarly, for each $i' \notin \{i : z_i = j\}$, there exist j' and $i_{j'}$ satisfying that $i' \in \{i : z_i = j'\}$ and $\widehat{z}_{i_{j'}}(\mathbf{I}_{\widetilde{\mathbf{C}}}) = \widehat{z}_{i'}(\mathbf{I}_{\widetilde{\mathbf{C}}})$, so $\widehat{z}_{i_j}(\mathbf{I}_{\widetilde{\mathbf{C}}}) \neq \widehat{z}_{i'}$. Hence,

$$L(\widehat{\boldsymbol{z}}(\mathbf{I}), \boldsymbol{z}) = L(\widehat{\boldsymbol{z}}(\mathbf{I}_{\widetilde{\mathbf{C}}}), \boldsymbol{z}) = 0.$$

Thus,

$$\begin{aligned}\mathbb{E}\{L(\widehat{\boldsymbol{z}}(\widehat{\mathbf{I}}),\boldsymbol{z})\} =&\mathbb{E}\{L(\widehat{\boldsymbol{z}}(\widehat{\mathbf{I}}),\boldsymbol{z})|\widehat{\mathbf{I}}=\mathbf{I}\}\mathbb{P}(\widehat{\mathbf{I}}=\mathbf{I})+\mathbb{E}\{L(\widehat{\boldsymbol{z}}(\widehat{\mathbf{I}}),\boldsymbol{z})|\widehat{\mathbf{I}}\neq\mathbf{I}\}\mathbb{P}(\widehat{\mathbf{I}}\neq\mathbf{I})\\ \leqslant&10\exp\left\{-\frac{\min_{b_{ik}\neq0}|b_{ik}|}{\delta\log\{m^{-1}\log^{-1}(p)pT\}(T^{-1/2}\sqrt{\log(p)}+p^{-1/2})}\right\}\{m^{-1}\log^{-1}(p)pT\}^{-1}\\ \leqslant&\frac{10\exp\{-(CD_A\sqrt{r_0})^{-1}d_B\sqrt{\min_j r_j}\min_{b_{ik}\neq0}|b_{ik}|\}m\log(p)}{pT}.\end{aligned}$$

The conclusion follows.

□

Proof of Corollary 4.4.2. (i) By Corollary 4.4.1, with probability at least $1-10e^{-s}$,

$$\begin{aligned}\max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{ij}-\boldsymbol{\Sigma}_{ij}| =&\max_{i,j}\left|\sum_{k=1}^{K}\widehat{c}_{ik}\widehat{c}_{kj}-\sum_{k=1}^{K}c_{ik}c_{kj}\right|\leqslant 2K\max_{i,k}|c_{ik}|\max_{i,k}|\widehat{c}_{ik}-c_{ik}|\\ \leqslant&\frac{2K\max_{ik}|c_{ik}|D_A\sqrt{r_0}}{d_B\sqrt{\min_j r_j}}\left\{\sqrt{\frac{\log(p)}{T}}+\frac{1}{\sqrt{p}}\right\}s.\end{aligned}$$

Then, the conclusion follows from similar discussion in the proof of Theorem 4.4.4 and that $K=m$ under the conditions of Corollary 4.4.2.

(ii) Without loss of generality, we assume that $z_1\leqslant\cdots\leqslant z_p$ and the loadings for the variables in first cluster is monotone increasing. Then,

$$\begin{aligned}\mathrm{sCOD}(1,2) =&\max_{\ell\neq1,2}\left|\frac{\boldsymbol{\Sigma}_{1\ell}-\widehat{\boldsymbol{\Sigma}}_{2\ell}}{\sqrt{(\boldsymbol{\Sigma}_{11}+\boldsymbol{\Sigma}_{22}-2\boldsymbol{\Sigma}_{12})\boldsymbol{\Sigma}_{\ell\ell}}}\right|\\ \geqslant&\left|\frac{\boldsymbol{\Sigma}_{13}-\boldsymbol{\Sigma}_{23}}{\sqrt{(\boldsymbol{\Sigma}_{11}+\boldsymbol{\Sigma}_{22}-2\boldsymbol{\Sigma}_{12})\boldsymbol{\Sigma}_{33}}}\right|\\ =&1\neq0.\end{aligned}$$

Thus, by Algorithm 5, variable 1 is put into a cluster with a single variable itself. Similar result holds for other variables in the first cluster. Thus, estimated clustering assignments for

variables in the first cluster are all wrong, so

$$L(\hat{z}, z) \geqslant \frac{p_1}{p}.$$

The conclusion follows from Condition 4.2.1 that $p_1 \asymp p$.

□

Proof of Theorem 4.4.5. Recall that

$$\frac{1}{T}\mathbf{Y}\mathbf{Y}^\top = \mathbf{C}\mathbf{C}^\top + \frac{1}{T}\mathbf{C}\mathbf{F}^\top\mathbf{U}^\top + \frac{1}{T}\mathbf{U}\mathbf{F}\mathbf{C}^\top + \frac{1}{T}\mathbf{U}\mathbf{U}^\top$$

and

$$\frac{1}{T}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top) = \mathbf{C}\mathbf{C}^\top + \frac{1}{T}\mathbb{E}(\mathbf{U}\mathbf{U}^\top).$$

Also, note that $\mathbf{C}\mathbf{C}^\top$, $T^{-1}\mathbf{C}\mathbf{F}^\top\mathbf{U}^\top$ and $T^{-1}\mathbf{U}\mathbf{F}\mathbf{C}^\top$ have rank K, so for each $k = K+1, \ldots, \min(p, T)$, $\widehat{\lambda}_k \asymp \lambda_k(T^{-1}\mathbf{U}\mathbf{U}^\top)$ and $\lambda_k \asymp \lambda_k(\mathbb{E}(T^{-1}\mathbf{U}\mathbf{U}^\top))$.

(i) If $p < T$, by Condition 4.2.4 and Theorem 5.58 in Vershynin (2010), with probability at least $1 - e^{-2s}$,

$$|\lambda_k(T^{-1}\mathbf{U}\mathbf{U}^\top) - \lambda_k(\mathbb{E}(T^{-1}\mathbf{U}\mathbf{U}^\top))| \leqslant \max(\zeta, \zeta^2),$$

for each $k = 1, \ldots, p$, where $\zeta = \sqrt{C}T^{-1/2}\sqrt{p} + \sqrt{c}T^{-1/2}\sqrt{s}$ and C and c are positive constants only depending on $\boldsymbol{u}_t$. Thus, with probability at least $1 - e^{-2s}$,

$$|\lambda_k(T^{-1}\mathbf{U}\mathbf{U}^\top) - \lambda_k(\mathbb{E}(T^{-1}\mathbf{U}\mathbf{U}^\top))| \leqslant C\sqrt{\frac{p}{T}} + \frac{c}{\sqrt{T}}\sqrt{s},$$

(ii) If $p > T$, note that the first T largest eigenvalues of $T^{-1}\mathbf{U}\mathbf{U}^\top$ are the same as those of $T^{-1}\mathbf{U}^\top\mathbf{U}$. By Condition 4.2.4 and Theorem 5.38 in Vershynin (2010), for each $k = 1, \ldots, T$,

with probability at least $1 - e^{-2s}$,

$$\sqrt{\frac{p}{T}} - C - \frac{c}{\sqrt{T}}\sqrt{s} \lesssim \lambda_k(T^{-1}\mathbf{U}^\top\mathbf{U}) \lesssim \sqrt{\frac{p}{T}} + C + \frac{c}{\sqrt{T}}\sqrt{s}.$$

By Condition 4.2.7, for each $k = 1, \ldots, r_0$, $\lambda_k(\mathbf{C}\mathbf{C}^\top) = O(p)$ and for each $k = r_0 + 1, \ldots, K$, $\lambda_k(\mathbf{C}\mathbf{C}^\top) = O(p^{1-\gamma})$. Also, by Condition 4.2.4, $p^{-1}\lambda_k(T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top))$ are bounded and by the discussion above, $p^{-1}\lambda_k(T^{-1}\mathbf{U}\mathbf{U}^\top)$ is bounded. Thus, for each $k = 1, \ldots, r_0$, with probability at least $1 - e^{-2s}$,

$$\begin{aligned}\frac{\widehat{\lambda}_k}{\lambda_k} &\leqslant \frac{\lambda_k(\mathbf{C}\mathbf{C}^\top) + \lambda_{\max}(T^{-1}\mathbf{C}\mathbf{F}^\top\mathbf{U}^\top) + \lambda_{\max}(T^{-1}\mathbf{U}\mathbf{F}\mathbf{C}^\top) + \lambda_{\max}(T^{-1}\mathbf{U}\mathbf{U}^\top)}{\lambda_k(\mathbf{C}\mathbf{C}^\top) + \lambda_{\min}(T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top))} \\ &\leqslant 1 + \frac{C}{\sqrt{T}} + \frac{c}{\sqrt{pT}}\sqrt{s}\end{aligned}$$

and

$$\begin{aligned}\frac{\widehat{\lambda}_k}{\lambda_k} &\geqslant \frac{\lambda_k(\mathbf{C}\mathbf{C}^\top) + \lambda_{\min}(T^{-1}\mathbf{C}\mathbf{F}^\top\mathbf{U}^\top) + \lambda_{\min}(T^{-1}\mathbf{U}\mathbf{F}\mathbf{C}^\top) + \lambda_{\min}(T^{-1}\mathbf{U}\mathbf{U}^\top)}{\lambda_k(\mathbf{C}\mathbf{C}^\top) + \lambda_{\max}(\mathbb{E}(\mathbf{U}\mathbf{U}^\top))} \\ &\geqslant 1 - \frac{C}{\sqrt{T}} - \frac{c}{\sqrt{pT}}\sqrt{s}.\end{aligned}$$

Thus, if $p < T$, for $s < c^{-2}(\sqrt{T} - C\sqrt{p})^2$, with probability at least $1 - e^{-s}$,

$$\begin{aligned}\frac{\widehat{\lambda}_k}{\widehat{\lambda}_{k+1}} &\leqslant \frac{\lambda_k}{\lambda_{k+1}}\left\{1 + \frac{C}{\sqrt{T}} + \frac{c}{\sqrt{pT}}\sqrt{s}\right\}^2, \qquad k = 1, \ldots, K-1, \\ \frac{\widehat{\lambda}_K}{\widehat{\lambda}_{K+1}} &\geqslant \frac{\lambda_K}{\lambda_{K+1}}\left\{1 - \frac{C\sqrt{p}}{\sqrt{T}} - \frac{c}{\sqrt{T}}\sqrt{s}\right\}\left\{1 - \frac{C}{\sqrt{T}} - \frac{c}{\sqrt{pT}}\sqrt{s}\right\}, \\ \frac{\widehat{\lambda}_k}{\widehat{\lambda}_{k+1}} &\leqslant \frac{\lambda_k}{\lambda_{k+1}}\left\{1 + \frac{C\sqrt{p}}{\sqrt{T}} + \frac{c}{\sqrt{T}}\sqrt{s}\right\}^2, \qquad k = K+1, \ldots, L,\end{aligned}$$

Thus, for $s < c^{-1}T + c^{-1}Cp$, with probability at least $1 - e^{-s}$,

$$\frac{\widehat{\lambda}_K/\widehat{\lambda}_{K+1}}{\max_{k\neq K}\widehat{\lambda}_k/\widehat{\lambda}_{k+1}} \geqslant \frac{\lambda_K/\lambda_{K+1}}{C_3}\left\{1 - \frac{C\sqrt{p}}{\sqrt{T}} - \frac{c}{\sqrt{T}}\sqrt{s}\right\}^4$$

where $C_3 = \max(\max_{1\leqslant k\leqslant K-1} \lambda_k/\lambda_{k+1}, \max_{K+1\leqslant k\leqslant L} \lambda_k/\lambda_{k+1})$. Thus,

$$\mathbb{P}\left(\widehat{K} = K\right) \geqslant 1 - 2\exp\left\{-(C_1\sqrt{T} - C_2\sqrt{p})^2\right\},$$

where

$$C_1 = \frac{1}{c}\left[1 - \left\{\frac{\lambda_{K+1}}{\lambda_K}\max\left(\max_{1\leqslant k\leqslant K-1}\frac{\lambda_k}{\lambda_{k+1}}, \max_{K+1\leqslant k\leqslant L}\frac{\lambda_k}{\lambda_{k+1}}\right)\right\}^{1/4}\right]$$

and $C_2 = c^{-1}C$.

Also, if $p > T$, for $s < c^{-2}(\sqrt{p} - C\sqrt{T})^2$, with probability at least $1 - e^{-s}$,

$$\begin{aligned}
\frac{\widehat{\lambda}_k}{\widehat{\lambda}_{k+1}} &\leqslant \frac{\lambda_k}{\lambda_{k+1}}\left\{1 + \frac{C}{\sqrt{T}} + \frac{c}{\sqrt{pT}}\sqrt{s}\right\}^2, \qquad k = 1, \ldots, K-1,\\
\frac{\widehat{\lambda}_K}{\widehat{\lambda}_{K+1}} &\geqslant \frac{T\lambda_K}{p\lambda_{K+1}}\left\{1 - \frac{C\sqrt{T}}{\sqrt{p}} - \frac{c}{\sqrt{p}}\sqrt{s}\right\}\left\{1 - \frac{C}{\sqrt{T}} - \frac{c}{\sqrt{pT}}\sqrt{s}\right\},\\
\frac{\widehat{\lambda}_k}{\widehat{\lambda}_{k+1}} &\leqslant \frac{\lambda_k}{\lambda_{k+1}}\left\{1 + \frac{C\sqrt{T}}{\sqrt{p}} + \frac{c}{\sqrt{p}}\sqrt{s}\right\}^2, \qquad k = K+1, \ldots, L.
\end{aligned}$$

Thus, for $s < c^{-1}p + c^{-1}CT$, with probability at least $1 - e^{-s}$,

$$\frac{\widehat{\lambda}_K/\widehat{\lambda}_{K+1}}{\max_{k\neq K}\widehat{\lambda}_k/\widehat{\lambda}_{k+1}} \geqslant \frac{T\lambda_K/(p\lambda_{K+1})}{C_3}\left\{1 - \frac{C\sqrt{T}}{\sqrt{p}} - \frac{c}{\sqrt{p}}\sqrt{s}\right\}^4.$$

Thus,

$$\mathbb{P}\left(\widehat{K} = K\right) \geqslant 1 - 2\exp\left\{-(C_4\sqrt{p} - C_2\sqrt{T})^2\right\},$$

where

$$C_4 = \frac{1}{c}\left[1 - \left\{\frac{p\lambda_{K+1}}{T\lambda_K}\max\left(\max_{1\leqslant k\leqslant K-1}\frac{\lambda_k}{\lambda_{k+1}}, \max_{K+1\leqslant k\leqslant L}\frac{\lambda_k}{\lambda_{k+1}}\right)\right\}^{1/4}\right]$$

The conclusion follows. □

Note that the conditions of the case m diverges are the same as those of the case m is finite except Conditions 4.2.1 and 4.5.1, i.e. the condition of m. Thus, the proof for Theorems 4.5.1 to 4.5.3 can be derived using essentially the same argument as those of Theorems 4.4.1 to 4.3.1. Since we show the non-asymptotic results in Theorems 4.4.1 to 4.3.1, the results hold for any m. Hence, by Condition 4.5.1, it is straight forward to replace m by p^γ in Theorems 4.4.1 to 4.3.1 to get Theorems 4.5.1 to 4.5.3. It is easy to see that the parameter space is larger if the number of groups diverges. Consequentially, the convergence rate is smaller with respect to p. Details of the proof are omitted.

C.2 Auxiliary Lemmas

In this section, we denote the first r_0 columns of $\mathbf{F}_0$ as $\mathbf{F}_0$ and the rest columns as $\mathbf{F}_u$.

Lemma C.2.1 (Lemma B.1 in Fan et al. (2016))**.** $\mathbb{E}(\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}}^2) = O(pT)$.

Lemma C.2.2 (Lemma C.1 in Wang and Fan (2017))**.** *(i)* $\mathbb{E}(\|\mathbf{U}\|_2^2) = O(p)$*;*

(ii) $\|\mathbf{B}^\top \mathbf{U}\|_{\max} = O(T\sqrt{\lambda_{\max}(\mathbf{B}\mathbf{B}^T)})$ *for any* $p \times K$ *matrix* $\mathbf{B}$*;*

(iii) $\mathbb{E}(\|\mathbf{U}^\top \mathbf{U}\|_{\max}) = O(\sqrt{p}T + p)$.

Lemma C.2.3. *Under Conditions 4.2.4-4.2.7,*

(i) $\mathbb{E}(\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}}^2) = O(pT)$, $\mathbb{E}(\|\mathbf{F}_u^\top \mathbf{U}^\top\|_{\mathbb{F}}^2) = O(pTm)$, $\mathbb{E}(\|\mathbf{U}\|_2^2) = O(p)$, $\mathbb{E}(\|\mathbf{A}^\top \mathbf{U}\|_{\mathbb{F}}^2) = O(D_A^2 r_0 pT)$, $\mathbb{E}(\|\mathbf{A}^\top \mathbf{U}\mathbf{F}_0\|_{\mathbb{F}}^2) = O(D_A^2 r_0 pT)$ *and* $\mathbb{E}(\|\mathbf{C}^\top \mathbf{U}\mathbf{F}\|_{\mathbb{F}}^2) = O(D_A^2 r_0 pTm)$.

(ii) With probability at least $1 - 6e^{-s}$, $\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}} \lesssim (pT)^{1/2}\sqrt{s}$, $\|\mathbf{F}_u^\top \mathbf{U}^\top\|_{\mathbb{F}} \lesssim (pT/m)^{1/2}\sqrt{s}$, $\|\mathbf{U}\|_2 \lesssim p^{1/2}\sqrt{s}$, $\|\mathbf{U}^\top \mathbf{U}\|_{\max} \lesssim (\sqrt{p}T + p)s$, $\|\mathbf{A}^\top \mathbf{U}\|_{\mathbb{F}} \lesssim D_A(r_0 pT)^{1/2}\sqrt{s}$, $\|\mathbf{A}^\top \mathbf{U}\mathbf{F}_0\|_{\mathbb{F}} \lesssim D_A(r_0 pT)^{1/2}\sqrt{s}$ *and* $\|\mathbf{C}^\top \mathbf{U}\mathbf{F}\|_{\mathbb{F}} \lesssim D_A(r_0 pTm)^{1/2}\sqrt{s}$.

Proof. (i) By Lemmas C.2.1 and C.2.2, $\mathbb{E}(\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}}^2) = O(pT)$, $\mathbb{E}(\|\mathbf{F}_u^\top \mathbf{U}^\top\|_{\mathbb{F}}^2) = O(pTm)$ and $\mathbb{E}(\|\mathbf{U}\|_2^2) = O(p)$. In addition,

$$\mathbb{E}(\|\mathbf{A}^\top \mathbf{U}\|_{\mathbb{F}}^2) = \sum_{t=1}^{T}\sum_{k=1}^{r_0} \mathbb{E}\left\{\left(\sum_{i=1}^{p} a_{ik}u_{it}\right)^2\right\} = \sum_{t=1}^{T}\sum_{k=1}^{r_0}\sum_{i=1}^{p}\sum_{i'=1}^{p} a_{ik}a_{i'k}\mathbb{E}(u_{it}u_{i't})$$

$$\leqslant pr_0 \max_{k\leqslant r_0, i\leqslant p} a_{ik}^2 \sum_{t=1}^{T} \max_{i'\leqslant p} \sum_{i=1}^{p} |\mathbb{E}(u_{it}u_{i't})| = O(D_A^2 r_0 pT)$$

The remaining bounds can be derived similarly.

(ii) For any $x > 0$, it holds

$$\begin{aligned}
\mathbb{P}(\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}}/\sqrt{C_0 pT} > M) &\leqslant \exp(-xM)\mathbb{E}[\exp\{x\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}}/\sqrt{C_0 pT}\}] \\
&\leqslant \exp(-xM)\mathbb{E}\Big[1 + x\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}}/\sqrt{C_0 pT} \\
&\quad + x^2\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}}^2/\{2C_0 pT\} + o(x^2\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}}^2/\{2C_0 pT\})\Big] \\
&\leqslant \exp\{-xM + x + x^2/2 + o(x^2)\}
\end{aligned}$$

since $\mathbb{E}(\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}}^2) \leqslant C_0 pT$ for some $C_0 > 0$. The minimum of right hand side is $\exp\{-(M-1)^2/2\}$. Letting $s = 2^{-1}(M-1)^2$, we have with probability at least $1 - e^{-s}$, $\|\mathbf{F}_0^\top \mathbf{U}^\top\|_{\mathbb{F}} \lesssim \sqrt{pTs}$. The remaining bounds can be derived similarly.

□

Denote $\mathbf{K}_1$ a $r_0 \times r_0$ diagonal matrix with diagonals equal to the first r_0 eigenvalues of $(pT)^{-1}\mathbf{Y}^\top\mathbf{Y}$. Then $(pT)^{-1}\mathbf{Y}^\top\mathbf{Y}\widehat{\mathbf{F}_0} = \widehat{\mathbf{F}_0}\mathbf{K}_1$. Let

$$\mathbf{H}_1 = (pT)^{-1}\mathbf{A}^\top\mathbf{A}\mathbf{F}_0^\top\widehat{\mathbf{F}_0}\mathbf{K}_1^{-1}.$$

By model (4.2.7), we have $\widehat{\mathbf{F}_0} - \mathbf{F}_0\mathbf{H}_1 = \left(\sum_{i=1}^{3}\mathbf{M}_i\right)\mathbf{K}^{-1}$ where

$$\mathbf{M}_1 = \frac{1}{pT}\mathbf{F}_0\mathbf{C}^\top\mathbf{U}\widehat{\mathbf{F}_0}, \quad \mathbf{M}_2 = \frac{1}{pT}\mathbf{U}^\top\mathbf{C}\mathbf{F}_0^\top\widehat{\mathbf{F}_0}, \quad \mathbf{M}_3 = \frac{1}{pT}\mathbf{U}^\top\mathbf{U}\widehat{\mathbf{F}_0}.$$

Then, we will provide a bound on $\|\mathbf{H}_1 - \mathbf{I}_{r_0}\|_{\mathbb{F}}$ using Lemmas C.2.4 to C.2.8.

Lemma C.2.4. *Under Conditions 4.2.4-4.2.7,, with probability at least* $1 - 7e^{-s}$, $\|\mathbf{K}^{-1}\|_2 \lesssim (d_A\sqrt{r_0})^{-1}(1 + D_A\sqrt{r_0}T^{-1/2}\sqrt{s})$.

Proof. The r_0 largest eigenvalues of $(pT)^{-1}\mathbf{Y}^\top\mathbf{Y}$ are the same as those of $\mathbf{W} = (pT)^{-1}\mathbf{Y}\mathbf{Y}^\top$. As $\mathbf{Y} = \mathbf{C}\mathbf{F}^\top + \mathbf{U}$, we have $\mathbf{W} = \sum_{i=1}^5 \mathbf{W}_i$ where

$$\mathbf{W}_1 = \frac{1}{p}\mathbf{C}\mathbf{C}^\top, \quad \mathbf{W}_2 = \frac{1}{pT}\mathbf{C}\mathbf{F}^\top\mathbf{U}^\top, \quad \mathbf{W}_3 = \mathbf{W}_2^\top,$$
$$\mathbf{W}_4 = \frac{1}{pT}\mathbf{U}\mathbf{U}^\top, \quad \mathbf{W}_5 = \frac{1}{p}\mathbf{C}\left(\frac{1}{T}\mathbf{F}^\top\mathbf{F} - \mathbf{I}_K\right)\mathbf{C}^\top.$$

By Lemma C.2.3, with probability at least $1 - 6e^{-s}$,

$$\|\mathbf{W}_2\|_2 \leqslant (pT)^{-1}\|\mathbf{C}\|_2(\left\|\mathbf{F}_0^\top\mathbf{U}\right\|_{\mathbb{F}} + \left\|\mathbf{F}_u^\top\mathbf{U}\right\|_{\mathbb{F}}) \lesssim D_A\sqrt{r_0}T^{-1/2}\sqrt{s},$$

and

$$\|\mathbf{W}_4\|_2 \leqslant (pT)^{-1}\|\mathbf{U}\|_2^2 \lesssim T^{-1}s.$$

By Condition 4.2.5, with probability at least $1 - e^{-s}$, $\|\mathbf{W}_5\|_2 \lesssim T^{-1/2}\sqrt{s}$. For $k = 1, \ldots, K$, $|\lambda_k(\mathbf{W}) - \lambda_k(\mathbf{W}_1)| \leqslant \|\mathbf{W} - \mathbf{W}_1\|_2$. This implies, with probability at least $1 - 6e^{-s}$, $|\lambda_k(\mathbf{W}) - \lambda_k(\mathbf{W}_1)| \lesssim T^{-1/2}\sqrt{s}$ for each $k = 1, \ldots, K$. Note that the r_0 largest eigenvalues of $\mathbf{W}_1$ is also the r_0 largest eigenvalues of $p^{-1}\mathbf{C}^\top\mathbf{C}$. Thus, with probability at least $1 - 7e^{-s}$, $\|\mathbf{K}_1^{-1}\|_2 \lesssim (d_A\sqrt{r_0})^{-1}(1 + D_A\sqrt{r_0}T^{-1/2}\sqrt{s})$. □

Lemma C.2.5. *Under Conditions 4.2.4-4.2.7, with probability at least* $1 - 8e^{-s}$*,*

$$\frac{1}{T}\|\widehat{\mathbf{F}}_0 - \mathbf{F}_0\mathbf{H}_1\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2}{d_A^2}\left(\frac{1}{p} + \frac{1}{T^2}\right)\left(1 + \frac{s}{T}\right)s^2.$$

Proof. Note that $\|\widehat{\mathbf{F}}_0\|_{\mathbb{F}} = \sqrt{r_0 T}$ with probability 1 and by Condition 4.2.5,

$$\|\mathbf{F}_0\|_{\mathbb{F}} \lesssim \sqrt{T}\{1 + T^{-1/2}\sqrt{s}\}$$

with probability at least $1 - e^{-s}$. Then, by Lemma C.2.3, with probability at least $1 - 7e^{-s}$, $\|\mathbf{M}_1\|_{\mathbb{F}}, \|\mathbf{M}_2\|_{\mathbb{F}} \lesssim D_A\sqrt{r_0 p^{-1}T}s$ and $\|\mathbf{M}_3\|_{\mathbb{F}} \lesssim T^{-1/2}s$. Then, the results follows Lemma C.2.4.

□

Lemma C.2.6. *Under Conditions 4.2.4-4.2.7, with probability at least* $1 - 7e^{-s}$*,*

(i) $T^{-1}\|\mathbf{M}_1\|_{\mathbb{F}}^2 \lesssim D_A^2 d_A^{-2}(p^{-2} + p^{-1}T^{-1})(1 + T^{-1}s)s^3$*,*

(ii) $T^{-2}\|\mathbf{F}_0^\top \mathbf{M}_2\|_{\mathbb{F}}^2 \lesssim D_A^2 d_A^{-2}(pT)^{-1}(1 + T^{-1}s)s$*,*

(iii) $T^{-2}\|\mathbf{F}_0^\top(\widehat{\mathbf{F}}_0 - \mathbf{F}_0\mathbf{H}_1)\|_{\mathbb{F}}^2 \lesssim D_A^2 d_A^{-2}(p^{-2} + p^{-1}T^{-1})(1 + T^{-1}s)s^3$*,*

(iv) $T^{-2}\|\widehat{\mathbf{F}}_0^\top(\widehat{\mathbf{F}}_0 - \mathbf{F}_0\mathbf{H}_1)\|_{\mathbb{F}}^2 \lesssim D_A^2 d_A^{-2}(p^{-2} + p^{-1}T^{-1})(1 + T^{-1}s)s^3$*.*

Proof. (i) With probability at least $1 - 7e^{-s}$,

$$\|\mathbf{H}_1\|_2 \leqslant (pT)^{-1}\|\mathbf{A}\|_{\mathbb{F}}^2\|\mathbf{F}_0\|_{\mathbb{F}}\|\widehat{\mathbf{F}}_0\|_{\mathbb{F}}\|\mathbf{K}_1^{-1}\|_2 \lesssim \frac{D_A^2}{d_A^2}(1 + D_A^2 r_0 s T^{-1})$$

by Lemma C.2.4. Then by Lemmas C.2.3 and C.2.5, with probability at least $1 - 7e^{-s}$,

$$\begin{aligned}\|\mathbf{C}^\top\mathbf{U}\widehat{\mathbf{F}}_0\|_{\mathbb{F}}^2 &\leqslant 2\|\mathbf{C}^\top\mathbf{U}(\widehat{\mathbf{F}}_0 - \mathbf{F}_0\mathbf{H}_1)\|_{\mathbb{F}}^2 + 2\|\mathbf{C}^\top\mathbf{U}\mathbf{F}_0\mathbf{H}_1\|_{\mathbb{F}}^2 \\ &\lesssim \frac{D_A^2}{d_A^2}pT\left(\frac{T}{p} + 1\right)s^3 + \frac{D_A^2}{d_A^2}pTs\left(1 + \frac{s^2}{T^2}\right) \\ &\lesssim \frac{D_A^2}{d_A^2}(T^2 + pT)s^3.\end{aligned}$$

The result follows that $\|\mathbf{F}_0\|_{\mathbb{F}} \lesssim \sqrt{T}\{1 + T^{-1/2}\sqrt{s}\}$ with probability at least $1 - e^{-s}$.

(ii) Similar to (i), with probability at least $1 - 7e^{-s}$,

$$\frac{1}{T^2}\|\mathbf{F}_0^\top\mathbf{M}_2\|_{\mathbb{F}}^2 \leqslant \frac{1}{p^2T^4}\|\mathbf{F}_0^\top\mathbf{U}^\top\mathbf{A}\|_{\mathbb{F}}^2\|\mathbf{F}_0\|_{\mathbb{F}}^2\|\widehat{\mathbf{F}}_0\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2 s}{d_A^2 pT}\left(1 + \frac{s}{T}\right).$$

(iii) Combining (i) and (ii), the result follows from the proof of Lemma C.2.5.

(iv) The result follows from $\|\widehat{\mathbf{F}}_0^\top(\widehat{\mathbf{F}}_0 - \mathbf{F}_0\mathbf{H}_1)\|_{\mathbb{F}} \leqslant \|\widehat{\mathbf{F}}_0 - \mathbf{F}_0\mathbf{H}_1\|_{\mathbb{F}}^2 + \|\mathbf{H}^\top\mathbf{F}_0^\top(\widehat{\mathbf{F}}_0 - \mathbf{F}_0\mathbf{H})\|_{\mathbb{F}}$.

□

Lemma C.2.7. *Under Conditions 4.2.4-4.2.7, with probability at least* $1 - 7e^{-s}$*,*

$$\|\mathbf{H}_1^\top \mathbf{H}_1 - \mathbf{I}_{r_0}\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2}{d_A^2}\left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right)^3.$$

Proof. By Condition 4.2.5, $\|T^{-1}\mathbf{F}_0^\top \mathbf{F}_0 - \mathbf{I}_{r_0}\|_{\mathbb{F}} \lesssim T^{-1/2}\sqrt{s}$ with probability at least $1 - e^{-s}$. Also, $\widehat{\mathbf{F}}_0^\top \widehat{\mathbf{F}}_0 = T\mathbf{I}_{r_0}$. Thus,

$$\mathbf{H}_1^\top \mathbf{H}_1 - \mathbf{I}_{r_0} = \mathbf{H}_1^\top \left(\mathbf{I}_{r_0} - \frac{1}{T}\mathbf{F}_0^\top \mathbf{F}_0\right)\mathbf{H}_1 + \frac{1}{T}(\mathbf{F}_0\mathbf{H} - \widehat{\mathbf{F}_{00}})^\top \mathbf{F}_0\mathbf{H}_1 + \frac{1}{T}\widehat{\mathbf{F}}_0^\top(\mathbf{F}_0\mathbf{H}_1 - \widehat{\mathbf{F}}_0).$$

The result follows from Lemma C.2.6. □

Lemma C.2.8. *Under Conditions 4.2.4-4.2.7, with probability at least* $1 - 7e^{-s}$*,*

$$\|\mathbf{H}_1 - \mathbf{I}_{r_0}\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2}{d_A^2}\left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right).$$

Proof. Note that $p\mathbf{H}_1\mathbf{K}_1 = \mathbf{A}^\top\mathbf{A}\left(T^{-1}\mathbf{F}_0^\top\widehat{\mathbf{F}}_0 - \mathbf{H}_1\right) + \mathbf{A}^\top\mathbf{A}\mathbf{H}_1$. By Lemma C.2.6, with probability at least $1 - 7e^{-s}$,

$$\begin{aligned}
&\left\|\frac{1}{p}\mathbf{A}^\top\mathbf{A}\left(\frac{1}{T}\mathbf{F}_0^\top\widehat{\mathbf{F}}_0 - \mathbf{H}_1\right)\right\|_{\mathbb{F}}^2 \\
\leqslant&\frac{1}{p^2}\|\mathbf{A}^\top\mathbf{A}\|_{\mathbb{F}}^2\frac{1}{T^2}\|\mathbf{F}_0^\top(\widehat{\mathbf{F}}_0 - \mathbf{F}_0\mathbf{H}_1)\|_{\mathbb{F}}^2 + \frac{1}{p^2}\|\mathbf{A}^\top\mathbf{A}\|_{\mathbb{F}}^2\left\|\mathbf{I}_{r_0} - \frac{1}{T}\mathbf{F}_0^\top\mathbf{F}_0\right\|_{\mathbb{F}}^2\|\mathbf{H}\|_2^2 \\
\lesssim&\frac{D_A^2}{d_A^2}\left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right).
\end{aligned}$$

Therefore, with probability at least $1 - 7e^{-s}$,

$$\left\|\frac{1}{p}\mathbf{A}^\top\mathbf{A}\mathbf{H}_1 - \mathbf{H}_1\mathbf{K}_1\right\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2}{d_A^2}\left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right).$$

This implies that with probability at least $1-7e^{-s}$, $\mathbf{H}_1$ (up to an error term) is a matrix consisting of eigenvectors of $p^{-1}\mathbf{A}^\top\mathbf{A}$. By Condition 4.2.5, $\mathbf{A}^\top\mathbf{A}$ 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 $1-7e^{-s}$,

$$\|\mathbf{H}_1-\mathbf{J}_1\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2}{d_A^2}\left(\frac{s}{T}+\frac{s^2}{T^2}+\frac{s^3}{p^2}+\frac{s^3}{pT}\right)\left(1+\frac{s}{T}\right)$$

for some diagonal matrix $\mathbf{J}_1$. By Lemma C.2.7, with probability at least $1-7e^{-s}$, for each $k=1,\ldots,r_0$,

$$|\lambda_k(\mathbf{H}_1)-\eta|^2 \lesssim \frac{D_A^2}{d_A^2}\left(\frac{s}{T}+\frac{s^2}{T^2}+\frac{s^3}{p^2}+\frac{s^3}{pT}\right)\left(1+\frac{s}{T}\right)$$

where η is either 1 or -1. Without loss of generality , we can assume that all entries of $\mathbf{H}_1$ is positive (otherwise we can multiply the corresponding columns of $\widehat{\mathbf{F}}_0$ and $\widehat{\mathbf{A}}$ by -1). Hence, denoting $\mathbf{H}_1=\{h_{ij}^{(1)}\}_{i,j=1}^{r_0}$, with probability at least $1-7e^{-s}$,

$$\|\mathbf{H}_1-\mathbf{I}_{r_0}\|_{\mathbb{F}}^2=\sum_{i\neq j}(h_{ij}^{(1)})^2+\sum_{i=1}^{r_0}(h_{ii}^{(1)}-1)^2 \lesssim \frac{D_A^2}{d_A^2}\left(\frac{s}{T}+\frac{s^2}{T^2}+\frac{s^3}{p^2}+\frac{s^3}{pT}\right)\left(1+\frac{s}{T}\right).$$

□

Recall that $\widehat{\mathbf{A}}=T^{-1}\mathbf{Y}\widehat{\mathbf{F}}_0$. We have $\widehat{\mathbf{A}}-\mathbf{A}\mathbf{H}_1=\sum_{i=1}^3\mathbf{E}_i$ where

$$\mathbf{E}_1=\frac{1}{T}\mathbf{A}\mathbf{F}_0^\top(\widehat{\mathbf{F}}_0-\mathbf{F}_0\mathbf{H}_1),\quad \mathbf{E}_2=\frac{1}{T}\mathbf{U}\mathbf{F}_0\mathbf{H}_1,\quad \mathbf{E}_3=\frac{1}{T}\mathbf{U}(\widehat{\mathbf{F}}_0-\mathbf{F}_0\mathbf{H}_1).$$

Lemma C.2.9. *Under Conditions 4.2.4-4.2.7, with probability at least* $1-7e^{-s}$, $p^{-1}\|\widehat{\mathbf{A}}-\mathbf{A}\mathbf{H}_1\|_{\mathbb{F}}^2 \lesssim d_A^2(T^{-1}s+T^{-2}s^2+p^{-1}T^{-1}s^3+p^{-2}s^3)(1+T^{-1}s)$.

Proof. By Lemmas C.2.3 and C.2.5, with probability at least $1-7e^{-s}$, $\|\mathbf{E}_1\|_{\mathbb{F}}^2 \lesssim D_A^2d_A^{-2}(p^{-1}+T^{-1})(1+T^{-1}s)s^3$, $\|\mathbf{E}_2\|_{\mathbb{F}}^2 \lesssim T^{-1}p(1+T^{-1}s)^2s$ and $\|\mathbf{E}_3\|_{\mathbb{F}}^2 \lesssim D_A^2d_A^{-2}(T^{-1}+pT^{-3})(1+T^{-1}s)s^3$. So $p^{-1}\|\widehat{\mathbf{A}}-\mathbf{A}\mathbf{H}_1\|_{\mathbb{F}}^2 \lesssim D_A^2d_A^{-2}(pT^{-1}s+pT^{-2}s^2+T^{-1}s^3+p^{-1}s^3)(1+T^{-1}s)$.

□

Then, we will provide similar results corresponding to the first K eigenvalues of $(pT)^{-1}\mathbf{Y}^\top\mathbf{Y}$. Denote $\mathbf{K}_2$ a $K\times K$ diagonal matrix with diagonals equal to the first K eigenvalues of $(pT)^{-1}\mathbf{Y}^\top\mathbf{Y}$. Then $(pT)^{-1}\mathbf{Y}^\top\mathbf{Y}\widehat{\mathbf{F}}=\widehat{\mathbf{F}}\mathbf{K}_2$. Let $\mathbf{H}_2=(pT)^{-1}\mathbf{C}^\top\mathbf{C}\mathbf{F}^\top\widehat{\mathbf{F}}\mathbf{K}_2^{-1}$. Using our central model (4.2.7)

in the paper (*i.e.* $\mathbf{Y} = \mathbf{C}\mathbf{F}^\top + \mathbf{U}$), we have $\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2 = \left(\sum_{i=1}^3 \mathbf{N}_i\right)\mathbf{K}_2^{-1}$ where

$$\mathbf{N}_1 = \frac{1}{pT}\mathbf{F}\mathbf{C}^\top\mathbf{U}\widehat{\mathbf{F}}, \quad \mathbf{N}_2 = \frac{1}{pT}\mathbf{U}^\top\mathbf{C}\mathbf{F}^\top\widehat{\mathbf{F}}, \quad \mathbf{N}_3 = \frac{1}{pT}\mathbf{U}^\top\mathbf{U}\widehat{\mathbf{F}}.$$

Then, we will provide a bound on $\|\mathbf{H}_2 - \mathbf{I}_K\|_\mathbb{F}$ using Lemmas C.2.10 to C.2.14.

Lemma C.2.10. *Under Conditions 4.2.4-4.2.7, with probability at least* $1 - 7e^{-s}$, $\|\mathbf{K}_2^{-1}\|_2 \lesssim (d_B\sqrt{\min_j r_j})^{-1}(1 + D_A\sqrt{r_0}T^{-1/2}\sqrt{s})$.

Proof. The proof is similar as that of Lemma C.2.4 □

Lemma C.2.11. *Under Conditions 4.2.4-4.2.7 in the main paper, with probability at least* $1-7e^{-s}$,

$$\frac{1}{T}\|\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2\|_\mathbb{F}^2 \lesssim \frac{D_A^2 r_0}{d_B^2 \min_j r_j}\left(\frac{m}{p} + \frac{1}{T^2}\right)\left(1 + \frac{s}{T}\right)s^2.$$

Proof. Note that $\|\widehat{\mathbf{F}}\|_\mathbb{F} = \sqrt{KT}$ with probability 1 and by Condition 4.2.5 in the main paper, $\|\mathbf{F}\|_\mathbb{F} \lesssim \sqrt{T}\{1 + T^{-1/2}\sqrt{s}\}$ with probability at least $1 - e^{-s}$. Then, by Lemma C.2.3, with probability at least $1 - 6e^{-s}$, $\|\mathbf{N}_1\|_\mathbb{F}, \|\mathbf{N}_2\|_\mathbb{F} \lesssim D_A\sqrt{r_0 pT/ms}$ and $\|\mathbf{N}_3\|_\mathbb{F} \lesssim T^{-1/2}s$. Then, the results follows Lemma C.2.10. □

Lemma C.2.12. *Under Conditions 4.2.4-4.2.7 in the main paper, with probability at least* $1-7e^{-s}$,

(i) $T^{-1}\|\mathbf{N}_1\|_\mathbb{F}^2 \lesssim D_A^2 r_0 d_B^{-2}(\min_j r_j)^{-1}(p^{-2}m + p^{-1}T^{-1}m)(1 + T^{-1}s)s^3$,

(ii) $T^{-2}\|\mathbf{F}^\top\mathbf{N}_2\|_\mathbb{F}^2 \lesssim D_A^2 r_0 d_B^{-2}(\min_j r_j)^{-1}p^{-1}T^{-1}m(1 + T^{-1}s)s$,

(iii) $T^{-2}\|\mathbf{F}^\top(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2)\|_\mathbb{F}^2 \lesssim D_A^2 r_0 d_B^{-2}(\min_j r_j)^{-1}(p^{-2}m + p^{-1}T^{-1}m)(1 + T^{-1}s)s^3$,

(iv) $T^{-2}\|\widehat{\mathbf{F}}^\top(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2)\|_\mathbb{F}^2 \lesssim D_A^2 r_0 d_B^{-2}(\min_j r_j)^{-1}(p^{-2}m + p^{-1}T^{-1}m)(1 + T^{-1}s)s^3$.

Proof. (i) With probability at least $1 - 7e^{-s}$, $\|\mathbf{H}_2\|_2 \leqslant p^{-1}T^{-1}m\|\mathbf{C}\|_\mathbb{F}^2\|\mathbf{F}\|_\mathbb{F}\|\widehat{\mathbf{F}}\|_\mathbb{F}\|\mathbf{K}_2^{-1}\|_2 \lesssim D_A^2 r_0 d_B^{-2}(\min_j r_j)^{-1}m(1 + d_B^2 \min_j r_j sT^{-1})$ by Lemma C.2.10. Then by Lemmas C.2.3 and C.2.11, with probability at least $1 - 7e^{-s}$,

$$\|\mathbf{C}^\top\mathbf{U}\widehat{\mathbf{F}}\|_\mathbb{F}^2 \leqslant 2\|\mathbf{C}^\top\mathbf{U}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2)\|_\mathbb{F}^2 + 2\|\mathbf{C}^\top\mathbf{U}\mathbf{F}\mathbf{H}_2\|_\mathbb{F}^2$$

$$\lesssim pTm\left(\frac{T}{p}+1\right)s^3+pTms\left(1+\frac{s^2}{T^2}\right)$$
$$\lesssim \frac{D_A^2 r_0}{d_B^2 \min_j r_j} m(T^2+pT)s^3.$$

The result follows that $\|\mathbf{F}\|_{\mathbb{F}} \lesssim \sqrt{T}\{1+T^{-1/2}\sqrt{s}\}$ with probability at least $1-e^{-s}$.

(ii) Similar to (i), with probability at least $1-7e^{-s}$,

$$\frac{1}{T^2}\|\mathbf{F}^\top \mathbf{M}_2\|_{\mathbb{F}}^2 \leqslant \frac{1}{p^2T^4}\|\mathbf{F}^\top\mathbf{U}^\top\mathbf{C}\|_{\mathbb{F}}^2\|\mathbf{F}\|_{\mathbb{F}}^2\|\widehat{\mathbf{F}}\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2 r_0 ms}{d_B^2 \min_j r_j pT}\left(1+\frac{s}{T}\right).$$

(iii) Combining (i) and (ii), the result follows from the proof of Lemma C.2.11.

(iv) The result follows from $\|\widehat{\mathbf{F}}^\top(\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H}_2)\|_{\mathbb{F}} \leqslant \|\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H}_2\|_{\mathbb{F}}^2 + \|\mathbf{H}_2^\top\mathbf{F}^\top(\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H}_2)\|_{\mathbb{F}}$.

□

Lemma C.2.13. *Under Conditions 4.2.4-4.2.7 in the main paper, with probability at least* $1-7e^{-s}$*,*

$$\|\mathbf{H}_2^\top\mathbf{H}_2-\mathbf{I}_K\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2 r_0}{d_B^2 \min_j r_j}\left(\frac{s}{T}+\frac{s^2}{T^2}+\frac{s^3}{p^2}+\frac{s^3}{pT}\right)\left(1+\frac{s}{T}\right)^3.$$

Proof. By Condition 4.2.5 in the main paper, $\|T^{-1}\mathbf{F}^\top\mathbf{F}-\mathbf{I}_K\|_{\mathbb{F}} \lesssim T^{-1/2}\sqrt{s}$ with probability at least $1-e^{-s}$. Also, $\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}} = T\mathbf{I}_K$. Thus,

$$\mathbf{H}_2^\top\mathbf{H}_2-\mathbf{I}_K = \mathbf{H}_2^\top\left(\mathbf{I}_K-\frac{1}{T}\mathbf{F}^\top\mathbf{F}\right)\mathbf{H}_2+\frac{1}{T}(\mathbf{F}\mathbf{H}_2-\widehat{\mathbf{F}})^\top\mathbf{F}\mathbf{H}_2+\frac{1}{T}\widehat{\mathbf{F}}^\top(\mathbf{F}\mathbf{H}_2-\widehat{\mathbf{F}}).$$

The result follows from Lemma C.2.12. □

Lemma C.2.14. *Under Conditions 4.2.4-4.2.7 in the main paper, with probability at least* $1-7e^{-s}$*,*

$$\|\mathbf{H}_2-\mathbf{I}_K\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2 r_0}{d_B^2 \min_j r_j}\left(\frac{s}{T}+\frac{s^2}{T^2}+\frac{s^3}{p^2}+\frac{s^3}{pT}\right)\left(1+\frac{s}{T}\right).$$

Proof. Note that $p\mathbf{H}_2\mathbf{K}_2 = \mathbf{C}^\top\mathbf{C}\left(T^{-1}\mathbf{F}^\top\widehat{\mathbf{F}} - \mathbf{H}_2\right) + \mathbf{C}^\top\mathbf{C}\mathbf{H}_2$. By Lemma C.2.6, with probability at least $1 - 7e^{-s}$,

$$\begin{aligned}
&\left\|\frac{1}{p}\mathbf{C}^\top\mathbf{C}\left(\frac{1}{T}\mathbf{F}^\top\widehat{\mathbf{F}} - \mathbf{H}_2\right)\right\|_{\mathbb{F}}^2 \\
\leqslant &\frac{1}{p^2}\|\mathbf{C}^\top\mathbf{C}\|_{\mathbb{F}}^2\frac{1}{T^2}\|\mathbf{F}^\top(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2)\|_{\mathbb{F}}^2 + \frac{1}{p^2}\|\mathbf{C}^\top\mathbf{C}\|_{\mathbb{F}}^2\left\|\mathbf{I}_K - \frac{1}{T}\mathbf{F}^\top\mathbf{F}\right\|_{\mathbb{F}}^2\|\mathbf{H}_2\|_2^2 \\
\lesssim &\frac{D_A^2 r_0}{d_B^2\min_j r_j}\left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right).
\end{aligned}$$

Therefore, with probability at least $1 - 7e^{-s}$,

$$\left\|\frac{1}{p}\mathbf{C}^\top\mathbf{C}\mathbf{H}_2 - \mathbf{H}_2\mathbf{K}_2\right\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2 r_0}{d_B^2\min_j r_j}\left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right).$$

This implies that with probability at least $1 - 7e^{-s}$, $\mathbf{H}_2$ (up to an error term) is a matrix consisting of eigenvectors of $p^{-1}\mathbf{C}^\top\mathbf{C}$. By Condition 4.2.5 in the main paper, $\mathbf{C}^\top\mathbf{C}$ 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 $1 - 7e^{-s}$,

$$\|\mathbf{H}_2 - \mathbf{J}_2\|_{\mathbb{F}}^2 \lesssim \frac{D_A^2 r_0}{d_B^2\min_j r_j}\left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right)$$

for some diagonal matrix $\mathbf{J}_2$. By Lemma C.2.13, with probability at least $1 - 7e^{-s}$, for each $k = 1, \ldots, K$,

$$|\lambda_k(\mathbf{H}_2) - \eta|^2 \lesssim \frac{D_A^2 r_0}{d_B^2\min_j r_j}\left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right)$$

where η is either 1 or -1. Without loss of generality , we can assume that all entries of $\mathbf{H}_2$ is positive (otherwise we can multiply the corresponding columns of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{C}}$ by -1). Hence,

denoting $\mathbf{H}_2 = \{h_{ij}^{(2)}\}_{i,j=1}^K$, with probability at least $1 - 5e^{-s}$,

$$\|\mathbf{H}_2 - \mathbf{I}_K\|_{\mathbb{F}}^2 = \sum_{i \neq j}(h_{ij}^{(2)})^2 + \sum_{i=1}^K (h_{ii}^{(2)} - 1)^2 \lesssim \frac{D_A^2 r_0}{d_B^2 \min_j r_j}\left(\frac{s}{T} + \frac{s^2}{T^2} + \frac{s^3}{p^2} + \frac{s^3}{pT}\right)\left(1 + \frac{s}{T}\right).$$

□

Recall that $\widehat{\mathbf{C}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}$. We have $\widehat{\mathbf{C}} - \mathbf{C}\mathbf{H}_2 = \sum_{i=1}^3 \mathbf{G}_i$ where

$$\mathbf{G}_1 = \frac{1}{T}\mathbf{C}\mathbf{F}^\top(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2), \quad \mathbf{G}_2 = \frac{1}{T}\mathbf{U}\mathbf{F}\mathbf{H}_2, \quad \mathbf{G}_3 = \frac{1}{T}\mathbf{U}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}_2).$$

Lemma C.2.15. *Under Conditions 4.2.4-4.2.7, with probability at least* $1-7e^{-s}$, $p^{-1}\|\widehat{\mathbf{C}}-\mathbf{C}\mathbf{H}_2\|_{\mathbb{F}}^2 \lesssim D_A^2 r_0 d_B^{-2}(\min_j r_j)^{-1}(T^{-1}s + T^{-2}s^2 + p^{-1}T^{-1}s^3 + p^{-2}s^3)(1 + T^{-1}s)$.

Proof. By Lemmas C.2.3 and C.2.11, with probability at least $1-7e^{-s}$, $\|\mathbf{G}_1\|_{\mathbb{F}}^2 \lesssim D_A^2 r_0 d_B^{-2}(\min_j r_j)^{-1}$ $(p^{-1}m+T^{-1})(1+T^{-1}s)s^3$, $\|\mathbf{G}_2\|_{\mathbb{F}}^2 \lesssim T^{-1}pm(1+T^{-1}s)^2 s$ and $\|\mathbf{G}_3\|_{\mathbb{F}}^2 \lesssim D_A^2 r_0 d_B^{-2}(\min_j r_j)^{-1}(T^{-1}+$ $pT^{-3})(1 + T^{-1}s)s^3$. So $p^{-1}\|\widehat{\mathbf{C}} - \mathbf{C}\mathbf{H}_2\|_{\mathbb{F}}^2 \lesssim D_A^2 r_0 d_B^{-2}(\min_j r_j)^{-1}(pT^{-1}s + pT^{-2}s^2 + T^{-1}s^3 +$ $p^{-1}ms^3)(1 + T^{-1}s)$.

□

Lemma C.2.16. *With probability at least* $1 - 3e^{-s}$,

(i) $\|\mathbf{N}_1\|_{\max}, \|\mathbf{N}_2\|_{\max} \lesssim D_A\sqrt{r_0}p^{-1/2}\{\log(T)\}^{2/r_2}s$;

(ii) $\|\mathbf{N}_3\|_{\max} \lesssim p^{-1/2}\{\log(T)\}^{1/r_2}s$.

Proof. By Lemma C.2.3, with probability at least $1 - 6e^{-s}$, $\|\mathbf{U}^\top\mathbf{U}\|_{\max} \lesssim (\sqrt{p}T + p)s$. Also, by Lemma C.2.2, $\|\mathbf{U}^\top\mathbf{C}\|_\infty \lesssim D_A\sqrt{r_0 p}T$. Hence, with probability at least $1 - e^{-s}$, $\|\mathbf{U}^\top\mathbf{C}\|_\infty \lesssim$ $D_A\sqrt{r_0 p}Ts$. Then, the results follow from that $\|\mathbf{F}\|_{\max} \lesssim \{\log(T) + s\}^{1/r_2}$ with probability at least $1 - e^{-s}$. □

Recall that

$$\mathbf{\Sigma}(\boldsymbol{z}) = \mathbf{C}_0(\boldsymbol{z})\mathbf{C}_0(\boldsymbol{z})^\top + \sigma_u^2\mathbf{I}_p.$$

where

$$\mathbf{C}_0(\boldsymbol{z}) = \begin{bmatrix} D_A\mathbf{A}_1^{(0)} & d_B\mathbf{B}_1^{(0)} & & \\ \vdots & & \ddots & \\ D_A\mathbf{A}_m^{(0)} & & & d_B\mathbf{B}_m^{(0)} \end{bmatrix},$$

$$(\mathbf{A}_j^{(0)}, \mathbf{B}_j^{(0)}) = \begin{bmatrix} \mathbf{D}_j^{(1)} \\ \mathbf{D}_j^{(2)} \end{bmatrix},$$

$$\mathbf{D}_j^{(1)} = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ -1 & 1 & \dots & 1 & 1 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -r_j+1 & 1 \end{bmatrix}$$

is a $r_j \times (r_0 + r_j)$ matrix and

$$\mathbf{D}_j^{(2)} = \begin{bmatrix} \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}$$

is a $(p_j(\boldsymbol{z}) - r_j) \times (r_0 + r_j)$ block diagonal matrix consisting of $r_0 + r_j$ vectors of $d_0\mathbf{1}$ for $j = 1, \dots, m$.

Lemma C.2.17. *For $\boldsymbol{\Sigma}(\boldsymbol{z})$ defined above, any $\ell \neq \ell'$ and $\boldsymbol{z}^{(\ell)}, \boldsymbol{z}^{(\ell')} \in \mathcal{T}$,*

$$\begin{aligned} \log\left(\frac{|\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell)})|}{|\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell')})|}\right) =& r_0 \log\left(\frac{D_A^2 r_0 \sum_{k=1}^m \frac{p_k(\boldsymbol{z}^{(\ell)})}{r_0+r_k} + \sigma_u^2}{D_A^2 r_0 \sum_{k=1}^m \frac{p_k(\boldsymbol{z}^{(\ell')})}{r_0+r_k} + \sigma_u^2}\right) \\ &+ \sum_{k=1}^m r_k \log\left(\frac{d_B^2 \min_j r_j p_k(\boldsymbol{z}^{(\ell)}) + \sigma_u^2(r_0+r_1)}{d_B^2 \min_j r_j p_k(\boldsymbol{z}^{(\ell')}) + \sigma_u^2(r_0+r_1)}\right). \end{aligned}$$

Proof. By definition of $\mathbf{C}_0(\boldsymbol{z})$ in (C.1.1) and definition of $\boldsymbol{\Sigma}(\boldsymbol{z})$ in (C.1.3), the eigenvalues of $\mathbf{C}_0(\boldsymbol{z})^\top \mathbf{C}_0(\boldsymbol{z})$ is

$$\left(\underbrace{D_A^2 r_0 \sum_{k=1}^m \frac{p_k(\boldsymbol{z})}{r_0+r_k}, \ldots, D_A^2 r_0 \sum_{k=1}^m \frac{p_k(\boldsymbol{z})}{r_0+r_k}}_{r_0}, \underbrace{\frac{d_B^2 \min_j r_j p_1(\boldsymbol{z})}{r_0+r_1}, \ldots, \frac{d_B^2 \min_j r_j p_1(\boldsymbol{z})}{r_0+r_1}}_{r_1}, \ldots, \right.$$
$$\left. \underbrace{\frac{d_B^2 \min_j r_j p_m(\boldsymbol{z})}{r_0+r_m}, \ldots, \frac{d_B^2 \min_j r_j p_m(\boldsymbol{z})}{r_0+r_m}}_{r_m} \right).$$

Since $\mathbf{C}_0(\boldsymbol{z})^\top \mathbf{C}_0(\boldsymbol{z})$ and $\mathbf{C}_0(\boldsymbol{z}) \mathbf{C}_0(\boldsymbol{z})^\top$ share the same non-zero eigenvalues,

$$|\boldsymbol{\Sigma}(\boldsymbol{z})| = \left(D_A^2 r_0 \sum_{k=1}^m \frac{p_k(\boldsymbol{z})}{r_0+r_k} + \sigma_u^2 \right)^{r_0} \Pi_{k=1}^m \left(\frac{d_B^2 \min_j r_j p_k(\boldsymbol{z})}{r_0+r_k} + \sigma_u^2 \right)^{r_k} \sigma_u^{2(p-\sum_{k=1}^m r_k)}.$$

Note that by definition of $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)}\}$, $p_k(\boldsymbol{z}^{(j)})$ and $p_k(\boldsymbol{z}^{(\ell)})$ are the same for $m-2$ pairs and with difference 1 for 2 pairs. Then for any $\ell \neq \ell'$,

$$\begin{aligned} \log \left(\frac{|\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell)})|}{|\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell')})|} \right) =& r_0 \log \left(\frac{D_A^2 r_0 \sum_{k=1}^m \frac{p_k(\boldsymbol{z}^{(\ell)})}{r_0+r_k} + \sigma_u^2}{D_A^2 r_0 \sum_{k=1}^m \frac{p_k(\boldsymbol{z}^{(\ell')})}{r_0+r_k} + \sigma_u^2} \right) \\ &+ \sum_{k=1}^m r_k \log \left(\frac{d_B^2 \min_j r_j p_k(\boldsymbol{z}^{(\ell)}) + \sigma_u^2 (r_0+r_1)}{d_B^2 \min_j r_j p_k(\boldsymbol{z}^{(\ell')}) + \sigma_u^2 (r_0+r_1)} \right). \end{aligned}$$

□

Lemma C.2.18. *For $\boldsymbol{\Sigma}(\boldsymbol{z})$ defined above, any $\ell \neq \ell'$ and $\boldsymbol{z}^{(\ell)}, \boldsymbol{z}^{(\ell')} \in \mathcal{T}$,*

$$\operatorname{tr}\left(\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell)})^{-1} \boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell')}) \right) = p.$$

Proof. Note that

$$\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell)})^{-1} = \begin{bmatrix} E & -\frac{1}{1+\sigma_u^2}\boldsymbol{c}_1(\ell)^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_0^{-1}(\ell) \\ -\frac{1}{1+\sigma_u^2}\boldsymbol{\Sigma}_0^{-1}(\ell)\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell) & \boldsymbol{\Sigma}_0^{-1}(\ell) \end{bmatrix},$$

where $E = (1+\sigma_u^2)^{-1} + (1+\sigma_u^2)^{-2}\boldsymbol{c}_1(\ell)^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_0^{-1}(\ell)\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell)$, $\boldsymbol{\Sigma}_0(\ell) = \boldsymbol{\Sigma}_{-1} - (1+\sigma_u^2)^{-1}\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell)\boldsymbol{c}_1(\ell)^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top$ and $\boldsymbol{\Sigma}_{-1} = \sigma_u^2 \mathbf{I}_{p-1} + \mathbf{C}_{-1}(\boldsymbol{z}_{-1})\mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top$. Then,

$$\begin{aligned}
&\mathrm{tr}(\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell')})^{-1}\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell)})) \\
=&1 + \frac{1}{1+\sigma_u^2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\ell')\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top\right) \\
&- \frac{1}{1+\sigma_u^2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\ell')\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell)\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top\right) \\
&- \frac{1}{1+\sigma_u^2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\ell')\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')\boldsymbol{c}_1(\ell)^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top\right) - \mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\ell')\boldsymbol{\Sigma}_{-1}\right)
\end{aligned}$$

Since

$$\begin{aligned}
&\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\ell')\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top\right) \\
=&\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}(\ell)\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell') \\
&- \frac{\frac{1}{1+\sigma_u^2}\{\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')\}^2}{1 + \frac{1}{1+\sigma_u^2}\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')}, \\
&\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\ell')\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell)\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top\right) \\
=&\,\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\ell')\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')\boldsymbol{c}_1(\ell)^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top\right) \\
=&\boldsymbol{c}_1(\ell)^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}(\ell)\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell') \\
&- \frac{\frac{1}{1+\sigma_u^2}\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')\boldsymbol{c}_1(\ell)^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}(\ell)\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')}{1 + \frac{1}{1+\sigma_u^2}\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')}, \\
&\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\ell')\boldsymbol{\Sigma}_{-1}\right) = p - 1 - \frac{\frac{1}{1+\sigma_u^2}\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}(\ell)\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')}{1 + \frac{1}{1+\sigma_u^2}\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}(\ell)\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')},
\end{aligned}$$

we have

$$\mathrm{tr}(\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell')})^{-1}\boldsymbol{\Sigma}(\boldsymbol{z}^{(\ell)})) = -\frac{\frac{2}{1+\sigma_u^2}\boldsymbol{c}_1(\ell)^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}(\ell)\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')}{1+\frac{1}{1+\sigma_u^2}\boldsymbol{c}_1(\ell')^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}(\ell)\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell')}.$$

Note that since

$$\boldsymbol{\Sigma}_{-1}^{-1} = \frac{1}{\sigma_u^2}\mathbf{I}_{p-1} - \frac{1}{\sigma_u^2}\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\left\{\mathbf{I}_K + \frac{1}{\sigma_u^2}\mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})\right\}^{-1}\mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top,$$

we have

$$\boldsymbol{c}_1(\ell)^\top \mathbf{C}_{-1}(\boldsymbol{z}_{-1})^\top \boldsymbol{\Sigma}_{-1}^{-1}(\ell)\mathbf{C}_{-1}(\boldsymbol{z}_{-1})\boldsymbol{c}_1(\ell') = 0$$

for $\ell \neq \ell'$. The conclusion follows. □